Information Retrieval - Search Engines
Prior-Art Search Assignment
Part II

Simon Platiotis, Undergraduate Student, IHU

***Abstract*—This document is the second of the two parts of the mandatory assignment on 'Information Retrieval – Search Engines' course of IEE department at IHU. It's goal is the familiarization of the student with the process of query generation in TREC_format, the consumption of the query by a search engine and the result evaluation using TREC_eval.**

***Keywords*—information retrieval, search engine, terrier, trec, topic, query, evaluation, precision, mean average precision, recall**

I. Introduction

Based on one of the indices created on the first part of the assignment, supplied with a subset of the collection [1] and each subset object's content organized in .xml format [2], students are requested to create 30 queries for the search engine of their choice. After 3 seperate runs, search engine results for every query must be evaluated using TREC_eval and the subset's Qrel file [3].

II. Query Generation

A. Fields

Initial objective of the second part of this project is query generation. Firstly the fields contained in each query have to be chosen. All of the available fields can be identified from any xml files inside PAC_topics directory. We decided on the fields UDIC, TITLE, APPLICANTS, INVENTORS, ABSTRACT. And each unique Run included the following fields,

Fields Per Run Table

Run #	*Doc Tag*	*Id Tag*	*Field 1*	*Field 2*	*Field 3*	*Field 4*
1	TOP	UCID	TITLE	ABSTRACT	- - -	- - -
2				APPLICANTS	INVENTORS	- - -
3				APPLICANTS	INVENTORS	ABSTRACT

B. Field Selection Reasoning

The process behind our choises was the following,
TOP is Terrier's default doctag, so followed the stabdard for consistency.
UCID is the unique identifier for each patent of the collection and so was used as the query's IdTag.
TITLE is the invention's title which we determined to be a great candidate for one of the search fields for its distinctive value. Specificly we selected only the English title of each invention.
APPLICANTS is a collection of every 'name' or 'last-name' field under the field 'adressbook' of every Applicant of the patent.
INVENTORS follow the same schema as the applicants but for every Inventor of the patent.
ABSTRACT is the patent's abstract field containing a brief summary of the patent. Finaly we excluded the field **COPYRIGHT** since it's a template terms and conditions agreement used across each patent and provides no information or diversity for the patents.

C. Query Scripts

After the field selection we created a python script which initially reads the PAC_topic name from a file containting a list of the names inside the subset. Then it uses the name to find and traverse the xml topic file, locate the desired fields and save their contents in a temporary dictionary. Having completed this seperation the script uses the dictionary data, parses them to TREC format and saves them inside the query file. This process is repeated for every file of the subset inside the PAC_topics file.

III. Query Execution

With the query generation complete we can now use them on Terrier and get the results. To achieve this we have to modify the terrier.properties file to match our needs. Specificly TrecQueryTags.doctag is the start of each topic, TrecQueryTags.idtag is the unique identifier of the topic, TrecQueryTags.process are the fields of the TREC topic which are to be processed, TrecQueryTags.skip are the TREC topic fields which will be skipped and finaly trec.topics is the full path to the file containing the topics. The values for this parameters for each run are the following,

A. Run 1

TrecQueryTags.doctag=TOP
TrecQueryTags.idtag=UCID
TrecQueryTags.process=TOP,UCID,TITLE,ABSTRACT
TrecQueryTags.skip=DESC,NARR,COPYRIGHT
trec.topics=Path_to_Run_1/queries.txt

B. Run 2

TrecQueryTags.doctag=TOP
TrecQueryTags.idtag=UCID
TrecQueryTags.process=TOP,UCID,TITLE,ABSTRACT
TrecQueryTags.skip=DESC,NARR,COPYRIGHT
trec.topics=Path_to_Run_1/queries.txt

C. *Run 3*

TrecQueryTags.doctag=TOP
TrecQueryTags.idtag=UCID
TrecQueryTags.process=TOP,UCID,TITLE,ABSTRACT
TrecQueryTags.skip=DESC,NARR,COPYRIGHT
trec.topics=Path_to_Run_1/queries.txt

Now that terrier.properties is correctly set up we can run **terrier br**, sorthand for batch retrieval, from the bin directory to execute the queries and get the results. The results are saved in the var directory of terrier under the directory results.

IV. Result Evaluation

The evaluation of the results was calculated using **trec eval**. The trec_eval command is located in the bin directory of Terrier and requires 2 parameters for it's execution, Qrel file path and the results from a batch retrieval command. In addition using -m option allows the user to choose the output metrics. Example execution for all the available metrics *trec_eval -m all_metrics path_to_qrl_file path_to_result_file* The metrics we will focus on for the sake of the assignment are Precision, Recall and Mean Averate Precision and are displayed on the following tables.

Evaluation Table I (Precision)

Run #	*MAP*	*P_R**	*P_10*	*P_100*	*P_500*
1	0.366	0.371	0.260	0.050	0.012
2	0.211	0.217	0.147	0.021	0.006
3	0.379	0.365	0.263	0.047	0.012

*R is the number of relevant documents

Evaluation Table II (Recall)

Run #	*Set_Recall*	*R_10*	*R_100*	*R_500*
1	0.856	0.390	0.655	0.788
2	0.514	0.229	0.304	0.428
3	0.822	0.401	0.606	0.742

A. *Data Analysis*

Precision is the fraction of (total number of every relevant document retrieved) to (total number of all the documents retrieved) [4].
Recall is the fraction of (total number of every relevant document retrieved) to the (total number of relevant documents)[4].
Average precision(AP) is the mean of the precision scores after each relevant document retrieval [5].
Mean Average Precision(mAP) is simply the mean of all the Average Precisions for every query.

B. *Data Driven Conclusion*

As a reminder all Runs included the fields UCID and TITLE but Run #1 used only the ABSTRACT field as an extra, Run #2 used the APPLICANTS and INVENTORS fields and Run #3 used all of the available fields.

Referencing the tables and the above inforamation, it is clear that the best results were from Run #3 which contained all the fields. Additionally, comparing Run #1 and Run #3 we can identify that their precision results across the board are almost identical throught the increasing cut-off level. This fact paired with the poor result of Run #2 we can conclude that the fields APPLICANTS and INVENTORS played a minor to almost irrelevant role in the effectivness of the IRS. And even though overall the mAP of Run #3 is higher than the mAP of Run #1, Run #1 had better results in average cut-off levels and also in relative precision. This is an effect of the wider spectrume of relevancy created by the 2 extra fields, which for some documents was not the actual case and didn't overlap with the relevant documents.

References

[1] M. Salabasis, "Collection Subset Query Targets", Web Information Retrieval M.Sc., 2020.

[2] M. Salabasis, "Patent Topics", Web Information Retrieval M.Sc., 2020.

[3] M. Salabasis, "Collection Subset Qrels", Web Information Retrieval M.Sc., 2020.

[4] Ting K.M. (2011) Precision and Recall. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.

[5] Zhang E., Zhang Y. (2009) Average Precision. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

[6] Ren Jie Tan "Breaking Down Mean Average Precision(mAP)", Towards Data Science, March 2019.