Ad-hoc Retrieval on FIRE Data Set with TF-IDF and Probabilistic Models

Chandra Shekhar Jangid
M.Tech, CSE
Sir Padampat Singhania University
Udaipur, Rajasthan

Santosh K Vishwakarma
Ph.D. Scholar
Sir Padampat Singhania University
Udaipur, Rajasthan

Kamaljit I Lakhtaria
Assistant Professor
Sir Padampat Singhania University
Udaipur, Rajasthan

ABSTRACT
Information Retrieval is finding documents of unstructured nature which should satisfy user's information needs. There exist various models for weighting terms of corpus documents and query terms. This work is carried out to analyze and evaluate the retrieval effectiveness of various IR models while using the new data set of FIRE 2011. The experiments were performed with tf-idf and its variants along with probabilistic models. For all experiments and evaluation the open search engine, Terrier 3.5 was used. Our result shows that tf-idf model gives the highest precision values with the news corpus dataset.

General Terms
Information Retrieval, IR Models, Weighting Schemes

Keywords
TF-IDF, BM25, DFR, Retrieval Effectiveness, Precision

1. INTRODUCTION
Information retrieval is finding documents of unstructured nature which should satisfy user information need [1]. The objective of information retrieval system is to retrieve the documents from the collection of documents that fulfill the user need which is express in terms of query. Information retrieval system uses many models to understand the query of user, according to that gives the ranks to the all documents and bring the top relevant documents from data set. For this paper we use static data set to evaluate the result of different models. Before assigning ranks to the documents, information retrieval system goes through some preprocessing steps that will be discussed in the next section of the paper.

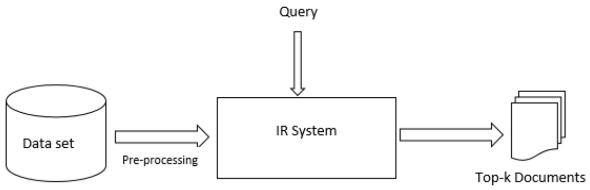

Fig. 1 Information Retrieval

2. PRE-PROCESSING
Every document must go through some preprocessing steps, to make it simple for search engines to take them as a document and run them on their algorithms[2]. Steps like tokenization, stemming, normalization, stop word removing etc. are applied on each document.

2.1 Tokenization
Tokenization is define as the process of breaking text documents into words, phrases, numbers, symbols etc. all these are called tokens. Tokenization faces issues like language recognition, as this process is language dependent. User data, Meta data, machine learning methods are useful to determine information about document language.

2.2 Stemming
Stemming is the process in which reducing the words by removing letters to their root word. A root word can represent many words that might have different meaning and root word may be doesn't make any meaning some times. Avoiding the many original words and using stem words helps to reduce the size of dictionary, that contain all words of document collection. In other sense stemming help user by providing the choices of related options of user query, just after typing the few letter in query box.

2.3 Stop-words
A document contains hundreds or thousands words, for the user perspective every word is not equally important. Generally this word appears many times in the document and doesn't contribute any information for the document makes it informative. As we think practically, people don't search words like 'the', 'a', 'of' and many other words these words called stop words. Stop-words list are can different for different document collection based on the purpose.

2.4 Inverted Index
Every documents collection contains many documents and a document has many different words. Now question is how we make search very fast and how to know which document contain query terms. The answer is inverted index, it is kind of link list in which every word from whole collection connected with the nodes that represent documents numbers in which that particular term appears.

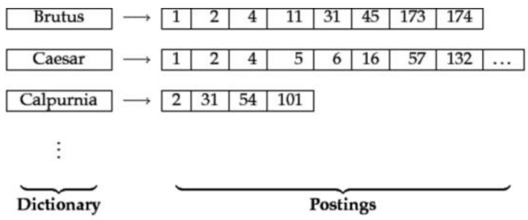

Fig 2 Inverted Index [1]

Generally we use stemmed words in the index, like compute for computer, computation, computations and many other similar words as well doesn't include stop word in index, reason is just to avoid space problem for system. In above figure, collection of all searchable words called dictionary and link of documents IDs called posting.

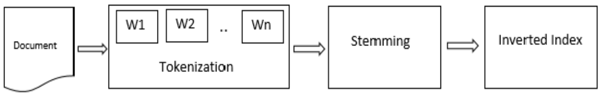

Fig. 3 Pre-processing

3. REVIEW OF IR MODELS

Information Models define the way to represent the document text and the query as well. Other objective of IR models is to compare the document and query, assign ranks to documents. We are using many version of different model such as Probabilistic models, TF-IDF weighting model and Divergence from randomness.

3.1 TF-IDF Model

Generally Term frequency(TF) define as how many times a term appears in a document and document frequency known as in how many documents a term appears.TF-IDF model is describe in simple form as

$$TF - IDF_i = 1 + \log(tf_i) * log\frac{N}{df_i}$$

Where tf_i is term frequency of term i, df_i is document frequency and N is total number of documents in data set.

3.2 BM25 Model

BM25 is probabilistic model that is developed by Stephen E. Robertson, Karen Spärck Jones, and others. BM25 model doesn't use a single function, it use set of functions.

$$Score(D,Q) = \sum_{i=1}^{n} IDF(q_i) * \frac{f(q_i,D) * (a_1 + 1)}{f(q_i,D) + a_1 * (1 - b + b * \frac{|D|}{D_l})}$$

Where $IDF(q_i) = \frac{N-n(q_i)+0.5}{n(q_i)+0.5}$, N total number of documents in data set, $f(q_i, D)$ denote term frequency of query term in the document, a_1 and b is free parameter $a_1 \in [1.2, 2.0]$ and b= 0.75, $|D|$ is length of a document and D_i is average length of text documents in data set.

3.3 Divergence from Randomness

Divergence from randomness has an idea that says if the term frequency of a term within a document has more divergence than it's frequency within the data set than the term is more important for that document. Divergence from randomness is a generalization form of harter's 2-Possion indexing model. DFR Models that based on DFR frameworks are BB2, DFI0, In_expB2, In_expC2, PL2, InL2, DFR_BM25 and many other variants, in this paper we use most of them for experiment purpose. DFR Models use first normalization and term normalization.

4. EXPERIMENTAL EVALUATION

Evaluation is always an important part of any research of any area in all around the world, same it useful in context of information retrieval. Evaluation in the simple mean is how effective a system performs and produces valuable result with accuracy.

4.1 Evaluation Measures

In Information Retrieval, to measure the effectiveness of the system our requirement is a data set, a set of queries and some function to judge relevance factor between document and queries. Simple IR system just fetches the best relevant documents that are related to the query and assign ranks to them. Now effectiveness depends on measurements used for evaluation, better measurements give better ranked list of documents. In this paper we use very common measurements such as precision and recall that are discussed in next section.

4.1.1 Precision

In simple words, precision can be defined as the ratio of number of relevant retrieved documents to the number of retrieved documents [8].

$$Precision = \frac{No.of\ relevant\ retrieved\ documents}{No.of\ retrived\ documents}$$

Precision generally mention in form of percentage and as the number of retrieved documents increase, the precision of system will decrease.

4.1.2 Recall

Recall is another measure for information Retrieval model, which can be described as a ratio of number of relevant retrieved documents to the number of relevant documents [8].

$$Recall = \frac{No.of\ relevant\ retrieved\ documents}{No.of\ relevant\ documents}$$

Recall and precision is inter depended, recall will increase when relevant retrieved documents increase. Recall and precision are inversely related.

4.1.3 Mean Average Precision

In our paper we used MAP to evaluate our results. Mean average precision is standard measure and accepted by TCER community for their evaluation [9]. MAP defined as the average of precision of all top-k documents and this value again averaged over information needs.

Let $[d_1, d_2, d_{3.....}d_{nj}]$ are relevant documents for query $q_j \in Q$, Q is set of queries for data set and R_{jk} is top relevant retrieved documents for the query q_j.

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

4.2 Description of Data Set

The experiments has been carried out on the data set of FIRE 2011[http://www.isical.ac.in/~fire/] for English. The data set contains various documents from English news domain-The Telegraph. These news articles are extracted from 2001 to 2010 and contain 303,292 documents. We just took a sample form this data set for our experiment. The task of corpus creation was carried out to support experiments for research purpose in information retrieval domain.

4.2.1 FIRE and Document Format

FIRE is stands for form of Information Retrieval and Evaluation. It's an India based organization for research on information retrieval. FIRE works on languages of South Asian contraries.

Document format that used in FIRE collection follow the standard representation of TREC collection. Documents contain tags like DOC, DOCNO and TEXT. DOCNO is unique number for every document in the data set. Text field contains the actual news article in plain text. The example of a text file is shown below.

```
<DOC>
<DOCNO>doc_03/0003</DOCNO>
<TEXT>
Americans used more health services and
spent more on prescription drugs in 2013,
reversing a recent trend, though greater use of
cheaper generic drugs helped control
spending, according to a report issued on
Tuesday by a leading healthcare information
company.
</TEXT>
</DOC>
```

Fig 4. Document Format

4.2.2 Topic File

Topics file contain some pre-fixed queries for the data set, these queries almost cover every document within the data set. According to our sampled data of FIRE data collection we take 9 queries. Example of our topic file is shown in figure 5.The topic file format contain tags such as top, num and title. Title is the query and number is assign to every topic.

```
<topics>
<top>
<num>1</num>
<title>shikhar dhawan</title>
</top>
<top>
<num>2</num>
<title>icc cricket world cup 2015 </title>
</top>
```

Figure 5 Topic File

Fig 5 Topic File

4.2.3 Qrels File

Qrels file format describes the presence and absence of the every query terms in the document. Format of qrels shown in figure 6 and description of format is like this, a first column show the query ID that is according to the topic file, second place show iteration, third place the document ID that is mention in document format and last column shows the presence and absence of that query in document by 0 or 1.

```
1 Q0 doc_13/0013 1
1 Q0 doc_14/0014 0
1 Q0 doc_15/0015 0
1 Q0 doc_16/0016 0
1 Q0 doc_17/0017 0
1 Q0 doc_18/0018 0
1 Q0 doc_19/0019 1
1 Q0 doc_20/0020 0
```

Fig. 6 View of qrel file

5. RESULT AND ANALYSIS

We performed our experiments in Terrier 3.5. It has all the necessary codes to support experiments for FIRE dataset. We make some changes in terrier. Properties file. There is many Information model already supported by the terrier-3.5. Initially we are showing the result of all version of Divergence from randomness.

We just use some of DFR models from the list, models are BB2, DFI0, In_expC2, InL2 and PL2. Letter we use two model of TF-IDF, TF_IDF and LemurTF_IDF and a probabilistic model BM25.We use two measures for all model that is MAP and R-Precision. In figure 7 example of eval file that generated for every model by terrier- 3.5 and it shows information about retrieved and relevant documents.

```
Number of queries = 9
Retrieved         = 103
Relevant          = 45
Relevant retrieved = 32
Average Precision: 0.6247
R Precision      : 0.6000
```

Fig. 7 Eval File

We applied various models in our dataset and compare the results. Table 1 illustrates the result of comparisons. TF_IDF gives the MAP value of 0.6481 and it is highest in the class of all its variants.

Table 1: Models

Models	BB2	DFI0	In_expC2	InL2	PL2	TF_IDF	LemurTF_IDF	BM25
MAP	0.6247	0.6115	0.6420	0.6455	0.6266	0.6481	0.6393	0.6467
R-Precision	0.6000	0.5578	0.6222	0.6444	0.6000	0.6444	0.6222	0.6444

We plot the Precision values of all the implemented models in Figure 8 as shown following.

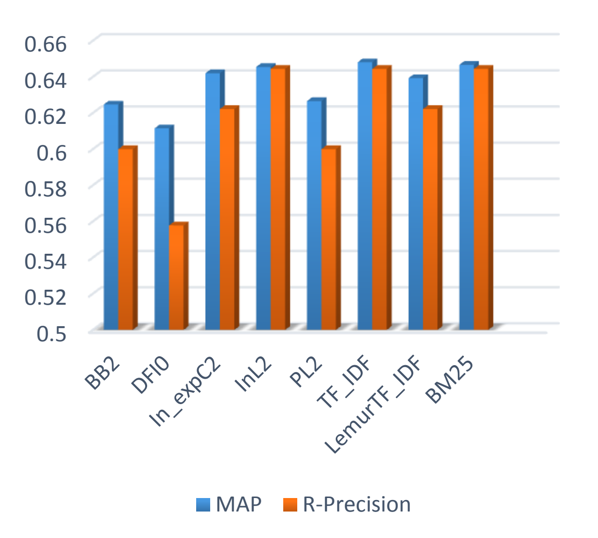

Fig. 8 Comparision of different models

6. CONCLUSION

This work has been carried out to analyze the performance of various Information Retrieval Models with the FIRE dataset which contains corpus of various newspapers. We implemented the tf-idf model and its variants and compare the results with the probabilistic model. Based on our results we conclude that tf-idf produces the best results for all the topics file. The results were evaluated and successfully compared with Terrier, the open source search engine.

7. REFERENCES

[1] An Introduction to Information Retrieval Christopher D. Manning Prabhakar Raghavan Hinrich Schütze.

[2] Sager, Juan C. *A practical course in terminology processing*. John Benjamins Publishing, 1990.

[3] Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. Modern information retrieval. Vol. 463. New York: ACM press, 1999.

[4] Frakes, William B. "Stemming Algorithms." (1992): 131-160.

[5] Patel, B. N., Prajapati, S. G., & Lakhtaria, K. I. (2012). Efficient Classification of Data Using Decision Tree. Bonfring International Journal of Data Mining, 2(1), 06-12.

[6]Xia, Tian, and Yanmei Chai. "An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm." Journal of Software (1796217X) 6.3 (2011).

[7]Alvarez, Sergio A. "An exact analytical relation among recall, precision, and classification accuracy in information retrieval." Boston College, Boston, Technical Report BCCS-02-01 (2002): 1-22.

[8] Akhilesh Sharma, Kamaljit Lakhtaria, Santosh Vishwakarma, "Data Mining Based Predictions For Employees Skill Enhancement Using Pro-Skill-Improvement Program & Performance Using Classifier Scheme Algorithm", International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Vol. 4, No. 3, March 2013, Page No. 102 – 107

[9] Robertson, Stephen. "Understanding inverse document frequency: on theoretical arguments for IDF." Journal of documentation 60.5 (2004): 503-520.

[10] Santosh K. Vishwakarma, Kamaljit I Lakhtaria, Divya Bhatnagar, Akhilesh Sharma (2014). "An efficient approach for inverted index pruning based on document relevance" Conference Proceeding of Fourth International Conference on Communication Systems and Network Technologies, Page No. 487-490. DOI 10.1109/CSNT.2014.103

[11] Lakhtaria, Kamaljit I., Bhaskar N. Patel. "Implementing R-Tree Index Optimizatioin in Core Banking system." International Journal of Research in Management, Economics & Commerce, 2(3) (2012), 42-48

[12] Saracevic, Tefko. "Evaluation of evaluation in information retrieval." *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1995.

[13] Amati, Giambattista. Probability models for information retrieval based on divergence from randomness. Diss. University of Glasgow, 2003.

[14] Robertson, Stephen, and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc, 2009.

[15] Lakhtaria, Kamaljit I. Technological Advancements and Applications in Mobile Ad-hoc Networks: Research Trends. Information Science Reference, 2012.

[16] Lakhtaria, K. I., Patel, P., & Gandhi, A. (2010). Enhancing Curriculum Acceptance among Students with E-learning 2.0. arXiv preprint arXiv:1004.2560.

[17] www.terrier.org

[18] Sharma, Akhilesh K., Kamaljit I. Lakhtaria, Avinash Panwar, and Santosh K. Vishwakarma. "An efficient approach using LPFT for the karaoke formation of musical song." In Advance Computing Conference (IACC), 2014 IEEE International, pp. 601 - 605. IEEE, 2014.