International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

Probabilistic Modelling in Information Retrieval on Ohsumed Dataset

Sajal Kaushik[1], Rahul vats[2], Narina Thakur[3]

[1,2,3]*Bharati vidyapeeth's College of Engineering,Paschim-Vihar east,Delhi-110063*

Abstract - *This paper presents a comparison of various probabilistic models and our normalized optimized formula of probabilistic models in information retrieval for improving various parameters. This paper analyzes the performance of various probabilistic models on the ohsumed dataset. Improving interpolated recall precision has been challenging on various datasets for a given set of queries.This paper presented normalized formula which marginally improved the interpolated precision recall.*

KeyWords:-**BM25,TF-idf,Poisson,Dfr,BM25f.**

1.INTRODUCTION

This paper analyzes the impact of various models on [1] Ohsu Trec dataset on the interpolated precision recall and various other parameters. BM25 is arguably one of the most important and widely used information retrieval functions. BM25F is an extension of BM25 that prescribes how to compute BM25 across a document description over several fields. [2] Pl2 model is one the normalized form of the okapi-poisson model. Finally, this paper introduces a new variant of DFR framework which have improved the accuracy for interpolated precision recall.

2. LITERATURE REVIEW

We have reviewed various paper which have analysed various probabilistic models in information retrieval. Every model gives accuracy which is quite dataset specific. BM25,tf-idf and cosine model has been implemented on cosine model.[3]This paper analyzed the interpolated precision recall of various models on ohsumed dataset.In the given below section there is a detailed analysis of various models.

2.1 Standard Models

(A) BM25

In information retrieval, Okapi BM25 (BM stands for Best Matching) is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones, and others. to push out any text that may try to fill in next to the graphic.[4]The Okapi BM25 retrieval function has been the state-of-the-art for nearly two decades.

(B) PL2 MODEL

In the field of information retrieval, [5]divergence from randomness,one of the very first models, is one type of probabilistic model. It is basically used to test the amount of information carried in the documents. It is based on Harter's 2-Poisson indexing-model.[6] The 2-Poisson model has a hypothesis that the level of the documents is related to a set of documents which contains words occur relatively greater than the rest of the documents.It is not really a 'model', but a framework for weighting terms using probabilistic methods, and it has a special relationship for Term weighting based on notion of eliteness.Term weights are being treated as the standard of whether a specific word is in that set or not. Term weights are computed by measuring the divergence between a term distribution produced by a random process and the actual term distribution.DFR models set up by instantiating the three main components of the framework: first selecting a basic randomness model, then applying the first normalization and at last normalizing the term frequencies.

(C) TF-IDF

In information retrieval, [7]tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Nowadays, tf-idf is one of the most popular term-weighting schemes; 83% of text-based recommender systems in the domain of digital libraries use tf-idf.

Variations of the tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. tf–idf can be successfully used for stop-words filtering in various subject fields, including text summarization and classification.

One of the simplest ranking functions is computed by summing the tf–idf for each query term; many more sophisticated ranking functions are variants of this simple model.

3. NEW PROPOSED MODEL

The DFR[10] models are based on this simple idea: "The more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in the document d". In other words the term-weight is inversely related to the probability of term-frequency within the document d obtained by a model M of randomness:

weight (t|d) ∝ -log ProbM (t d| Collection)

where the M stands for the type of model of randomness employed to compute the probability. In order to choose the appropriate model M of randomness, we can use different urn models. IR is thus seen as a probabilistic process, which uses random drawings from urn models, or equivalently random placement of coloured balls into urns. Instead of urns we have documents, and instead of different colours we have different terms, where each term occurs with some multiplicity in the urns as anyone of a number of related words or phrases which are called tokens of that term. There are many ways to choose M, each of these provides a basic DFR model.

Let basic model is G and computes the value:

-log Prob (t d | C= -log((1/(1+)*(/(1+))tf))

where:

- *TF* is the term-frequency of the term *t* in the collection
- *tf* is the term-frequency of the term *t* in the document *d*
- *N* is the number of documents in the Collection
- λ = T*F*/*N*.
- C is Collection.

(A)First Normalisation

When a rare term does not occur in a document then it has almost zero probability of being informative for the document. On the contrary, if a rare term has many occurrences in a document then it has a very high probability (almost the certainty) to be informative for the topic described by the document. Similarly to Ponte and Croft's language model, we include a risk component in the DFR models. [9]If the term-frequency in the document is high then the risk for the term of not being informative is minimal. In such a case Formula gives a high value, but a minimal risk has also the negative effect of providing a small information gain. Therefore, instead of using the full weight provided by the Formula , we tune or smooth the weight of Formula by considering only the portion of it which is the amount of information gained with the term:

Where:

$$P_{risk} = 1/(tf+1) \qquad \textit{(Laplace model L)}$$

4. DATASET USED

The OHSUMED[11][12]test collection is a set of 348,566 references from MEDLINE, the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). The available fields are title, abstract, MeSH indexing terms, author, source, and publication type. The National Library of Medicine has agreed to make the MEDLINE references in the test database available for experimentation.

5. RESULTS

Experimentally, the newly proposed technique have marginally outperformed standard IR models on this particular data set.

Results of various models is shown in given below table which is showing increase in interpolated precision recall of our proposed model.

	NEW_TECHN	PL2	BM25	TF_IDF
ircl_prn.0.00	**0.6011**	0.5652	0.5779	0.3741
ircl_prn.0.10	**0.5509**	0.5253	0.5474	0.3406
ircl_prn.0.20	**0.4172**	0.3892	0.4081	0.2252
ircl_prn.0.30	**0.3375**	0.3217	0.3296	0.166
ircl_prn.0.40	**0.2831**	0.2787	0.2794	0.1227
ircl_prn.0.50	**0.2361**	0.24	0.241	0.1015
ircl_prn.0.60	**0.1447**	0.1382	0.1324	0.0542
ircl_prn.0.70	**0.0667**	0.0551	0.0567	0.0247
ircl_prn.0.80	**0.0309**	0.0279	0.0296	0.0138
ircl_prn.0.90	**0.003**	0.0058	0.0056	0.0012
ircl_prn.1.00	**0.003**	0.0019	0.0019	0.0012

Table-(i)

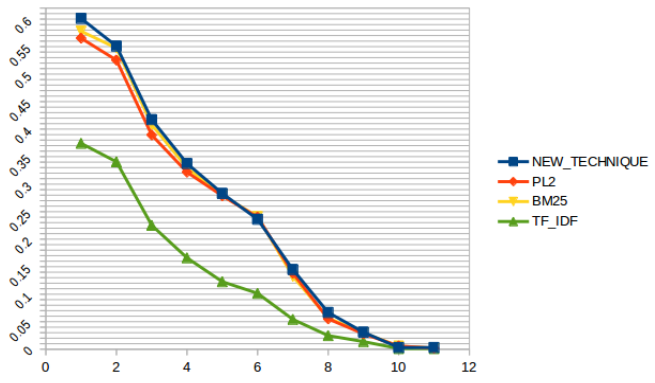

figure-(i)

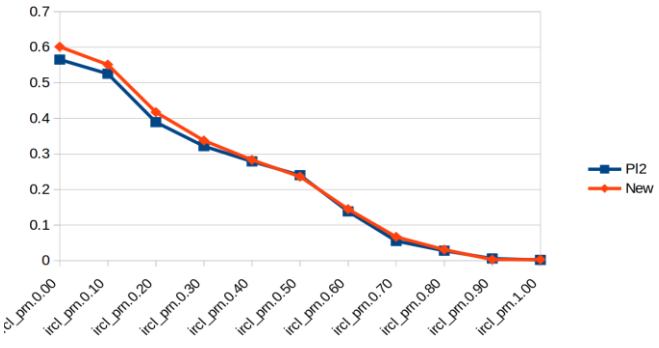

figure-(ii)

6. CONCLUSION

Interpolated precision recall is an important factor in determining the accuracy on ohsu trec dataset. In this paper We proposed a model of geometric distribution in order to improve the results.The future scope is to improvise further to achieve state of art model.

7. REFERENCES

[1]G. O. Young, "Information retrieval analysis on ohsu trec ," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2]W.-K. Chen, "qualitative analysis of models" (Book style).Belmont, CA: Wadsworth, 1993, pp. 123–135.H. Poor, [3] Dfree Model-a study using maximum likelihood". New York: Springer-Verlag, 1985, ch.

[4].B. Smith, "An approach to probabilistic models (Unpublished work style)," unpublished.

[5] Gianni Amati and Cornelis Joost Van Rijsbergen.Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst., 20:357–389, October 2002.

[6] Kenneth W. Church and William A. Gale. Poisson mixtures. Natural Language Engineering, 1:163–190, 1995.

[7] Hui Fang, Tao Tao, and ChengXiang Zhai. A formal study of information retrieval heuristics. In SIGIR '04, pages 49–56, 2004.

[8] Ben He and Iadh Ounis. On setting the hyper-parameters of term frequency normalization for information retrieval. ACM Trans. Inf. Syst., 25, July 2007.

[9] Jaakko Hintikka. On Semantic Information. In J. Hintikka and P. Suppes, editors, Information and Inference, pages 3–27. D. Reidel Pub., 1970.

[10] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. In Information Processing and Management, pages 779–840, 2000.

[11] Yuanhua Lv and ChengXiang Zhai. Lower-bounding term frequency normalization. In CIKM '11, 2011.

[12] Yuanhua Lv and ChengXiang Zhai. When documents are very long, bm25 fails! In SIGIR '11, pages 1103–1104, 2011.

[13] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In SIGIR '94, pages 232–241, 1994.

[14]Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Simple bm25 extension to multiple weighted fields. In CIKM '04, pages 42–49, 2004.

[15]Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In TREC '94, pages 109–126, 1994.

[16]Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In SIGIR '96, pages 21–29, 1996.