**ES倒排索引介绍**

## 【说明】仅记录实验室测试过程，不作为官方文档使用，可能会有很多地方未能

## 验证，因此无法进行技术兜底，需使用方多加验证测试，涉及到高危需走变更。

## 倒排索引

索引是构成搜索引擎的核心技术之一，索引在日常生活中其实也是非常常见的，比如当我们看一本书的时候，我们首先会看书的目录，通过目录可以快速定位到某一章节的页码，加快对内容的查询速度。

倒排索引，也常被称为反向索引，是一种索引方法，被用来存储在全文搜索下某个单词在一个文档或者一组文档中的存储位置的映射，它是文档检索系统中最常用的数据结构。

下面我们通过具体实例深入理解倒排索引，通过简单文档以小见大，体验倒排索引的建过程。

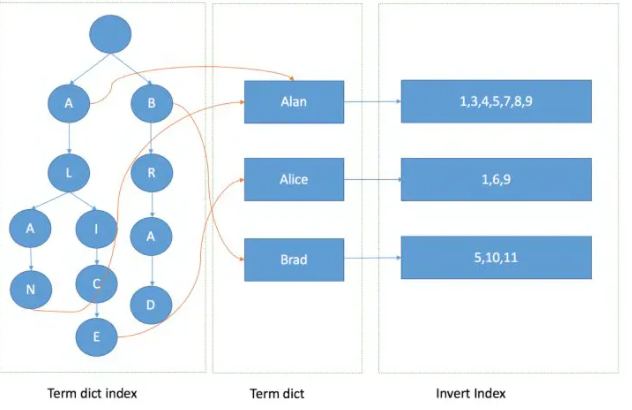
| 文档ID | 文档内容 |
| --- | --- |
| 1 | 人工智能成为互联网大会焦点 |
| 2 | 谷歌推出开源人工智能系统工具 |
| 3 | 互联网的未来在人工智能 |
| 4 | 谷歌开源机器学习工具 |

对于文档内容，先要经过词条化处理。与英文不同的是，英文通过空格分隔单词，中文的词与词之间没有明确的分隔符号，经过分词系统进行中文分词以后把矩阵切分成一个个的词条。

| 词项 | 文档频率 | 倒排记录表 |
| --- | --- | --- |
| 人工 | 3 | 1,2,3 |
| 智能 | 3 | 1,2,3 |
| 成为 | 1 | 1 |
| 互联网 | 2 | 1,3 |
| 大会 | 1 | 1 |
| 焦点 | 1 | 1 |
| 谷歌 | 2 | 2,4 |
| 推出 | 1 | 2 |
| 开源 | 2 | 2,4 |
| 系统 | 1 | 2 |
| 工具 | 2 | 2,4 |
| 的 | 1 | 3 |
| 未来 | 1 | 3 |
| 在 | 1 | 3 |
| 机器 | 1 | 4 |
| 学习 | 1 | 4 |

然而这套索引的实现，要上生产环境，那还远着。这个世界上那么多单词，中文、英文、日文、韩文 … 每次搜索一个单词，都要全局遍历一遍，很明显不行。

于是有了排序，我们需要对单词进行排序，像 B+ 树一样，可以在页里实现二分查找。



Lucene 的倒排索引，增加了最左边的一层「字典树」term index，它不存储所有的单词，只存储单词前缀，通过字典树找到单词所在的块，也就是单词的大概位置，再在块里二分查找，找到对应的单词，再找到单词对应的文档列表。

Lucene 的实现会要更加复杂，针对不同的数据结构采用不同的字典索引，使用了FST模型、BKDTree等结构。

真实的倒排记录也并非一个链表，而是采用了SkipList、BitSet等结构。