Course Project Report (Jul-Dec 2020)

Precision Information Retrieval of Clinical Trials using both Structured and Unstructured Data

Submitted By

Arpitha Raghunandan (171IT211)
Nikitha K M (171IT128)
Neha Chauhan (171IT127)

as part of the requirements of the course

Information Retrieval (IT458)

under the guidance of

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

NOVEMBER 2020

Department of Information Technology

National Institute of Technology Karnataka, Surathkal

C E R T I F I C A T E

This is to certify that the Course project Work Report entitled **"Precision Information Retrieval using both Structured and Unstructured Data"** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature with Date
Arpitha Raghunandan	171IT211	
Nikitha K M	171IT128	
Neha Chauhan	171IT127	

as the record of the work carried out by them as part of the course **Information Retrieval (IT458)** during the semester **Jul - Dec 2020**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology.**

(Name and Signature of Course Instructor)
Dr. Sowmya Kamath S

D E C L A R A T I O N

We hereby declare that the project work report entitled **"Precision Information Retrieval using both Structured and Unstructured Data"** submitted by us for the course **Information Retrieval (IT458)** during the semester **July - Dec 2020**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Arpitha Raghunandan	171IT211	
2. Nikitha K M	171IT128	
3. Neha Chauhan	171IT127	

Place: NITK, Surathkal
Date: 22nd November 2020

Precision Information Retrieval of Clinical Trials using both Structured and Unstructured Data

Arpitha Raghunandan[1], Nikitha K M[1] and Neha C[1]

Abstract— **This project aims at Precision Information Retrieval using both structured and unstructured data. It uses the clinical trial dataset of the precision medicine track of TREC 2017. We propose a pipeline to retrieve a set of relevant ranked clinical trials for a given condition/disease that could not be cured using any of the traditional treatments suggested. Since this involves suggesting treatments to patients, the proposed model must be extremely accurate and precise because it can affect the health of other human beings. We use various popular techniques in information retrieval, all tied together in a pipeline, with the aim of extracting highly accurate and relevant trials. This involves use of probabilistic models like bm25, techniques like query reformulation and re-ranking using deep learning methods like learning to rank on unstructured data. This is followed by extracting relevant documents using exact match on the structured data. The final result is a set of accurate ranked relevant clinical trials that may benefit a patient.**

Keywords: ***information retrieval, clinical trials, structured, unstructured, bm25, query reformulation, learning to rank, exact match, precision, recall***

I. INTRODUCTION

Precision medicine can result in better treatment outcomes than using the same strategy for everyone. Advancement in the sequencing technology have led to the development of genetic testing for the molecular diagnosis of diseases, particularly cancers. Genetic variants have been shown to be factors implicated in various cancers, such as breast cancer. The genetic variant indicates the risk level of a cancer, knowing the specific genetic variant of a cancer should be informative for prevention and treatment. Finding the most relevant and recent research can be quite challenging due to the high volume of scientific literature.

Information retrieval (IR) provides an efficient and effective way to retrieve relevant documents from a large corpus. This is extremely important in situations where doctors are looking for suitable treatments for patients based on their condition. If no existing treatment is found to be helpful, patients may opt to take part in clinical trials. Given the patients condition and demographic information like their age and gender, it is possible to retrieve suitable clinical trials. With the increasing number of clinical trials, it is essential that highly relevant clinical trials are retrieved, such that it may actually benefit the patient. This requires precision information retrieval.

This project involves the precision information retrieval of clinical trial information given the patient's conditions and demographic information as a query. It uses the clinical trial dataset of the precision medicine track of TREC 2017 to create a suitable information retrieval model. The model is then evaluated using the test topics and the relevance judgements of the same track of TREC 2017.

The unstructured data used includes a summary and description of each clinical trial, along with other information like MeSH (Medical Subject Headings) terms. The structured data is the demographic information, like gender and age. A detailed pipeline using several popular information retrieval techniques was developed to leverage both structured and unstructured data to achieve good results as will be explained in the later sections.

II. RELATED WORK

(Wang et al., 2017) outlines the Mayo Clinic NLP team's involvement in the Text retrieval Conference (TREC) 2017 Precision Medicine track that uses structured and unstructured data for Information Retrieval. Markov models is used for retrieval using unstructured data. MeSH on Demand is used due to its efficacy for mentioning pertinent PubMed articles.

Several approaches have been made to achieve better Precision-Recall Performance using a variety of methods. A binary classification and Learning-to-Rank approach was used in (Shenoi et al., 2020). LSTM and CNN was used to boost the retrieved results. Query reformulation techniques followed by filtering and ranking was suggested in (Agosti et al., 2018). This involved both query expansion and reduction.

(Qu and Wang, 2019) outlines TREC's (2019) Precision Medicine Track that carries out Query Expansion using an API called Lexigram instead of MeSH and uses Learn-to-rank strategy to improve Precision-Recall. Here, only structured data was used.

There is very little work that leverages both unstructured data and structured data. We propose to work with both and achieve precision information retrieval using various existing models tied together in a pipeline as shall be discussed in the later sections. We also propose to use query reformulation techniques to improve results.

III. PROPOSED METHODOLOGY

We propose a pipeline to extract accurate relevant clinical trials given information about the patient like their disease, gene, demographic and other relevant information as a query. This pipeline can be seen in Fig. 1.

As can be seen in the figure, the pipeline leverages both structured and unstructured data to create a ranked list of

[1]Student, Department of Information Technology, National Institute of Technology Karnataka, Surathkal

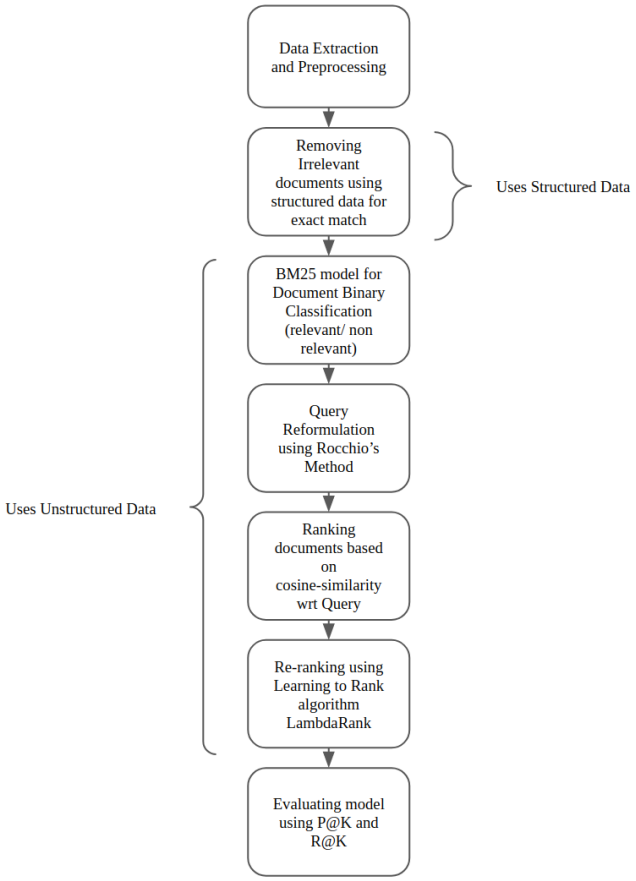

Fig. 1: Methodology Pipeline

relevant clinical trials. Every clinical trial has the following information:

1) Title
2) Summary
3) Description
4) Criteria
5) Mesh Terms
6) Keyword
7) Gender
8) Minimum Age
9) Maximum Age

The content of the first 6 parts of the clinical trial information can be used as unstructured data in the pipeline. The last three components, that is, the age range and gender can be used as structured data. Checking if the patient fits these structured criteria can be used to eliminate certain clinical trials in the ranking suggested for the given query. The demographic content of the query includes this data. Each query was expanded by finding MeSH terms using MeSH on Demand. These mesh terms also served as additional structured data.

The entire pipeline of the proposed methodology involves 7 steps:

A. Data Extraction and Preprocessing

The first step is to extract all the required data for each clinical trial and preprocess it so that the proposed precision information retrieval model can use it. The entered query will also have to be preprocessed accordingly to extract useful information. Like in any other information retrieval system, the preprocessing step will involve steps like removing accents and other punctuation to normalize the text, removing stopwords, converting text to lower case, stemming, lemmatization, etc. Since our query also involves information about *genes*, it will help expand and include extra gene information in the question to improve the quality of the retrieved trials and their ranking. This can be done using Mesh or Lexigram.

B. Leveraging Structured Data

The next step involves eliminating irrelevant clinical trials based on structured data information, demographic information like age and gender. Suppose the patient in the given query is of age, which is not in the range of ages of patients on whom a clinical trial is being tested, the clinical trial is removed from the ranked list of trials. The same is done if the clinical trial requires patients of a particular gender not satisfied by the patient. After this, clinical trials with MeSH terms completely different from the MeSH terms found for a query were removed. If there was even a single match in MeSH terms between the clinical trial and the query, the trial was retained.

C. Binary Classification using BM25

Next, the BM25 probabilistic information retrieval model will be applied on the extracted trial corpus, using only unstructured data. Given a query, each clinical trial will be given a score. Higher the score, higher the relevancy of the clinical trial with respect to the query. One of the features of this BM25 model is that it scores a document as 0 if it is considered irrelevant. This is exploited to use this model to now classify the clinical trials as relevant or not for a given query. A trial is classified relevant if it has a score greater that 0, and irrelevant if it has a score of 0. The formula used to calculate the BM25 score can be seen in Fig. 2 and Fig. 3.

$$\mathcal{B}_{i,j} = \frac{(K_1+1)f_{i,j}}{K_1\left[(1-b)+b\frac{len(d_j)}{avg\_doclen}\right]+f_{i,j}}$$

where b is a constant with values in the interval $[0, 1]$

Fig. 2: B[i,j] Factor used in BM25

D. Query Reformulation

The next step involves query reformulation. This is done to modify the query such that it moves the query closer to the

$$BM25(d_j, q) \sim \sum_{k_i[q,d_j]} \mathcal{B}_{i,j} \times \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

Fig. 3: BM25 Formula to calculate the score of a document wrt query

set of relevant documents and away from a set of irrelevant documents. We have proposed to use Rocchio's method for query reformulation. As can be inferred, this method will require information about a set of relevant and irrelevant clinical trials for a give query. This information is extracted in the previous step and is used here. The formula for query reformulation using Rocchio's method can be seen in Fig. 4. The term being added is considered positive feedback and the term being subtracted is negative feedback. This step also makes use of unstructured data, and is run on the query n times, where n is chosen such that the final retrieved and ranked clinical trials are as accurate as possible.

$$Q_1 = Q_0 + \frac{\beta}{n_1}\sum_{i=1}^{n_1} R_i - \frac{\gamma}{n_2}\sum_{i=1}^{n_2} S_i$$

where

Q_0 = the vector for the initial query

R_i = the vector for the relevant document i

S_i = the vector for the non-relevant document i

n_1 = the number of relevant documents chosen

n_2 = the number of non-relevant documents chosen

β and γ tune the importance of relevant and nonrelevant terms

Fig. 4: Rocchio's method for query reformulation

E. Ranking Clinical Trials

The similarity between the reformulated query and the clinical trials is then calculated using cosine-similarity. This similarity is used to rank the clinical trials for the given query. Here, again, higher the similarity, higher the ranking. The formula used to calculate cosine similarity between a query and a clinical trial can be found in Fig. 5.

$$Cos - sim(q, d_i) = \cos(\alpha) = \frac{(d_i.q)}{|d_i|.|q|}$$

$$Cos - sim(q, d_i) = \cos(\alpha) = \frac{\sum_{j=1}^{t} w_{qj} w_{ij}}{\sqrt{\sum_{j=1}^{t}(w_{ij})^2 \sum_{j=1}^{t}(w_{qj})^2}}$$

Fig. 5: Cosine-Similarity Formula

F. Re-rank

The next step involves using unstructured data to re-rank the clinical trials using Learning to Rank algorithm like LambdaRank. All steps leading to this produced a ranked list of clinical trials for a given query. In this step, relevance grades will be assigned to each clinincal trial for a query. If the trial-query pair have a relevance grade in the relevance judgements file provided in the contest, that is considered. Else, each trial-query pair is assigned a relevance based on the rank of the trial for the given query. This data is used to train the LambdaRank model, and to finally provide a ranked list of clinical trials relevant to a query.

G. Evaluation Techniques

Finally we evaluate our proposed model for precision information retrieval using nDCG. The formula for this can be seen in Fig. 6.

$$DCG_n = \sum_{i=1}^{n} \frac{rel_i}{\log_2^{i+1}},$$
$$NDCG_n = \frac{DCG_n}{IDCG_n},$$

Fig. 6: nDCG Formula

Higher the nDCG value better the model. The best nDCG value in case of perfect ranking is 1 and the least is 0.

IV. IMPLEMENTATION SPECIFICS

The implementation involved using the Clinical Trial dataset of TREC 2017 Precision Medicine track. This was evaluated using queries and relevance judgements as given for the contest which can be found here.

The Clinical Trial dataset contains around 12,000 XML files. Each file had *id_info*, *brief_title*, *lead_sponsor*, *source*, *brief_summary*, *detailed_description*, *overall_status*, *phase*, *study_type*, *has_expanded_access*, *study_design_info*, *condition*, *intervention*, *eligibility*, *gender*, *age*, *location*, *location_countries*, *verification_data*, *lastchanged_date*, *firstreceived_date*, *condition_browse*.

There were 30 queries or topics considered for this project. This was in XML format. Each query had *disease*, *gene*, *demographic* and *other* as fields. We expanded each query by adding an additional field for MeSH terms which was found using MeSH on Demand.

The implementation specifics of each step in the pipeline is discussed in their following sections:

A. Data Extraction and Preprocessing

The clinical trial information was in XML format, which was extracted to a json object. Each part of it was preprocessed. Similarly, the queries were also in XML format, and information was extracted to a python dictionary for each

query. Some parts of the query were preprocessed. The data preprocessing step involved:

1) Tokenizing text
2) Removing punctuation and non alphanumeric characters
3) Normalizing tokens
4) Converting tokens to lower case
5) Removing accents
6) Removing stopwords
7) Lemmatization

The lemmatization was done using the WordNetLemmatizer of Python's NLTK library.

Apart from these standard techniques of preprocessing, each query was expanded to include MeSH terms. MeSH terms were found for a query using MeSH on Demand. Each query was divided into a stuctured data component and an unstructured data component. The structured component included the disease, gene and MeSH terms. The age, gender and again MeSH terms were considered as structured data.

B. Leveraging Structured Data

The structured data leveraged here include the age, gender and MeSH terms. Documents are considered relevant if the patient in the query is of age within the range specified in the clinical trial, and of gender applicable to be a part of the trial. At least one MeSH term must be common between the clinical trial and the query. If any one of these conditions was not satisfied, the clinical trial became irrelevant, and was not considered for the later sections.

C. Binary Classification using BM25

Binary classification involves classification of clinical trials into two categories: relevant and non-relevant. This was done using the BM25 probabilistic model. The BM25Okapi() method of the rank_bm25 python library was used for this step. Each clinical trial was given a score with respect to a query. Higher the score, higher its relevancy. If the score is 0, it is not relevant. This fact was used in classification at this step. All trials with score greater than 0 were considered relevant and the rest irrelevant.

D. Query Reformulation

Query reformulation step involved modifying the query to make it more similar to the relevant documents and move it away from the irrelevant ones. Rocchio's method was used here. The algorithm uses = 0.75 and = 0.25. It was run 10 times to produce a new reformulated query.

E. Ranking Clinical Trials

The reformulated query was then compared with each of the clinical trials to find similarity. Higher the similarity, more relevant the trial. Cosine similarity was used here, which was implemented using functions from python's Numpy library. Once the similarity for all clinical trials were found, the trials were ranked from highest to lowest similarity or in other words from most relevant to least relevant. This was done using a heap.

F. Re-rank

The Re-rank step helps us re arrange queries with better ranking. Here we have used Lambda Rerank method to re-rank the documents using query relevance judgements. We have applied a simple Neural Network with (16, 8) hidden layers with relu activation function and adam optimizer. Number of epochs is set to 5 and batch size to 64. This was done using the LambdaRankNN library in Python. Using relevance judgements we trained the LambdaRankNN.

G. Evaluation Techniques

We have used nDCG technique as a metric for our pipeline evaluation

V. EXPERIMENT RESULTS AND ANALYSIS

Leveraging structured data as well as unstructured data produced best results. The current nDCG score for our pipeline is 0.58.

The use of strucured data, that is, the age and gender of a patient to find the relevant clinical trials helped improve this nDCG score. The use of MeSH terms further produced a significant increase in the nDCG score.

Table I shows how the nDCG value varies as the number of query reformulations increases. This is represented graphically in Fig. 7.

Table II compares our results with those of other papers working on precision information retrieval using the TREC clinical trial dataset. Our work shows slight improvement in results from the other papers.

Rochchio Reformulation Iterations	nDCG
0	0.18175295903539446
1	0.40657359638272916
2	0.020438239758848613
3	0.3528020232905472
4	0.2
5	0.4581539706806682
6	0.32043823975884866
7	0.392456853815636
8	0.4924568538156361
9	0.5440372281135749
10	0.5839441578296375

TABLE I: Change in nDGC after reformulation

Paper	year	nDCG
Y Wang, R Komandur, M Rastegar-Mojarad, and H Liu	2017	0.3
Shenoi SJ, Ly V, Soni S, Roberts K	2020	0.46
M Agosti, G Nunzio, S Marchesin	2018	0.49
Jiaming Qu and Yue Wang	2019	0.46
Current Paper	2020	0.58

TABLE II: Overall comparison of results between previous paper and current pipeline

VI. DISCUSSION

We started by initially only leveraging unstructured data for precision information retrieval. This resulted in low nDCG score. Next, we eliminated some irrelevant articles considering the age and gender of the patient considered

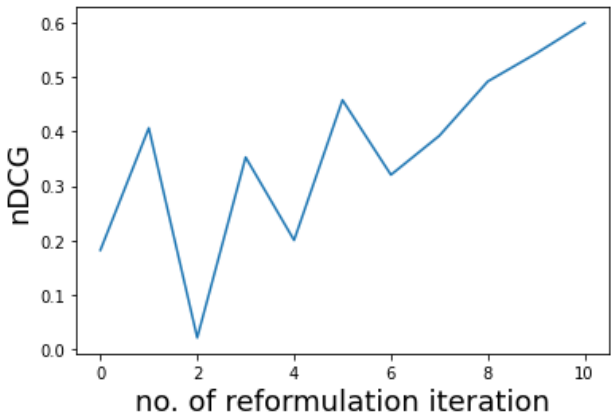

Fig. 7: nDGC v/s query reformulation graph

in the query. This was done before applying the IR model pipeline using unstructured data. There was improvement in the results with by leveraging the structured data and we found the nDCG score to improve. Next, we also considered using MeSH terms. Clinical trials which did not have any matching MeSH term with a query were not considered for ranking with unstructured data for that query. This further improved the nDCG score to 0.58. Hence, leveraging both structured and unstructured data was necessary to achieve best results. Using MeSH terms further helped boost the ranking of relevant trials.Finally we can see the query reformulation improves the model.

VII. CONCLUSION AND FUTURE WORK

This project worked with the clinical trial data of the Text retrieval Conference (TREC) 2017 Precision Medicine track. We aimed to develop a better performing precision information retrieval model for clinical trials. This is essential since finding a suitable clinical trial for a patient (on whom no traditional treatments have proved useful) can turn out to be life saving.

We have worked on vigorous precision information retrieval using structured and unstructured data to retrieve a relevant ranked list of clinical trials for a given query. Structured data and MeSH terms were used to eliminate irrelevant clinical trials. Further, a bm25 model to classify clinical trials to relevant or irrelevant, Rocchio's method for query reformulation, and LambdaRank for re-ranking the clinical trials using unstructured data was developed. This was done with the aim of achieving a better performing precision information retrieval methodology.

While we have received a decent nDCG score, there is always room for improvement. As we advance, we hope to measure better ranking quality by tuning the hyperparameters. We can change hyperparameters used by LambdaRank and try out a different neural network architecture. This includes tinkering with the number of hidden layers, the optimizer used, etc to find the best working model that when integrated in our pipeline provides best ranking results. An alternative would be to use LambdaMART.

VIII. INDIVIDUAL CONTRIBUTIONS

- Arpitha - BM25, Leveraging structured data, Query Reformulation
- Nikitha - Data Extraction, LambdaRerank, Results and Analysis
- Neha - Data preprocessing, Extracting Mesh Terms

REFERENCES

Agosti, M., Di Nunzio, G. M., and Marchesin, S. (2018). An analysis of query reformulation techniques for precision medicine.

Qu, J. and Wang, Y. (2019). Unc sils at trec 2019 precision medicine track.

Shenoi, S. J., Ly, V., Soni, S., and Roberts, K. (2020). Developing a search engine for precision medicine. *AMIA Summits on Translational Science Proceedings*, 2020:579.

Wang, Y., Elayavilli, R. K., Rastegar-Mojarad, M., and Liu, H. (2017). Leveraging both structured and unstructured data for precision information retrieval.

11.pdf

ORIGINALITY REPORT

13%	10%	2%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	trec.nist.gov Internet Source	5%
2	Submitted to Sandra Day O'Connor High School Student Paper	4%
3	Submitted to National Institute of Technology Karnataka Surathkal Student Paper	1%
4	www.nitk.ac.in Internet Source	1%
5	Shihchieh Chou, Zhangting Dai. "Construction and application of specialty-term information for document re-ranking", Online Information Review, 2016 Publication	<1%
6	www.i-scholar.in Internet Source	<1%
7	Submitted to Yonkers High School Student Paper	<1%

pdfs.semanticscholar.org

8	Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		