Technologically assisted systematic reviews in empirical medicine

CS522 - Advanced Data Mining – Final Project Report

Bhargav Mardimani, A20381095, Illinois Institute of Technology

***Abstract*—Evidence-based Medicine is a vital part of modern health care. This methodology makes use of the systematic review articles that have summarized the evidences of a certain medication or treatment or a diagnostic test. Researchers need to search and retrieve all the information/documents that is related to a certain topic in order to write useful articles of this kind.**

In this project we make the search and retrieval process of relevant documents more efficient and optimized, using Elastic Search and we search through PubMed Central library. We perform Boolean search and use techniques to attain higher recall.

Keywords—EBM, Elastic Search, Boolean Search, Information retrieval;

I. INTRODUCTION

Searching the relevant documents from large source of data collection is important task of any information retrieval domain. Evidence based medicine is method of using systematic reviews which have the articles with the methodology of treatment or medication of a particular disease. Medical researchers in order to write new systematic review, would need to have an efficient search operation which retrieves all the relevant documents of a topic and Implementing a high recall search is essential for automating this process of finding the relevant documents from Medical libraries such as PubMed which is expanding rapidly which makes the search an expensive operation.

This project aims at developing an efficient higher recall search which finds relevant document using the Elastic Search engine and perform the search of the documents based on different fields and evaluate to get higher recall.

A. Data

As a part of CLEF eHealth 2017, We are given the development dataset and it consists of 20 topics for Diagnostic Test Accuracy (DTA) reviews a two sets of qrels, (a) a qrel at the abstract level dictating which PubMed IDS (PIDs) were excluded after the abstract screening and which ones were further processed, and (b) a qrel at the document level dictating which PubMed IDS (PIDs) were excluded after the abstract or/and document screening and which ones were included in the review written.

The PubMed Central library has full text of all articles is publicly available and a large part of PubMed Central (1,227,716 out of 1,317,348 articles) is also published in the important MEDLINE library. All documents are downloaded, and the meta fields such as title, publish date, keywords, body, and abstract are extracted from the raw XML format. The files which are in the qrels format have sections like TOPIC, DOCUMENT#, RELEVANCE which are defined as topic number, document number in PID in PubMed and relevancy is indication of the document in Boolean form.

Table1 lists the Topic ID's, PID's (documents) for each topic ID given in the data.

Topic ID	No. of PIDS	TOPIC ID	No. of PIDS
CD010438	3249	CD011975	8227
CD007427	1469	CD009323	3857
CD009593	15076	CD009020	1576
CD011549	12704	CD011548	12706
CD011134	1952	CD011984	8221
CD010409	43484	CD007394	2542
CD010771	316	CD009944	1225
CD009591	8082	CD008643	15078
CD008691	1322	CD008686	3963
CD010632	1508	CD00804	3148

Table1: **Topics and documents of the data**

The Elastic search 5.2.2 engine was used to index and search through the documents. Word tokenizer, Stemming, PubMed, Elastic search, text mining, Pandas, NumPy, matplotlib, NLTK packages are being made use in python 3.

The raw data was in text format and was preprocessed where each section of the qrels were split and stored separately. The below three steps were performed:

- Tokenization: Removing the non alpha numeric characters like symbols, were removed by using regular expressions '\d+' and '\w+'.

- Normalization of the data was done to remove the stopwords with the help of natural language toolkit.

B. Research Question

The goal of this project is defined by the task 2 of the CLEF eHealth 2017 which is to conduct experiments and share results for a total recall task that specialises in the medical domain, and release a reusable test collection that can be used as a reference for comparing different retrieval approaches in the field of medical systematic reviews Method.

C. General Setup

1) The library

The PubMed Central library was used. This library was chosen because the full text of all articles is publicly available. A large part of PubMed Central (1,227,716 out of 1,317,348 articles) is also published in the important MEDLINE library. All documents were downloaded, and the meta fields (title, publish date, keywords), body, and abstract were extracted from the raw XML.

PubMed library is the source of articles which correspond to the PID's that are filtered based on the query, all documents were downloaded, and the meta fields (title, publish date, keywords), body, and abstract were extracted from the raw XML. The ElasticSearch 5.2.2 engine was used to index and search through the documents. The data load processes are as follows:

- Starting Elastic Search Server.
- Creating index for storing data – indexing enables faster search.
- Storing the data fetched and discarding the irrelevant data.

2) The search engine

The Elasticsearch 2.3 engine [2] (running on Ubuntu 14.04) was used to index and search through the documents.

D. Evaluation

1)Ranking of the documents

ElasticSearch ranks the documents based on the value of the score which is calculated by the with the relevance of the document as per the query as it parses the query into key words and calculates the scores and ranking. The scoring function is based on Term Frequency (tf) and Inverse Document Frequency (idf) that also uses the Vector Space Model (vsm) for multi-term queries. We implement Boolean search algorithm but we process with ElasticSearch for query search because the Query DSL (Domain Specific Language) in ElasticSearch is included in the Search APIs and provides a robust, flexible interface to query. Score is calculated by

score(q,d) = queryNorm(q) * coord(q,d) *SUM (tf(t in d), idf(t)², t.getBoost(), norm(t,d)) (t in q)

q- query; d- document; queryNorm- normalization factor; coord- coordination factor; sum of weights of each term t in doc d; tf- term frequency; idf- inverse doc frequency; t.get.boost- boost on query; norm

2)Measures

A search returns a (ranked) list of documents, and there are several metrics to quantify the accuracy of the fit. The measures used in this article are described and motivated below which makes use of the confusion matrix components.

Recall expresses the proportion of documents that are correctly retrieved. This measure is also known as sensitivity in the systematic review domain, defined by recall = **tp / (tp + fn)**
(# relevant documents retrieved)/(# all relevant documents)

Precision expresses the proportion of the retrieved documents that are correct. It is defined by precision = **tp / (tp + fp)**
(# relevant documents retrieved)/(# all documents retrieved)

F1 defines the geometric mean between recall and precision. It is defined as F1 = **2 (recall*precision)/ (recall + precision)**

Fβ In areas like systematic reviewing, recall may be more important than precision. The Fβ measure allows to put more weight on one of the two. It is defined so that a β value of 10 means that recall is 10 times as important as precision.

It is defined by **Fβ**= (1 + β²) ***** **(**recall *precision)/recall + (**β**² *precision)

MAP The measures above define the performance of the search engine at a specific rank. As a result, each query has as many precision, or **Fβ**-values as the number of documents that it retrieves. This makes it difficult to combine the performance for different queries. The average precision is a measure of performance over the entire rank list. It equals to the area under the precision-recall plot. The average precision can in turn be averaged over queries resulting in the Mean Average Precision (MAP).

All the above parameters were implemented. For each query, precision/recall plots and F1 plots were constructed and manually inspected. Recall measured at the rank at which the last document was retrieved which seemed to be a good descriptor of search performance.

3)Query Expansion and Mining

Each review article in the PubMed central library contained a detailed description of the query used to search through PubMed. Because the features of the PubMed search engine do not match completely with their implementation in Elasticsearch, some changes had to be made. The general approach was to adjust the query in such a way that recall was maintained.

- This experiment allows fine tuning by normalizing the query keywords and returning the relevant documents

where the query which contains the key words of the search, which are synonyms and may have words have many other words which mean the same and hence we make use of the **'*most_field*'** function for elastic search which is similar to the stemming process and the fetches the content with same meaning words.

- We used the split parts of the qrels file to search documents that are highly relevant.

II. EXPERIMENTS

1)Title vs Query

In the first experiment, we used tokenized keywords from the title and searched for the relevant documents pids in the different sections of the documents and calculated the recall for the retrieved documents and preformed similar experiment by using the keywords of the Query.

2)Boolean Search

At the first stage a Boolean query expressing what constitutes relevant information is built. The query is then submitted to a medical database containing titles and abstracts of medical studies. We developed a function which scans the titles and the abstract of the articles and then searches the keyword from the query which expresses what constitutes the relevant information and has a count of the words which are the exact match of it and returns the relevant sentence/paragraph or the document id which has the set count of the key words. This function is utilized to retrieve the relevant documents related to the query, in order to have an efficient Boolean Query over the articles we used Apache Lucene based text search engine Elastic Search which converts the raw data files into simpler structure such as JSON object.

3)Title vs Abstract vs Title-Abstract Screening

In the abstract condition, the query was applied to the abstract and meta fields of the article. In the title condition, the keywords are searched through the Title field of the document In the Title-Abstract condition, the combination of the Title and abstract was also searched through and for all the above procedures we calculated the recall parameter with the keywords from title and the query.

4)Clustering

Clustering methods such as K-means and Mean shift are implemented in order to group the documents based on the score which is measure of the relevance of the document based on the query, so that we can filter out the unwanted documents by choosing the higher scored cluster of documents.

III. RESULTS

1)Documents Ranking

We see that the documents are being ranked based on the score measure which is based on the tf- idf factor and from the figure below we see that higher the relevance of the document we get high score and a low value rank indicating it to be more relevant for the topic CD010632

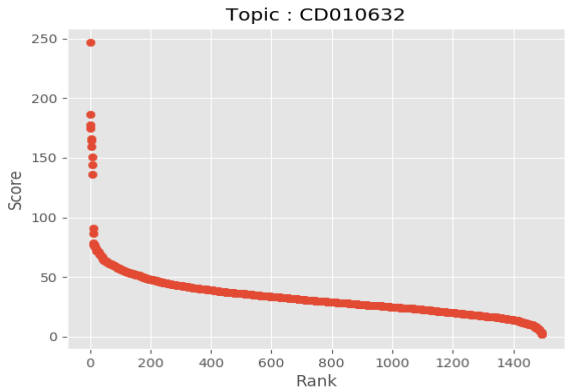

Figure1. **Score vs Rank plot** indicates higher scored documents have lower ranks which are more relevant documents

2)Title vs Query

The keywords of the title were tokenized and the set of words were used to search through different fields like Abstract and other section and measuring parameters were calculated and analyzed. From the table 1 we see that that keywords from the query find more relevant document when we search in the title and Abstract filed as we can see the higher recall value. However, we observe that the keywords from the title of the document gives higher recall that is it retrieves more relevant documents when we search in the Title-Abstract field together.

3)Boolean Search vs Best Match

The search algorithm used here is Boolean search which performs very effectively when compared to the Best match as Boolean search gives higher precision which is the measure of the accuracy of the search.

4)Clustering Techniques

We observe from the figure 5 that the with the K means clustering, there is better cluster which is formed based on the score of the documents which are more relevant to the query and the mean shift algorithm from figure 6 that it creates smaller clusters based on the score values hence refining the granularity of the relevance of the documents for the topic CD010632.

Search on Title Field of the Document				
Using	**Topic ID**	**Recall**	**Precision**	**F1 Score**
Title	CD011975	0.85	0.1346	0.2325
	CD011548	0.664	0.0168	0.0329
	CD009944	0.812	0.0986	0.17593
Query	CD011975	0.948	0.081562	0.15021
	CD011548	0.912	0.010959	0.02166
	CD009944	0.829	0.093901	0.1687
Search on Abstract Field of the Document				
Factor	**Topic ID**	**Recall**	**Precision**	**F1 Score**
Title	CD011975	0.869	0.137772	0.23784
	CD011548	0.876	0.022272	0.04344
	CD009944	0.821	0.099688	0.17778
Query	CD011975	0.89	0.07656	0.14099
	CD011548	0.885	0.010639	0.02103
	CD009944	0.821	0.092933	0.16696
Search on Title & Abstract Field (FullText)of the Document				
Factor	**Topic ID**	**Recall**	**Precision**	**F1 Score**
Title	CD011975	0.95	0.150576	0.25994
	CD011548	0.938	0.023847	0.04651
	CD009944	0.838	0.101765	0.18148
Query	CD011975	0.971	0.083507	0.15379
	CD011548	0.938	0.011278	0.02229
	CD009944	0.838	0.094869	0.17044

Table2. **Result values of Recall, Precision, F1**

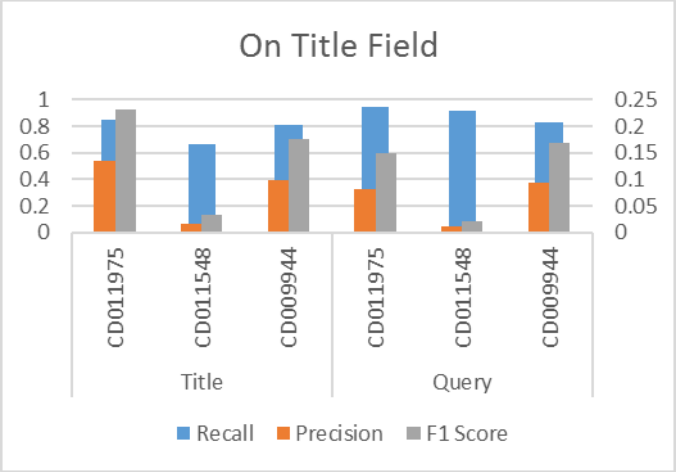

Figure2. **Search results for keywords on Title field** indicating that the recall ranges from 0.66 which is the lowest meaning that the less relevance of the document and the query.

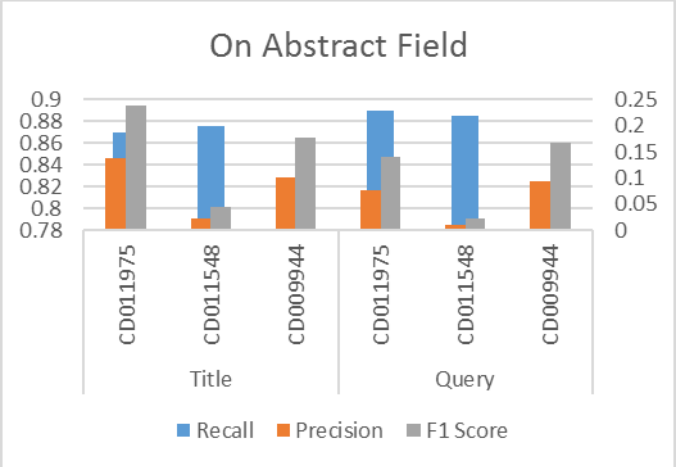

Figure3. **Search results for keywords on Abstract field** indicates that the average recall value of 0.85 and is better than Title as there are more words in Abstract field

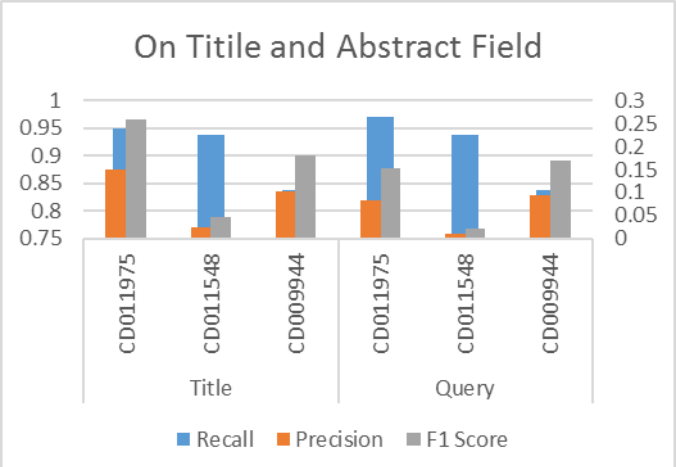

Figure4. **Search results for keywords on Title-Abstract field** Indicates highest recall value with an average of 0.94 meaning the documents retrieved are highly relevant.

5)Title vs Abstract vs Title-Abstract Fields

Figure 2 shows the performance of the Boolean query which is comprised of keywords from query gives higher recall when compared to the keywords from title searched on Title section, Figure 3 shows the performance of the Boolean query which is comprised of the keywords from query gives higher recall than keywords from title Abstract and Title-Abstract. As expected, searching on Title-Abstract improves recall as compared with search on abstract and meta only as shown in Figure 4. However, the cost of this is decreased precision when keywords from query is used, meaning that more documents have to be examined in order to find the same number of relevant documents. Hence the keywords of title searched on Title-Abstract filed gives more accurate results that is higher relevant documents.

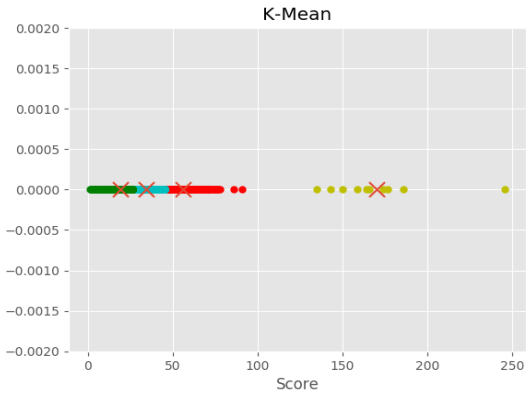

Figure5. **K-Means Clustering of Topic CD010632**

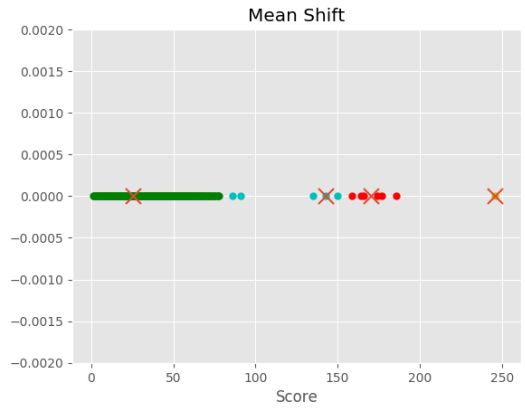

Figure6. **Mean Shift Clustering of Topic CD010632**

5)Recall vs Precision

Recall and precision are the most important factors of any document retrieval process and on calculating plotting the recall and precision on the search performed on sections like title text, abstract and entire text of the document we observe from Figure 7 is the search operation using the keywords from title gives better results than from Figure 8 which is the search performed using the keywords from the query and when scanned through the entire text gives us higher relevant documents as the average value for precision and recall is higher in Full text search and using the title keywords.

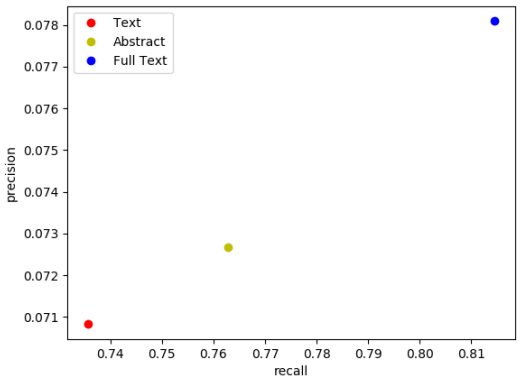

Figure7. **Average Recall and Precision using Title**

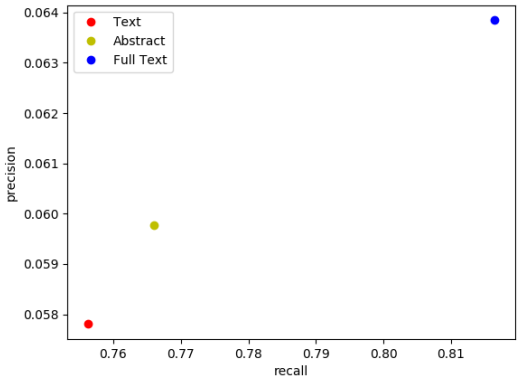

Figure8. **Average Recall and Precision using Query**

Using	Search on	MAP	F-β
Title	Title	0.0271	0.670443
	Abstract	0.0167	0.625825
	Full text	0.0221	0.60205
Query	Title	0.0274	0.596124
	Abstract	0.0168	0.607194
	Full text	0.0211	0.649437

Table3. **Values of MAP and F-β**

IV. CHALLENGES

Cleaning the data was difficult as the knowing the useful parts of the data was important before any preprocessing was performed.

Knowing the idle number of clusters was difficult as the data distribution is non uniform that is the number of the documents per topic is not equal as seen in the Table1.

Finding the threshold value for the functions of Elastic search was another difficult task of filtering the documents like the *most_fields, minimum_should_match* functions which makes the data sparse.

The results on using the title keywords and query keywords almost matched which makes the efficiency low.

V. Conclusion

The enhancements to document retrieval in systematic reviewing are based on the effectiveness of each experiment performed and it varies, but in general there seem to be a lot of chance for improvement. Title-Abstract, as opposed to abstract search increases recall, but does so at the cost of decreasing precision. Investigating best-match search is difficult, because an alternative query needs to be formulated. Of the search algorithm, the Boolean query is the most effective and also we see that there is a tradeoff between the expense and the accuracy of the search.

VI. Work done

I worked on the developing search algorithm that is the Boolean search which scans the keywords and fetches the documents relevant to it which was done in incremental level like by sentence level, paragraph level and then on to the document level and expanding it to the Elastic Search which is a text based search engine works on the tf-idf factor making the search more efficient as it uses the index mechanism, worked on implementing compatibility of the PubMed library and Elastic search tool, implementing the query expansion part by using the *most_fields* part of the elastic search, on the tokenization of the title and search on title, abstract and title abstract fields and calculating the result measures like recall and precision for various cases as calculating using title and query keywords on sections like title, abstract, full text and analyzing the improvement in the values of recall and precision concluding the best of all.

Appendix

Source code and supporting materials can be found here : GoogleDrive

References

[1] https://sites.google.com/site/clefehealth2017/

[2] Elastic revealing insights from data. https://www.elastic.co/. Accessed: 2016-06-20.

[3] Ftp service - ncbi - national institutes of health. http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/. Accessed: 2016-06-20.

[4] James Thomas, John McNaught, and Sophia Ananiadou. Applications of text mining within systematic reviews. Research Synthesis Methods, 2(1):1–14, 2011 and Ties van work.

[5] Mariska MG Leeflang, Jonathan J Deeks, Yemisi Takwoingi, and Petra Macaskill. Cochrane diagnostic test accuracy reviews. Systematic reviews, 2(1):1, 2013.

[6] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.