信息检索与Web搜索

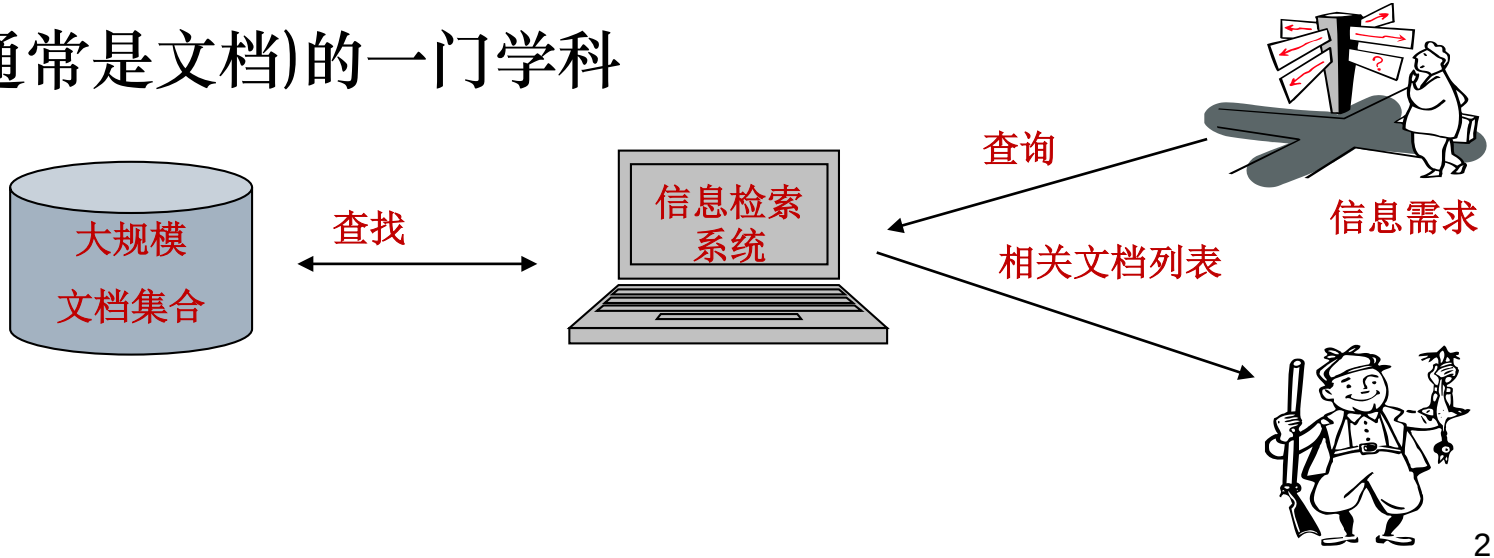
第1讲 概述

授课人：高曙明

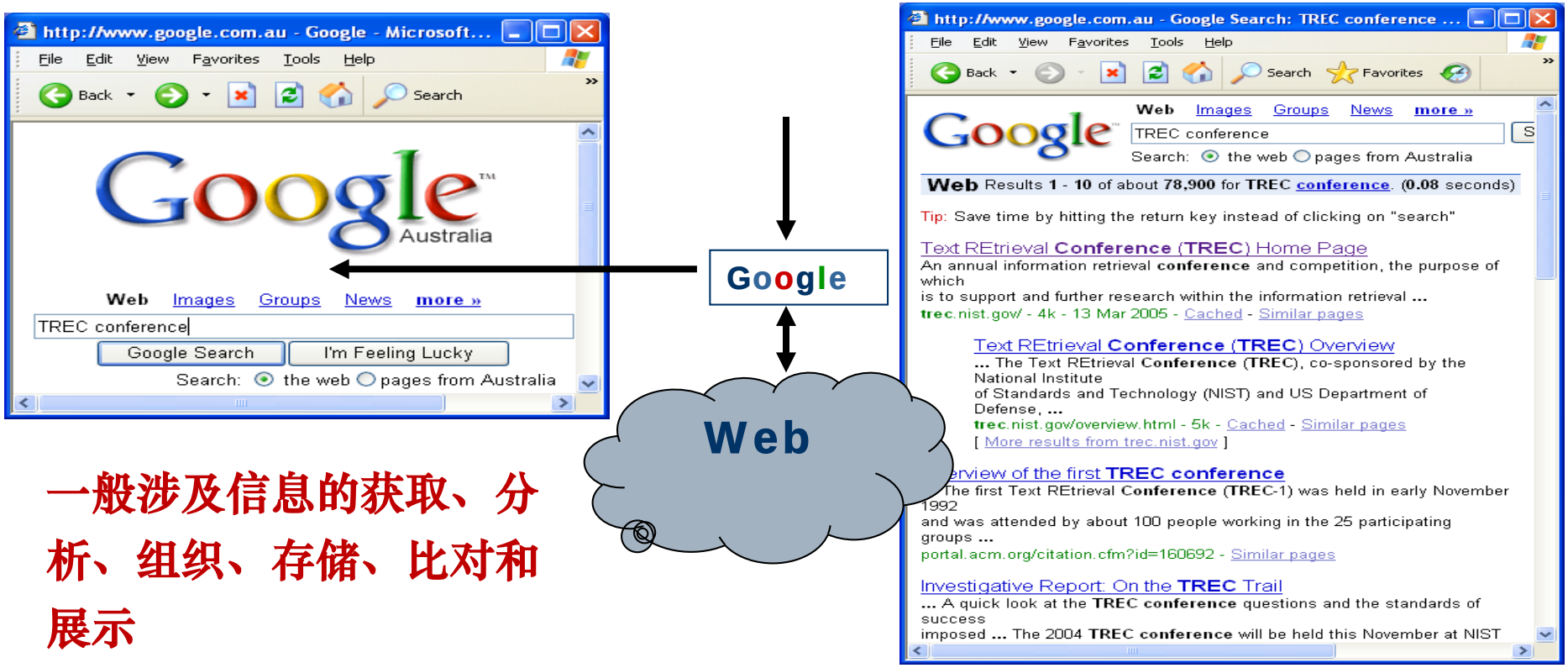
*改编自“现代信息检索”网上公开课件（http://ir.ict.ac.cn/~wangbin）

信息检索概念

- 从大规模的具有非结构化特性(通常是文本)的资料集合(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)的一门学科



信息检索概念

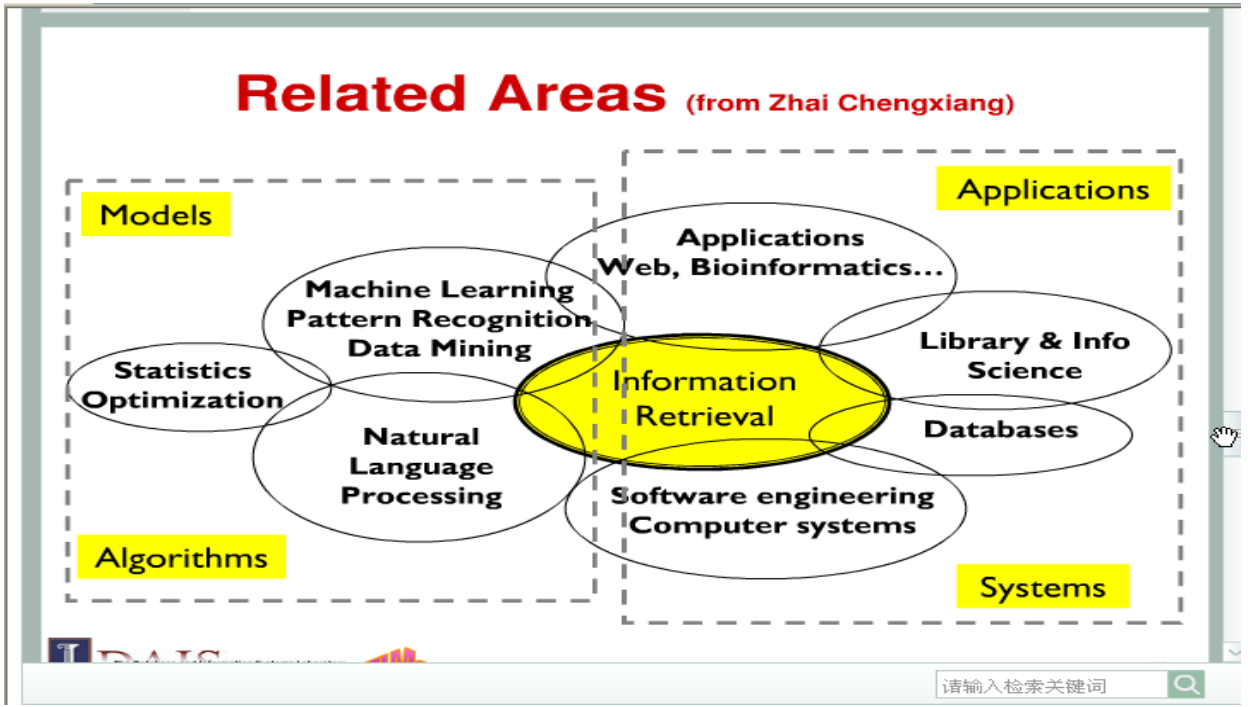


一般涉及信息的获取、分析、组织、存储、比对和展示

信息检索概念

- **文档（Document）：**指以文本内容为主的信息源，如纯文本、网页、邮件、论文、专利、图书等
- **非结构化文档：**指没有清晰和明显结构的文档，主要是纯文本
- **半结构化文档：**指带有简单结构表示的文档，如网页 <title>李甲主页</title>
  <body>…</body> …
- **结构化文档：**指带有复杂结构表示的文档，如XML文档

信息检索 VS.相关学科


Related Areas (from Zhai Chengxiang)
Models
Applications
Applications
Web, Bioinformatics...
Machine Learning
Pattern Recognition
Data Mining
Statistics
Optimization
Information
Retrieval
Library & Info
Science
Databases
Natural
Language
Processing
Software engineering
Computer systems
Algorithms
Systems
请输入检索关键词

信息检索 VS. 关系数据库

- IR系统主要用于查询文档
- RDB系统主要用于查询结构化数据，即记录集合，这些记录中包含预先定义的语义属性及属性值，如一本书的作者、标题、出版年份等

	RDB搜索	非结构化检索	结构化检索
对象	记录	非结构化文档	以文本为叶节点的树
模型	关系模型	向量空间或其他	?
主要数据结构	表格	倒排索引	?
查询语言	SQL查询	自由文本查询	?

基于规模的信息检索分类

- **个人信息检索**：个人相关文档的搜索，如桌面搜索(Desktop Search)，属小规模
- **企业级信息检索**：企业内部文档的搜索，行业文档的搜索等，属中大规模
- **Web信息检索**：数万亿网页的搜索，属超大规模。

信息检索技术的重要性

- **用户需要信息检索技术：**信息时代的信息量爆炸式增长、噪音太多，寻找所需要的信息非常不容易
  - 使用搜索引擎寻找所需要的信息已经成为很多人的日常行为
  - 使用专业信息检索系统（如专利、法律条文、科技论文等检索系统）则是专业人员的经常行为
- **内容提供者需要信息检索技术**
- **目前的搜索引擎和专业信息检索系统还不尽如人意**

信息检索技术的重要性

- **公司需要信息检索技术**
  - 搜索引擎公司需要：Yahoo、Google、Baidu，还有Microsoft、Sina、Sohu、Tencent、Netease、360、Facebook等也都加入到搜索技术的竞争
  - 电子商务(如亚马逊网站、阿里巴巴)、社交网(微博、Facebook、twitter、校内网)、数字图书馆、大规模数据分析等都需要信息检索技术
- **搜索是未来操作系统的重要组成部分**

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

内容安全

facebook Tencent 腾讯

挖掘

信息检索技术的发展历史

- **1960-70's:**
  - 开始探索使用计算机为一些小规模科技、法律和商业文献的摘要建立文本检索系统
  - 形成最基本的概念、模型和算法
  - Salton教授是奠基人
- **1980's:**
  - 由公司主导开发大规模文档数据库系统，如Lexis-Nexis, Dialog, MEDLINE

信息检索技术的发展历史

1990's:

- **第一个网络搜索工具**：1990年加拿大McGill大学开发的FTP搜索工具Archie
- **第一个WEB搜索引擎**：1994年美国CMU开发的Lycos
- **Yahoo搜索引擎**：1995斯坦福大学博士生开发
- **开始进行IR软件评测**：NIST TREC
- **推荐系统的出现**：Ringo, Amazon

信息检索技术的发展历史

2000's:

- **Google搜索引擎**：斯坦福大学博士生开发，采用链接分析技术
- **跨语言IR**: DARPA Tides
- **移动搜索**：语音搜索、位置搜索
- **个性化搜索**：自动分析用户所在的位置、身份、习惯、当前行为、个人偏好等个人信息，并利用这些信息提供个性化搜索结果

信息检索技术的发展历史

2010's:

- **基于语义的智能搜索**：一切皆可搜索且搜索必达
  - **问答系统**：IBM Waston、微软小冰、百度小度
  - **知识图谱的研发和使用**：支持用户按语义对象而不是字符串检索，实现在语义层面上进行信息检索。应用案例：梁思成的儿子是谁；梁思成是谁的儿子；谁是梁思成的父母
  - **深度学习与IR的结合**：基于深度学习确定合理的权重因子；基于深度学习确定语义特征；基于深度学习进行文档分类

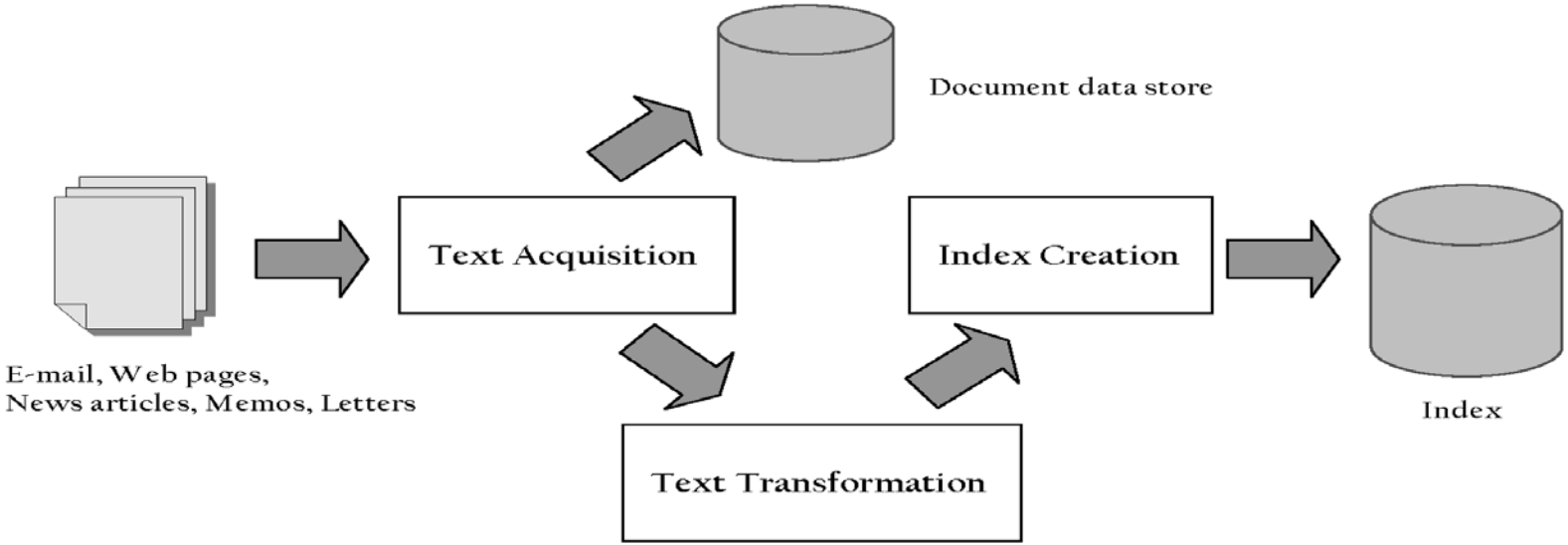
信息检索技术的发展历史

□ 未来：

更准、更快、更大规模、更全、
用户体验更好

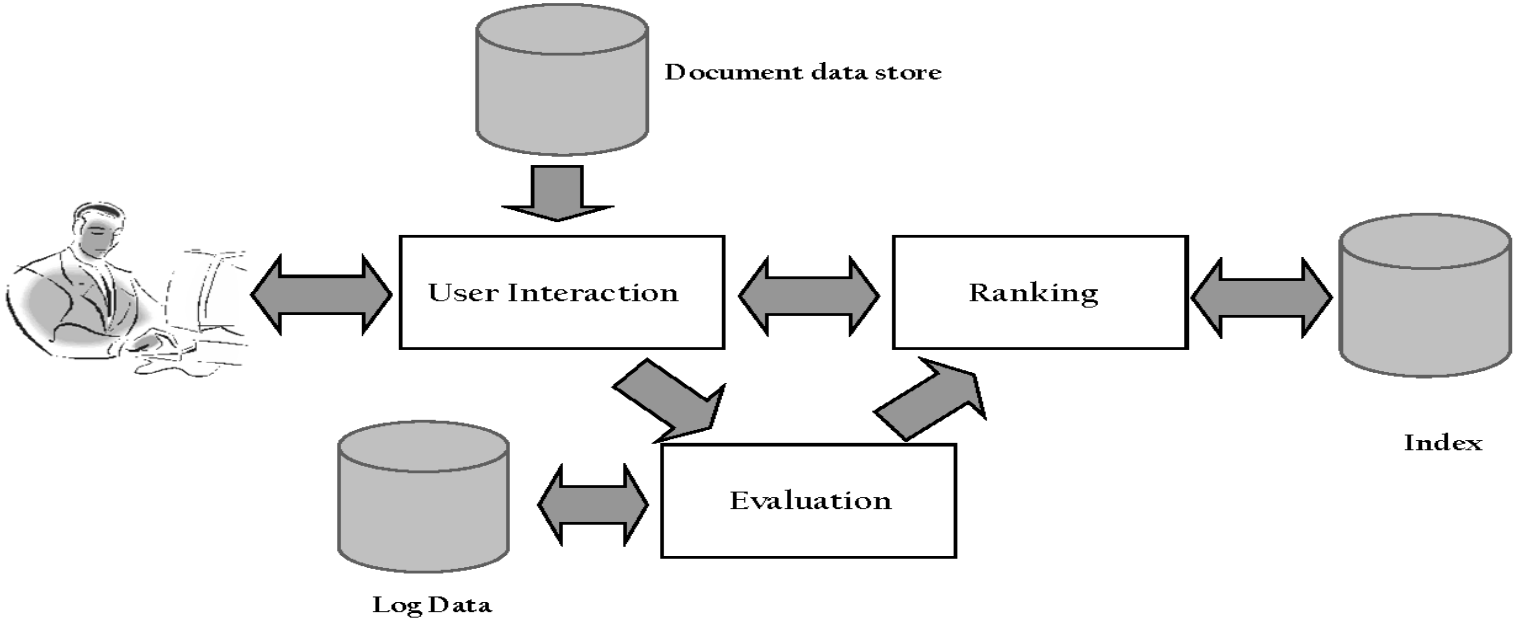
信息检索的基本内容

- 信息检索原理图



信息检索的基本内容

- 信息检索原理图



信息检索的基本内容

- 文档采集
  - **功能**：自动获取有用的文档，用于建立文档库
  - **主要内容**：Web采集器（web crawler）
- 文本分析
  - **功能**：文档预处理，将文档转化成索引词项或特征
  - **主要内容**：词条化、去除停用词、词项归一化、词形还原和词干归并、链接分析等

信息检索的基本内容

□ 索引构建

- **功能**：创建索引数据结构，用于支持快速搜索
- **主要内容**：倒排索引、词典索引、基于块排序的索引构建、单遍内存式扫描构建、分布式(MapReduce)及动态索引构建

信息检索的基本内容

□ 索引压缩

- **功能**：对索引数据结构进行压缩表示，用于节省磁盘空间，提高检索系统效率
- **主要内容**：词项的统计特性(Heaps定律、Zipf定律)、词典的压缩、倒排记录表的压缩

信息检索的基本内容

□ 检索模型与排序算法

- **功能**：用于判断查询和文档之间的关联性
- **主要内容**：布尔检索模型、向量空间模型、概率检索模型、语言模型、TF-IDF词项权重计算机制以及基于TF-IDF 的文档排序算法、概率排序原理、PageRank算法、HITS算法、基于向量空间模型的XML文档排序算法

信息检索的基本内容

用户交互

- **功能**：支持用户创建和精化查询，允许容错查询，支持检索结果的展示
- **主要内容**：查询输入、查询变换、相关反馈和伪相关反馈、查询扩展及重构、检索结果展示等

信息检索的基本内容

□ 检索评价

- **功能**：对检索系统的效果和效率进行评价
- **主要内容**：正确率、召回率、正确率-召回率曲线、标准测试集及评测会议、用户体验及结果摘要等

课程目标

- 通过本课程的学习，使同学们能够**掌握信息检索和Web搜索的基本思想和基础知识**，包括基本的概念、原理、模型和算法，并具备一定的信息检索系统和搜索引擎研发能力

- 不是教同学们怎么使用信息检索工具，而是了解信息检索工具背后的基本原理和技术，**为今后能够从事与信息检索和Web搜索相关的研发工作打好基础**

老师介绍

- **主讲高曙明：** 浙江大学应用数学系博士毕业，教授，博士生导师。现为浙江大学CAD&CG国家重点实验室CAD方向学术带头人
  - 电话：88206081-514，13858053730
  - Email：smgao@cad.zju.edu.cn
  - 办公地点：紫金港校区图书信息B楼525室
  - 个人主页：http://mypage.zju.edu.cn/smgao
- **助教：** 陈浩，浙大计算机学院博士研究生
  - 983702292@qq.com，15850750168

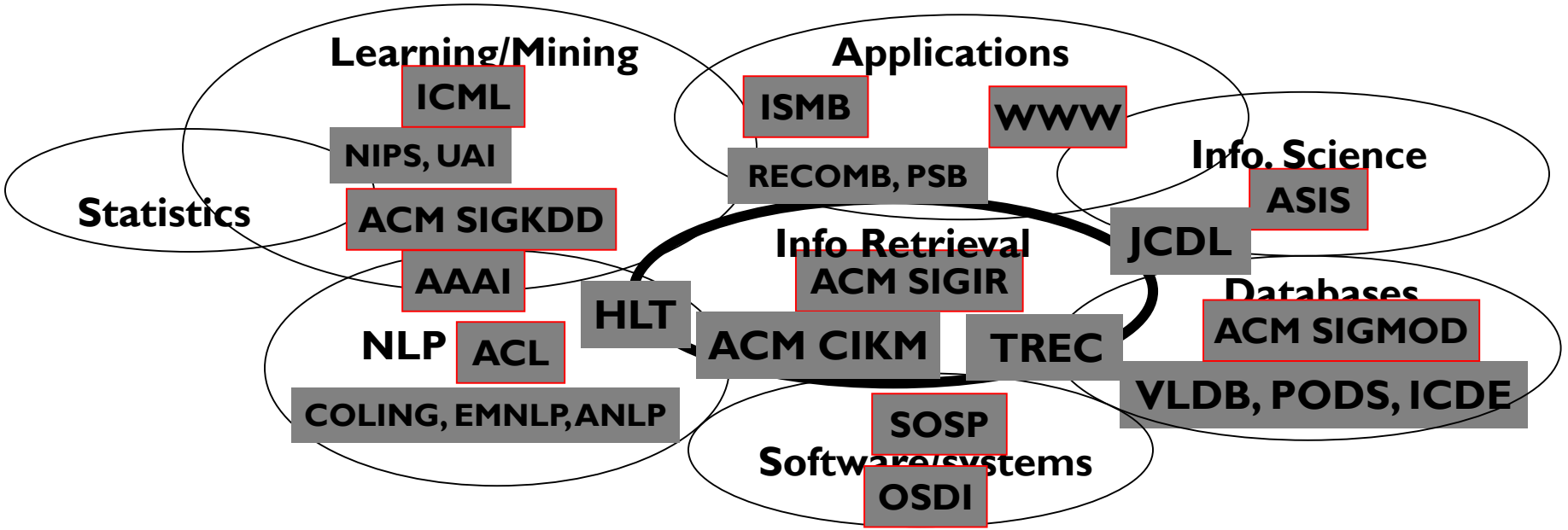
课程基础

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

考核方式

- 平时作业+大程展示
  - 不定期考勤和课堂发言　10%
  - 若干小作业　30%
  - 简单搜索引擎开发（小组大程项目）　60%
  - 具体参考课程网站
    http://10.76.3.31

IR相关的重要会议



*本页借用了Chengxiang Zhai的课件

ACM SIGIR

- ACM：美国计算机学会
- SIGIR：special interest group on information retrieval，特定兴趣组
- ACM SIGIR Conference：IR领域的最重要会议，起始于1971年，2014年是第37届。

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
- **国内:**
  - 中文信息学报
  - 情报学报

重要工具

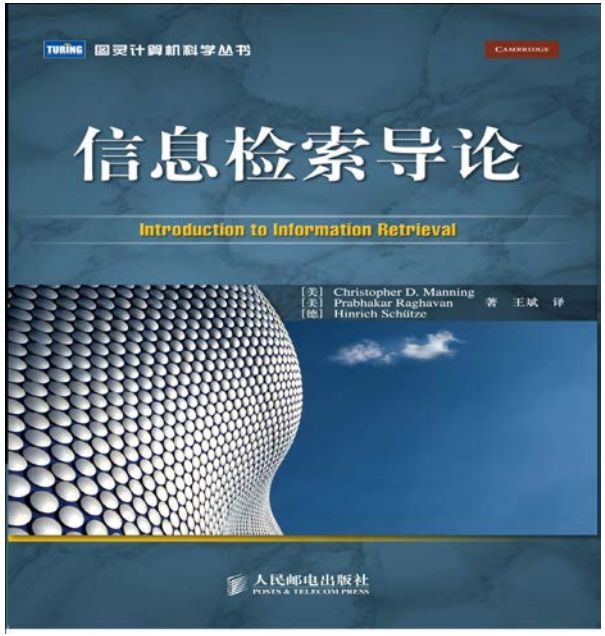
- SMART：向量空间模型工具, C编写
- Lemur、Indri：包含各种IR模型的实验平台，C++，可以直接对TREC语料进行处理，CMU&Umass联合开发
- Terrier：格拉斯哥大学开发的IR实验平台，除其他IR模型外，还包含该组倡导的DFR模型，Java
- Lucene：开源检索工具，Java版本受维护，存在其他各种版本，主要是向量空间模型
- Sphinx： C++检索工具，实现了BM25概率模型，和MySQL集成较好，据说不要定制

重要工具

- **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

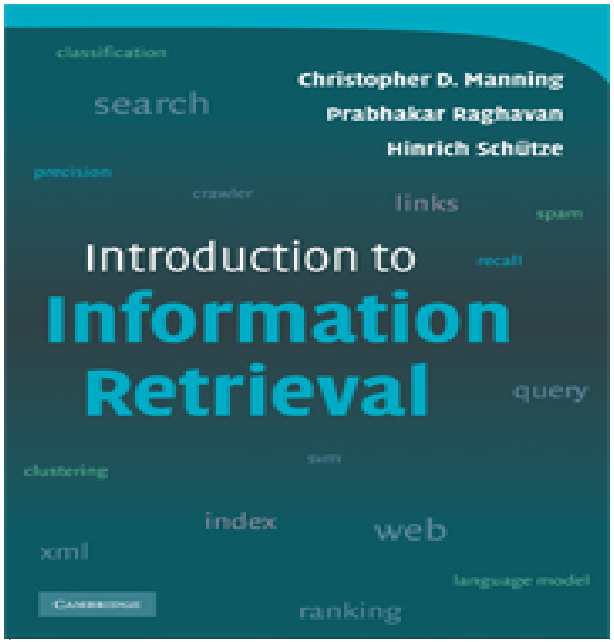
- 最好选择最近一次印刷的版本（目前是第五次印刷）
- 网上有英文电子版(可供对照阅读) http://nlp.stanford.edu/IR-book/
- 勘误表: http://www.ituring.com.cn/book/127



原书

- **Stanford大学信息检索课程教材：**2008年7月出版，在Amazon信息检索类排名第一，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)

- **特色：**内容比较新、例子多、有相关最新算法的介绍、有实现相关的内容



参考书籍及文献--1

- 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

参考书籍及文献--2

- 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