目录

第一章 信息检索概述

1.1 信息的概念、特征及类型

1.1.1 信息概念

信息(information,资讯[港台地区])无时无处不在，与人类发展历史不可离，从日常生活到科学研究，人们都在自觉不自觉地利用信息。信息是普遍存在的，一切信息来源于自然界，来源于人类社会，人们的生产、生活、学习、科研以及社会活动都是信息产生的来源。

从不同的角度对信息的解释：

- 香农(信息论创始人，美国数学家)

用来清除随机事件的形式的不定性的东西，信息就是不定性减少的量，是两次不定性之差。

- 哲学家

信息是物质存在的方式和运动所蕴涵的间接存在物的标志。

- 经济学家

信息是与物质、能源相并列的客观世界的三大要素之一，是为管理和决策提供依据的有效数据。

- 图书情报学家

信息是文献、资料、情报、知识、数据、消息、新闻的总称。

- 计算机专家

信息是数据处理的最终产品，是经过收集、记录、处理、以能检索形式存储的事实或数据。

1.1.2 信息的特征

- 普遍性 事物存在就存在信息
- 客观性 信息不是虚无缥缈的东西，是现实中各种事物的状态与方式的客观反映
- 抽象性 看不见，摸不着，人们见的是信息载体，如：语言、文字、图画、符号、光盘等
- 可加工性（可处理） 一次、二次、三次信息等
- 传递性 信息在运动中产生，在传递中发挥价值
- 共享性 信息能够通过时空进行传递，可被人类共享
- 可存储性 存储在多种存储介质上
- 时效性

1.1.3 信息的类型

(1) 根据信息的性质分

◇语法信息　◇语义信息　◇语用信息

(2) 根据信息的应用部门分

◇工业信息　◇农业信息　◇政治信息

◇科技信息　◇文化信息　◇经济信息

(3) 根据信息的记录符号分

◇语声信息　◇图像信息　◇文字信息

◇数据信息

(4) 根据信息的内容分

- 主观信息

是指依据事实和分析说明个人的观点和见解。主观信息在对一个事件、论题进行评估时能提供很多有价值的信息。

- 客观信息

一般是指不加主观判断的如实反映客观的信息，全面客观地描述一个问题的各个方面，使人们对问题有一个全面的概念。

(5) 按信息的传播渠道分

◇口传(口语)信息　◇体语信息　◇实物信息

◇文献信息　◇电子信息

1.1.4 信息的作用

- 信息是人类社会生存的条件，信息是人类社会发展的资源。
- 信息是主客体的中介，信息是思维的材料。
- 信息是组织的保证，信息是管理的基础。
- 信息是决策的依据，信息是控制的灵魂。

1.1.5 信息资源的类型

(1) 按信息的性质

- 社会信息

是人类在社会实践活动中，为生存、生产和社会发展而产生、处理和利用的信息。

- 自然信息

自然界中的事物变化、特征及事物之间的内在关系的反映。如自然景观。

(2) 按信息的载体

- 印刷型信息
- 缩微型信息

- 声像型信息
- 电子型信息

(3) 按信息内容的表现形式

- 文献型信息 信息内容以语言文字存储在不同的载体上
- 数据型信息 信息内容以数据形式出现并存储在不同的载体上
- 声像型信息 信息内容以声音或图像形式出现
- 多媒体型信息 集文字、声音、图像于一体

(4) 按保密程度

- 公开信息
- 半公开信息 在一定范围内
- 非公开信息（保密信息）

(5) 按信息加工处理程序

- 零次信息

口头携带、传播的信息。如：交谈、聚会，产生于交流过程。

特点：未经过纪录加工，不便于积累检验，增加获取难度。

- 一次信息

未经过加工或粗加工的原始信息资源（原始信息），是人们在社会实践活动中直接产生或得到的各种数据、概念、知识、经验及总结。

如：著作、报纸、期刊、会议资料、研究报告、政府出版物、专利说明书、产品样本、标准文献、学位论文等。

特点：价值高、数量大，是最基本的信息，具有重要参考和使用价值。

- 二次信息（检索工具）

是以一次信息为依据加工整理而形成的信息，是对一次信息的浓缩或有序化产物。

如：目录、文摘、索引等。

特点：为查找一次信息提供线索，具有系统性、工具性特点。

- 三次信息

对零次信息、一次信息、二次信息、进行分析研究，加工提炼和概括综合而形成的信息。

如：综述、述评、进展报告、学科年度总结等。

特点：信息量大、综合性强、系统性好。

  - 综述（综合性叙述）：将大量分散的有关特定课题的文献、事实和数据进行归纳、分析、综合、筛选，以精炼的文字扼要地叙述出来，内容十分概括。撰写综述一般要求：述而不作。

    特点：客观全面地整理、分析和总结现有信息，不加评论。

  - 述评：针对某一学科或某一问题，全面系统地总结各种情况、观点和数据，并给予精辟的分析评价。

    特点：有述有评。

- 综述和述评能够帮助人们用较少的精力和较短的时间，对有关课题的内容、意义，以及历史、现状有一个简明的了解。

(6) 依据信息内容

- 经济信息
- 科技信息
- 政务信息
- 文化信息
- 教育信息
- 军事信息

(7) 按信息出版发行特点

- 正式出版信息
  - 图书：有国际标准书号 ISBN（International Standard Book Number），由 13 位数字组成。如：ISBN 978-7-300-09671-1(前缀号-地区组号-出版者号-书序号-校验码)

- 期刊（杂志）

  有国际连续出版物标准刊号 ISSN(International Standard Serial Number)，由 8 位数字组成。

- 报纸

  也属于连续出版物，具有周期性短，时效性强的特点。

- 非正式出版信息（特种文献）：不经过公开出版，不大量发行，为一部分用户使用的内部文献信息资料

  - 会议文献：国内外各种学术会议上产生的文献包括：会前文献、会中文献、回后文献

    - 会前文献：会议通知、会议议程、论文摘要
    - 会中文献：开幕词、闭幕词、讨论记录、大会提案、决议
    - 会后文献：经过整理出版的专门会刊(Transactions)、会议录(Proceedings)、会议论文(Conference Papers)等

  - 学位论文：高校和研究机构的学生为获得某种学位而撰写的科学论文。

    如：学士、硕士、博士论文

  - 政府出版物

    各国政府部门及其所属机构出版的文件

  - 研究报告

    报道研究工作和开发调查工作的成果或进展情况的一种文献

  - 档案

    具有保存价值的历史记录

  - 专利文献

    专利说明书及相关文献

  - 标准文献

    记录技术标准、管理标准和其他具有标准性质的文件

1.2 信息检索的基本概念

1.2.1 信息检索

- Information Retrieal (简称 IR)
- Information Storage and Retrieal（广义）
- 较规范、正式的学术术语，最早由美国学者穆尔斯(C.W.Mooers)于 1949 年提出并使用
- 广义理解
  - 将信息按照一定的方式组织和存储起来，并根据用户的需要找出其中相关信息的过程。(包括存、取环节)
- 狭义理解
  - 如何从存储的信息集合中快速获取各种需要的信息。(信息查询)

1.2.2 信息检索的类型

早期分类方法

- 文献检索

以文献（包括文摘、题录或全文）为检索对象的一类信息查询活动。是一种相关性检索，不直接回答用户所提的问题本身，只是提供有关的文献供参考。

- 典型的文献检索
  - 为了解某一理论、方法的具体内容或技术细节，研究人员需要查找能提供相关知识的文献；
  - 为了编写教材或撰写综述性论文，作者需要对论述相关问题的大量文献进行搜集及阅读；

- 为了审查某项专利发明的新颖性与先进性，审查员需要在规定的“新颖性调查范围”内查阅有关的专利说明书及其他资料。

- 事实检索

  针对从文献中提取出来的各种事实（或知识项）所进行的检索活动。

  - ✓ 例如
  - 查找某出版社 2008 年出版图书的信息；
  - 查找某公司在全球哪些地区设立了分公司、分公司地址、员工数、主要负责人等。

- 数据检索

  针对经过选择、整理、鉴定的各种数据信息为检索对象。

  - ✓ 例如
  - 人口、国民生产总值，建筑材料的各种性能参数等，以此作为主要检索对象的一类检索操作。

  最新分法

- 文本检索

  以各种自然语言符号系统所表示的信息作为主要检索对象的信息检索。

如：结构化书目信息检索、无结构或半结构化的自由文本检索、关键词检索、语义检索等

- 数值检索

  针对数值型数据的查询发展起来的一类较有特色的信息检索活动

- 音频、视频检索

  新兴的信息检索操作，前沿领域。

1.3 信息检索的基本原理

信息检索基本原理：对信息资源集合与信息需求集合的匹配与选择。

为了顺利实现这种匹配，两者必须依赖统一的交流“语言”，以此来描述文献信息内容的特征，同时也以此来描述用户需求特征。只有两者采用共同的“语言”，才能把文献特征的标识与需求特征的标识彼此对应，完成检索的标识匹配过程。

- 这种信息交流中沟通双方的“语言”就是检索语言

- 信息资源集合

  有关某一领域的、经选择性采集和组织加工的信息集合体。

- 信息需求集合

  当人们为完成某一任务时，经常觉得缺少的某些知识，即信息需求。

- 匹配与选择

  需要一种匹配机制。

匹配机制的主要功能：能够把信息需求集合与信息资源集合依据某种相似性标准进行比较与判断，选择出符合用户需要的信息。

1.4 信息检索的主要研究问题

研究问题涉及的理论、技术和应用

1.4.1 信息检索理论

- 检索语言
  - 分类语言
  - 主题语言
  - 引文语言

- 代码语言

- 检索模型

  是各类实用检索服务系统开发设计的基础框架

  - 已提出的模型

    - ✓ 集合论模型
    - ✓ 代数论模型
    - ✓ 概率论模型

- 标引理论

关于匹配标准的理论，即如何判定信息资源与用户提问之间的相关性，是检索系统开发设计、性能评价分析等诸多环节需要解决和明确的重要理论问题。

- 知识组织与表示理论

  - 知识表示问题是概念检索、语义检索与推理需要解决的理论问题。
  - 涉及到知识的形式化表示方法、知识单元之间的语义关联和逻辑推理等。
  - 是目前信息检索理论研究领域的前沿

1.4.2 信息检索工具/系统

- 信息检索工具/系统是由有序化的信息资源、设备、检索技术和检索方法组成的有机整体，构成了实现信息检索活动的物质基础，是信息检索的现实研究对象。

- 研究内容

  - ✓ 信息检索系统的结构；
  - ✓ 信息检索系统的功能；
  - ✓ 信息检索系统的设计开发；
  - ✓ 信息检索系统的管理运营；
  - ✓ 信息检索系统的应用评价。

1.4.3 信息资源及其收集、加工

- 主要涉及存储问题
- 目的是建立可供检索的数据库
  - 各类数据库生产技术标准的制定
  - 数据库信息组织模型
  - 数据库文档结构设计
  - 数据库的更新与维护
  - 有关新兴技术（如数据仓库）

1.4.4 检索技术与方法

- 文本检索技术
  - 比较成熟
  - 如：布尔检索、截词检索、文字检索、加权检索、聚类检索等
- 数值检索技术
- 音频、视频信息索技术
  - 新兴的研究领域
  - 例如
    - ✓ 音乐信息的旋律检索；
    - ✓ 语音信息的自动识别与检索；
    - ✓ 图象信息的颜色检索、纹理检索与形状检索；
    - ✓ 视频信息的运动目标检索、关键帧提取与镜头检索等。
- 网络搜索技术
  - 网络信息自动采集技术

- 超链接分析技术
- 搜索结果排序技术
- 元搜索技术
- 网络挖掘与个性化服务技术

1.4.5 用户研究与检索策略

- 研究用户的查询心理、检索需求、查询信息的行为特征
- 检索策略研究
  - 用户信息需求分析
  - 检索式构造
  - 相关反馈方法
  - 检索过程调整与控制
- 其他密切相关的自动化处理技术
  - 自动聚类与分类
  - 自动摘要
  - 信息可视化
  - 信息过滤
  - 信息提取
  - 机器翻译
  - 人机交互

1.4.6 相关学科几领域

信息检索是一个典型的交叉研究领域

- 计算机科学

信息检索的技术核心

- 涉及到的计算机学科基础知识
  - ✓ 程序设计语言
  - ✓ 算法与数据结构
  - ✓ 数据库原理
  - ✓ 系统分析与设计
  - ✓ 网络原理与技术

- 数学

信息检索研究的主要理论工具之一

- 数学的贡献与价值
  - ✓ 信息检索模型的创建
  - ✓ 检索算法的设计
  - ✓ 检索系统的评价分析

- 系统科学

一门具有广泛适应性及应用指导价值的学科

- 语言学与计算语言学
  - 计算语言学：由计算机科学和语言学交叉形成的计算语言学
  - 目的：建立形式化的数学模型来分析、理解人类自然语言

1.5 信息检索的发展历史

1.5.1 手工检索时期

- 1830 年，柏林科学院出版著名文摘刊物《药学总览》，标志手工信息检索活动的正式开始。
- 1876 年，美国图书馆协会(ALA)成立并召开第一届大会
- 1883 年，美国波士顿公共图书馆设立第一个专职的参考咨询职务—参考馆员。

1.5.2 计算机化检索时期

- 早期脱机批处理检索（1954-1964）
- 联机实时检索（1965-1975）
- 联机网络化和多元化信息检索（1975-1990）
- 利用数据通信网络

1.5.3 网络化检索时期

- 网络搜索引擎的兴起与发展
- 传统联机检索系统的网络化改造
- 网络化时期面临的主要研究问题
  - 超文本/超媒体技术应用
  - 多媒体信息检索
  - 自然语言理解
  - 海量信息资源的组织和检索
  - 检索可视化
  - 知识检索与语义检索

第二章 信息检索技术

2.1 信息检索基本技术

- 通用的检索功能
  - 浏览
    - 由信息工作者将各种信息按一定的方式组织起来
    - 按信息的主题、分类等方式编制成树状结构体系
    - 用户层层点击，进入不同分支查看检索结果列表
  - 简单检索
    - 利用检索词(检索式)进行检索
  - 高级检索
    - 利用检索词(检索式)进行检索

2.1.1 布尔逻辑检索

- 运用布尔逻辑运算符对检索进行逻辑组配，表达两个检索词之间的逻辑关系。
- 常用的组配符：AND、ＯＲ、ＮＯＴ

2.1.2 截词检索

- 截词符：“?”、“*”或“$”、“!”
- 是指检索者将检索词在被认为合适的地方用截词符进行截断的方法。可分为前截词、中间截词和后截词。

- 将截词符加在检索词的前后或中间，以扩大检索范围，计算机在查找过程中如遇截词符，不进行匹配对比，只要其他部位字母相同，即算命中。
- 前方截词：截词符放在词根前边

  如：“?ware”可以包含 software，hardware
- 后方截词：截词符放在词根后面

  如：“comput?”可以包含 compute,computer
- 中间截词：截词符放在检索词中间

  如：“colo?r”可以变换 colour，color

2.1.3 限定字段检索

- 指定检索词在记录中出现的字段，检索时，计算机只在限定字段内进行匹配运算，可以提高检索效率和查准率。
- 数据库中常见的字段和代码
  - 基本字段
  - 辅助字段
- 基本字段
  - TI Title 题目
  - AB Abstract 文摘
  - DE Descriptor 叙词
  - ID Identifier 标题词
- 辅助字段
  - DN Document Number 记录号
  - AU Author 作者
  - CS 作者单位

- JN Journal 期刊名称
- PY Public Year 出版年月
- CO Country 出版国
- DT Document Type 文献类型
- TR Treatment Code 文献性质
- LA Language 语种

2.1.4 限定范围检索

- 限制数字信息的检索范围
- 常用限定符
  - ":"或"-"表示包含，如 PY = 1998:2008
  - ">"表示大于，如 SA>300
  - "<"表示小于，如 SA<300
  - "="表示等于，如 SA=200
  - ">="表示大于等于，如 SA>=200
  - "<="表示小于等于，如 SA<=200
  - !:表示范围之外，如 SA>=200

2.2 信息检索策略

2.2.1 检索策略的制定

检索策略就是在分析课题内容基础上，确定检索系统、检索途径、并科学安排各词之间的位置关系、逻辑关系和查找步骤。

1) 信息需求分析

2) 选择数据库

3) 确定检索词

4) 编制检索式、执行检索

5) 调整检索式，优化策略

2.2.2 检索途径

- 内容特征途径
  - 主题途径
    - ✓ 按文献信息的内容主题进行检索的途径
    - ✓ 确定主题词、关键词、叙词或标题词
  - 分类途径
    - ✓ 按文献信息所属学科(专业)类别进行检索的途径
    - ✓ 按分类法进行分类，依据获取的分类号检索
  - 代码途径
    - ✓ 许多文献信息具有唯一或一定的代码
- 外表特征途径
  - 题名途径
    - ✓ 根据文献信息的题名查找文献
    - ✓ 题名包括
      - ■ 书刊名称
      - ■ 论文名称
      - ■ 专利名称
      - ■ 标准名称
  - 责任者途径

- ✓ 根据已知文献责任者查找文献
- ✓ 文献责任者包括
  - ■ 个人责任者
  - ■ 团体责任者
  - ■ 专利发明人
  - ■ 专利申请人

➢ 机构名称途径

- ✓ 根据机构名称检索该机构出版或发表的文献信息情况

➢ 编号途径

- ✓ 根据文献信息出版社或发布时给出的编号检索信息
- ✓ 编号包括
  - ■ 图书 ISBN 号
  - ■ 连续出版物 ISSN 号
  - ■ 专利申请号
  - ■ 专利号
  - ■ 标准编号
  - ■ 报告合同号
  - ■ 论文存取号

2.2.3 检索效果的评价

- 反映检索系统的检索性能和检索能力
- 常用的指标
  - ✓ 收录范围
  - ✓ 查全率

查全率 = (检索出相关文献量 / 系统中的相关文献量)*100%=a/(a+c)

注：参加检索的全部文献分为有关、无关和查出、未查出四种检索效果评估相关数据的关系

- ✓ 查准率

查准率 = (检索出相关文献量 / 检索出的文献总量)*100%=a/(a+b)

- ✓ 响应时间
- ✓ 输出形式

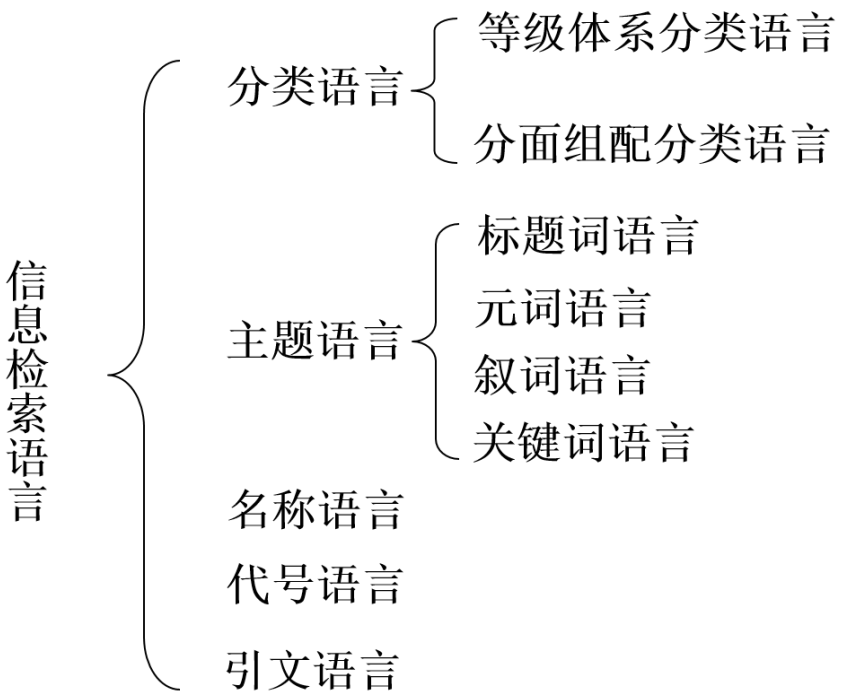
2.3 信息检索语言

2.3.1 信息检索语言

- 是从自然语言中精选出来的一整套词汇、符号，用来对文献内容和信息特征进行概括和规范的一种人工语言。它是文献信息工作者用来描述文献特征，检索者用来表达检索提问的语言，是沟通存储过程和检索过程的桥梁，是信息检索全过程得以顺利实现的保证。

2.3.2 信息检索语言分类

(1) 按结构原理划分



- 分类语言
  - 是以数字、字母为检索标识，作为有关类目的代号，便于信息存储与信息检索双方进行交流的一种检索语言。
  - 历史最为悠久，最为常见的是等级体系分类语言，至今仍然是世界上各种图书馆组织和检索藏书的主要依据。
  - 分面组配分类语言较为少见。
- 主题语言
  - 一种描述性语言。
  - 以自然语言中的名词、名词性词组或句子来描述文献所述或研究的事物概念，这些作为标识的语词按字顺(或音顺)排列并使用参照系统直接表达概念之间的关系。
- 名称语言
  - 是以人名(作者、译者、编者等)、机构名、地名、书名、刊名、篇名等能够代表信息特征的名称为检索标识，作为标引文献和检索文献双方共同采用的交流语言。

- 各种数据库中设置的作者检索途径、机构检索途径、出版物检索途径等均用名称语言对信息的特征予以描述和展开的结果。

- 代号语言

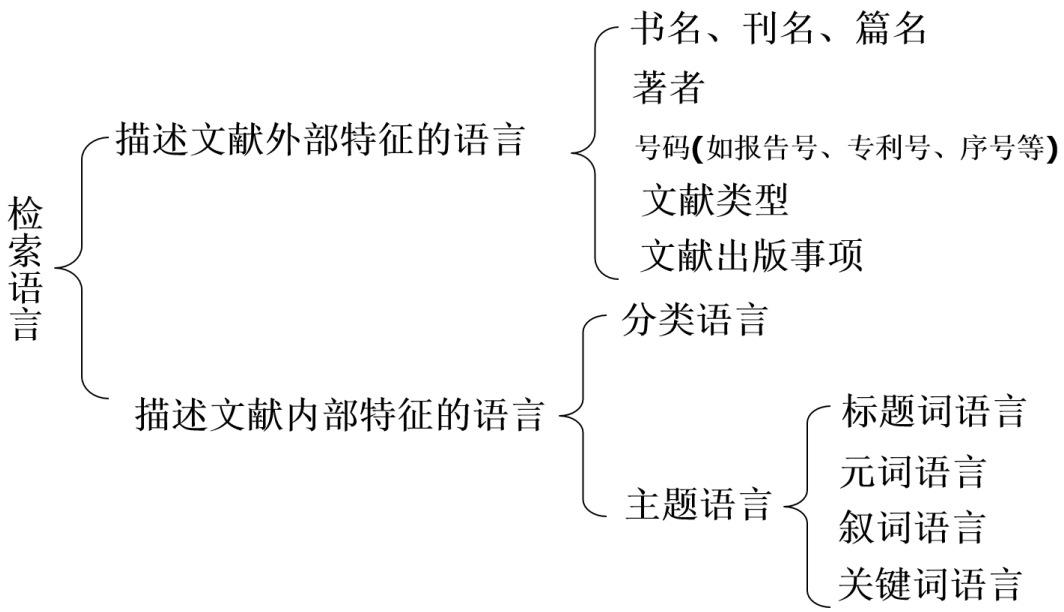
  - 以文献特有的顺序号(如专利号、标准号、化学物质登记号、合同号等)为检索标识，作为标引具有特指性序号文献和检索该类文献双方共同采用的交流语言。

- 引文语言

  - 利用文献之间的引用与被引用关系，来表达检索文献主题之间的相互关系，无需词表，也不必标引文献，检索简单有效。

  - 引文语言诞生于 20 世纪 60 年代，现广泛应用于数据库文献信息检索中。

(2) 按信息特征的描述划分



- 描述文献外部特征的语言

  - 描述文献外部特征的语言包括：

    - 书名、刊名、篇名等

    - 作者、译者、编者等

    - 号码（如报告号、专利号、序号、标准号、合同号等）

- 文献类型
- 文献出版事项

- 描述文献内部特征的语言
  - 描述文献内部特征的语言包括：
    - 分类语言（包括等级体系分类语言、分面组配分类语言）
    - 主题语言（包括标题词语言、元词语言、叙词语言、关键词语言）

2.3.3 信息检索语言的构成及其要求

(1)信息检索语言的构成

- 词表(检索语言的构成主体)
  - 规范检索语言中各个标识的概念意义及其使用，是检索语言的典据和依据。
- 从语言学角度看，信息检索语言的构成
  - 用于组成词汇的形式化符号，如字母、数字、文字。
  - 表达基本概念意义的词汇。
  - 语法，如各种标引规则、组配规则、引用次序等。

(2)对信息检索语言的要求

- 能够描述文献和提问的特征
  - 专指性
    - 基本词汇和词组有足够的专指度和语义区分能力，能够识别和区分不同的信息主题内容。
  - 唯一性

- 检索语言的基本词汇和词组，与概念意义之间应该能达到一一对应，应尽可能减少同义和多义现象，以免因表达含糊而引起标引和检索的失误。

- 灵活性
  - 检索语言的基本词汇有限，应该充分利用词汇之间的灵活组合创造出几乎无限的表达能力。

- 能够联系和沟通标引与检索
  - 易用性
  - 严谨性
  - 文献保障和用户保障

2.4 信息检索工具

2.4.1 检索工具的定义

- 检索工具就是人们用来报道、存储和查找文献信息的工具。
  - 二次、三次印刷型检索工具
  - 缩微阅读检索工具
  - 基于计算机的光盘检索系统
  - 联机检索系统
  - 基于 Internet 的网络信息检索系统
  - 搜索引擎
  - 各种信息检索系统

2.4.2 检索工具的类型

- 按照检索手段分
  - 手工检索工具

    各种类型的工具书
    - ✓ 检索类：目录、题录、文摘、索引
    - ✓ 词语类：各类字典、词典
    - ✓ 资料类：百科全书、年鉴、手册、名录、类书、政书
    - ✓ 表谱类：年表、历表、专门性历史表谱
    - ✓ 图书类：地图、历史图集、人物图录、文物图录
    - ✓ 边缘类：资料汇编、史志
  - 机械检索工具

    是手工检索工具向计算机检索系统过渡的中间检索工具，已被淘汰
  - 计算机检索工具
    - ✓ 光盘检索系统
    - ✓ 联机检索系统
    - ✓ 网络检索系统
- 按照著录形式分
  - 目录型检索工具
  - 题录型检索工具
  - 索引型检索工具
  - 文摘型检索工具
  - 全文型检索工具
- 按照载体形式分

- 书本式检索工具
- 卡片式检索工具
- 缩微式检索工具

第三章 搜索引擎

3.1 网络与网络信息资源

3.1.1 网络

网络也就是人们所称的计算机网络。是指将分散在各处，却具有独立功能的多台计算机终端及其附属设备，通过通信设备和线路联结起来，运用功能完善的通信软件按照网络协议进行数据通信，以实现资源共享的系统。

3.1.2 网络信息资源

网络信息资源是指以电子数据的形式将文字、图像、声音、动画等多种形式的信息存放在光、磁等非印刷质的载体中，并通过网络通信、计算机或终端等方式再现出来的信息资源总和。

3.1.3 网络信息资源的类型

- 按照所采用的网络传输协议划分
  - Web 信息资源(WWW 信息资源)

    采用超文本传输协议 HTTP 在 WWW 客户端和服务器端之间传输的网页
  - Telnet 信息资源

    采用远程登录协议 Telnet,在权限允许的范围内检索和使用远程计算机系统中的软硬件资源。

- FTP 信息资源

  借助文件传输协议,以文件方式在联网计算机之间传输的信息资源。

- 用户服务组信息资源

  如：新闻组、电子邮件组、专题讨论组等

- 按照网络信息资源的组织方式划分
  - 文件方式
  - 超文本/超媒体方式
  - 数据库方式

    必须利用各个数据库的专用检索系统，如《中国学术期刊全文数据库》
  - 网站

3.1.4 网络信息资源的特性

- 海量规模
- 分散无序
  - ✓ 局部有序、整体无序
- 动态更新(不稳定性)
  - ✓ 据统计，web 网络资源的每月变化量约占总量的 40%。
- 种类/形式多重多样(多媒体性)
  - ✓ 有链接、融合大量图形、图像、音频、视频等的信息。
- 非结构化或半结构化
- 语义冗余和质量控制缺乏
- 需求与使用方式个性化

3.2 网络信息检索的特点

- 1.检索范围覆盖整个 Internet
- 2.传统检索方法与全新网络检索技术相结合
- 3.用户界面友好且操作方便
- 4.用户透明度高
- 5.信息检索效率不高

3.3 网络信息检索的一般方法

- 浏览方式
  - 随意浏览
  - 分类体系浏览

  通过浏览网页资源指南的分类体系获得相关信息
- 查询方式

  通过输入检索条件

3.4 搜索引擎的概念

- 搜索引擎(Search Engine)
  - ✓ 搜索引擎是基于 Web 平台提供网络信息检索服务的工具或系统。
  - ✓ 实际是个专用的 WWW 服务器，它存有庞大的索引数据库，收集了全世界上百万甚至上千万个 WWW 主页的文字信息。
- 又称：网络搜索引擎、网络检索引擎
- 广义上是指一种基于 Internet 的信息查询系统，包括信息搜集、整理与检索。

- 搜索引擎提供给用户的结果也是文献线索（链接及其简介），只不过采用了超文本技术，单击链接即可见原文。
- Yahoo 是 Internet 上的第一个搜索引擎。是一种基于分类目录的主题指南

3.5 搜索引擎的功能

- 搜索信息建立索引数据库，并自动跟踪信息源的变动，不断地更新索引记录，定期维护数据库。
- 提供网络的导航与检索服务
- 为用户提供多种信息服务

  如：广告、免费的电子邮件、聊天室、地图等。

3.6 搜索引擎的分类

3.6.1 按检索语言的不同划分

- 基于分类类目体系的搜索引擎
  - Yahoo (http://www.yahoo.com/; http://cn.yahoo.com/)
  - 搜狐 (http://www.sohu.com/)
- 基于关键词检索的搜索引擎
  - Google (http://www.google.cn/)
  - Lycos (http://www.lycos.com/)
  - 百度 (http://www.baidu.com/)

3.6.2 根据搜索内容划分

- 综合搜索引擎 如：google
- 专业搜索引擎 如：google 网上论坛

3.6.3 根据搜索引擎的集成程度划分

- 独立搜索引擎（单一搜索引擎）

  有自己的数据库，使用自己的数据库为用户提供信息检索服务

- 元搜索引擎

是建立在多个独立搜索引擎之上的一种搜索引擎，通常没有自己数据库和搜索机器人，但集成了多个独立搜索引擎或元独立搜索引擎，其搜索结果来源于它所管理的独立搜索引擎。

3.6.4 根据搜索引擎界面语种划分

- 中文搜索引擎
- 英文搜索引擎

3.6.5 按其工作方式划分

- 全文搜索引擎（Full Text Search Engine）

全文搜索引擎是名副其实的搜索引擎，国外代表有 Google，国内则有著名的百度搜索。它们从互联网提取各个网站的信息（以网页文字为主），建立起数据库，并能检索与用户查询条件相匹配的记录，按一定的排列顺序返回结果。

- 根据搜索结果来源的不同，全文搜索引擎可分为两类
  - ✓ 一类拥有自己的检索程序（Indexer），俗称“蜘蛛”（Spider）程序或“机器人”（Robot）程序，能自建网页数据库，搜索结果直接从自身的数据库中调用。Google 和百度就属于此类；

- ✓ 另一类则是租用其他搜索引擎的数据库，并按自定的格式排列搜索结果，如 Lycos 搜索引擎。

- 目录索引

目录索引虽然有搜索功能，但在严格意义上算不上是真正的搜索引擎，仅仅是按目录分类的网站链接列表而已。用户完全可以不用进行关键词（Keywords）查询，仅靠分类目录也可找到需要的信息。

目录索引中最具代表性的搜索引擎：

- Yahoo 雅虎
- 其他著名的还有 Open Directory Project（DMOZ）
- LookSmart
- About
- 搜狐
- 新浪
- 网易

- 元搜索引擎 (META Search Engine)

元搜索引擎接受用户查询请求后，同时在多个搜索引擎上搜索，并将结果返回给用户。在搜索结果排列方面，有的直接按来源排列搜索结果，如 Dogpile；有的则按自定的规则将结果重新排列组合，如 Vivisimo。

著名的元搜索引擎：

- InfoSpace（http://www.infospace.com/）
- Dogpile（http://www.dogpile.com/）
- Vivisimo（http://vivisimo.com/）
- 万纬搜索引擎(中文元搜索引擎) http://www.widewaysearch.com/

3.7 常用的搜索引擎

3.7.1 Google

- 网址： http://www.google.cn/, 创建于 1998 年
- 搜索功能

可搜索网页、图片、资讯、地图、学术资源、图书等。

(1)基本搜索

在首页输入框中输入检索词或检索式，单击 Google 搜索按钮即可。

基本搜索主要功能特点：

- 搜索范围限定。可选择所有网页、中文网页、简体中文网页、中国的网页
- 手气不错。单击“手气不错”按钮直接进入最符合搜索条件的网站。
- 多关键词搜索。多个关键词之间是 and 关系。
- 词组搜索。用引号将搜索词括住，即可得出包含某个完整词组的结果。如：“泡沫陶瓷”。
- 限定题名与网站检索。如“intitle:西湖”，“股票 site:sina.com.cn”
- 否定字词。若搜索字词具有多种含义，可在需要排除的含义相关字词前加一个“-”号，并在减号前加一空格，以便排除这些结果。
- 按链接搜索。在搜索词前加“link:”,显示所有指向该网址得网页。如：“link:www.google.cn”
- 指定文件搜索。如：查找 word 文件，只需搜索“关键词 filetype:doc”
- 查询天气。如：天气（或 tq 或 TQ）地名（或地名的汉语拼音），tq changchun
- 查询金融信息。在检索框中输入股票或基金的名称或代码。

- 查询邮政编码区号及相关信息。输入关键词（“邮编”、yb 和 YB 任选其一，“区号”、qh 和 QH 任选其一）和要查询的城市地名或邮政编码或电话区号
- 查询手机号码归属地。输入手机号码即可。
- 查询定义。查看字词或词组的定义，输入“define 关键词”。

(2)高级搜索

单击 google 首页输入框右侧的“高级搜索”即可，进入高级搜索界面。

网页目录

除关键词搜索外，google 建立了一个网页目录。利用网页目录可以从分类途径浏览相关网站。从 google 首页的“更多”展开，选择“更多”即可进入。

其他功能

(1)网页快照

google 存储的网页，当存有网页的服务器速度较慢或出现故障时，可以使用。

(2)类似网页

利用搜索结果条目后的“类似网页”链接，可以查询与这一网页相关的网页。

(3)相关搜索

google 搜索结果清单下方，提供一些与原搜索相关的搜索词，这些相关的搜索词比原搜索词更常用，并且可能产生相关的结果。

(4)错别字改正

如检索词输入有误，系统会提示，并自动给出正确检索词。

(5)翻译功能

利用 google 的段落与网页翻译界面，可以进行翻译。

(6)计算器

在首页的输入框直接输入算术表达式，单击回车键即可。

(7)大学搜索

可以将搜索限定在某个大学的网站内。

(8)学术搜索

在 google 网页目录首页，单击“学术搜索”，进入 google 学术搜索功能，可以搜索学术研究成果。

(9)地图搜索

在 google 首页，单击“地图”，进入 google 地图搜索功能，可以查询地址、搜索周边线路。

(10)图书搜索

在 google 网页目录首页，单击“图书搜索”，进入 google 图书搜索功能，可以搜索图书全文。

3.7.2 百度

- 网址： http://www.baidu.com/ ,创建于 2000 年。全球最大的中文搜索引擎。
- 搜索功能

除提供搜索功能外，还提供新闻、网页、贴吧、知道、MP3、文档、地图、图片、视频等。

(1)基本搜索

在首页输入框中输入检索词或检索式，单击“百度一下”按钮即可。

基本搜索主要功能特点：

- 多个关键词之间是“与”的关系，没有“或”的关系。
- 可以限定题名与网站检索。

(2)高级搜索

- 主题式目录

除关键词搜索外，还提供一个主题式目录—百度网站。从百度首页的“更多”链接展开

- 其他主要功能

(1)相关搜索

(2)拼音提示

若只知道某词的发音，不知道如何写，输入查询词的汉语拼音，百度就将最符合要求的对应汉字提示出来。

(3)错别字提示

如检索词输入有误，系统会提示，并自动给出正确检索词。

(4)百度快照

百度为每一个收录的网页，存储一个纯文本的备份。当存有网页的服务器速度较慢或出现故障时，可以使用。但非文本内容无法显示。

(5)国学

提供上至先秦，下至清末历代文化典籍的检索和阅读。

(6)专业文档搜索

(7)百度常用搜索

在百度首页单击“更多”链接，再单击“常用搜索”链接即可。可以搜索天气、地图、火车、航班、电视预告、常用电话号码等。

3.7.3 Dogpile

- 网址： http://www.dogpile.com
- 元搜索引擎
- 提供以下检索
  - 网页、图片、音视频、新闻、黄页、白页
- 网页检索由 google、Yahoo! 、AskJeeves、About、LookSmart、OpenDirectory、Overture、FindWhat 提供
- 图片检索由 Yahoo! Image 和 Ditto 提供
- 音频文件检索由 Yahoo!Audio 和 SingingFish 提供

- 新闻检索由 google!新闻、Topix、Fox News、ABC News
- 支持关键词检索和目录浏览检索。
- Dogpile 提供三类检索
  - 页面检索
    - 简单检索
    - 高级检索
  - 黄页查询
    - 单击 Dogpile 首页的“Yellow Pages”链接，即可进行黄页查询。
    - 商业公司的信息查询
  - 白页查询
    - 单击 Dogpile 首页的“White Pages”链接，即可进行白页查询
    - 个人信息查询
    - 只输入姓就可以查询，也可输入人名、城市、州等信息限定检索

第四章 图书、期刊检索

4.1 图书信息检索

4.1.1 概述

- 古代的“图书”是指地图和文书档案
- 现代的图书是指以传播知识为目的，将文字、符号或图形记载于某种载体上并有一定形式的著作物

4.1.2 图书检索

- 图书信息的检索途径
  - 分类检索
  - 关键词检索
- 图书检索工具
  - 一次检索工具：可以获得图书全文信息
    - ✓ 光盘版的全文图书库
    - ✓ 网络全文数字图书网

  如：超星数字图书馆、书生数字图书馆、方正数字图书馆，Spring-Link 等
  - 二次检索工具（先获得图书概要信息，在查找原文）
    - ✓ 书目工具
    - ✓ 信息机构的图书馆藏目录数据库
    - ✓ 联机图书目录查询系统
    - ✓ 出版机构的图书书目查询系统

  如：网上书店、读者俱乐部等

4.1.3 中文图书检索

- 联机馆藏目录检索系统
  - 单一馆藏目录与查询系统
    - 如：各个高校、单位、地区的图书馆
    - 清华大学：http://www.lib.tsinghua.edu.cn/
    - 吉林大学：http://lib.jlu.edu.cn/
    - 北京大学：http://www.lib.pku.edu.cn/portal/index.jsp

- 香港大学：http://lib.hku.hk/

- 联合目录
  - 某个较大结构，与某一类相近或有共性的图书馆结合形成的统一界面检索目录
  - CALIS 联合书目
    - “211 工程”100 所高校图书馆馆藏联合目录数据库
    - 主页：http://www.calis.edu.cn/

- 电子图书检索
  - 电子书 eBOOK
  - 网上中文电子图书系统
    - 超星数字图书馆
      - 网址 http://www.ssreader.com/
      - 国家“863”计划中国数字图书馆示范工程
      - 目前最大中文在线图书馆
      - 需要安装阅读器 ssreader
      - 使用方式
        - 购买超星图书卡
        - 集体设置镜像站点
  - 书生之家数字图书馆
    - 网址 http://www.21dmedia.com/
    - “书生之家数字图书馆”由北京书生数字技术有限公司于 2000 年创办，目前可提供 23 多万种图书全文在线阅读。其中大部分为近几年出版的新书，侧重教材教参与考试类、文学艺术类、经济金融与工商管理类图书。在线阅读“书生之家”电子图书全文请先下载并安装“书生阅读器”，然后返回首页，单击“登录”按钮即可在线浏览全文。

- 中国数字图书馆
  - 网址 http://www.d-library.com.cn/
- 方正中文电子书网
  - 网址 http://www.apabi.com/
- 中国电子图书网
  - 网址 http://www.cnbook.com.cn/
  - 辽宁出版集团主建，国内出版业第一家电子图书

- 网上图书检索
  - 网上书店

1. 当当网上书店
   - 网址 http://www.dangdang.com/
   - 全球最大中文网上书店
   - 由美国 IDG 集团、卢森堡剑桥集团、日本软库、中国科文公司投资
2. 800 网上书店
   - 网址 http://www.book800.com/
   - 由北京八维在线电子商务公司创办,为公众提供最快的新书资讯，将全国 500 多家出版社每年出版的 10 多万种图书，以最快的速度上网，为各地读者提供 12 大类、3 个层次共计 700 多个子类 20 多万种图书的基本信息。
3. 亚马逊网上书店
   - 网址 http://www.amazon.com/
   - 1995 年 7 月开办，总部在美国华盛顿州的西雅图市
   - 综合销售
4. 中国图书网

- 网址 http://www.bookschina.com/
- 中国图书网创建于 1998 年，有 11 年的历史，是国内最早的网上图书销售平台之一。是国内图书品种最全的网上书店。
- 中国图书网是北京英典电子商务有限责任公司的主要网站

5. 亚马逊网上书店
   - 网址 http://www.amazon.com/
   - 1995 年 7 月开办，总部在美国华盛顿州的西雅图市
   - 综合销售
6. 华储网上书店
   - 网址 http://www.huachu.com.cn/

4.1.4 外文图书检索

- 联机馆藏目录检索系统
  - 美国国会图书馆联机目录数据库(LC)
    - 美国国会图书馆(The Library of Congress),世界上最大图书馆之一
    - 网址：http://www.loc.gov
    - 拥有藏书 1200 万册
    - 包括图书、期刊、计算机文档、手稿、音乐、录音、视频资料
    - 通过主题、著者姓名、题名、图书登记号、关键字等途径检索
  - WorldCat
    - About WorldCat
      - Worldcat 是 OCLC 公司（联机计算机图书馆中心）的在线编目联合目录，Worldcat 是世界范围图书馆和其他资料的联合编目库，同时也是世界最大的联机数目数据库。

- Worldcat 目前可以搜索 112 个国家的图书馆，包括近 9000 家图书馆的书目数据。Worldcat 可以让你搜索书籍、期刊、光盘等等的书目信息和馆藏地址。你可以搜索当地所有的图书馆，只要输入你想搜索的内容和 pin 代码即可，就会有一个列表将你所需求信息显示出来。
- 人们不愿意去图书馆的一个理由是，他们不能发现他们所要找的东西，因为里面的资料实在太庞大。要解决这个问题可以去 Worldcat，这是一个公开的图书馆搜索网站，你可以搜索世界上的图书馆所收藏的资料信息。取决于不同图书馆的规定，也许你可以直接查看一本书。如果你要常常搜索的话可以创建一个免费账户，它可以让你建立一个个人列表，写下观感和从 Amazon 上直接购买。

- WorldCat
  - 全球统一检索目录
  - 网址：http://www.worldcat.org/
  - 包含 4100 万条记录
  - 覆盖了从公元 1000 年至今的资料

- 电子图书检索
  - OCLC Net library 电子图书
    - 网址： http://www.netlibrary.com
  - SpringerLink
    - 德国施普林格科技出版集团
    - 网址 http://link.springer-ny.com/
    - 提供学术期刊及电子图书的在线服务
    - 分 11 个学科数据库，覆盖了生命科学、化学、地球科学、计算机、数学、医学、物理与天文学、工程学、环境科学、经济学、法律等。

- 可免费查阅文摘，获取全文必须是注册用户或期刊订购用户。

- John Wiley
  - 知名的学术和专业出版机构
  - 网址 http://www.interscience.com/onlinebooks
  - 提供化学、电子工程和通信、生命科学和医学、数理和统计四个领域
- Ebrary 外文电子图书数据库
  - 网址： http://www.igroup.ebrary.com
  - Ebrary 公司提供，150 多家出版社的电子版图书
  - 覆盖经济、计算机、技术工作、语言文字、社会科学、医学、科技、哲学等领域
  - 大部分内容是近 3 年的较新作品
- 世纪顶新外文数字图书馆
  - 网址： http://www.3xebooks.com/help/faq/jsp
  - 世纪顶新教育集团于 2003 年开创
  - 中国第一家原版引进外文图书的数字图书馆

网上图书检索

- 国外大出版商自己的网站
  - Wiley 约翰·威利父子公司
    - 网址 http://www.Wiley.com/
  - Macmillan 麦克米伦计算机出版社
    - 网址 http://www.mcp.com/
    - 计算机方面的书籍
  - Prentice Hall

- 网址 http://www.prentice.hall.com/
- 大学书籍

- Springer-Verlag
  - 网址 http://www.springer.com/
  - 科技类图书

- 网上书店
  - 巴诺网上书店
    - 网址 http://www.Bn.com/
    - 主要销售图书、音乐作品、软件、杂志、印刷品等
    - 现货图书 75 万种之多
  - 贝塔斯曼在线（中国）
    - 网址 http://www.Boclchina.com/
    - 德国在线书店，全球联网
  - 鲍德斯网上书店
    - 网址 http://www.borders.com/
    - 成立于美国密歇根州
    - 国际上著名的精品书店
  - 沃兹沃思网上书店
    - 网址 http://www.wordsworth.com/
    - 可以获得任一本在美国出版过的图书，也可检索任一本已绝版的图书信息

4.2 期刊文献信息检索

4.2.1 概述

- 印刷型期刊（杂志）
  - 定期或不定期出版的有固定名称的连续型出版物
  - 以纸张为载体
  - 有连续的卷、期或年、月顺序号
- 电子期刊
  - 以数字形式出版发行
  - 存储在光、磁介质上（目前以数据库中存储居多）
- 核心期刊
  - 某学科学术论文较多的、受读者重视的、能反映该学科当前研究状态的、最为活跃的期刊。
  - 集中某学科的大部分重要文献，能反映某学科当前的研究状况和发展方向
  - 学术性强，研究成果新颖、专题集中、系统
  - 核心期刊的作用
  - 可以为图书馆期刊采购提供依据
  - 可以为图书馆导读工作和参考咨询提供依据
  - 可以为数据库建设提供支持
  - 可以为期刊扩大影响，提高学术水平服务
  - 可以为我国学术论文统计分析提供依据
  - 可以为读者投稿提供参考
  - 核心期刊名称查阅 2004 年推出的《中文核心期刊目录总览》

- 期刊影响因子（Impact Factor --IPF）
- 期刊质量的测评重要指标
- 是从引文角度测度期刊重要性及影响的一项重要指标
- 通常表示为某种期刊中论文的平均被引用次数
- IPF=S/M
- S:源期刊(统计来源期刊)引用该刊前两年论文的次数
- M:该刊前两年发表论文的总篇数
- 影响因子越大，表明期刊被引用的程度越高，其影响力和学术作用越大

4.2.2 国内期刊论文及其检索

1.中国期刊全文数据库(CNKI)

- 中国知网 China national Knowledge Infrastructure
- 由清华大学、清华同方股份有限公司、中国学术期刊(光盘版)电子杂志社、光盘国家工程研究中心共同开发
- 目前世界上最大的连续动态更新的中国期刊全文数据库
- 包括：
  - ✓ 8700 多种重要期刊
  - ✓ 中国优秀博硕学位论文全文数据库
  - ✓ 中国重要报纸全文数据库
  - ✓ 中国重要会议论文全文数据库
  - ✓ 中国专利产品数据库
- 网址 http://www.cnki.net
- 需要用户名和密码
- 检索需要选择数据库，即可进入检索界面。

- 各个数据库的检索界面基本相同，分为初级检索、高级检索、专业检索。不同的数据库，检索项目有所不同，检索结果有所差异。
    - ✓ 期刊全文数据库提供论文全文的 CAJ 格式和 PDF 格式下载。
    - ✓ 博硕论文据库的原文下载可选择在线浏览整篇论文、整本下载、章节下载、分页下载 4 种方式。

2.中文科技期刊全文数据库

- 重庆维普咨讯有限公司信息资源系统中的数据库
- 包括 3 个数据库：
    - ✓ 中文科技期刊全文数据库
    - ✓ 中文科技期刊引文数据库
    - ✓ 外文科技期刊文摘数据库
- 网址 http://www.tydata.com
- 各个数据库的检索界面基本相同，分为分类检索、初级检索、高级检索三种方式。

3.万方数据资源系统

- 万方数据有限公司建立的信息服务平台
- 包括 5 个数据库：
    - ✓ 中国学位论文数据库
        - ✓ 收录 1980 年以来我国自然科学领域博士后、博士、硕士论文
    - ✓ 中国学术会议论文数据库
    - ✓ 企业、公司及产品数据库
    - ✓ 中国科技成果数据库
- 网址 http://www.wanfangdata.com.cn
- 检索分为简单检索、单库高级检索、跨库高级检索、跨库专业检索四种方式。

4.2.3 国外期刊论文及其检索

1.SDOS 期刊全文数据库

- 荷兰著名学术期刊出版商 Elsevier Science 公司的电子期刊
- 包括：
  - ✓ 数学、物理、生命科学、计算机科学、医学、环境科学、材料科学、航空航天、工程与能源、地球科学、天文、商业管理、社会科学等领域 1700 多种电子期刊
- 检索方式包括期刊浏览、分类浏览、快速检索

2.Springer 外文期刊数据库

- 德国著名科技出版集团
- 包含 500 余种学术期刊，其中近 400 种为英文期刊
- 包括：
  - ✓ 数学、物理、生命科学、化学、计算机科学、医学、环境科学、工程学、地球科学、天文学等领域
- 检索方式包括期刊浏览、按学科浏览、篇目检索
- 网址：http://www.springer.lib.tsinghua.edu.cn

3.美国工程索引(EI,The Engineering Index)

- 美国工程信息公司出版
- Ei 公司始建于 1884 年，作为世界领先的应用科学和工程学在线信息服务提供者，一直致力于为科学研究者和工程技术人员提供专业化、实用化的在线数据信息服务。
- 《工程索引》（The Engineering Index），简称 Ei，是世界上著名的检索工具之一，在世界的学术界、工程界、信息界中享有盛誉。它是检索工程技术领域文献的最主要的工具之一。
- EI 来源

- 全选期刊
- 选收期刊
- 扩充期刊 Ei Page One
- EI 收录期刊 3500 余种，会议录 2000 余种
- 其中 10%来自非英语国家
- 收录我国出版的期刊 300 多种

- 检索方式包括简单检索、快速检索、专家检索
- 网址：http://www.engineeringvillage2.org.cn

4.科学引文索引(SCI,Science Citation Index)

- 美国科技信息所（ISI-Institute for Scientific Information）著名的科学引文索引数据库(SCI：Science Citation Index)，历来被公认为世界范围最权威的科学技术文献的索引工具，能够提供科学技术领域最重要的研究成果。SCI 引文检索的体系更是独一无二，不仅可以从文献引证的角度评估文章的学术价值，还可以迅速方便地组建研究课题的参考文献网络。发表的学术论文被 SCI 收录或引用的数量，已被世界上许多大学作为评价学术水平的一个重要标准。
- 美国费城科学情报所出版
- 学术界公认权威的科技文献检索工具
- 包括：
  - 自然科学、工程技术、生物医学等 150 多个学科
- 检索方式包括引文检索、作者检索、主题检索、机构检索、循环检索

第五章 学术论文与学位论文的撰写

5.1 学术论文的写作

5.1.1 学术论文含义

- 学术论文也称科学论文、科研论文或研究论文
- 定义
  - 国家标准《GB7713-1987 科学技术报告、学位论文和学术论文的编写格式》
  - 某一学术课题在实验性、理论性或观测性上具有新的研究成果或创新见解和知识的科学记录；或是某种已知原理应用于实际中取得新的进展的科学总结，用以提供学术会议上宣读、交流或讨论；或在学术刊物上发表；或作其他用途的书面文件。
- 学术论文是对某一学科领域中的问题进行探讨与研究后，将研究成果总结表述而成的文章。

5.1.2 学术论文的特点

- 学术性
  - 将专门性的知识系统化，加以探讨、研究
  - 学术论文的基本要求
  - 学术性要求材料选择、用词和言语表达的专业性，推理论证的逻辑性与表达的简洁性。
- 科学性
  - 论文内容客观真实，数据准确可靠，方法切实可行，论证严谨，观点前后一致，表述全面清晰，研究成果，能够经得起实践得重复实验

- 创新性
  - 指创造性与新颖性

5.1.3 学术论文的类型

按出版形式划分：

- 期刊论文
  - 发表在科学期刊上的学术论文。
  - 篇幅大多不长，一般在 3000-5000 字，多者 6000-10000 字。选题不能太大。
- 会议论文
  - 为参加国内外的各学科专业的学术会议撰写的论文，供学术会议上宣读、交流、讨论。
  - 篇幅与期刊论文类似。
- 学位论文
  - 为申请学位而撰写的论文。
  - 有规范的操作程序，写作之前要作开题报告，中期检查，论文写作
  - 篇幅较长

5.1.4 学术论文的基本结构

- 题名（Title 标题或题目）
  - 题名要求准确、精练和新颖，对全文起画龙点睛得作用
  - 中文一般不超过 20 个汉字
  - 题名位于论文得最前面一行得居中位置
- 作者姓名和单位(author and department)

- 作者姓名位于题名下一行，位置居中
- 另起一行居中位置标明单位、所在城市及邮编
- 若有多个作者，按其对研究工作与论文撰写得贡献大小降序排列，在下一行注明各个作者的单位、所在城市及邮编
- 若多个作者为同一单位，则不需要分别注明

- 摘要(abstract)
  - 摘要是论文的内容不加注释和评论的简短陈述。
  - 摘要能使读者不用阅读全文，就能获得必要的信息，决定是否需要阅读全文。
  - 摘要结构要严谨，表达要简洁，语义要确切。
  - 要用第三人称的写法。如采用"对……进行了研究"、"报告了……现状"等。
- 关键词(key word)
  - 关键词是为文献标引工作从报告、论文中选取出来用以表示全文主题内容信息款目的单词或术语。
  - 3~8 个关键词，排在摘要的左下方
  - 关键词可以从学术论文的题名、摘要和正文中的各级标题与全文中提取，有时需要综合全文内容提出论文涉及主题的上位概念或相关概念做关键词。
- 正文(main body)
  - 学术论文的原文
  - 一般由引言、本论、结论三部分组成
  - 引言（绪论或序论）
    - 简要说明为什么要研究这个题目
    - 解释这一论题讨论、研究的意义，言简意赅
  - 本论
    - 论文的核心内容，占全文的三分之二左右。

- 要详细阐述所研究的成果，包括：调查对象、实验方法、仪器设备、材料原料、实验结果，计算方法、数据资料、经过加工整理的图表、论证的过程、形成的论点和导出的结论等。
- 正文中的图或表应有自明性。

- 结论是学术论文最终的、总体的结论。不是正文中各段的小结的简单重复。结论应该准确、完整、明确、精练。若不可能导出应有的结论，也可以没有结论而进行必要的讨论。可以在结论或讨论中提出建议、研究设想、尚待解决的问题等。

- 参考文献(references)
  - 学术论文中引用的有关文献信息资源。
  - 参考文献的著录方法有顺序编码制和著者-出版年两种，前者居多。
  - 在正文中标注引用的文献时，按出现的先后顺序从 1 开始连续编码，并将序号置于方括号中，然后设成上标。如：信息检索[1]。

5.1.5 学术论文的写作步骤

- 选题
  - 选定学术论文所要研究的主要问题
- 资料的收集与整理
- 确定主题，初步确定论文的题名
  - 主题是指作者在一篇论文中提出的基本观点或中心论点。
- 拟定写作提纲
  - 包括：题名、中心论点、内容提要、章节标题
- 撰写初稿
  - 引言—本论---结论
- 修改定稿

5.1.6 学术论文的投稿

- 要了解学术期刊的详细情况
- 尽量投正刊，不要投增刊
- 不要一稿多投

5.2 学位论文的写作

5.2.1 学位论文的特点

除具有学术论文的学术性、创新性、科学性的特点外，还具有其独自的特性。

- 有规范的操作程序
    - 写作之前要作开题报告，中期检查，论文写作
- 篇幅较长
        - 学士论文：1 万字左右
        - 硕士论文：2-4 万字左右
        - 博士论文：5 万字以上
- 格式、装订与版式有特殊要求

5.2.2 学位论文的组成部分

- 前置部分
    - ✓ 封面
        - 论文题名

- 论文作者
- 导师姓名

✓ 封二

- 学位论文使用声明
- 版权声明
- 作者及导师签名

✓ 题名页

- 中图分类号
- 学校代码

✓ 英文题名页

✓ 内容提要

- 500 字左右
- 包括关键词

✓ 目录

- 主体部分
  - ✓ 各个章节；
  - ✓ 另页开始，每一章另起页；
  - ✓ 一般从引言开始，以结论或总结结束；
  - ✓ 引言（绪论）应包括论文的研究目的，流程和方法；论文研究领域的历史回顾，文献回溯，理论分析等；
  - ✓ 主体部分由于涉及的学科、选题、研究方法、结果表达方式等有很大差异，不能统一规定；
  - ✓ 图、表应该有统一编号，从 1 开始；
- 参考文献
  - ✓ 要求有固定的格式；

- 致谢
- 中英文摘要
  - ✓ 2000 字左右
- 附录
  - ✓ 可有可无

5.2.3 学位论文的开题与写作部分

- 选题
  - ✓ 是学位论文写作的关键性一步，是撰写学位论文的基础；
  - ✓ 选题应该具有新颖性与创新性，不能有歧义，以免产生误解；
  - ✓ 应该根据自己的专业特点、研究条件与科研能力，选择大小适中、难度得当的课题
  - ✓ 工程硕士选题应该具有工程背景；
  - ✓ 选题可以从以下几方面考虑：
    - 所学的专业课中选；
    - 结合导师的科研项目；
    - 与自己工作相关的内容；
    - 当前的研究热点、前沿问题；
  - ✓ 查阅文献资料
    - 以免进行毫无意义的重复研究；同时可以启发思路、借鉴方法；
  - ✓ 选题一般要查询以下几类文献：
    - 学位论文
    - 学术论文
    - 科研成果、专利与产品数据库

- 撰写开题报告
  - ✓ 是对论文选题的系统总结；
  - ✓ 开题报告的质量直接影响学位论文的写作与质量；
  - ✓ 开题报告包括的内容：
    - 论文选题的理由与意义，说明课题的来源，理论和实际意义、价值与可能达到的水平；
    - 国内外关于该论题的研究现状及发展趋势（文献综述）；
    - 研究内容、方法与技术路线。包括研究目标、内容、拟突破的难题或攻克的难关、论文的创新点或实用价值，拟采用的额研究方法、实验方案或可行性分析；
    - 研究计划与进度安排（包括中期报告及答辩时间）；
    - 主要参考文献（有的单位要求文献不少固定的数量）；

第六章 信息检索的数学模型

6.1 信息检索系统的形式化表示

❖ 信息检索的基本原理概括：

检索系统在用户信息需求集合与系统存储的信息资源集合之间所进行的某种匹配与选择。

实现信息检索涉及对以下的三个关键要素的处理

✓ 信息资源集合

原始信息一般不能直接进行信息检索，需要从原始信息文档中抽取其逻辑视图。

✓ 用户信息需求

进行查询的依据，系统将据此搜索文档集合。

- ✓ 匹配选择

一种相似形的匹配，查询结果需要按照某种相似形排序算法有序输出。

6.1.1 信息检索系统的形式化表示

四元组：System=$(D,Q,F,R(d_j,q))$

(1)信息资源集合—D

$D=\{d_1,d_2,...,d_n\}(N>=0)$

若以文本信息为例，集合 D 表示 N 篇文档

(2)用户信息需求集合—Q

$Q=\{q_1,q_2,...,q_m\}$

$q_i(I=1,2,...,m)$表示一个具体的用户提问，提问式可以理解为用户信息需求的一种逻辑视图表示。

(3)信息资源与信息需求的匹配处理框架—F

F 是寻求在 D 与 Q 之间建立一种沟通与联系机制，提供对文档视图、提问式以及它们之间关系进行模型化处理的框架与规则。

(4)匹配函数—$R(d_j,q)$

用于计算任一文档 $d_j(d_j \in D)$与任一提问 $q\ (q\in Q)$形成的文档-提问对(d_j,q)之间的相似度大小，

一般情况下， $R(d_j,q)$的函数值为一实数，其取值区间为[0，1]。

匹配函数具备以下特点：

- ✓ 计算方法简单，计算量小；
- ✓ 函数值在取值区间均匀分布；
- ✓ 针对某一提问所获取的相关文档集合，能够实现合理的排序输出。

6.2 集合论检索模型

6.2.1 布尔检索模型

建立在经典集合论和布尔代数的基础上。

(1)布尔模型的基本原理

布尔模型在解释信息检索处理过程时，主要遵循以下两基本规则：

a.系统索引词集合*(K)*中的每一索引词在一篇文档中只有两种状态：出现或不出现。则每个索引词的权值 $w_{ij} \in \{0,1\}$;

b.用户提问式 q 由 3 种布尔运算符“and”、“or”、“not”连接索引词构成。如：$q=k_1$ and (k_2 or not k_3)

布尔模型对于任一篇文档 $d_j \in D$，定义 d_j与用户提问 q 的匹配函数为

$$sim(d_j, q) = \begin{cases} 1 & \text{若存在} q_{cc} | (q_{cc} \in q_{dnf}) \text{且对于任意} k_i, \text{有} g_i(d_j) = g_i(q_{cc}) \\ 0 & \text{其他情况} \end{cases}$$

(2)布尔模型的分析与评价

布尔模型具有简单、容易理解、简洁的形式化优点。

主要问题：

(a)精确匹配策略问题；

(b)布尔逻辑表达用户需求的能力问题。

6.2.2 模糊集合模型

系统中每一个检索词对应一个模糊的命中文档集合，而每一文档对于这个命中集合而言，都具有各自不同的隶属度值。这种信息检索过程的解释成为各种模糊检索模型建立的共同基础。

这里涉及：

(1)模糊集合论的基本知识

- ✓ 模糊集合的定义；
- ✓ 模糊集合的基本运算；
- ✓ 模糊关系；

(2)模糊检索模型

- ✓ 索引词关联矩阵；
- ✓ 文档的隶属度；
- ✓ 用户提问及表示；

6.2.3 扩展布尔模型

一种基于布尔逻辑框架的、混合有布尔向量特性的检索模型。

这里涉及：

(1)扩展布尔模型的基本原理

(2)扩展布尔模型的主要特点

- ✓ 与传统布尔检索中的倒排文档技术相兼容，支持使用标准布尔逻辑表达式的提问式结构；
- ✓ 允许在文档和提问式中进行词加权处理；
- ✓ 支持按相似度的大小排序输出检索结果；
- ✓ 通过调整参数的取值，可以灵活选择并得到不同的检索结果。

6.3 代数论检索模型

代数论检索模型以线性代数、矩阵计算等数学理论为基础，利用代数论知识揭示信息间关系的检索模型。

6.3.1 向量空间模型（Vector Space Model-VSM）

(1)向量空间模型的基本原理

- ✓ 文档向量的构造；
- ✓ 提问向量的构造；

(2)向量空间模型技术

- ✓ 采用部分匹配策略，实现多值相关性的判断；
- ✓ 采用基于统计学方法的词加权处理模式，使检索效率得到显著改善；
- ✓ 采用对检索结果排序输出的策略，对检索结果数量的控制与调整具有相当大的自由度。

(3)向量空间模型的应用

引入量化处理思想充分发挥计算机的计算特长。

(4)典型的基于 VSM 理论的文本信息处理主要包括以下几个分支领域：

- ✓ 文本检索(Text retrieval)
- ✓ 文本分类(Text Classification)
- ✓ 文本过滤(Text Filtering)
- ✓ 文本挖掘(Text Mining)
- ✓ 文本浏览与可视化(Text Browsing and Visualization)

存在的问题：

- ✓ 处理结果的可解释性较差；
- ✓ 大规模和超大规模真实文本环境中有效性需要验证；
- ✓ 如何与自然语言理解技术进行融合…

6.3.2 潜在语义索引模型(Latent Semantic Indexing-LSI)

基于 VMS 理论框架，提出的一种新的信息检索模型，源于自然语言中词语的多义性和同义性现象。

潜在语义模型的基本原理主要建立在对索引词-文档矩阵的奇异值分解计算上。主要涉及奇异值矩阵的分解等。

6.3.3 神经网络模型

信息检索处理中需要具体定义一个人工神经网络模型来模拟文档、用户提问及其匹配操作。

6.4 概率论检索模型

概率论检索模型主要基于概率论原理来解决信息检索问题。

6.4.1 经典概率模型

基本原理：给定一个用户提问，则信息检索系统中存在一个与该提问相关的理想命中结果集合 R，若能已知集合 R 的主要特征及其描述，则用户的检索得以实现。但在用户提出检索要求时，不知道结果集合的特征。需要在检索开始时对 R 的特性进行某种猜测，据此得到一个初步的命中结果集合。在此基础上，用户可以对初始检索结果集合中文档相关与否进行判断，根据这些反馈信息，在后续的检索处理中不断优化与改进。

- ✓ 经典概率模型的基本原理；
- ✓ 经典概率模型的分析与评价。

6.4.2 基于 Bayesian 网络的检索模型

Bayesian 网络是概率理论的一个主要分支。通常 Bayesian 网络可以看作一个有向无环图(DAG)。图中的节点一般用来表示随机变量，有向边表示随机变量之间的因果关系，它由表示原因的随机变量指向结果的随机变量，因果关系影响力的大小（权值）用条件概率表示。

- ✓ 推理网络模型

用数学方法从文档文本内容推理得出该文档满足用户信息需求得概率，将这个概率值作为文档与用户查询提问得匹配程度，并根据匹配程度得大小对文档进行排序。以上需要得出一个推理网络模型，文档部分和提问部分在网络中不分离。

- ✓ 信念网络模型

信念网络模型与推理网络模型具有共同得理论基础，不同点是二者得网络拓扑结构不同。文档部分和提问部分在网络中是被分离的。

6.5 其他信息检索模型与数学理论

- 集合论模型、代数论模型、概率论模型的共同点
  - ❑ 信息内容特征的提取
- 新型信息检索系统模型的特点
  - ❑ 信息的结构特征及其提取

6.5.1 结构化检索模型

- 是一种基于信息的结构特征匹配的检索模型。

如：某文档的某页内容中有一幅图片，图的标题文字中包含“earth”一词，而围绕该图的文字内容还包含字符串“our family”。

- 代表性的模型
  - ✓ 基于非重叠链表的模型

- ✓ 基于邻近节点的模型

6.5.2 其他新型数学理论

- 遗传算法
- 粗糙集理论

第七章 信息检索系统及其构成

7.1 信息检索系统的及其类型

7.1.1 信息检索系统的定义

(1)系统的概念及特征

- 概念

  三元组：System=（Input,Processing,Output）

  其中：

  Input=$\{i_1,i_2,...,i_m\}$ $(m>=0)$;输入有限集合

  Processing= $\{p_1,p_2,...,p_k\}$ $(k>=0)$;处理函数集合

  Ouput=$\{o_1,o_2,...,o_n\}$ $(n>=0)$;输出有限集合

- 特征
  - ✓ 整体性
  - ✓ 关联性
  - ✓ 层次性

- ✓ 目的性
- ✓ 适应性

(2)信息检索系统的定义

- 定义

具有存储和信息查询功能的一类信息服务设施（或工具）。通常是人机交互信息系统。

- 系统基本要素
  - ✓ 明确的目标

    检索系统应具有明确的服务对象、专业范围及用途
  - ✓ 不可缺少的资源

    检索系统必须搜集、加工、存储一定数量的信息资源
  - ✓ 技术装备

    存储信息的载体、匹配选择、信息的输入/输出/显示/传递等设备
  - ✓ 方法与措施

    检索系统应提供一定的方法与措施，保证检索系统的查全率和查准率
  - ✓ 功能

    检索系统所应具有的检索及其他信息服务功能

7.1.2 信息检索系统的类型

- 按照设备划分
  - ✓ 书本式检索系统
  - ✓ 卡片式检索系统
  - ✓ 穿孔卡片式检索系统(机械化检索系统)

- ✓ 缩微式检索系统
- ✓ 计算机化检索系统
- ✓ 网络检索系统

- 按功照能划分
  - ✓ 文献检索系统(DRS)
  - ✓ 数据库管理系统(DBMS)
  - ✓ 自动问答系统(QAS)
  - ✓ 管理信息系统(MIS)
  - ✓ 决策支持系统(DSS)

7.2 信息检索系统的基本结构

7.2.1 信息检索系统的物理结构

- 系统的物理构成角度
  - ➢ 硬件部分
  - ➢ 软件部分
  - ➢ 信息资源集合
- 按物理空间分布情况
  - ➢ 集中式检索系统
  - ➢ 分布式检索系统

7.2.2 信息检索系统的逻辑结构

是指包括的功能模块（或子系统）及其相互关系。

- 信息存储

- 信息查询

7.3 信息存储功能模块

7.3.1 信息资源及其选择与采集模块

根据系统的经营方针和服务对象的需要，以快速、经济的手段，广泛、连续地从各种信息源或信息渠道完成信息资源的采集工作，为系统提供充足适用的数据来源。

7.3.2 信息标引处理模块

- 标引(Indexing)

是指对信息资源的各种检索特征进行分析并使之显形化，以便存储和检索这两个环节提供某种连接的一种重要的信息加工操作。

- 信息标引处理模块功能

对信息资源中具有检索价值的特征信息进行提取与标识，并组织成索引文档，为用户的查询和访问提供准确有效的入口。

- 标引处理的类型
  - ✓ 人工标引
  - ✓ 自动标引
- 自动标引
  - ✓ 全自动标引
  - ✓ 半自动标引
  - ✓ 自动抽词标引
  - ✓ 自动赋词标引

7.3.3 数据库创建与维护模块

- 数据库的设计
- 数据库创建与维护

7.4 信息查询功能模块

7.4.1 用户（检索）接口模块

- 基本构成
  - ✓ 用户模型
  - ✓ 信息显示
  - ✓ 交互语言
  - ✓ 反馈机制
- 用户接口设计的基本原则与技术
  - ✓ 基本原则
    - ➢ 提供反馈原则
    - ➢ 减轻记忆负担原则
    - ➢ 为不同用户提供不同接口原则
  - ✓ 技术
    - ➢ 字符用户界面
    - ➢ 图形用户界面
    - ➢ 多通道用户界面（目前处于探索与研究阶段）

7.4.2 提问处理与检索匹配模块

信息检索系统的核心模块。

- 功能

接受并处理用户输入的检索词或提问式，将它们与数据库倒排索引文档中存储的数据项进行匹配运算，然后把运算结果返回给用户。

- 提问处理与检索匹配模块的操作流程

1.接收用户提问

2.提问校验

3.提问加工 常用的加工方法：表展开法、逆波兰法、准波兰法、析取范式变换法

4.检索匹配