链滴

一行命令让ElasticSearch支持中文分词搜索

作者：BosonNLP

原文链接：https://ld246.com/article/1458887748087

来源网站：链滴

相信大家在开发博客，在线商城的时候会涉及到搜索功能。而近几年火起来的 ElasticSearch（ES）借其稳定、可靠、快速的实时搜索普遍受到大家的好评，连 Github、SoundCloud 也都将 ES 作为核心搜索组件。

但是 ES 本身对中文分词和搜索比较局限。因为内置的分析器在处理中文分词时，只有两种方式：一是单字（unigrams）形式，即简单粗暴的将中文的每一个汉字作为一个词（token）分开；另一种是字（bigrams）的，也就是任意相邻的两个汉字作为一个词分开。这两种方式都不能很好的满足现在中文分词需求，**进而影响了搜索结果。**

举个例子：

假设我们的 index 里面存储了3篇 documents 如下：

```
<table>
  <tr>
    <td>id</td><td>content</td>
  </tr>
  <tr>
    <td>1</td><td>美称中国武器商很神秘 花巨资海外参展却一言不发</td>
  </tr>
  <tr>
    <td>2</td><td>在第一界国际锦标赛中 国家代表李雷勇夺冠军</td>
  </tr>
  <tr>
    <td>3</td><td>国武公司近日上市</td>
  </tr>
</table>
```

● Case 1：查询“中国”，期望只得到 id 为1的 document。

用 unigram 的分析器（即默认的 Standard Analyzer）查询结果为 id 1和 id 2的content；bigram 分析器（名为cjk）的结果为id 1。Standard Analyzer 没有给出预期结果是因为它把“中国”切分为中”、“国”2个 token，因此误给出了 id 2的结果。

● Case 2：查询“国武”这一家公司，期望只得到 id 为3的 document。

Standard Analyzer 和 cjk 的查询结果都会同时给出 id 1和 id 3的 document，但是 id 1 的 documnt 中的“国武”并不是所指的公司。

（注：以上查询均用query_string）

因此我们可以发现内置的分析器有它的局限性，并不能满足复杂或者特定的搜索需求。为此，玻森数开发了一款基于玻森中文分词的 ES 插件（Elasticsearch-Analysis-BosonNLP），方便大家对中文据进行更精确的搜索。

现在已有一些成熟的 ES 中文分词插件，但在分词引擎准确率上，相信 BosonNLP 的中文分词能满足家不同领域上多样化的需求。有兴趣的朋友可以查看11款开放中文分词引擎大比拼。

接下来，3分钟教会大家如何安装使用玻森 ES 中文分词插件 Beta 版（以 ES 2.2.0 版本为例）：

● 安装

只需如下一个命令。

```
$ sudo bin/plugin install https://github.com/bosondata/elasticsearch-analysis-bosonnlp/relea
```

es/download/1.3.0-beta/elasticsearch-analysis-bosonnlp-1.3.0-beta.zip

注：对于其他不同版本的 ES，只需要在命令里更换对应的插件版本号即可。

- 使用

需要在elasticsearch.yml文件中的 analyzer 里配置好玻森 bosonnlp analyzer（需要配置 API_TOKE 以及分词的参数）。详情解释请查看 Github 上玻森 ES 中文分词的README。

```
bosonnlp:
    type: bosonnlp
    API_URL: http://api.bosonnlp.com/tag/analysis
    API_TOKEN: *PUT YOUR API TOKEN HERE*
```

完成以上步骤之后便可以使用玻森 ES 分词插件了。

对比之前 Case 2 的查询： 查询“国武”这一家公司，期望只得到 id 为3的 document。玻森ES分插件搜索结果：

```
{
  "took" : 70,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "failed" : 0
  },
  "hits" : {
     "total" : 1,
     "max_score" : 0.15342641,
     "hits" : [ {
        "_index" : "bosonnlp_test",
        "_type" : "text",
        "_id" : "3",
        "_score" : 0.15342641,
        "_source":
{
   "content":"国武公司近日上市"
}
    } ]
  }
}
```

当然，如果对分词有特定需求的小伙伴可以在配置里修改对应的参数。目前，玻森数据对于中文分词供了繁简转换、新词发现等功能，能满足不同领域的搜索需求。

希望这款插件能提升你的工作效率！

GitHub上有具体的说明。这里附上例子中索引 document 的 bash 文件以方便测试。