信息检索综合实践

第0讲：课程概述（信息检索简介）

苏州大学计算机科学与技术学院
贡正仙

本课程PPT改编自陈文亮老师分享的资料，感谢陈老师！

提纲

- 什么是信息检索？
- 为什么要学习信息检索？
- 课程内容

提纲

- **什么是信息检索？**
- 为什么要学习信息检索？
- 课程内容

新闻 网页 贴吧 知道 音乐 图片 视频 地图 百科 文库

Baidu 百科

信息检索

进入词条 搜索词条 帮助

首页 分类频道 特色百科 玩转百科 百科用户 百科校园 百科合作

信息检索（Information Retrieval，IR）

收藏 498 135

信息检索

编辑

信息检索（Information Retrieval）是指信息按一定的方式组织起来，并根据信息用户的需要找出有关的信息的过程和技术。狭义的信息检索就是信息检索过程的后半部分，即从信息集合中找出所需要的信息的过程，也就是我们常说的信息查寻（Information Search 或Information Seek）。

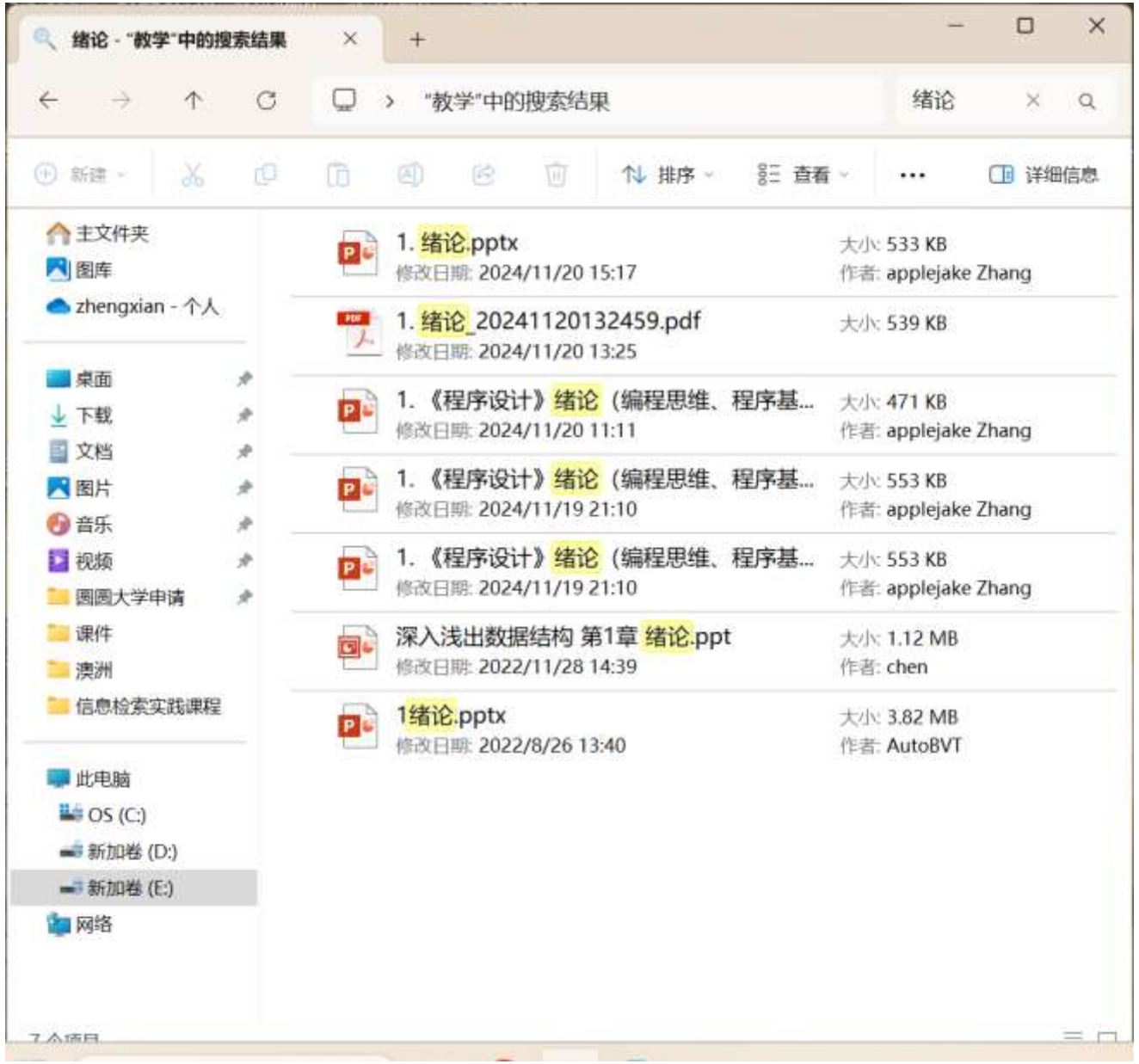
信息检索（Information Retrieval）是指从信息资源的集合中查找所需文献或查找所需文献中包含的信息内容的过程

匹 配 信息检索也是一个匹配过程。

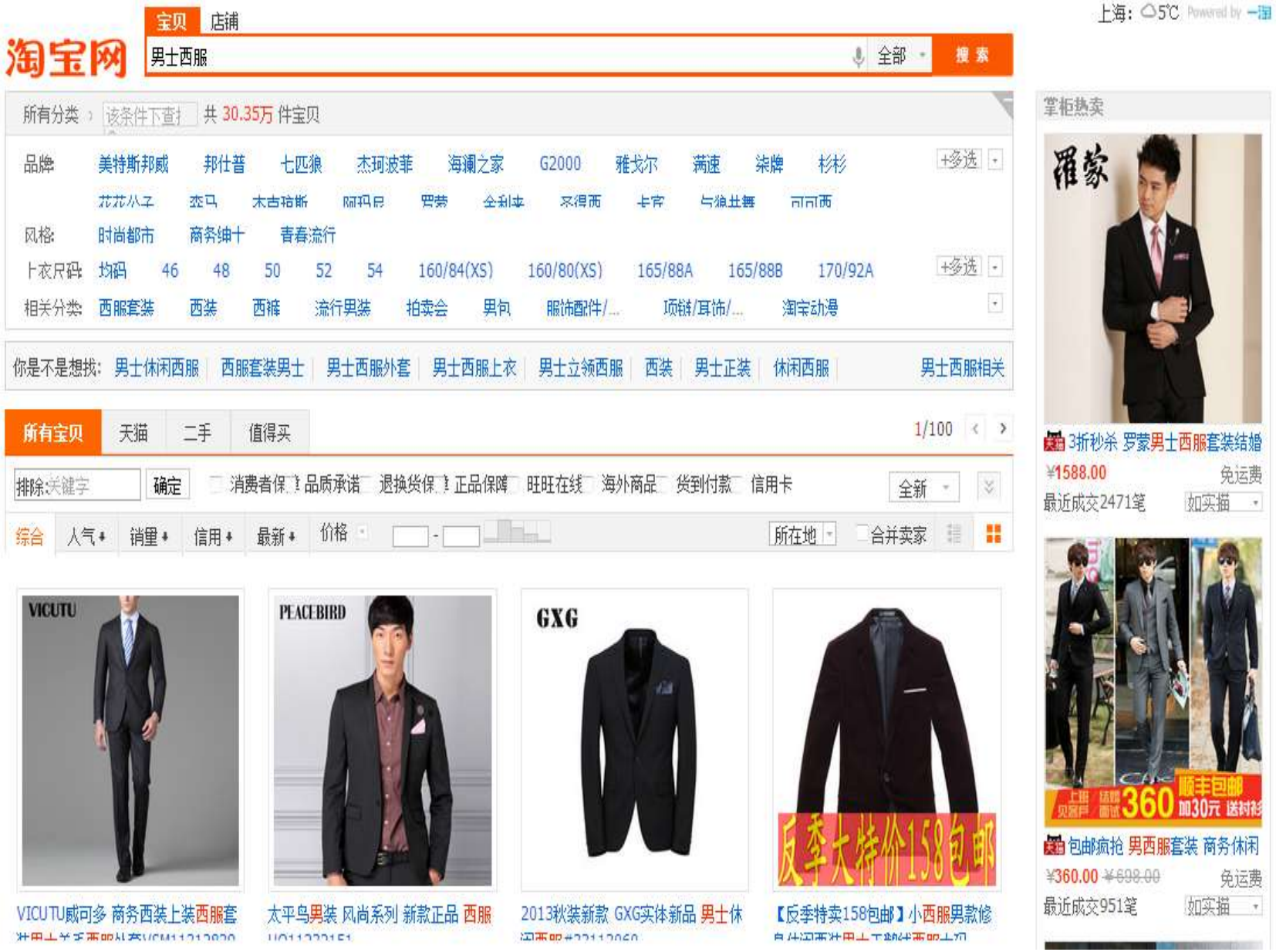
信息检索过程 包括信息处理和检索两个方面

目录

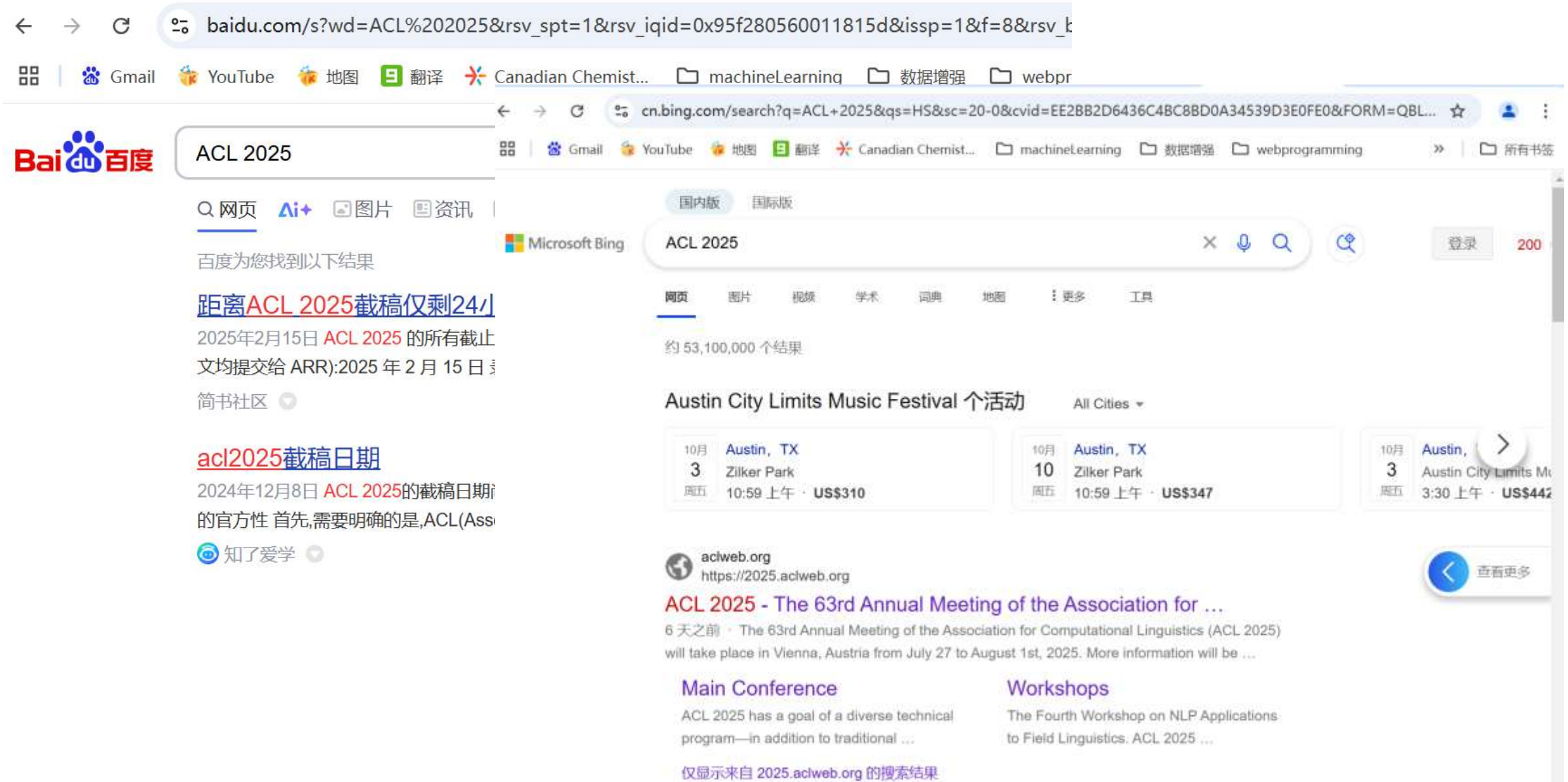
本地信息检索


绪论 - "教学"中的搜索结果
"教学"中的搜索结果
绪论
新建
排序
查看
详细信息
主文件夹
图库
zhengxian - 个人
桌面
下载
文档
图片
音乐
视频
圈圈大学申请
课件
澳洲
信息检索实践课程
此电脑
OS (C:)
新加卷 (D:)
新加卷 (E:)
网络
1. 绪论.pptx
修改日期: 2024/11/20 15:17
大小: 533 KB
作者: applejake Zhang
1. 绪论_20241120132459.pdf
修改日期: 2024/11/20 13:25
大小: 539 KB
1. 《程序设计》绪论（编程思维、程序基...
修改日期: 2024/11/20 11:11
大小: 471 KB
作者: applejake Zhang
1. 《程序设计》绪论（编程思维、程序基...
修改日期: 2024/11/19 21:10
大小: 553 KB
作者: applejake Zhang
1. 《程序设计》绪论（编程思维、程序基...
修改日期: 2024/11/19 21:10
大小: 553 KB
作者: applejake Zhang
深入浅出数据结构 第1章 绪论.ppt
修改日期: 2022/11/28 14:39
大小: 1.12 MB
作者: chen
1绪论.pptx
修改日期: 2022/8/26 13:40
大小: 3.82 MB
作者: AutoBVT

网站内部信息检索


淘宝网
宝贝
店铺
男士西服
全部
搜索
上海：5℃
所有分类
共 30.35万 件宝贝
品牌
美特斯邦威
邦仕普
七匹狼
杰珂波菲
海澜之家
G2000
雅戈尔
满速
柒牌
杉杉
+多选
风格：
时尚都市
商务绅士
青春流行
上衣尺码：
均码
46
48
50
52
54
160/84(XS)
160/80(XS)
165/88A
165/88B
170/92A
相关分类：
西服套装
西装
西裤
流行男装
拍卖会
男包
服饰配件/...
项链/耳饰/...
淘宝动漫
你是不是想找：
男士休闲西服
西服套装男士
男士西服外套
男士西服上衣
男士立领西服
西装
男士正装
休闲西服
男士西服相关
所有宝贝
天猫
二手
值得买
1/100
排除关键字
确定
消费者保障
品质承诺
退换货保障
正品保障
旺旺在线
海外商品
货到付款
信用卡
全新
综合
人气
销量
信用
最新
价格
所在地
合并卖家
VICUTU
PEACEBIRD
GXG
反季大特价158包邮
VICUTU威可多 商务西装上装西服套
太平鸟男装 风尚系列 新款正品 西服
2013秋装新款 GXG实体新品 男士休
【反季特卖158包邮】小西服男款修
掌柜热卖
罗蒙
3折秒杀 罗蒙男士西服套装结婚
¥1588.00
免运费
最近成交2471笔
如实描
顺丰包邮
加30元 送衬衫
360
包邮疯抢 男西服套装 商务休闲
¥360.00
¥698.00
免运费
最近成交951笔
如实描

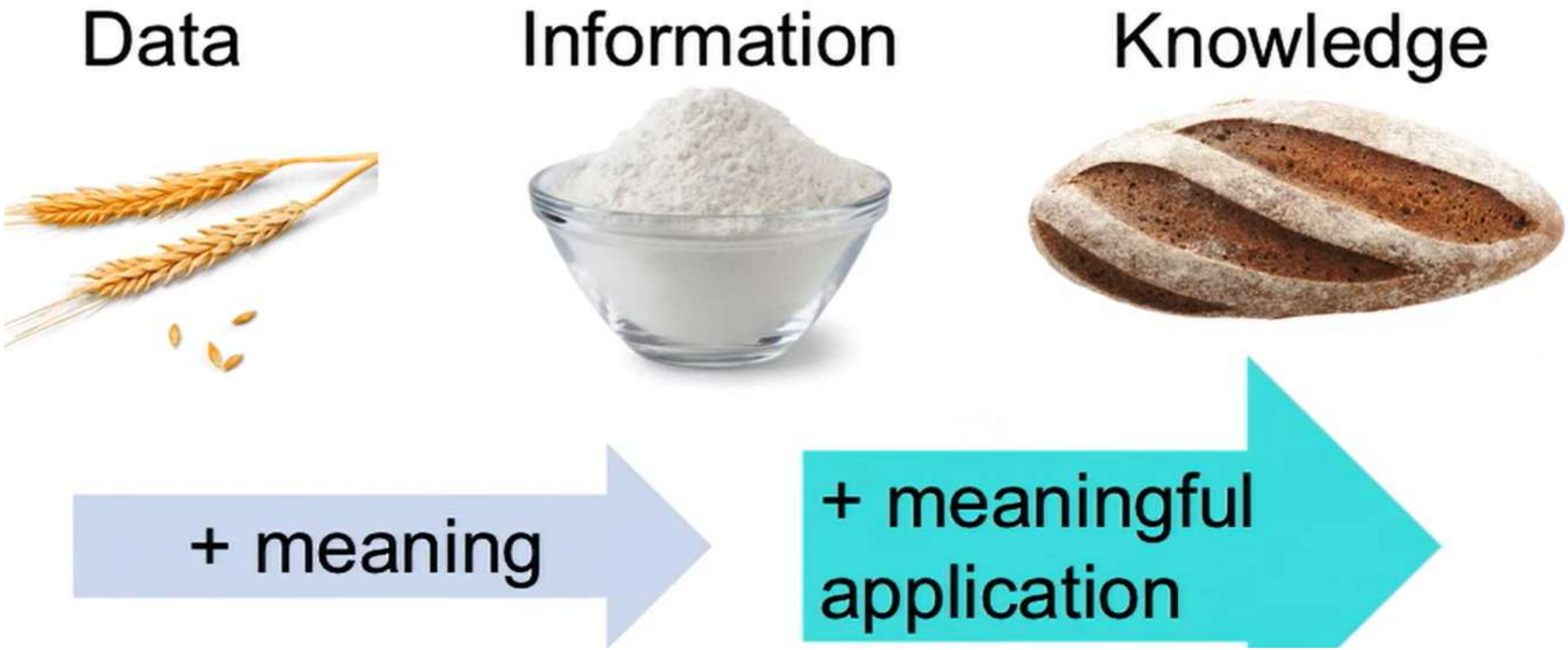
搜索引擎


baidu.com/s?wd=ACL%202025&rsv_spt=1&rsv_iqid=0x95f280560011815d&issp=1&f=8&rsv_b
Gmail
YouTube
地图
翻译
Canadian Chemist...
machineLearning
数据增强
webpr
Bai du 百度
ACL 2025
网页
图片
资讯
百度为您找到以下结果
距离ACL 2025截稿仅剩24小
2025年2月15日 ACL 2025 的所有截止
文均提交给 ARR):2025 年 2 月 15 日
简书社区
acl2025截稿日期
2024年12月8日 ACL 2025的截稿日期
的官方性 首先,需要明确的是,ACL(Ass
知了爱学
cn.bing.com/search?q=ACL+2025&qs=HS&sc=20-0&cvid=EE2BB2D6436C4BC8BD0A34539D3E0FE0&FORM=QBL...
Gmail
YouTube
地图
翻译
Canadian Chemist...
machineLearning
数据增强
webprogramming
所有书签
国内版
国际版
Microsoft Bing
ACL 2025
登录
200
网页
图片
视频
学术
词典
地图
更多
工具
约 53,100,000 个结果
Austin City Limits Music Festival 个活动
All Cities
10月 3 周五 Austin, TX Zilker Park 10:59 上午 · US$310
10月 10 周五 Austin, TX Zilker Park 10:59 上午 · US$347
10月 3 周五 Austin, Austin City Limits Mu 3:30 上午 · US$442
aclweb.org
https://2025.aclweb.org
查看更多
ACL 2025 - The 63rd Annual Meeting of the Association for …
6 天之前 · The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025) will take place in Vienna, Austria from July 27 to August 1st, 2025. More information will be …
Main Conference
ACL 2025 has a goal of a diverse technical program—in addition to traditional …
Workshops
The Fourth Workshop on NLP Applications to Field Linguistics. ACL 2025 …
仅显示来自 2025.aclweb.org 的搜索结果

三个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)
  - 本地信息检索：查找“绪论”
  - 淘宝的例子：对象 “男士西服”
  - Google/百度/Bing的例子：需求 “ACL 2025”
  - ……

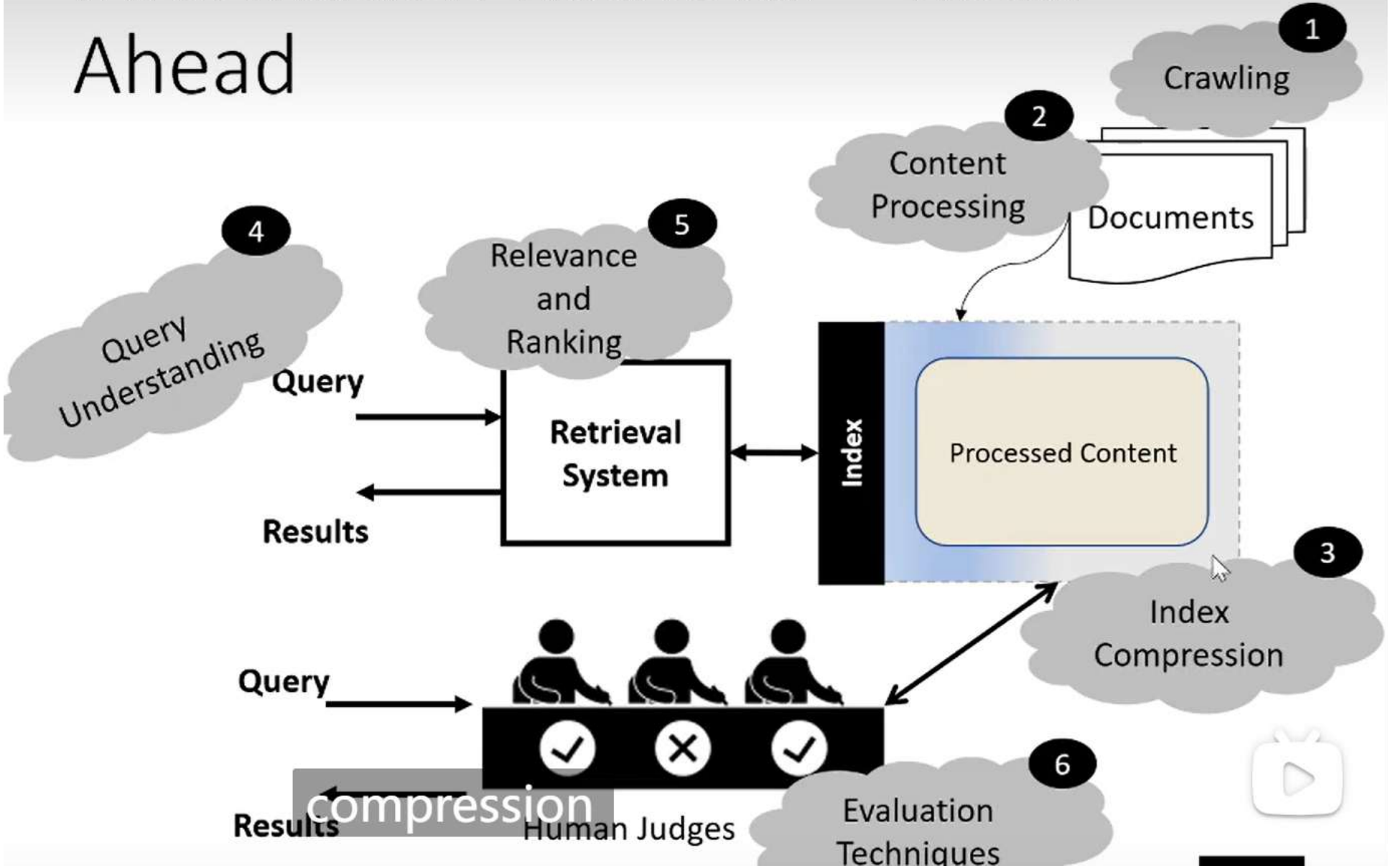
数据、信息和知识


Data
Information
Knowledge
+ meaning
+ meaningful application

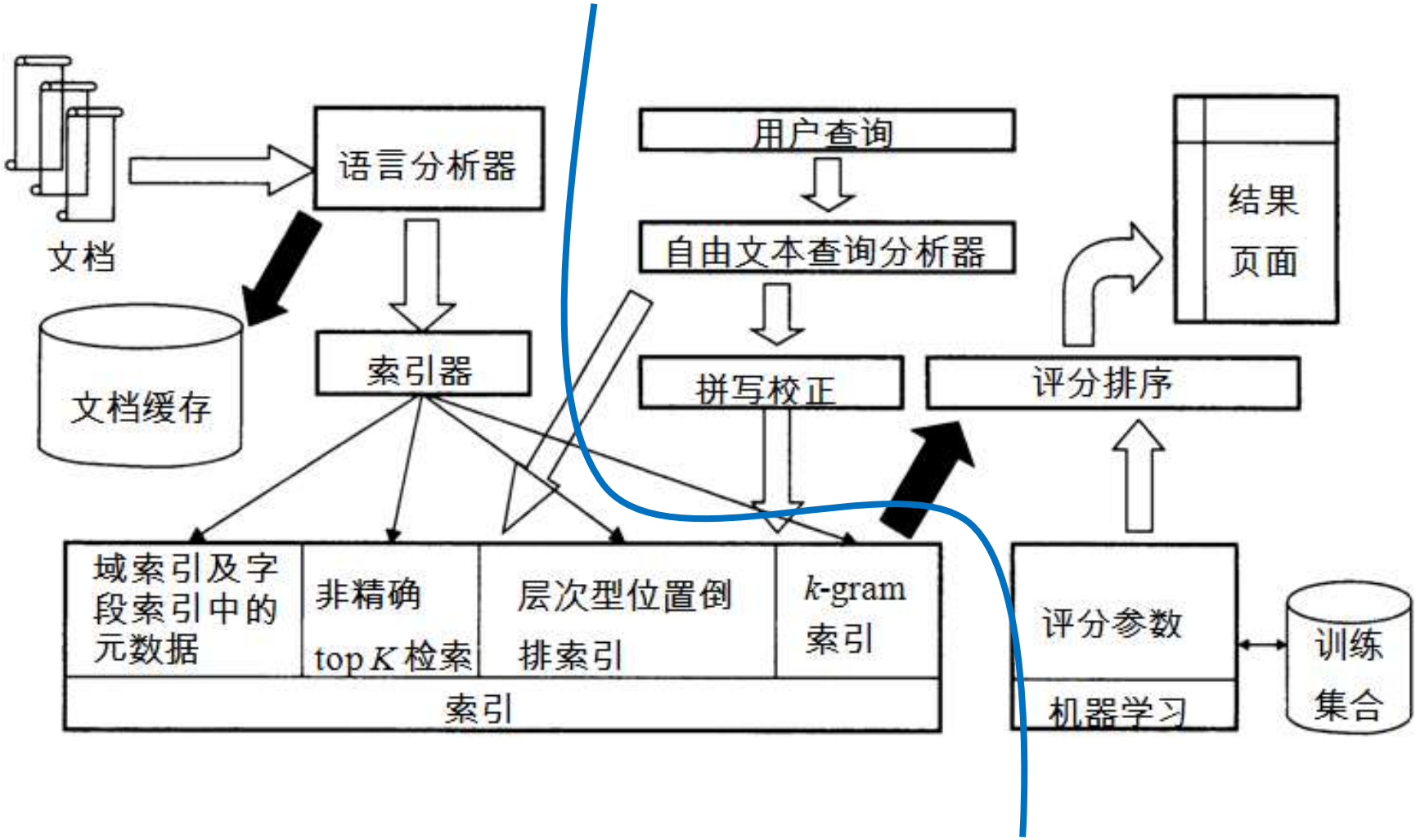
信息检索的一般定义

- 信息检索是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。
  - 结构化信息（数据库，SQL）
  - 非结构化信息（大规模文本，更一般，处理更复杂）
- 信息检索是从大规模非结构化数据（通常是文本）的集合中找出满足用户信息需求的资料的过程。

Information Retrieval – Road Ahead



详细的IR系统示意图



13

信息检索技术的应用

輿情分析

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

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。
- Web信息检索：在超大规模数据集上的检索。

提纲

- 什么是信息检索？
- **为什么要学习信息检索？**
- 课程内容

直接经济效益-能赚钱啊！

- 世界级牛公司
  - 很多互联网的公司：Google, baidu, … 高市值公司
- 软件工程师
  - 年薪高

市场发展的需求

- 用户需要信息检索技术：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，Google、Baidu，还有一些公司如Microsoft、Sina、Sohu、Tecent、Netease都加入到这个搜索技术的竞争。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、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
  - 人才的竞争：搜索相关人才人数出现缺口，他们非常抢手，待遇如日中天
  - 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？

几个应用需求

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

提纲

- 什么是信息检索？
- 为什么要学习信息检索？
- **课程内容**

课程目的

- 本课程面向目前主流的信息检索应用，指导学生进行实际的系统设计与开发。
- 使学生领会通用文本检索系统的主体架构、数据处理流程和核心模型
- 培养学生对特色模块的掌握与实践，包括文本预处理、基于统计和语义的特征建模，以及利用特征实现排序学习的机制。
- 同步开展程序开发与测试，督促学生成组进行分工协作，实现不同的检索系统。

回顾百度百科中的IR概念

一般情况下，信息检索指的是广义的信息检索。广义的信息检索是信息按一定的方式进行加工、整理、组织并存储起来，再根据信息用户特定的需要将相关信息准确的查找出来的过程。又称信息的**存储**与**检索**。。

课程内容

- **数据搜集和标注**
  - 网页下载（爬虫）
  - 网页信息提取
- **数据预处理**
  - 分词，去停用词
- **文档存储**
  - 倒排索引
- **查询表示和重构**
- **信息相关度计算**
- **文档排序**

信息存储

信息检索

课程内容

本课程**不包括**：

- 搜索引擎使用技巧
  - 复杂的检索表达式，文献检索、专利检索等
- 爬虫开发
- 非文本数据处理
  - 音频，图像
- PageRank等超链分析算法
- ……

Q&A

- 有什么问题？