搜索引擎每天都是在处理一个基本的需求：用户搜索一个关键词，搜索引擎马上找到相关的网页给用户。这个过程要怎么实现呢？

首先搜索引擎要尽可能多的把互联网上的网页搜集下来，这样能提供大量的网页给用户查询。这一部分由爬虫来解决，顺着互联网上的链接一个个往下抓取。最后就有了一堆记录着网页各种信息的资料库。

然后，搜索引擎拿到用户输入的这个关键词后，要从这个资料库里把相关的网页找出来给用户。这里就碰到好几个问题了：

1，要怎么快速的从千亿级的数量中找出匹配的网页的呢？

解决办法也倒简单，就是建立一份索引库。就像我们查《新华字典》一样，我们不会翻遍《新华字典》的每一页来查那个字在哪页，而是先去索引表那里找这个字，拿到页码后，直接翻到那页就可以了。搜索引擎也会为上百亿的网页建立一个索引库，用户查询信息的时候，是先到搜索引库里查一下要找的信息在哪些网页，然后就引导你去那些网页的。

（ppt出图）

2，索引库里用什么样的分类方式？

我们知道，《新华字典》的索引表是用字母列表或者偏旁部首的分类方式的。那么搜索引擎的索引库里是怎么分类的？是不是也可以用字母列表的方式？

最后找到一个解决办法：索引库里用词语来分类

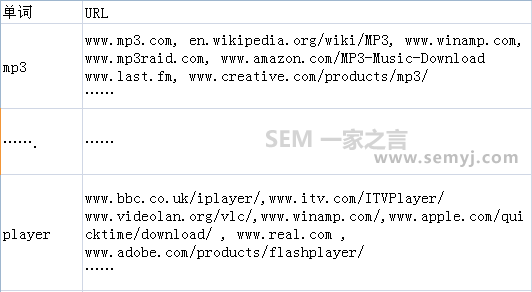
假设有100亿网页

搜索引擎如果以字母列表的方式排列索引库，那么平均每个字母下要查询的网页数量是  100亿÷26＝3.85亿 ，也还是一个很大的数字。而且搜索引擎上，今天是100亿个网页，过不了多久就是300亿个网页了。

因为尽管互联网上的网页是不断激增的，但是每一种语言里，词语的数量都是相对固定的。比如英语就是一百多万个单词，100亿 ÷ 1百万 =  1 万；汉语是8万多个词语，100亿÷8万=12万5千。都是计算机很容易处理得过来的。

用词语来分类还有一个好处，就是可以匹配用户查询的那个词语。本来用户就是要查这个词语的，那我就按这个词语去分类就是。

所以，搜索引擎的索引库，最后就是这个样子的：



当搜索引擎把一个网站抓取下来后，接着要做的事情就是把网页里的词语分开放到索引库里。分词在这个时候就要应用到了，所谓的分词，其实很简单，就是把词语分开而已。

英语的分词好处理一点，因为英语的每个单词之间是用空格分开的，基本上就只要处理一些虚词、介词，还有一些词语的单复数，变形词等等。

但是中文的分词就复杂很多了，句子中的每个字都连在一起，有时候即使是人来判断，都还有产生歧义的时候。

以百度中文分词为例，介绍：

字符串匹配的分词方法

这是种常用的分词法，百度就是用此类分词。

（1）正向最大匹配法

就是把一个词从左至右来分词。

举个例子：”不知道你在说什么”

这句话采用正向最大匹配法是如何分的呢？“不知道，你，在，说什么”。

（2）反向最大匹配法

"不知道你在说什么"反向最大匹配法来分上面这段是如何分的。“不，知道，你在，说，什么”，这个就分的比较多了，反向最大匹配法就是从右至左。

（3）最短路径分词法。

就是说一段话里面要求切出的词数是最少的。

“不知道你在说什么”最短路径分词法就是指，把上面那句话分成的词要是最少的。“不知道，你在，说什么”，这就是最短路径分词法，分出来就只有3个词了。

THANK YOU