南开大学
Nankai University

南　　开　　大　　学

计 算 机 学 院

信息检索系统原理实验报告

邮件检索系统实现

学号：1911475

姓名：　王隽毅

年级：2019 级

专业：计算机科学与技术

2023 年 3 月 11 日

摘要

本次实验内容为基于 ElasticSearch 和 Python 开发环境，对安然公司 150 位用户的 50W 封电子邮件进行检索系统实现。在本地计算机上运行 Elasticsearch 工具，通过 Python 的第三方包连接 Elasticsearch 环境，构建索引；将下载到本地的开源电子邮件文本数据使用 Python 读取并进行 json 格式化，将数据储存入索引中，对邮件数据在服务器后端进行相应的检索并通过网页前端将结果呈现。

关键字：信息检索，Python，Elasticsearch，**索引构建，检索实现**

目录

一、 索引设计和构建 1

（一） 索引介绍 1

（二） 索引设计 1

（三） 索引构建 2

二、 检索功能实现 4

（一） 检索参数构建 4

（二） 查询系统前端 5

（三） 查询系统后端 6

三、 ES **功能探索** 7

（一） 倒排索引 7

（二） 向量空间模型 8

（三） 文本分类分析 9

四、 实验总结 9

一、 索引设计和构建

(一) 索引介绍

Elasticsearch 是一个开源的搜索引擎，使用 Java 编写的，它的内部使用 Apache Lucene 做索引与搜索。它的目的是使全文检索变得简单，通过隐藏 Lucene 的复杂性，取而代之的提供一套简单一致的 RESTful API。[2]

本次实验中使用 Python 的 Elasticsearch 包内函数连接运行的 Elasticsearch 环境进行数据的索引和检索。

一个 Elasticsearch 的集群可以包含多个索引，相应的每个索引可以包含多个类型。这些不同的类型存储着多个文档，每个文档又有多个属性。将这些概念类比于数据库，一个索引相当于一个数据库，一种类型相当于数据库中的一张表，有一个文档相当于表中的一个条目，每个属性相当于各个条目的域。Elasticsearch 依赖于建立的索引中储存的各个属性对应的值进行检索。

(二) 索引设计

在 Python 脚本中连接上 Elasticsearch 后首先要建立新的索引。调用 indices 对象的 *create*() 函数建立新索引。需要传递 json 格式的文件指示函数建立什么格式的索引。先设定索引的配置 (setting) 字段。

number_of_shards 指定索引主分片的数量。一个分片是一个底层的工作单元，它仅保存了全部数据中的一部分。片是数据的容器，文档保存在分片内，分片又被分配到集群内的各个节点里，在本实验中即保存在本机中；主分片的数目决定着索引能够保存的最大数据量。默认将分片数目设为 5。

number_of_replicas 指定索引副本分片的个数，默认值 1。一个副本分片只是一个主分片的拷贝。副本分片作为硬件故障时保护数据不丢失的冗余备份，并为搜索和返回文档等读操作提供服务。

接下来指定索引的映射字段的值。所谓映射 (mapping) 是用来定义文档及其字段的存储方式和索引方式，可以为文档的每个字段指定字段类型、应用的分析器和是否可被索引到。[1] 观察要存储的邮件文本的格式，在正文之前总共有 15 种属性记录邮件相关信息。分析此 15 种字段的特点，决定 ID 号、版本号、编码格式等设为关键词 (keyword) 类型，即存储的时候不用分词器进行分词而是整

个字符串进行存储，因此检索时也使用整个字符串进行比对匹配；对日期字段指定为 Date 类型；对进行检索时使用的字段寄件人、收件人、主题和正文以及其余字段使用 text 类型，即存储前使用分析器进行分词处理，再对每个单词建立倒排索引，检索时也对检索值分词匹配；由于源数据邮件文本皆为英文文本，故分析器直接应用标准分析器即可。

(三) 索引构建

构建好指示索引建立方法的字典后作为参数传给 *create*() 函数的 body 参数即可。

需要注意之处为：在 Elasticsearch5.0 及之前的版本中，在指示建立索引的 JSON 格式参数中可指明当前建立的索引中类型（表）的名字，而在 7.0 之后的版本移除映射类型，为了防止发生各表之间的干扰并提高压缩数据的能力，因此在 JSON 中不能在 mapping 下写上类型名称，否则会发生报错。

建立好空索引后向索引中插入数据。调用 *walk*() 函数遍历邮件数据文件夹下的每一个文件，打开文件后对文件文本进行处理：读取储存了基本信息的 15 种属性，分别将属性名和属性的文本储存，在将正文每行读取并储存；

需要注意之处是：在进行文件打开并处理时，要进行异常处理，由于数据集本身存在的一些问题，使用 UTF-8 编码进行解码时在某个别文件会产生解码错误，此时读取的函数直接返回，放弃读取此编码有问题文件；以及部分文件的 Date 属性的描述也不符合日期格式要求，同样在遭遇转化错误 (parser error) 时将此文件抛弃。

将储存的各属性标签和属性值组合成用于批量写入索引的 JSON 格式字典，其中指明了要写入的索引名和类型名，并返回存入各个文件的 JSON 字典的列表中；当列表中满 1000 个元素后便调用 *bulk*() 函数进行批量写入文件，等待一定时间将所有文件写入指定索引。至此，索引的设计和构建完成。在网页上访问 Elasticsearch 服务器监察索引构建结果。

current: 最新动态
cur_url: https://cc.nankai.edu.cn/13291/list.htm
sequence: 1
title: 我校与百度公司签署人工智能人才培养合作协议
author: 孟璐
date: 2020-11-06
url: https://cc.nankai.edu.cn/2020/1106/c13291a315874/page.htm
main:
（通讯员：孟璐）11月4日，百度公司—南开大学人工智能人才培养合作协议签约仪式在计算机
副主任计湘婷、校企合作经理刘磊，百度飞桨运营负责人周奇出席。计算机学院部分教师参加签
我校与百度公司具有长期深厚的合作基础，在人才培养、学术研究、成果转化等方面已取得显著
进一步推动双方建立更广领域、更深层次的合作关系。李轩涯表示，百度将在案例共享、课程共
计划，致力于人工智能优秀人才的培养。
根据合作协议，双方将制定和实施人工智能“学、练、测、评”一体化创新实践计划，通过务实
新和人才输出平台，树立校企合作的典范。
sequence: 2
title: 计算机学院和网络空间安全学院开展 “筑牢理想信念 立志科研报国”实践交流活动
author: 高雨桐
date: 2020-11-06
url: https://cc.nankai.edu.cn/2020/1106/c13291a315864/page.htm
main:
sequence: 3
title: 我院两位教授获得2019年度高校计算机专业优秀教师奖
author: 祁晓飞
date: 2020-11-04
url: https://cc.nankai.edu.cn/2020/1104/c13291a315331/page.htm
main:
日前，中国教师发展基金会发来《关于邀请参加高校计算机专业优秀教师奖励计划颁奖典礼的函
金砖四国国际会议中心举办的2020中国计算机教育大会。
据悉，高校计算机专业优秀教师奖励计划是在教育部、国家自然科学基金委指导下，由图灵奖获
事长、中国工程院院士高文教授共同发起，由部分具有重要社会影响力的高科技企业共同出资设
教师，引导高校教师潜心投入本科教学，改革和探索教学方法，促进高校教师更加重视本科课堂
进行。
sequence: 4
title: 中国新一代人工智能发展战略研究院智能网络安全研究中心南开揭牌

图 1: 索引构建结果

二、 检索功能实现

(一) 检索参数构建

在 Elasticsearch 上进行检索要传递检索参数即查询表达式，同样是 JSON 格式字典，描述执行的检索类型和检索的属性。对于本次实验中可检索的四个属性，由于其都是 text 类型，故多数情况下可应用的检索类型即为 match 类型，为标准的查询类型；执行查询前，它将用正确的分析器去分析查询字符串，将分词得到各单词与待查询属性上建立的倒排索引上进行查找，再根据向量空间模型原理进行分数的计算，查询后结果显示按分数从高到低排序的前 10 个结果。在 keyword 等类型的属性上还可进行 term 精确查询，对查询字符串不分词，直接对原始文本进行精确值匹配；范围操作 range，找出落在指定区间内的数字或者时间；以及聚合操作，统计关键词属性上不同的属性的出现次数。

本次实验中由于要实现同时查询多个属性字段，故使用 bool 多元查询，其有以下参数：

1. must 文档必须匹配这些条件才能被包含进来
2. must_not 文档必须不匹配这些条件才能被包含进来
3. should 如果满足这些语句中的任意语句，将增加 _score，否则，无任何影响。它们主要用于修正每个文档的相关性得分
4. filter 必须匹配，但它以不评分、过滤模式来进行。这些语句对评分没有贡献，只是根据过滤标准来排除或包含文档

本次实验中使用 must 参数即可。在 must 键下添加要查询的 match 键值对即可。本次实验实现高亮显示搜索结果中匹配字段的功能，通过在查询表达式中插入 highlight 键值对即可。

总共 4 个可查询字段，每次确定有一个字段有查询需要，则在 JSON 字典中的 must 和 highlight 键下插入此属性名的键值对，若最后一个查询字段也没有，则在表达式中插入 match_all 键和空值，则默认查询所有结果。

(二) 查询系统前端

本次实验搭建了一个具有后端服务器和前端网页交互的查询系统。其中前端使用 HTML 编写的网页和基于 Javascript 的 Jquery 和 AJAX 对网页进行无需刷新的动态变化和与后端服务器传递信息。

网页上有 4 个输入文本框和提交按钮。在输入框中键入或不键入文本后点击提交按钮触发函数响应，函数内部用 AJAX 封装，实现按照参数指定的自动访问对应网页和按对应方式发送对应数据；AJAX 将输入框中的值发送给后端 Python 的 flask 服务器接受处理后返回搜索结果；AJAX 成功通信后接收到返回值字典取得各个键对应的值取得查询结果的原始数据全文或查询的高亮匹配域；根据网页上的单选框选择的值决定是只显示匹配查询值得查询字段或是查询结果的原始邮件文本全文；

网页界面如图：


```
{
  "took" : 3,
  "timed_out" : false,
  "_shards" : {
    "total" : 3,
    "successful" : 3,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 1149,
      "relation" : "eq"
    },
    "max_score" : 1.0,
    "hits" : [
      {
        "_index" : "database",
        "_type" : "_doc",
        "_id" : "CeHQqH0BATsytt2kuZMK",
        "_score" : 1.0,
        "_source" : {
          "current" : "最新动态",
          "cur_url" : "https://cc.nankai.edu.cn/13291/list.htm",
          "sequence" : "89",
          "title" : "我院获批天津市网络与数据安全技术重点实验室",
          "author" : "祁晓飞",
          "date" : "2018-12-28 00:00:00",
          "url" : "https://cc.nankai.edu.cn/2018/1228/c13291a147710/page.htm",
          "main" : "日前，天津市科技局与天津市教委联合发布天津市重点实验室认定名单，我院获批天津市网络与数据安全技术重点实验
        }
      },
      {
        "_index" : "database",
        "_type" : "_doc",
        "_id" : "DeHQqH0BATsytt2kupNS",
        "_score" : 1.0,
        "_source" : {
          "current" : "最新动态",
          "cur_url" : "https://cc.nankai.edu.cn/13291/list.htm",
          "sequence" : "93",
          "title" : "抢抓机遇 打造计算机学科新高地 ——南开大学计算机学科在智能计算领域取得系列进展",
          "author" : "刘菲",
          "date" : "2018-12-14 00:00:00",
          "url" : "https://cc.nankai.edu.cn/2018/1214/c13291a147700/page.htm",
          "main" : "南开新闻网记者 马超 乔仁铭\n华为Mate 10系列手机曾以其“大光圈智能拍照”功能赢得良好口碑，在国外DxOMark
程明明曾介绍，“显著性物体检测技术”对算法性能要求极高，其“幕后”的支撑研究是“图像内容理解与智能处理”，该研究几乎能与所有
多像程明明这样既瞄准世界科技前沿，又致力于技术应用的团队不断涌现出来，他们在科学研究、成果发表和课题申报等方面取得突破性进展
重大战略需求，同时将技术的福祉带进千家万户，真正实现计算机学科引领航向、创新发展。”计算机学院院长袁晓洁说。\n科研，对标前沿
术与实体经济的深度融合已成为国家的重大战略。2017年7月，国务院印发《新一代人工智能发展规划》，明确分析了当前的战略态势：人工
共中央政治局就人工智能发展现状和趋势举行的集体学习会议上，习近平总书记强调，人工智能是新一轮科技革命和产业变革的重要驱动力量
```


图 2: 查询网页界面

(三) 查询系统后端

其中后端使用 Python 脚本，建立服务器的功能使用第三方包 flask 实现。初始化一个 flask 变量，指定根目录为渲染显示查询网页的 HTML 文件；用于和前端 AJAX 通信的 submit 目录用于接收前端传输的数据，处理后作为调用的查询函数的参数；在查询参数中若该字段有查询值则插入到查询表达式的 must 和 highlight 键下，若无则不执行插入；执行 *search*() 函数与 Elasticsearch 通信执行查询操作；将查询得到的结果返回前端网页。查询示例网页界面如图：

```python
def GtoM(G, N):
    M = np.zeros((N, N))
    for i in range(N):
        D_i = sum(G[i])
        if D_i == 0:
            continue
        for j in range(N):
            M[j][i] = G[i][j] / D_i
    return M


def PageRank(M, N, T=300, eps=1e-6, beta=0.8):
    R = np.ones(N) / N
    teleport = np.ones(N) / N
    for time in range(T):
        R_new = beta * np.dot(M, R) + (1-beta)*teleport
        if np.linalg.norm(R_new - R) < eps:
            break
        R = R_new.copy()
    return R_new
```

图 3: 查询示例网页界面 1

图 4: 查询示例网页界面 2

如果要进行附件查询，由于邮件文本没有独立的附件域，所谓附件仅在正文中提到“带有某附件”之类的语句文本，则在正文查询的输入框中输入要查询的附件类型的文本，会显示正文中提到相应附件的邮件。

三、 ES 功能探索

(一) 倒排索引

Elasticsearch 使用一种称为倒排索引的结构，它适用于快速的全文搜索。一个倒排索引由文档中所有不重复词的列表构成，对于其中每个词，有一个包含它的文档列表。为了创建倒排索引，首先将每个文档的各属性的内容域拆分成单独的词 (tokens)，创建一个包含所有不重复词条的排序列表，然后列出每个词条出现在哪个文档。后续的检索便全部基于此倒排索引，但仅有此倒排索引不足以返回令用户满意的结果，需要一个评判标准决定在庞大的匹配文件中优先返回哪个文件。

(二) 向量空间模型

在 Elasticsearch 会对每个查询匹配结果进行评分，分数值即充当了查询结果匹配度评判的标准。在早期的版本中，使用的标准的算法是 Term Frequency/Inverse Document Frequency 即 (TF/IDF) 算法。分别计算每个词项对于每个文档的词项频率（某单个关键词在某个文档的某字段中出现的频率次数，各词条在整个 index 的所有 document 中出现的次数越多，则权重越高）和逆文本频率（统计各词条在所有文档中出现的次数，出现的次数越多，词条的特性越弱，该词条在后续用于评定相关度分数时，起到的作用也越低），根据统计得到两个数值 (TF/IDF）综合得到每一个词项在各个文档中的权重，形成一个文档的词项权重向量，再对查询条件也计算其文档权重向量，比较两个向量的相似度，向量空间模型认为，若两个向量之间的夹角越小，则相似度越高。根据相似度得到一个文档最终的分数值。后续版本更新的评分算法也是基于 TF/IDF 算法的改进版。

调用 explain 函数可查看具体的查询过程的分数计算公式。对其中一个文档进行查询结果解释请求返回值为：

图 5: 解释分数值文本

(三) 文本分类分析

根据邮件的文本内容可发挥 Python 服务端的优势对原始数据应用机器学习相关算法进行垃圾邮件分类或文本情感分析，但由于原始数据不具有分类标签，故目前无法进行分类学习，后续对数据集完善后可进行机器学习。

四、 实验总结

本次实验练习了 Elasticsearch 的使用相关操作，掌握了索引设计建立和文档检索的实现方法，并构建了基本功能较为完善的具有完整前后端的检索系统

实现输入检索条件返回检索结果，对后续进一步独立完善检索系统打好基础。

参考文献

[1] Es 中 mapping 详解. https://www.cnblogs.com/haixiang/p/12040272.html.

[2] Es 指导手册. https://www.elastic.co/guide/cn/elasticsearch/guide/current/index.html.