Foundation of Information Retrieval

Course: 2020-201600076-1A, Final Project, Group: 11, Submission Date: November 13,2020

Qiuke Li
q.li-7@student.utwente.nl
University of Twente
Enschede, The Netherlands

Sathvik Guru Rao
s.gururao@student.utwente.nl
University of Twente
Enschede, The Netherlands

ABSTRACT

Information Retrieval system is used in finding information for the users query. A document will be returned for the query which has the information. This document is called the relevant document. The main aim of this project is to increase the number of relevant documents for each query. Elasticsearch is a text search engine which is used for complex search features and requirements. The performance of the system can be improved by using different IR models, analyzers etc. The proposed system compares models with normalization and without the normalization and BM25 was used as an example in this project.

KEYWORDS

Elasticsearch, Information Retrieval System, TREC Genomics dataset, BM25

ACM Reference Format:
Qiuke Li and Sathvik Guru Rao. 2020. Foundation of Information Retrieval : Course: 2020-201600076-1A, Final Project, Group: 11, Submission Date: November 13,2020 . In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Elasticsearch is a distributed full-text search and analytics engine built on the Lucene StandardAnalyzer [1]. The elasticsearch uses an index to search for responses to a user's query. An index is a data structure which stores documents that have similar characteristics. The documents used for Elasticsearch are in json file formats. Indices are divided into multiple shards which help in horizontally dividing the data volume which in turn help in increasing the performance.
Information Retrieval system is used to find the right information for a user's query. Elasticsearch uses the query to search for the right document. These documents are called the relevant documents. While searching for the documents there can be some documents which could be missed because of no normalization, synonyms, uppercase words and mismatch in query. In this project, the performance of the IR system is improved by using lowercase words , removing the stop words and stemming the words to get the root word of inflected words. In this way, the expansion of original words can be stemmed by removing the derivation and inflections and used in searching.

2 PROBLEM STATEMENT

The search engines use the query from user to match with the document. Sometimes , there will be some documents which could be missed because of case sensitive , synonyms. To overcome this problem the proposed model is built by normalizing the text by removing the stop words, using stemmer to get the base form of words and then use for searching which will help in increasing the performance of the IR system.

3 DATASET DESCRIPTION

The Text REtrieval Conference(TREC) is a workshop on Information Retrieval research tracks which has been going on from 2003. The TREC goal is to encourage the retrieval research based on large text collections. In this project, the retrieval system is evaluated on the MEDLINE citations of the TREC genomics tracks. The dataset used for this project is in json file format which is convenient to use in elasticsearch. The documents are indexed with respect to the abstract(AB), Title(TI). The statistical information of this dataset is shown as Table 1, where N is the number of documents that contains the property. The 'training-queries-simple.txt' contains the topics and for each topic the search was done using elastic search.

Table 1: Dataset

Property	N	Word Count	String Length
TI	525937	11.71	88.93
AB	402369	171.28	1170.52

4 METHODOLOGY

We proposed a method to improve the performance of text retrieval by implementing text normalization including tokenization, lower-casing letters, removing stop words, and stemming. Based on Elasticsearch, we conducted BM25 similarity module with text preprocessing. Comparing the performance of BM25 model with and without text normalization by pairwise t-test, the significance of improvement can be evaluated.
Stemmer token filter was used in the model to carry out the stemming process. A stemmer will make sure the derivations and variations of a word will be matched during the process of searching [2]. stemmer token filter provide algorithmic stemming for many langauges and the English language stemmer was used in out model.

Conference'17, July 2017, Washington, DC, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

A standard tokenizer was used in the model which divides the text document into individual tokens according a boundary defined by the unicode text segmentation algorithm. This tokenizer removes most of the punctuation. The documents are indexed by frequency of words. There are some words like 'a', 'the'','and', etc, which have high frequency in a document which has no information. These kind of words are called stopwords . Removing stopwords can significantly decrease the size of the document which will help in increasing the speed of indexing. English stopwords was used in the model.

4.1 BM25 Similarity Model

This is a default similarity model in probabilistic models based on TF/IDF. The algorithm works by assigning a weight for each term as a product of tf and idk functions. This product is the score for the documents towards the query. The MB25 model has two parameters , 'k1' and 'b' . 'k1' was set to 1.2, which is the default value to control the saturation of non-linear term frequency. 'b' has the default value 0.75, which controls the degree to which the document length normalizes the tf values.

4.2 Analyzers

The analyzer of Elastic Search is a key point of our proposed method. To get better descriptions, also finer vectors of text, more text preprocessing methods should be implemented. After Tokenization, text sequences are divided into words or tokens. But some words are meaningless because most documents contain lots of that kind of words, such as "by" and "for". These words are called stop words [3], useless in information retrieval tasks, so we removed these words. Besides, The meanings of some words are related but not in the same form, so they would be recognized as different tokens, such as "troubling" and "troubled". This would affect the performance of retrieval. So, stemming is used to make them the root form. Thus they can be grouped as the same tokens.

5 EVALUATION MEASURES

The effectiveness of an Information Retrieval System can be known by using some evaluation measures which help in knowing how successful the system works. There are system oriented evaluations which focus on finding how well the system works in differentiating relevant and non-relevant documents according to the users query. This is a binary classification problem. A document is considered as relevant if it contains the correct information required by the user , or else it will be considered non-relevant. Some of the measures used in this project are discussed below.

5.1 Precision and Recall

Precision is one of the basic measurements used for the evaluation of information retrieval. It is given by a fraction of relevant documents among the retrieved documents.

$$Precision = \frac{number\ of\ documents\ retrieved\ that\ are\ relevant}{total\ number\ of\ retrieved\ documents}$$

Recall is the fraction of relevant documents that are successfully retrieved among all relevant documents.

$$Recall = \frac{total\ number\ of\ documents\ retrieved\ that\ are\ relevant}{total\ number\ of\ relevant\ documents}$$

5.2 Precision at k

When user search for a query he may not be interest in looking for all the relevant documents that were retrieved. So, the precision of k gives the number of relevant documents in the top k documents.

5.3 Interpolated precision at recall

When the precision-recall graph is plotted, when the retrieved document is relevant then both precision and the recall will increase. Where as , when the document retrieved is non-relevant, then the recall will be same for the top k documents. In order to rectify this , the the interpolated precision at a certain point will be calculated for different levels.Traditionally 11 points are considered.

$$P_{interp}(r) = \max_{r' \geq r} p(r')$$

5.4 Average Precision

The system gives the documents in a ranked sequence. Average precision can be calculated by taking the average of the precision for top k documents. The mean average precision score can be calculated using the average precision score for each query. This measure is popularly used in the TREC community which measure quality for all recalls.

$$MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1} m_j precision(R_{jk})$$

Where Q is the number of queries

$$R_{jk}$$

is the set of ranked retrieval results.

5.5 Significance Testing

Comparing a pair of observation is important. The significance test is a statistical method to check for the consistency between two IR models. This test will tell which of the two model is significantly better than the other, by calculating the t-score and p value, formulated as:

$$t = \frac{\bar{d} - \mu_0}{S_d / \sqrt{n}},$$

where $\bar{d}$ denotes the average of the differences between all pairs of values, S_d denotes the standard deviation of the differences between all pairs of values, n is the number of pairs.

6 RESULTS AND DISCUSSION

To evaluate the improvement of our methodology, we designed a comparative experiments between models with and without text normalization. Additionally, we use eleven measure methods to test the performance of these models. These results are presented in

Table 2: Comparison of Rebuilt Models and Default Models

Measures	Rebuilt BM25	BM25
Precision at 1	0.160000	0.160000
Precision at 5	0.088000	0.092000
Precision at 10	0.058000	0.054000
Precision at 50	0.022800	0.020400
Precision at 100	0.015200	0.016000
Interpolated precision at Recall 0.0	1.000000	1.000000
Interpolated precision at Recall 0.1	0.198899	0.196191
Interpolated precision at Recall 0.2	0.134443	0.133639
Interpolated precision at Recall 0.3	0.113077	0.109714
Interpolated precision at Recall 0.4	0.110575	0.108500
Interpolated precision at Recall 0.5	0.107767	0.106424
Interpolated precision at Recall 0.6	0.074328	0.072570
Interpolated precision at Recall 0.7	0.066311	0.064269
Interpolated precision at Recall 0.8	0.057523	0.054665
Interpolated precision at Recall 0.9	0.053307	0.046344
Interpolated Precision at Recall 1.0	0.053229	0.046216
Mean Average Precision	0.179042	0.176230

Table 2. "Rebuilt BM25" denotes the BM25 model with text normalization and "BM25" means the simple model without text normalization.

Comparing measure scores of these two models, we can find that the results of Rebuilt BM25 are better in most measures, but still worse in several measures. To make sure the significance of the differences, we conducted pairwise t-tests at 5% significance level. and result in $t = 3.2491, p = 0.00251548 < 0.05$, which means the Rebuilt BM25 model is significantly better than the simple BM25 model.

7 CONCLUSION

By the results found , the proposed model with normalization by using stemming and removing the stop words we can conclude that the model has performed well and has increased the performance compared to the baseline model.

REFERENCES

[1] Darshita Kalyani and Dr. Devarshi Mehta. Paper on Searching and Indexing Using Elasticsearch. *International Journal Of Engineering And Computer Science*, 6(6):21824–21829, 2017.

[2] Wahiba Ben Abdessalem Karaa. A New Stemmer to Improve Information Retrieval. *International Journal of Network Security & Its Applications*, 5(4):143–154, 2013.

[3] DrGSAnandha Mala. Text Preprocessing for the improvement of Information Retrieval in Digital Textual Analysis. In *International Conference on Mathematical Sciences*, number August 2016, 2014.