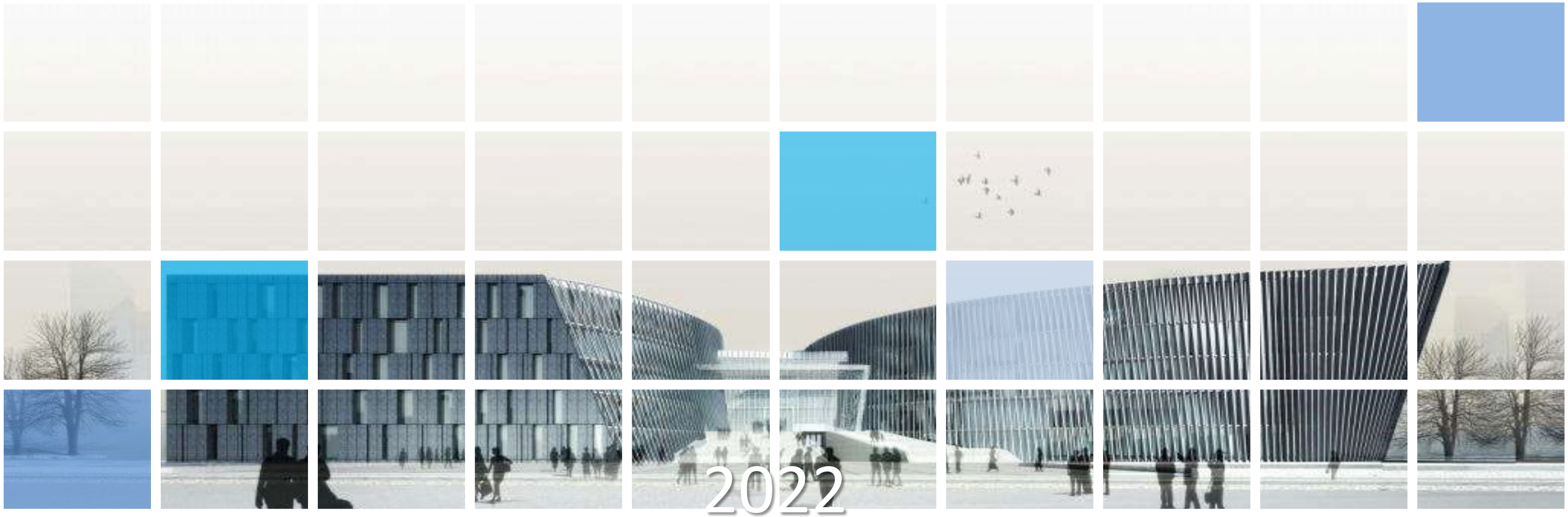


2022

南京大学信息管理学院

信息检索

邓三鸿

njuir@sina.com

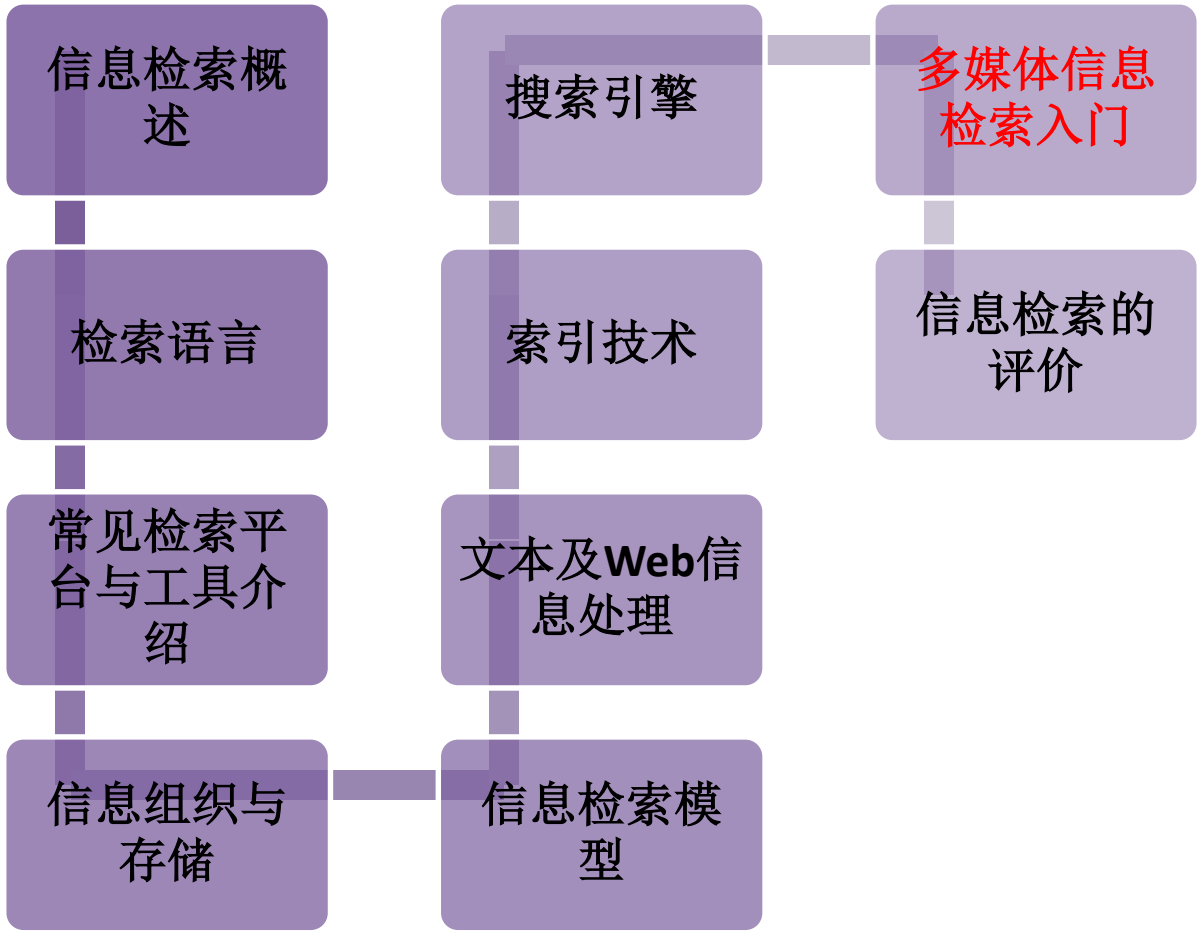
PART P.s

信息检索课程回顾

A Review of Information Retrieval

课程框架


信息采集
信息加工
信息组织
信息检索
信息分析
信息预测
信息服务


信息检索概述
检索语言
常见检索平台与工具介绍
信息组织与存储
搜索引擎
索引技术
文本及Web信息处理
信息检索模型
多媒体信息检索入门
信息检索的评价

教学与考核

教材： 暂无

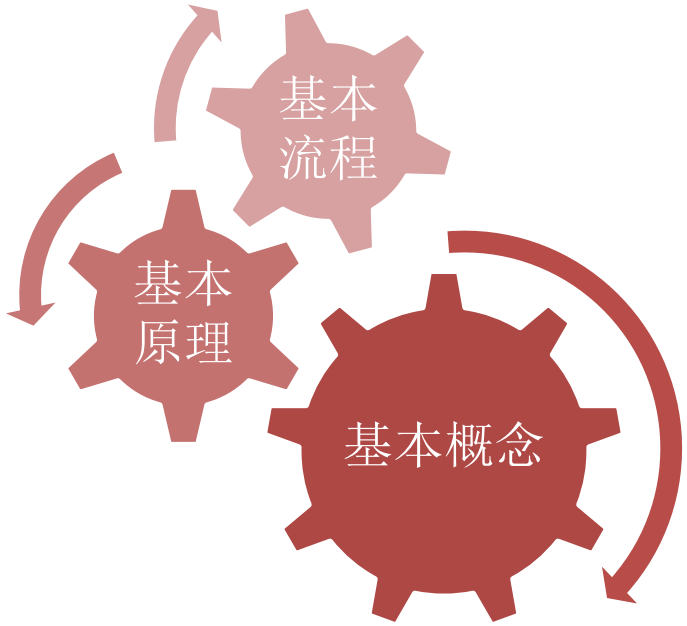
参考资料： 很多

考核： 平时作业（实践，10%）+综合实践（期末大作业，20%）+闭卷考试（70%）

课程邮箱： njuir@sina.com

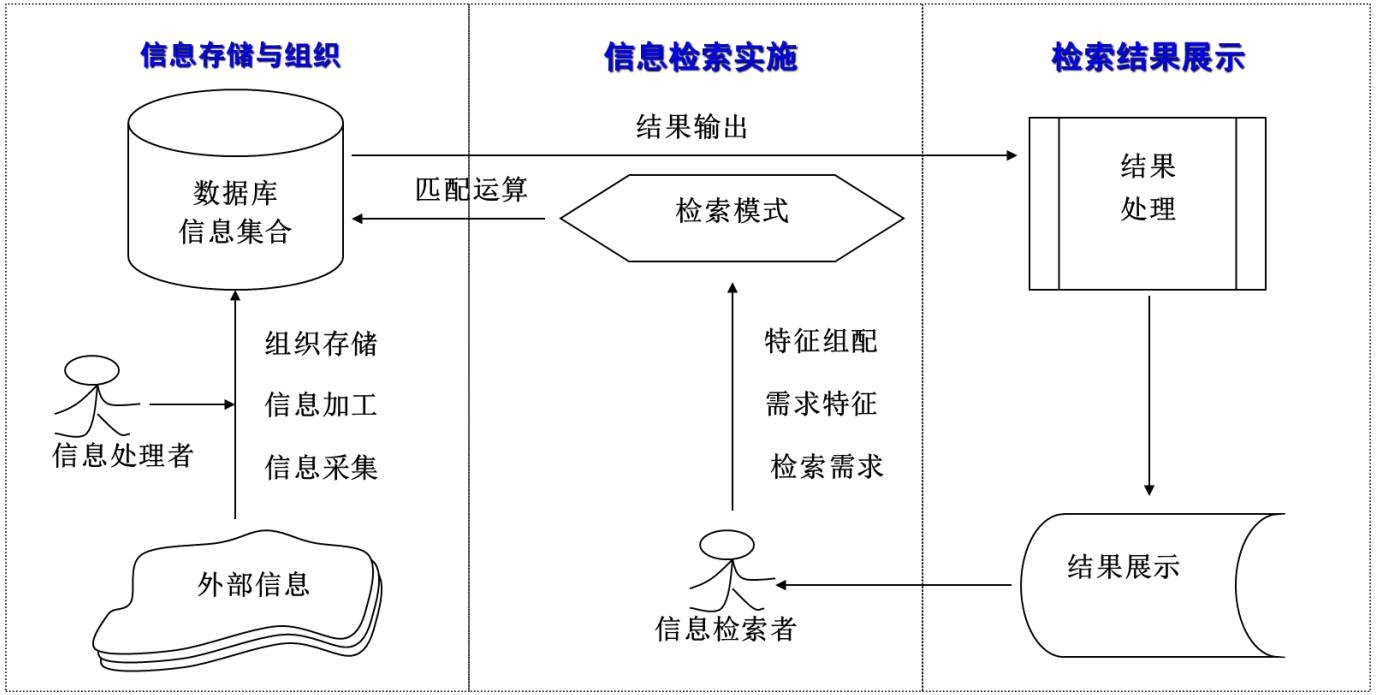
考试安排： 时间：2021.07.02 14:00-16:00 地点：仙I-207

重点关注



Ch1 概述

- 基本概念
  - 信息技术变迁、信息素养
  - 数据、信息、情报、文献
  - 信息检索
- 基本原理
  - 信息检索的原理
  - 信息检索的分类
- 基本流程
  - 信息管理流程

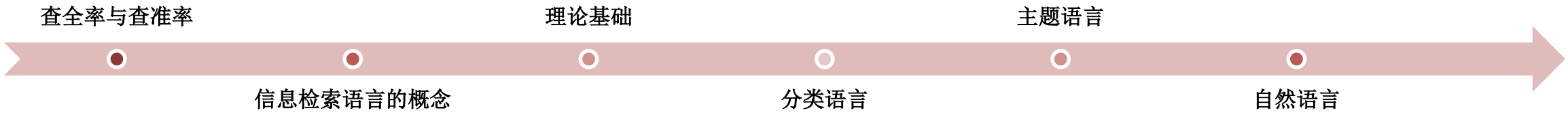


信息检索与信息素养 | 信息检索的流程 | 信息检索的发展

数据、信息、情报、知识 | 信息检索的研究内容

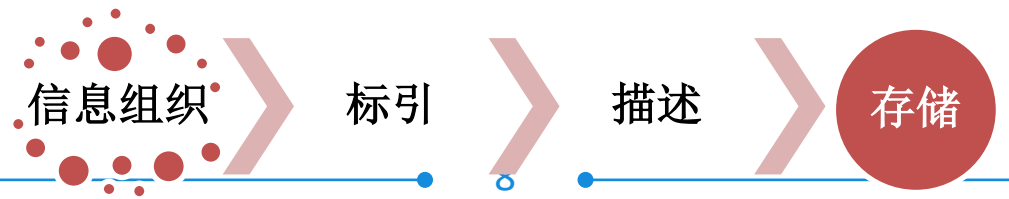
Ch2 信息检索语言

- 基本概念
  - 检索语言
  - 分类语言、主题语言
  - 术语、标题词、单元词、叙词、关键词
- 基本原理
  - 检索语言的分类
  - 中图法
- 基本流程
  - 无



Ch3 信息组织：标引、描述与存储

- 基本概念
  - 信息组织、描述、信息构建
  - 标引
  - 编目、Marc、元数据、标记语言
  - 信息存储
- 基本原理
  - 标引的分类
- 基本流程
  - 标引过程



Ch4 常见检索平台与工具

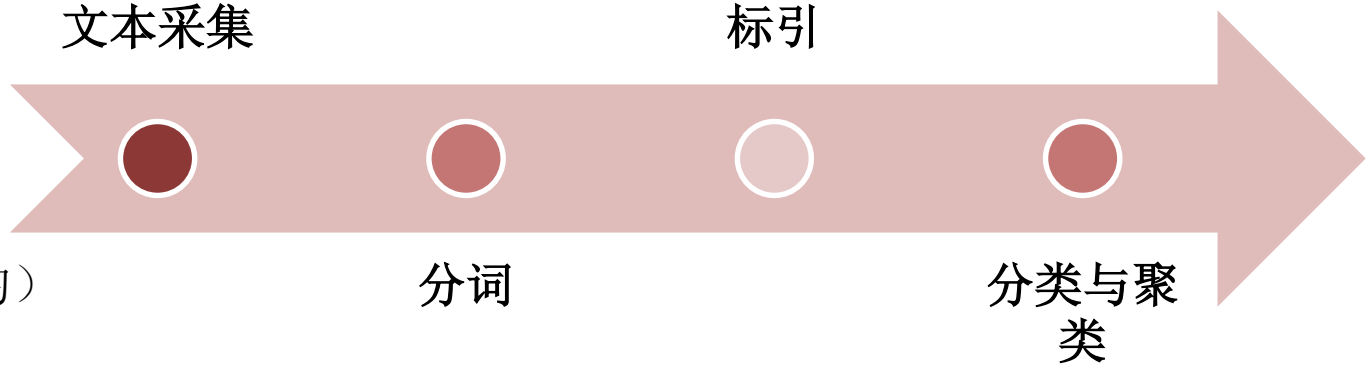
- 基本概念
  - 无
- 基本原理
  - 文献层次划分
  - 构建检索式
- 基本流程
  - 机检步骤

Ch5 信息检索模型

- 基本概念
  - 信息检索模型
  - 命题、联结词
- 基本原理
  - IR四元组
  - 布尔模型、VSM、概率模型
  - Tf-Idf、内积、Jaccard相似性
  - 条件概率
  - 联结词
- 基本流程
  - 析取范式

Ch6 文本信息处理

- 基本概念
  - 自动标引
  - 切分歧义、未登录词、停用词
  - Zipf定律
  - 自动分类、纯度
- 基本原理
  - 特征选择
  - 分类评测
  - 类间相似度（单链、全链、组平均）
  - 聚类的纯度
- 基本流程
  - FMM、BMM
  - KNN
  - HAC、K-means



Ch7 索引技术

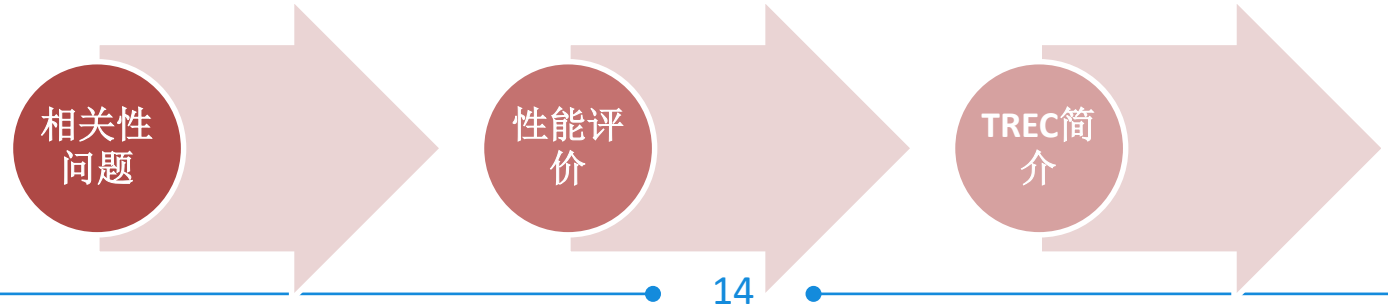
- 基本概念
  - 索引、索引的存储
  - 倒排文档
- 基本原理
  - 逆波兰表达式
  - 文本检索技术（截词、加权）
- 基本流程
  - BF、KMP、BM

Ch8 信息采集与搜索引擎

- 基本概念
  - 深网、搜索引擎
  - 锚文本、出链、入链
- 基本原理
  - 搜索引擎的分类
  - 爬行策略
  - HITS
- 基本流程
  - Pagerank

ChX 信息检索的评价

- 评价与相关性
  - 评价的概念与意义
  - 评价的基本条件，缓冲池
  - 统一评测
- 基本原理
  - 文档集的划分
  - 基本评价指标（P、R、F）
  - 单值评价指标（MAP、p@10、RP）



关于考试

考试时间：2022年6月14日 19:30-21:30

考试地点：仙I-107

题　　型：选择（30）、填空（20）、名词解释（20）、问答（30）

例：选择

在信息的五次技术革命中，以下哪种技术的出现使得知识开始能较大规模地进行传递？（　　）

A）文字　　B ）造纸与印刷术　　C ）电报电话电视　　D ）计算机及现代通讯

信息检索根据检索对象不同，一般分为(　　)。

A.　二次检索、高级检索　　B.分类检索、主题检索

C.数据检索、事实检索、文献检索　　D.计算机检索、手工检索

例：填空

制定“通用标记标言”的基本思想是把文档的_____与_____分开。

信息检索中的“相关性”主要是指检索系统针对用户的信息需求从文档集合中检出的与_____之间的一种匹配关系。

后缀表达式 abc*+de+- ，其转换成中缀表达式则为_____。

例：名词

- Marc

Machine Readable Catalog

“机器可读目录”，即以代码形式和特定结构记录在计算机存储载体上的、用计算机识别与阅读的目录。

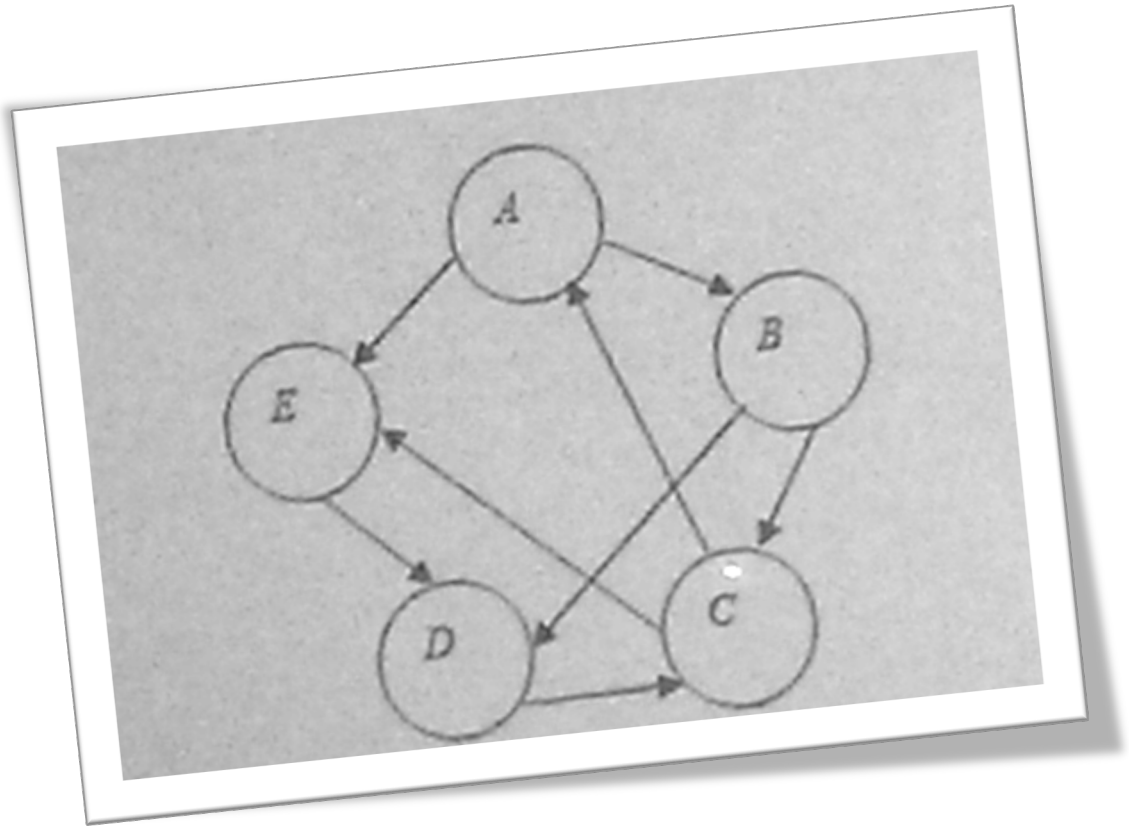
- 签名文件

签名文件（signature file）是基于散列（Hash）技术的面向单词的索引结构

在检索时需要顺序比较，适用于小规模的文本

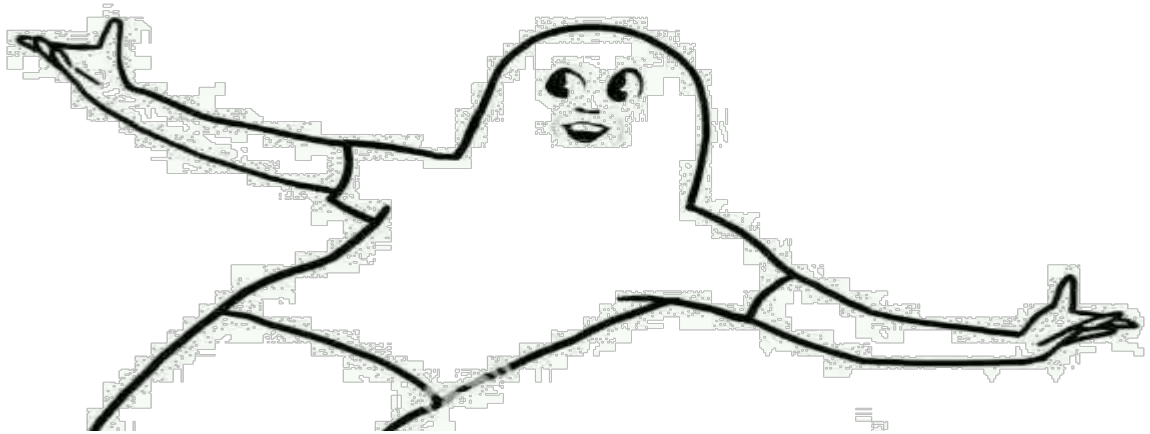
在大多数应用中，其性能不如倒排文件

例：问答



寄语

如果要在学习上加一个期限
我希望是一万年



世界上最远的距离
不是生与死
The farthest distance in the world
Is not the distance between
life and death
而是我在学习
你却不学习
But you don't study
when I am studying

后续课程

技术与理论基础（概率论、线性代数、数据库、程序设计、MIS）
自然语言处理
语义网技术及应用
信息检索实务
计算机图像处理
文本数据管理与分析
网络舆情分析