1、信息检索就是从信息集合中找出与用户需求相关的信息。被检索的信息除了文本外，还有图像、音频、视频等多媒体信息，我们只讨论文本的检索。

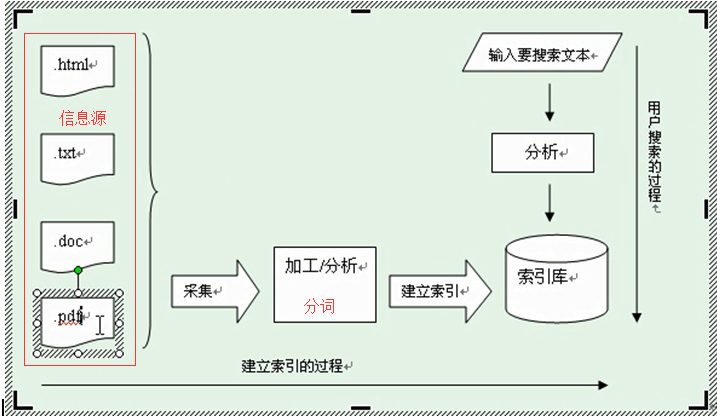
2、目前信息检索技术可分为3类：

1) 全文检索：只检索文本，把用户的查询请求和全文中的每一个词进行比较，不考虑查询请求与文本上的匹配。在信息检索工具中，全文检索是最具通用性和实用性的（我们只学习全文检索）。

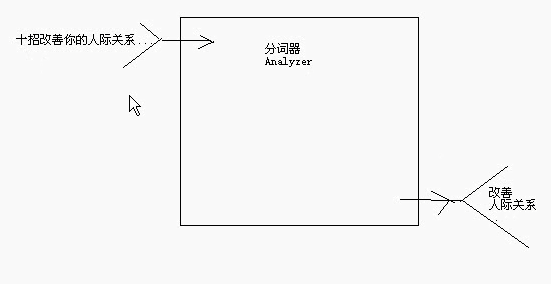
2) 数据检索：查询要求和信息系统中的数据遵循一定的格式，具有一定的结构，允许对特定的字段进行检索。其性能与使用有很大的局限性，并且支持语义匹配的能力也较差。

3) 知识检索：强调的是基于知识的、语义的匹配。

3、从信息源中拷到本地并进行加工，称为索引库(信息集合)，搜索时从本地的信息集合中进行搜索

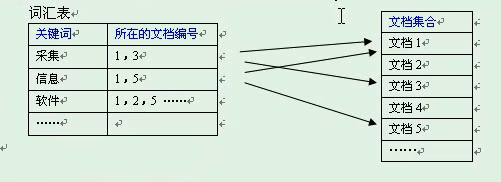


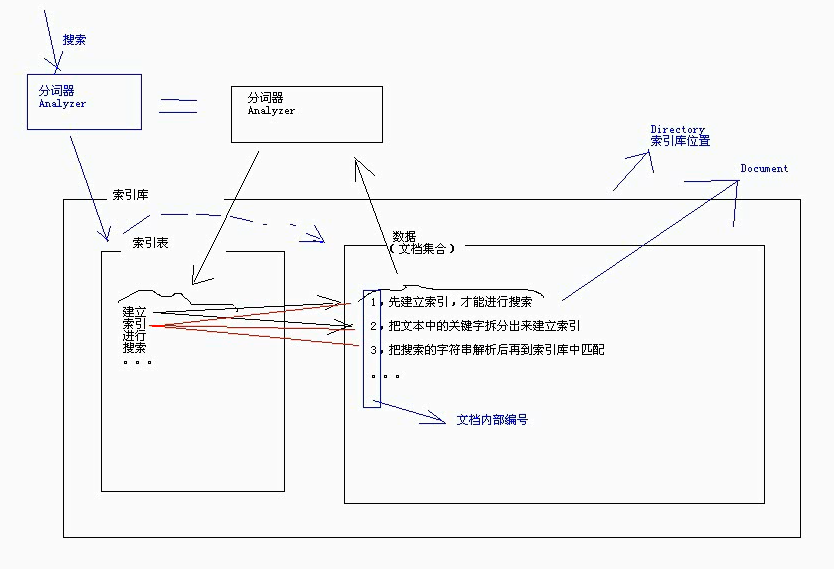
4、分词器(Analyzer)：将一句话分成几个单词。文本在建立索引和搜索时都会先使用分词器进行分词(注意：建立索引和搜索时需要用到同一个分词器)。



5、索引结构。

倒排索引，索引对象是文档中的单词，用来存储这些单词在一个文档中的位置。例如，有些书在最后提供的索引(“单词--页码”的对应表)，就可以看成是一种倒排索引。可以通过一些关键字，在全书中检索出与之相关的一部分。





6、在Lucene中索引库是一组文件的集合。Directory代表索引库的位置。里面的每一个数据都是一个Document，它是一个Field的集合。

7、是否进行索引

