现代信息检索
Modern Information Retrieval

第0讲 课程简介
About the course

提纲

① 什么是信息检索？

② 为什么要学习信息检索？

③ 课程情况

主讲老师介绍

- 主讲老师：何苯 国科大计算机学院教授，博士生导师。中科院软件所客座研究员
  - 研究方向：信息检索、自然语言处理
  - 邮件地址： benhe@ucas.ac.cn
- 曾主持国家自然科学基金、北京市自然科学基金、中科院重大战略专项课题等科研项目及多个企业横向项目
- 在SIGIR, ACL, EMNLP, TOIS, IPM等相关领域学术会议和期刊发表论文80余篇，Google Scholar引用3000余次
- SIGIR, ACL, EMNLP, AAAI, CIKM等顶会审稿人，IPM（一区期刊）副主编
- 曾获ECIR 2016 Test of Time award honorable mention 、SIGIR 2021 best short paper award等奖项

提纲

① 什么是信息检索？

② 为什么要学习信息检索？

③ 课程情况

- 从几个互联网应用说起……

网页 图片 视频 地图 新闻 音乐 购物 Gmail 更多 ▼

Google 现代信息检索 Google 搜索

获得约 272,000 条结果 （用时 0.19 秒） 高级搜索

所有结果
更多

网页
所有中文网页
简体中文网页

普通视图
神奇罗盘
时光隧道
更多描述

更多搜索工具

现代信息检索
本课程为计算机科学与技术、图书情报等相关学科研究生的专业基础课，本课程不是讲授如何利用检索工具进行情报检索，而主要以互联网内容应用为背景讲授和讨论现代信息 ...
ir.ict.ac.cn/ircourse/ - 网页快照 - 类似结果

《现代信息检索》图书详细资料信息/ China-Pub
2009年3月12日 ... 现代信息检索计算机_信息系统_综合教材_征订教材_高等理工教材_研究生/本科/专科教材_文法类_图书档案学教材_计算机教材_高职高专_计算机类计算机_ ...
www.china-pub.com › 计算机 › 信息系统 - 网页快照 - 类似结果

《现代信息检索（英文版）》图书详细资料信息/ China-Pub
2010年4月15日 ... 现代信息检索（英文版）计算机_信息系统_综合教材_研究生/本科/专科教材_文法类_图书档案学教材_计算机教材_高职高专_计算机类计算机_信息系统_管理 ...
www.china-pub.com/16606 - 网页快照 - 类似结果

现代信息检索——计算机科学丛书-图书-当当网
2010年5月26日 ... 本书介绍了现代信息检索的绝大部分研究领域，全面展示了现代信息检索的基础知识和高级主题，涉及该领域的各个方面。本书的两位主要作者是现代信息检索 ...
product.dangdang.com › 图书 › 计算机/网络 - 网页快照 - 类似结果

现代信息检索 - Google 图书结果
袁学松 - 2007 - Education, Higher - 182 页
21世纪高职高专规划教材.
books.google.com.hk/books?isbn=7508442997...

求 ”领导科学 ”和 ”现代信息检索 ”论文，各一篇！ 百度知道
2007年7月24日 ... 给你推荐一个网站，那里有不少相关论文，都是公开发表的专业论文，及博硕士毕业论文，你上去挑挑，应该能解决你的问题中国知网www.cnki.net 比如，你上去输入 ...
zhidao.baidu.com › 资源共享 › 文档/报告共享 - 网页快照 - 类似结果

京东

发现好货 全部分类

爆款台式机5折限量抢 搜索

我的购物车 0

潮流先锋 3C新品 农资绿植 服装城 狂欢序幕 999元抢洗碗机 苹果平板

发现好货

华为全面屏轻薄笔记本

采用全面屏设计带来沉浸式视觉体验，十点顺滑触控操作更轻松。选用金属材质坚固稳定经久耐用，轻薄机身易于携带方便出行。拥有隐藏式摄像头使用更安全，高低音四路输出带来清亮穿透的高音效果。

556

华为 无线充电 陶瓷手机

微晶陶瓷，256G内存。支持30W无线快充，充电更快速，不怕因忘记充电而影响工作进度。采用微晶陶瓷材质打造，机身细腻有光泽，彰显你的品味。配备256G大内存，能够容纳更多资料软件。

200

Apple A10芯片 视网膜...

A10仿生处理器，视网膜大屏幕。该系列搭载超强芯片，疾速散热操作性能稳定，高速网游拒绝卡顿现象。搭配护眼大屏幕，有力助于保护视觉，全天观看对眼睛没有压抑感，给人视觉带来流畅舒适。

193

联想 轻薄设计 办公本

时尚轻薄设计，防窥视屏幕。这款办公本专为办公人群设计，轻薄机身，薄至16.9mm，轻至1.3KG，携带轻巧外出方便；防窥视屏幕，防止窥视，让你有好的隐私感。

156

戴尔 28核16G 图形工作站

设计美观，性能强悍。【技术】内

联想 固态硬盘

此款硬盘为联想的SL700，轻便简

以上应用例子的共同特征

- 给定需求(或者是对象)，从信息库中找出与之最匹配的信息(或对象)
  - Google的例子：需求 “现代信息检索”
  - 京东的例子：对象 “近期搜索/浏览过的商品”

- 数据形式
  - 无固定结构的自由文本（谷歌搜索）
  - 结构化数据（京东商品）

信息检索(Information Retrieval, IR)

- 给定用户需求返回满足该需求信息的一门学科。通常涉及信息的获取、存储、组织和访问。
- 从大规模非结构化数据(通常是文本)的集合(通常保存在计算机上)中找出满足用户信息需求的资料(通常是文档)的过程。
- “找对象”的学科，即定义并计算某种匹配“相似度”的学科。

信息检索与其他学科领域的关系(非严格)

- 自然语言处理(Natural Language Processing, NLP)----对文本进行浅层、深层处理的学科(也称计算语言学)
- 数据挖掘(Data Mining, DM)----对结构化和非结构化信息进行分类、聚类、预测等分析处理的学科
- 机器学习(Machine Learning, ML)----从数据中学习到知识或规律的学科
- ……

信息检索技术的应用

舆情分析

Google 谷歌

Baidu 百度

搜索

amazon.com

Alibaba.com Global trade starts here.™

推荐

IR技术

情报处理

facebook

Tencent 腾讯

挖掘

内容安全

信息检索应用系统

- 搜索系统
  - Web搜索引擎(如Google)
  - IBM Waston问答系统、微软小冰
  - ……
- 推荐系统
  - 淘宝网
  - 豆瓣网
  - 微博推荐、好友推荐
  - ……

从信息规模上分类

- 个人信息检索：个人相关信息的组织、整理、搜索等。桌面搜索(Desktop Search)、个人信息管理(PIM = Personal Information Management)、个人数字记忆(Personal Digital Memory)

- 企业级信息检索：在企业内容文档的组织、管理、搜索等。企业级信息检索是内容管理(Content Management)的重要组成部分。局域网/内网搜索

- Web信息检索：在超大规模数据集上的检索。

提纲

① 什么是信息检索？

② 为什么要学习信息检索？

③ 课程情况

市场发展的需求

- 用户(国家、企业、个人等)需要信息检索技术：互联网的信息量太大、噪音太多，寻找所需要的信息非常不容易
- 公司需要信息检索技术：
  - 搜索引擎改变了很多传统的生活方式，Yahoo、Google、Baidu，还有一些公司如Microsoft、Sina、Sohu、Tencent、Netease、360、Facebook都加入到搜索技术的竞争。
  - 互联网五大盈利模式：1、（计算）广告，搜索广告、展示类广告、开屏广告、视频（流）广告；2、商品售卖（实物：京东、淘宝；虚拟：网课、地图API）；3、平台佣金（美团、滴滴）；4、增值服务（网盘、QQ会员）；5、金融服务
  - 或多或少都依赖信息检索技术的支撑
  - 目前搜索引擎公司甚至整个互联网正常运转的计算广告的核心技术是信息检索技术

市场发展的需求

- 用户(国家、企业、个人等)需要信息检索技术：互联网的不只是搜索引擎才需要信息检索技术，电子商务(如亚马逊网站、淘宝等)、社交网(微博、Facebook、twitter、校内网)、数字图书馆、大规模数据分析(金融证券行业等)等都需要信息检索技术

- 是不是泡沫：2000年左右出现的网络泡沫和现在的互联网有什么不同，搜索引擎在其中占什么位置？
  - 缺乏稳定的盈利模式，有不少是在烧钱，经常会传出亏损的报道
  - 搜索广告是一直以来互联网公司最稳定的也是最大的收入来源。搜索广告是持续数十年非常稳定的盈利模式

几个应用需求

- 移动搜索
- 产品搜索
- 专利搜索
- 广告推荐
- 社会网络分析
- 消费行为分析
- 网络评论分析
- SEO营销
- ……

提纲

① 什么是信息检索？

② 为什么要学习信息检索？

③ 课程情况

课程的宗旨

- 信息检索的基本原理、模型和方法(含部分机器学习、自然语言处理方法)
- 信息检索系统的基本实现方法

本课程的特点

- 不是教学生学怎么使用信息检索工具(另有课程)，而是了解信息检索工具背后的基本原理和技术，并且能够进行深层的研究或开发相关的应用。知其然知其所以然。

- 掌握原理+积极讨论+广泛阅读+深入实践

授课内容简介

- 基本内容
  - 布尔检索
  - 倒排及各种索引
  - 索引构建及压缩
  - 向量检索
  - 检索评价方法
- 高级内容
  - 概率模型
  - 语言模型
  - 分类算法
  - 分布式表示
  - 机器学习在IR上的应用（Learning to rank, neural IR models）
  - WEB采集、检索及链接分析

授课内容简介(另一个角度)

- 信息检索的基本概念
- 信息检索的评价
- 信息检索模型和算法
  - 模型(布尔模型、向量模型、概率模型、语言模型)
  - 文本处理技术
  - 文本分类
  - 信息组织和索引
- 信息检索的应用
  - WEB检索

授课方案

- 讲授内容既包含传统内容，也注意吸收最新研究成果
- 学术内容和业界进展相结合
- 既考虑一般学生普及入门的需求，也考虑相关专业学生更高的要求
- 与往年相比
  - 授课内容更新较多
  - 速度加快

课程基础

- 数学基础
  - 概率统计
  - 线性代数
- 计算机基础
  - 算法和数据结构
  - 程序设计

考核方式

- 作业+期末考试
  - 若干个小作业 20%
  - 大作业(课程项目)—30%
  - 期末考试 50%

国际著名研究机构和代表人物

- 美国康奈尔大学 Salton (1927-1995)
  - 现代信息检索的奠基人，倡导向量空间模型
  - SMART的完成人
  - 第一任Salton奖(1983年)得主，ACM Fellow

- 英国剑桥大学 Sparck Jones (1935-2007)
  - 概率检索模型的提出者之一
  - NLP和IR中的杰出先驱
  - 曾获ACL终身成就奖和1988年Salton奖

国际著名研究机构和代表人物

- 微软英国剑桥研究院、伦敦城市大学 Robertson，ACM Fellow
  - 概率检索模型的先驱和倡导者
  - 开发了OKAPI检索系统
  - 2000年Salton奖得主

- 美国 UMass CIIR W. B. Croft，ACM Fellow
  - 基于统计语言建模IR模型的提出者和倡导者
  - 和CMU共同开发了Lemur工具
  - 2003年Salton奖得主

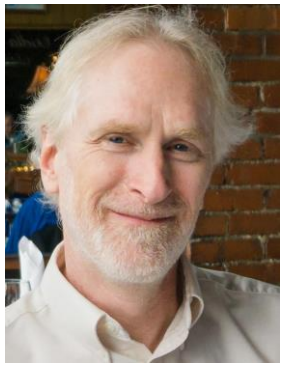
国际著名研究机构和代表人物

- 英国Glasgow大学 Rijsbergen，ACM Fellow
  - 信息检索逻辑推理学派的提出者和倡导者
  - 现在试图用量子理论解决IR问题
  - 2006年Salton奖得主

- 微软美国研究院 Susan Dumais
  - 隐性语义索引LSI的提出者
  - 2009年Salton奖得主

国际著名研究机构和代表人物

- 美国CMU大学 Jamie Callan教授
  - 早期致力于分布式检索研究
  - 现在试图用深度学习技术解决IR问题



- 中科院计算所郭嘉丰研究员
  - 基于深度学习技术的检索模型研究
  - 曾获得SIGIR / CIKM Best Paper Award

国际著名研究机构和代表人物

- 加拿大Waterloo大学 Jimmy Lin教授
  - 面向大规模预训练语言模型的排序算法
  - 开源工具Anserini作者

- 美国UIUC Chengxiang Zhai教授
  - 基于语言模型的排序算法的奠基人之一
  - 2021年获Salton 奖

重要会议

- 国际会议：
  - 信息检索：SIGIR，CIKM, WWW, WSDM…
  - 自然语言处理：ACL, EMNLP, COLING, NAACL…
  - 机器学习/人工智能：ICML, AAAI, IJCAI…
  - TREC、NTCIR评测会议
- 国内会议：
  - 全国自然语言处理会议 (NLPCC， 1年一届)
    - 计算机学会举办
  - 全国信息检索学术会议(CCIR, 1年一届)
  - 全国计算语言学会议(CCL, 1年一届)
    - 以上两个会议由中文信息学会举办

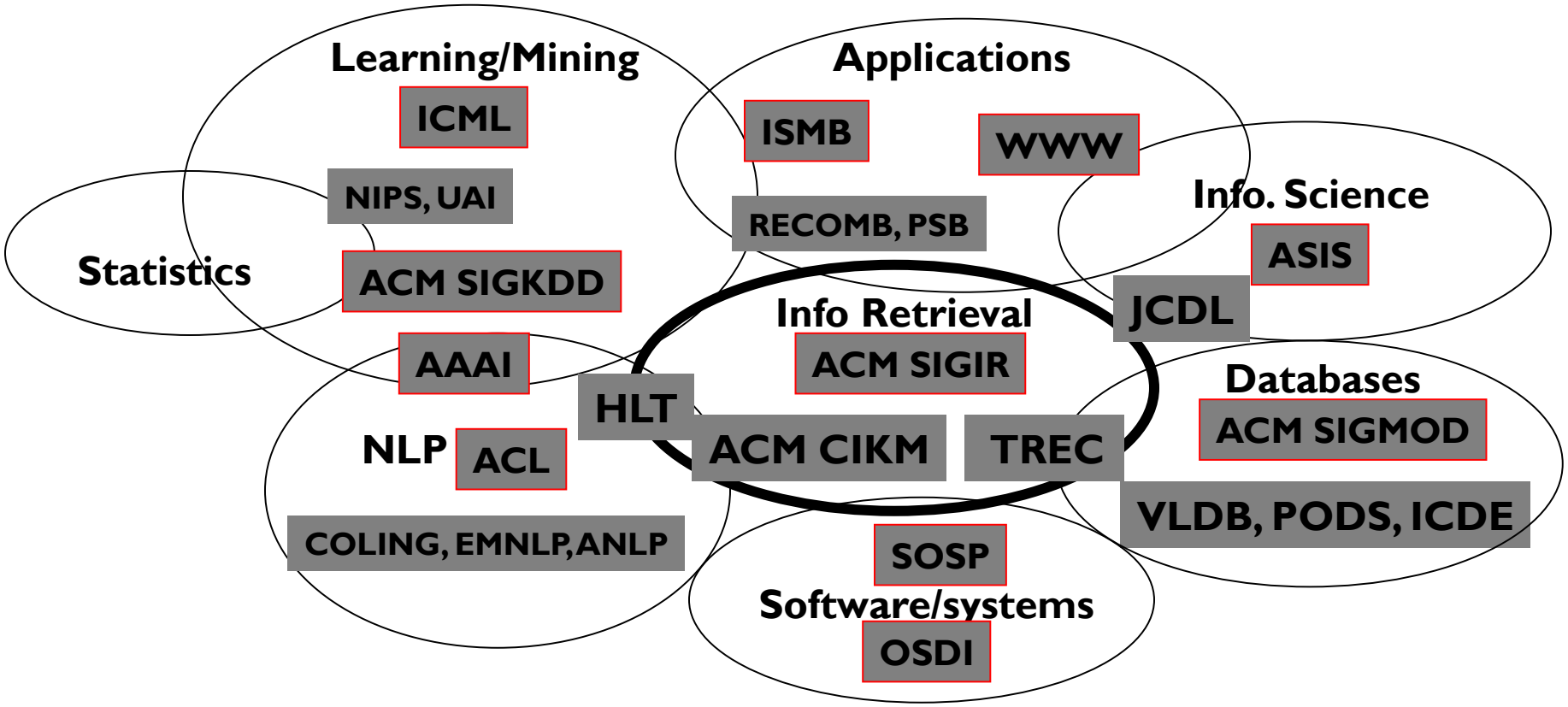
ACM SIGIR

- ACM：美国计算机学会
- SIGIR：special interest group on information retrieval
- ACM SIGIR Conference：IR领域的最重要会议，起始于1978年，2014年是第37届。
- 会议地点在美洲、欧洲和亚太三个地区轮换。
  - 2015，智利圣地亚哥　2016，意大利比萨
  - 2017，日本东京　2018，美国 Ann Arbor Michigan
  - 2019，法国巴黎　2020，中国西安
  - 2021，加拿大蒙特利尔　2022，西班牙马德里

SIGIR关于举办地原则的描述：

1. North and South America in 2021, 2024, 2027, and so on
2. Europe and Africa in 2022, 2025, 2028, and so on
3. Asia and Australia in 2023, 2026, 2029, and so on
4. SIGIR has no plans to include Antarctica in the rotation.

IR及相关研究领域重要会议



*本页借用了Chengxiang Zhai的课件

重要期刊

- 国际：
  - ACM Transactions on Information Systems (TOIS)
  - ACM Transactions on Asian Language Information Processing (TALIP)
  - Information Processing & Management (IP&M)
  - Information Retrieval
  - JASIST (美国情报学会会刊)
  - ……
- 国内
  - 中文信息学报
  - 计算机学报/软件学报/计算机研究与发展
  - ……

信息检索学科的特点

- 应用性
  - 目标非常实际，例如提升网络搜索引擎返回结果准确率、商品推荐转化率
- 经验性
  - 理论上漂亮的方法并不一定有用
  - 理论需要结合实践

重要工具

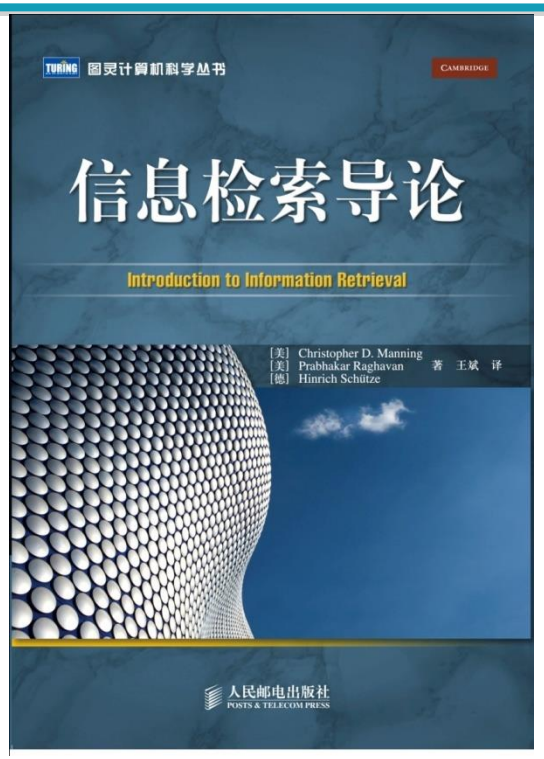
- 信息检索
  - SMART：向量空间模型工具，C编写
  - Lemur、Indri：包含各种IR模型的实验平台，C++，可以直接对TREC语料进行处理，CMU&Umass联合开发
  - Terrier：格拉斯哥大学开发的IR实验平台，除其他IR模型外，还包含该组倡导的DFR模型，Java
    - PyTerrier，PyTerrier_BERT：Terrier的Python版本，后者整合了近期提出的基于BERT的排序模型
- Anserini：标准语料实验工具，基于Python，强调“一键复现”
- 深度学习
  - TensorFlow：Google发布的深度学习开源工具平台
  - Theano：蒙特利尔大学开发的基于Python的深度学习工具
  - Keras：由Google工程师François Chollet将TensorFlow / Theano作为Backend的集成工具，近期微软也开发了Keras的Backend工具CNTK
  - Pytorch：Facebook发布的另一个基于Python的深度学习工具

重要检索工具平台

- Lucene，检索工具，Java版是维护版本，存在其他各种版本，主要是向量空间模型
- ElasticSearch：基于Lucene的搜索服务器，用Java开发，并作为Apache许可条款下的开放源码发布，是企业级搜索引擎
- Sphinx，C++检索工具，实现了BM25概率模型，和MySQL集成较好，据说不要定制
- Xapian， C++检索工具，实现了BM25概率模型，据说易定制
- Nutch，开源爬虫+Lucene
- Larbin：采集工具，C++
- Mahout：分布式数据挖掘平台 Java
- 更多：http://www.searchtools.com/tools/tools-opensource.html

教材

- 注意最好选(如果可选的话)最近一次(目前是**第五版**)印刷的版本
- 豆瓣上评价9.2分
- 网上有英文电子版(可供对照阅读)
- 勘误表：http://www.ituring.com.cn/book/127
- 但本课程内容与原书已有较大规模的更新



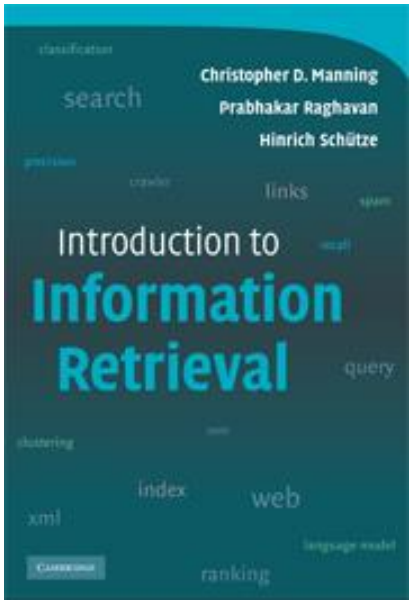
信息检索导论

原作名: Introduction to Information Retrieval，1E
作者: Christopher D.Manning / Hinrich Schütze / Prabhakar Raghavan
译者: 王斌
出版社: 人民邮电出版社
出版年: 201008
页数: 388
定价: 69.00元
装帧: 平装
ISBN: 9787115234247

9.2
(94人评价)
68.1%
27.7%
4.3%
0.0%
0.0%

原版

- Stanford大学信息检索课程教材Introduction to Information Retrieval，原书在Amazon信息检索类排名第一(2008年7月出版)。作者Chris Manning (Stanford大学教授、ACM Fellow)、Prabhakar Raghavan(前Yahoo!研究院院长，现Google)、 Hinrich Schütze(斯图加特大学教授)等。
  - 内容相对较新
  - 例子多
  - 有关NLP和分类聚类的内容较丰富
  - 有相关算法的介绍
  - 有系统实现相关的内容
  - 但是Stanford的教学内容已经有较大更新
    - 本课程也做了相应更新



参考书籍及文献

- Stanford课程网站 https://web.stanford.edu/class/cs276/
- Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press 2008 Electronic version (draft) can be downloaded from http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html
- B. Croft, D. Metzler, T. Strohman，Search Engine: Information Retrieval in Practice, Pearson Education, 2009 (国内机械工业出版社出版的影印版和哈工大刘挺等老师翻译的中文版)
- Baeza-Yates, R. & B. Ribeiro-Neto. eds. Modern Information Retrieval. ACM Press, 1999 (目前已出第二版，复旦黄萱菁等老师翻译的中文版)
- 李晓明，闫宏飞，王继民著，搜索引擎--原理、技术与系统，北京：科学出版社，2005
- Witten, Ian et al. Managing Gigabytes. Orlando, FL: Morgan Kaufmann Publishers Incorporated, 1999 (国内有清华梁斌的翻译版)
- William Frakes & Ricardo Baeza-Yates, Information Retrieval Data Structures and Algorithms. PrenticeHall, 1992
- Karen Sparck Jones & Peter Willet eds. Readings in Information Retrieval, Morgan Kaufmann, 1997
- 刘挺等著，信息检索系统导论，机械工业出版社，2008
- SIGIR/WWW/SIKDD/TREC/CIKM/ Proceedings
- More resources see: http://nlp.stanford.edu/IR-book/information-retrieval.html