搜索引擎是如何快速找到相关内容的

【马少平，王东】

在互联网上冲浪离不开搜索引擎，它可以帮助我们快速找到需要的内容。这是如何做到的呢？要知道，互联网上的内容多的吓人，想要在不到 1 秒钟内就把想要的东西找到，还同时服务大量的并发请求，想想都觉得困难。

我们通过一个小例子来理解搜索引擎背后的算法。假设有 1 万名学生，每个学生可以参加几个兴趣小组，每个兴趣小组由几十名学生组成。我们允许有大量的兴趣小组存在。现在我们想知道：对于任意给定的两个同学，如何找到他们共同参加的兴趣小组？

一种方法是，遍历查找每个兴趣小组，看看这两位同学是否在同一个小组里，并把他们共同的小组查找出来。由于兴趣小组数实在太多，这种方法会显然非常的慢，效率很低。
第二种方法，如果我们事先做好一个索引，记录好每个同学所参加的兴趣小组。这样根据同学的姓名，我们马上就知道两位同学各自参加了哪些兴趣小组，从而把共同参加的兴趣小组找到。由于每个同学参加的兴趣小组数并不是太多，这种方法速度要快的多。上述由学生到兴趣小组的索引称为“倒排索引。

回到搜索引擎上来。我们知道每篇文章是由一些单词及特殊的词串组成，这些单词和词串我们统称为词项。将词项当成是学生，而文章当作兴趣小组，并建立一个倒排索引来记录每个词项所在的文章。当用户向搜索引擎输入查询词时，搜索引擎把查询词划分为词项，再根据建立好的倒排索引找到包含所有查询词的文章，并按照一些原则计算出查询词与这些文章的相关性，最后按相关性大小排序后作为搜索引擎的输出。

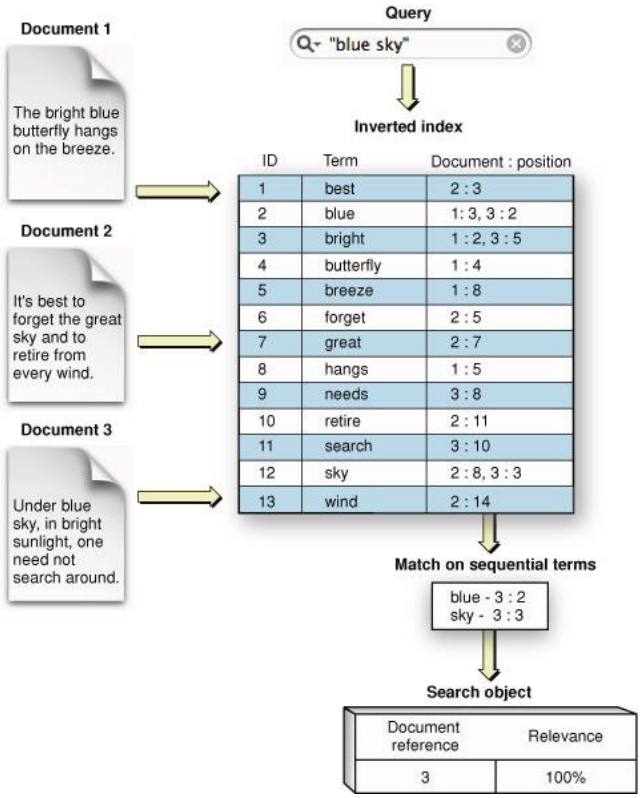

图 1：基于倒排索引的文章搜索过程

在实际搜索引擎中，倒排索引除了记录词项所在的文章外，还记录词项在该文章中的位置，以及词项的一些属性，比如是否在标题中、是否高亮词、字体大小等，这些属性可以用在相关性计算中，用以提高输出排序的合理性。