Information Retrieval in Nanoscience and Technology

- Dr. Wei Deng
- Institute of Functional Nano & Soft Materials
- College of Nano Science & Technology
- Soochow University
- Email: dengwei@suda.edu.cn

Storage procedure of literatures

- Information collection and selection
- Information description and processing
- Information indexing

Section 1: Information collection and selection

信息收集（Information Gathering）

- **Definition：** 通过各种方式获取所需要的信息。信息收集是信息得以利用的第一步，也是关键的一步。
- **Importance：** 信息收集工作的好坏，直接关系到整个信息管理工作的质量。

信息的搜集与选择

1、确定收集的原则

（Determine the principles of collection）

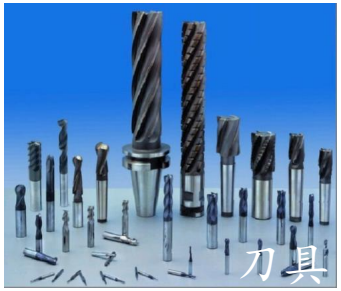
2、收集方法（Way of collection）

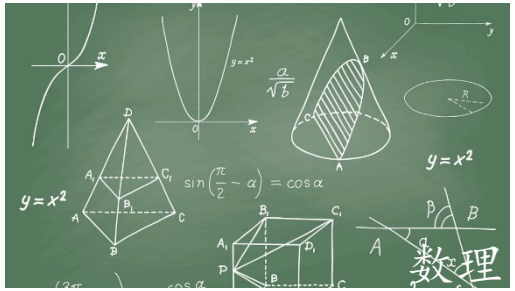
3、选择所需信息

（Choose the required information）

确定收集的原则：准确性、全面性、时效性

- 收集的范围（如学科范围：机械、物理、化学等）
- 主题范围（机床，刀具）
- 覆盖面
- 信息种类
- 文种
- 时间跨度
- 收集的数量
- 摘储率





收集的范围

1.内容范围

Definition： 是指根据信息内容与信息收集目标和需求相关性特征所确定的范围。

Classification： 本身内容范围、环境内容范围。

本身内容范围是由事物本身信息相关内容特征组成的范围；环境内容范围是由事物周边、与事物相关的信息的内容特征组成的范围。

收集的范围

2.时间范围

Definition： 是指在信息发生的时间上，根据与信息收集目标和需求具有一定相关性的特征所确定的范围，这是由信息的历史性和时效性所决定的。

3.地域范围

Definition： 是指在信息发生的地点上，根据与信息收集目标和需求具有一定相关性的特征所确定的范围。这是由信息的地域分布特征和信息收集的相关性要求所决定的。

收集方法 （Way of collection）

- 调查法 Investigation method
- 观察法 Observation method
- 实验法 Experimental method
- 文献检索 Literature retrieval method
- 网络信息收集 Network information collection

调查法 Investigation method

- **Definition：** 指通过考察了解客观情况直接获取有关材料，并对这些材料进行分析的研究方法。

- **Classification：普查和抽样调查**

普查是调查有限总体中每个个体的有关指标值。

抽样调查是按照一定的科学原理和方法，从事物的总体中抽取部分称为样本的个体进行调查，用所得到的调查数据推断总体。抽样调查是较常用的调查方法，也是统计学研究的主要内容。

观察法 Observation method

- **Definition**：是通过开会、深入现场、参加生产和经营、实地采样、进行现场观察并准确记录(包括测绘、录音、录相、拍照、笔录等)调研情况。
- **Characteristics**：应用广泛，常和询问法、搜集实物结合使用，以提高所收集信息的可靠性。
- **Classification**：

  一是对人的行为的观察；
  二是对客观事物的观察。

实验法 Experimental method

- **Definition：** 通过实验过程获取其他手段难以获得的信息或结论。
- **Characteristics：** 实验者通过主动控制实验条件，包括对参与者类型的恰当限定、对信息产生条件的恰当限定和对信息产生过程的合理设计，可以获得在真实状况下用调查法或观察法无法获得的某些重要的、能客观反映事物运动表征的有效信息，还可以在一定程度上直接观察研究某些参量之间的相互关系，有利于对事物本质的研究。
- **Classification：**

  实验室实验、现场实验、计算机模拟实验、计算机网络环境下人机结合实验等。

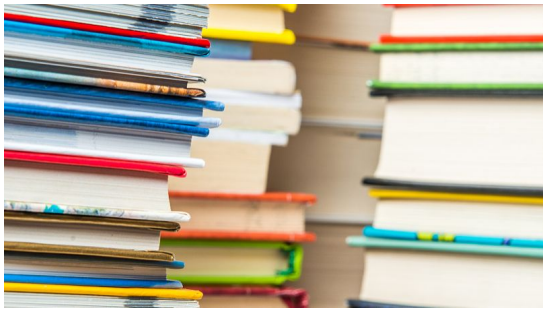
文献检索 Literature retrieval method

- **Definition**：从浩繁的文献中检索出所需的信息的过程。
- **Characteristics**：不同的检索方式具有不同的特点。
- **Classification**：

手工检索：通过信息服务部门收集和建立的文献目录、索引、文摘、参考指南和文献综述等来查找有关的文献信息；

计算机检索：是文献检索的计算机实现，其特点是检索速度快、信息量大，是当前收集文献信息的主要方法。

（后面详细介绍）



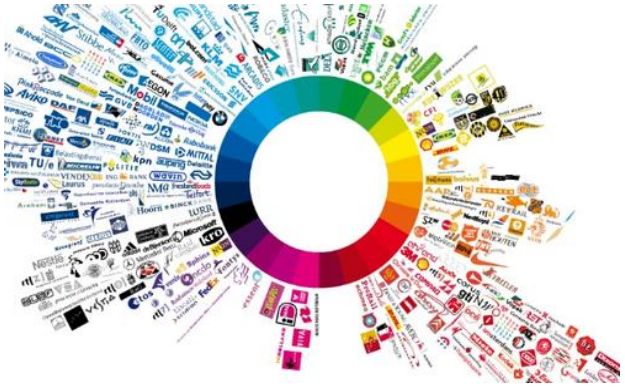
网络信息收集 Network information collection

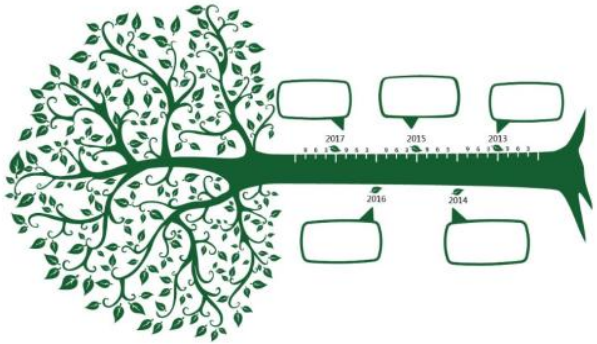
- **网络信息**是指通过计算机网络发布、传递和存储的各种信息。收集网络信息的最终目标是给广大用户提供网络信息资源服务，整个过程经过网络信息收集、整合、保存和服务四个步骤。
- 网络信息收集：按照用户指定的信息需求或主题，调用各种搜索引擎进行网页搜索和数据挖掘，将搜索的信息经过滤等处理过程剔除无关信息，从而完成网络信息资源的“收集”。

选择所需信息

- 1.要明确自己收集信息的目的，做到心中有数
- 2.从各个层次、各个角度、全方位的选择

使信息数量从多到少，信息质量从粗浅到精确，信息系统从杂乱无章到严密有序。





Section 2: Information description and processing

信 息(*Information*)

- 可以被看成是物质的一种属性，是对客观世界中各种事物的变化和特征的反应；是客观事物之间相互作用和联系的表现；是客观事物经过感知或认识的再现。
- 信即信号，息即消息，信息就是通过信号传递的消息。

信息的特性 Characteristics

- **客观普遍性：**信息无论是否被感知，它都是客观存在的；
- **依附性：**信息存在于口述、书面、广播、电视、存储设备、网络等载体中；
- **可传递性：**信息通过传递，产生作用，体现价值；
- **时效性：**信息有时效性，在一定时期内有效；
- **共享性：**信息可以被复制，为众人所拥有，共同享用；
- **可转换性：**信息可以在多种载体符号中进行转换；
- **可加工性：**信息可以进行加工，经过汇总、整理、归纳，去粗存精；
- **可存储性：**信息可存储于多种载体中。

信息素养 Information literacy

- **Definition：** 是一个内容丰富的概念。它不仅包括利用信息工具和信息资源的能力，还包括选择、获取、识别信息， 加工、处理、传递信息并创造信息的能力。

信息源 Information source

- **Definition：** 组织或个人为满足其信息需要而获得信息的来源。
- **Classification：**

  口头型信息源、实物型信息源、文献型信息源、
  电子型信息源、网络信息源。

信息源 Information source

- **Classification:**
- 1.**实物型信息源**，又称现场信息源，是指具体的观察对象在运动过程中直接产生的有关信息，包括事物运动现场、学术讨论会、展览会等。
- 2.**文献型信息源**，主要是指承载着系统的知识信息的各种载体信息源，包括图书、报纸、期刊、专利文献、学位论文、公文等。
- 3.**电子型信息源**是指通过使用电子技术实现信息传播的信息源，包括广播、电视、电子刊物等。
- 4.**网络信息源**是一种比较特殊的信息源，是指蕴藏在计算机网络，特别是因特网中的有关信息而形成的信息源。

信息、知识、文献以及情报之间的关系

- **Definition**

信息： 生物及具有自动控制系统的机器，通过感觉器官和相应的设备与外界进行交换的一切内容。

知识：人类对各种大量信息进行思维分析，加工提炼，并加以系统和深化而形成的结果。包括经验知识与理论知识。

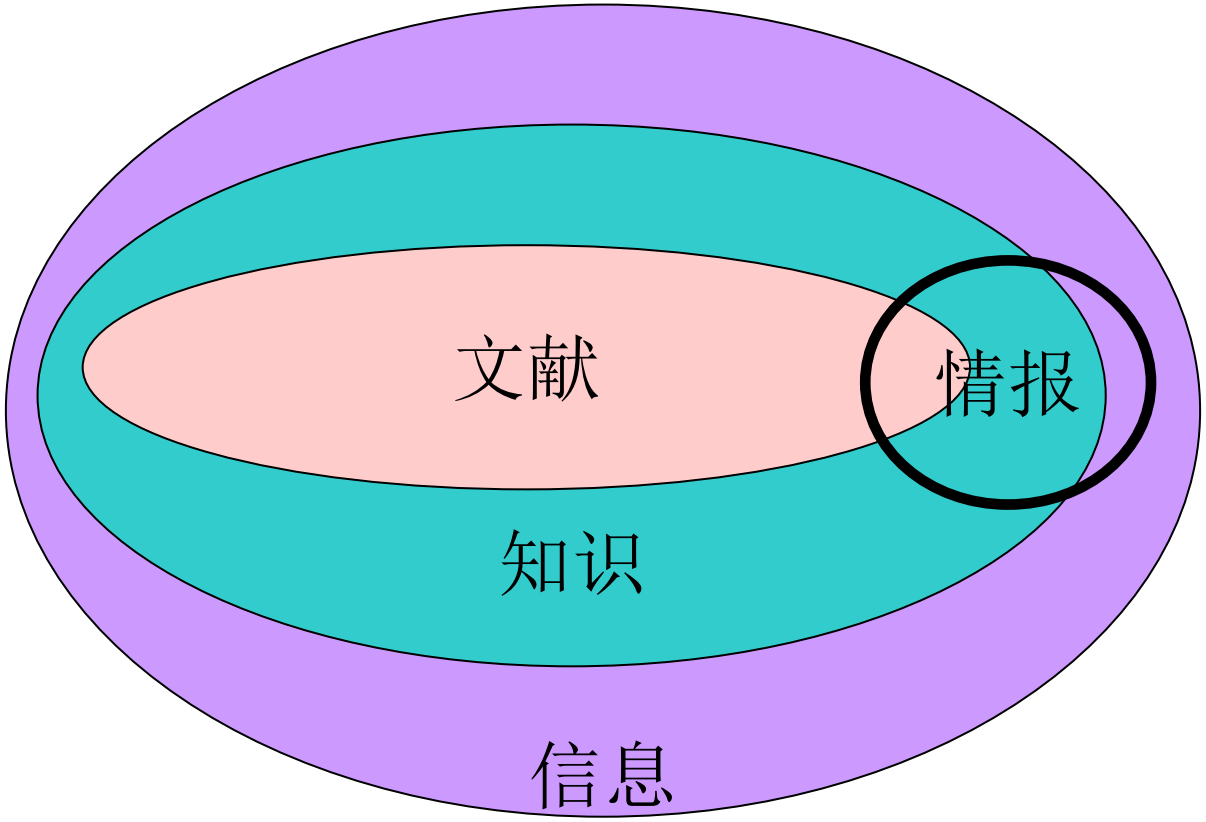
文献：记录有知识的一切载体，是传递知识和信息的工具。

情报：是指能为我们所用的知识和信息。

☞ Relationship：

知识来源于信息，理性化、优化和系统化了的信息；文献是它们的载体；知识和情报同属于信息的范畴，知识指的人类获得的自然界或人类社会信息。

信息、知识、文献以及情报之间的关系



一、信息的著录加工

1、著录目的：是把一篇文献变成一条著录，压缩后必须能体现文献的外表特征和内容特征。

- **外表特征：**指文献上显而易见的，一般情况下不反映文献实质意义的那些特征，如篇名、人名、各种符号标识（专利号，标准号，文献号等）、机构名等。
- **内容特征：**指表征文献实质意义的特征，如主题词（叙词，单元词、关键词）、分类号、化学符号等。

2、著录内容

外表特征：篇名，作者，工作单位，号码，文种等

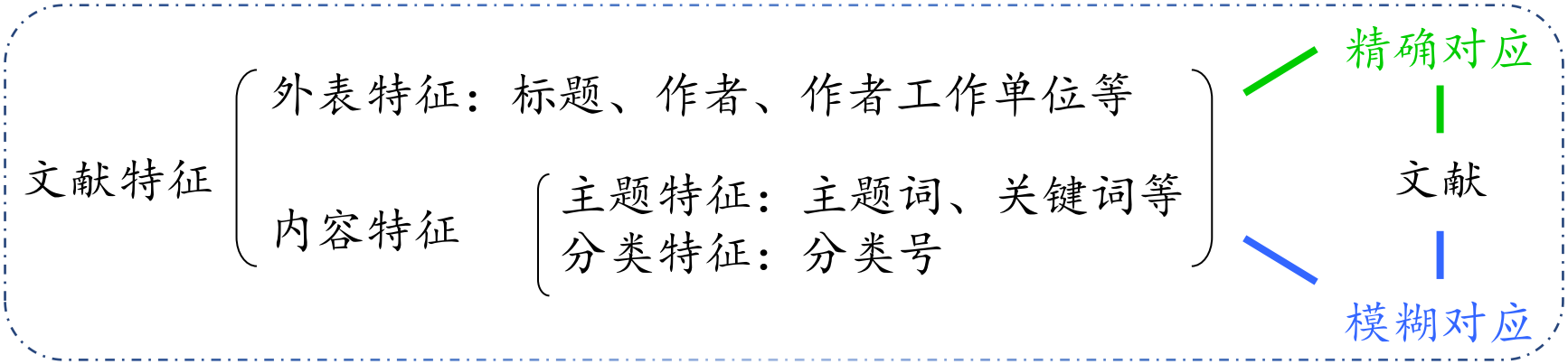
出处：发表在什么刊物，刊号，卷，期，页数等

内容特征：摘要等

3、著录格式（上一章有详细介绍）

文献特征与文献的对应关系

- 文献的外表特征与文献是一一对应的，即一组外表特征只对应一篇唯一的文献
- 文献的内容特征与文献却是一种模糊的对应关系，即一篇文献有多个主题词（关键词）或分类号，一个主题词（关键词）或分类号也可对应多篇甚至几百篇文献
- 利用外表特征只能检出较少的文献，有时只用于特定情况下（如已经知道作者名等）。利用内容特征一次则能检出一批文献。



二、信息的标引加工

- **Definition:** 把文献的主要内容用非常简明的标识（即标志）表示出来。标识可以是号码（分类号），也可以是科技名词或词组（主题词），也可以是其它的。
- **Characteristics**

  1）相同内容的文献集中在一起，不同内容的区分开来；

  2）形成有序的序列，即按一定规律排列，把存储进检索工具的著录按照一定的规律排列起来，形成有序的排检系统，这样可以提供检索途径。

- **Classification**

  分类法标引：用分类号作为标识；
  主题法标引：用代表文献主题内容的实质性的词汇作为标识。

三、信息的结构编排

- 一般有三种排列方式

1、按编码顺序排序：一条著录给一个顺序编码，且该编码唯一，其中编码可以表示存储地址，但体现不出文献的逻辑内容；

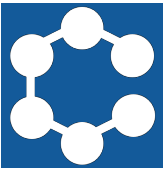
2、分类编排：按分类号的顺序；

3、按主题词的字母顺序。

例如：18位的数字组成的我国公民身份证号码

330382197702010101

户籍所在的省、市、地区信息　出生日期的信息　序列号与校验码



Section 3: Information indexing

信息索引技术

- **Introduction**
  - 信息索引技术是信息检索的关键技术，它的性能直接影响到搜索结果的准确性。
  - 通用信息索引的建立包括分析、索引和排序三个步骤。
  - 典型的信息索引技术有顺排索引技术和倒排索引技术，其中，倒排索引技术是信息检索系统中最普遍使用的索引机制。

信息索引的建立

- **Information index establishment**

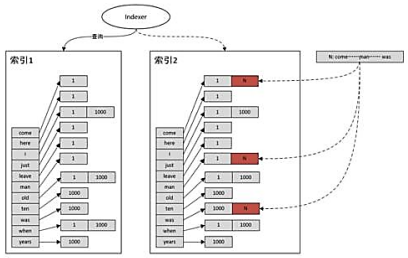
信息经过网页预处理后，可以建立索引数据库。但对于数目庞大的文档数据库使用简单匹配方法是不可行的，需要对文档的表示建立索引。为了提高检索效率，应该按照一定的规则建立索引。

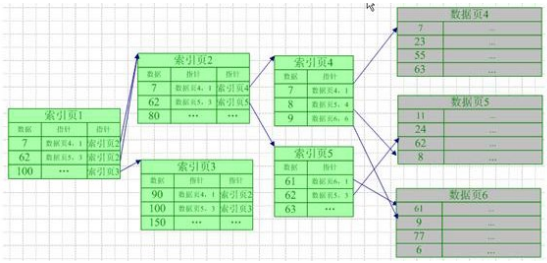
- **Steps**

(1) 分析：处理文件中可能的错误；

(2) 索引：完成分析的文件被编码存入索引数据库；

(3) 排序：将索引数据库按照一定的规则排序，产生全文索引。





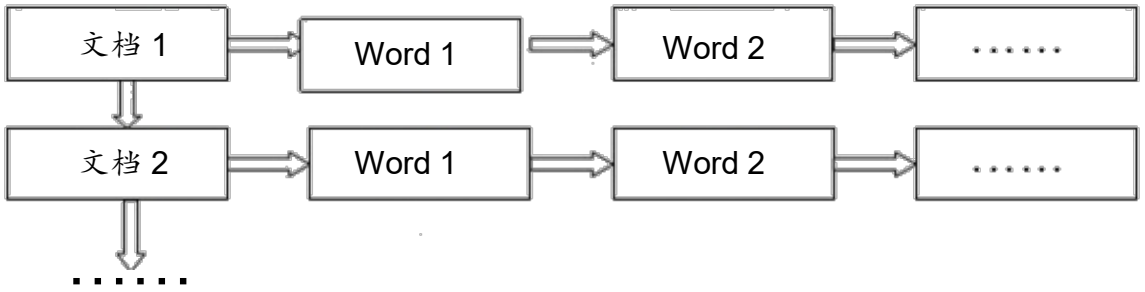
顺排索引 Forward index

- **Definition**：顺排索引用文档中的记录一条一条去匹配提问，是顺序对文档记录检索的方法，所以也称为顺排文档检索。
- **Tool**：正排表——顺排索引的关键技术

  以文档的ID为关键字，表中记录文档中每个字的位置信息，查找时扫描表中每个文档中字的信息直到找出所有包含查询关键字的文档。
- **Characteristics**

  Advantage：结构比较简单，建立比较方便且易于维护；

  Disadvantage：查询时需对所有的文档进行扫描，检索时间长，检索效率低下。

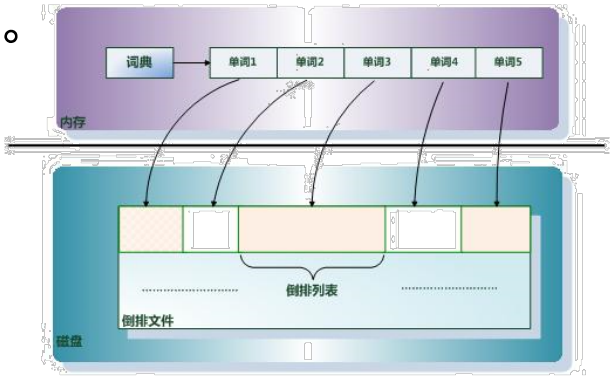


倒排索引 Inverted index

- **Definition：** 倒排索引是一种面向单词的索引机制，是将顺排文档中可检索字段的作者名、关健词、分类号等取出，按一定规则排序，归并相同词汇，并把在顺排文档中相关记录的记录号集合赋予其后，以保证通过某一特征词能够快速、方便地获取相关记录。

- **Tool：** 倒排表

  以字或词为关键字进行索引，表中关键字所对应的记录表项记录了出现这个字或词的所有文档，一个表项就是一个字表段，它记录该文档的ID和字符在该文档中出现的位置情况。



- **Characteristics**

  Advantage：效率高；
  Disadvantage：建立和维护都较为复杂。

建立倒排文档

- **Composition：** 关键字(作者、主题词、分类号等)、目长(含有该关键字记录的条数)、记录号集合(所有与该关键字有关的记录号)。
- **Establish：**

  倒排文档的建立是在顺排文档的基础之上提取可检索字段内容，或者采取自动从标题、文摘或全文中提取关键词，利用所得到的这些属性词来建立倒排文档。
- **Steps：**

  1）选择需要做索引的字段属性(如作者、关键词等)，抽出其中的内容，并在其后附上其记录号；
  2）对抽出的内容进行排序，使之便于归并相同内容；
  3）对相同内容进行归并，把合并后的内容放入倒排文档的主键字段(如标引词、作者等)，统计每一数据的频次作为目长，把每一内容后的记录号顺序放在记录号集合字段。

顺排索引与倒排索引

- 正排索引：文档 ---> 单词
- 倒排索引：单词 ---> 文档

文档ID	文档内容
1	elasticsearch是最流行的搜索引擎
2	php 是世界上最好的语言
3	搜索引擎是如何诞生的

单词	文档 ID列表
elasticsearch	1
流行	1
搜索引擎	1,3
php	2
世界	2
最好	2
语言	2
如何	3
诞生	3

索引的建立——简单法

- **Steps:**

  （1）将文档分析称单词term标记；

  （2）使用hash去重单词term；

  （3）对单词生成倒排列表。

  倒排列表就是文档编号DocID，没有包含其他的信息(如词频，单词位置等)，这就是简单的索引。

- **Advantage:** 简单索引可以用于小数据，例如索引几千个文档。
- **Disadvantages:**

  （1）需要有足够的内存来存储倒排表，对'J几搜索引擎来说，都是G级别数据，特别是当规模不断扩大时，难以提供这么多的内存。

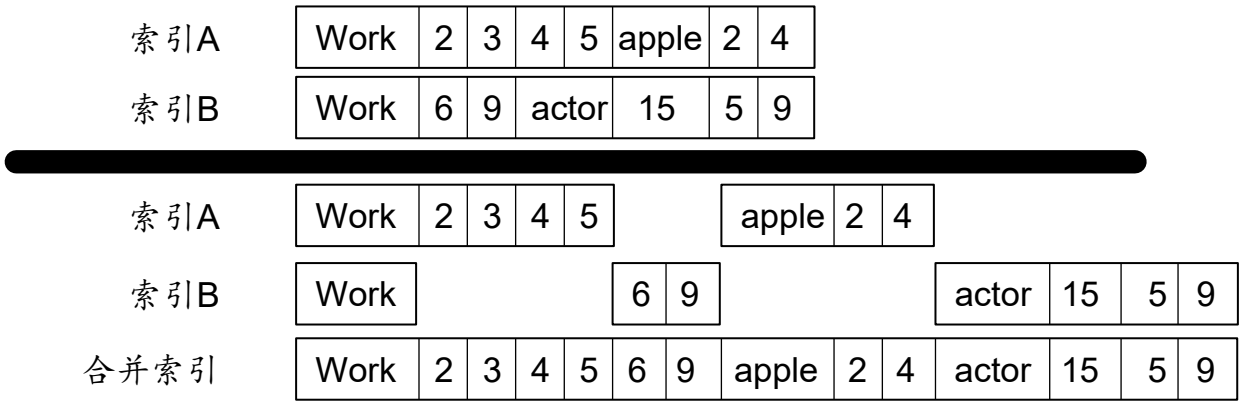
  （2）算法是顺序执行，不便于并行处理。

索引的建立——合并法

- **Steps:**

1）页面分析：生成临时倒排数据索引A、B，当临时倒排数据索引A、B占满内存后，将内存索引A、 B写入临时文件生成临时倒排文件；

2）对生成的多个临时倒排文件，执行多路归并，输出得到最终的倒排文件。



1. 搜索并下载导师的硕士/博士论文，并对其进行文献特征分析；
2. 寻找生活中的信息索引，拍照并留存（不少于3张，照片附上拍照日期）。
3. 检索自己感兴趣的老师的个人履历，不限于教育经历、工作经历、研究方向、联系邮箱等。