A Study on Retrieval Models and Query Expansion using PRF

Rekha Vaidyanathan,Sujoy Das, Namita Srivastava

Abstract— This article compares the state of the art retrieval models and reports how query expansion enhances the retrieval effectiveness. Five state- of-the- art retrieval models (parametric and non-parametric), three Query expansion Techniques Bo1, Bo2 and KL are selected and presented. A comparative study of the retrieval models, namely TF_IDF, DLH, DPH, I(n)L2 and PL2 enhanced with the mentioned QE models are experimentally shown using FIRE 2011 Adhoc data. This is an initial study carried out to understand how the performance of these approaches varies on multiple languages (English and Hindi).Furthermore, we explore the optimal parameter settings for the non-parameteric models incase of Short, Normal and Long queries. Results show that I(n)L2 performed well for Hindi dataset and BM25 and PL2 gave best MAP for English dataset. We use the Terrier, the information Retrieval framework for indexing, retrieval and evaluation. The models used for comparison are Terrier's DFR based weighting models.

Index Terms—pseudo relevance feedback, retrieval models, query expansion, FIRE, Terrier

1 INTRODUCTION

The purpose of retrieval models is to retrieve and rank the set of relevant documents based on the user's query. Boolean models, statistical models like the Vector Space Model, Probabilistic models and Language models have been developed for Information Retrieval. The Vector space and the probabilistic retrieval models give significantly good results in terms of Precision and recall compared to the exact match or Boolean retrieval models [1].

From the naïve method of classifying the documents as matching or otherwise, the next generation retrieval models have term weighting schemes where documents are ranked on their degree of relevance. Each document is given a score based on the words they contain pertaining to a given topic and ranked accordingly. These term weighting schemes can be parametric and non-parametric. Almost all term weighting models use the term frequency (tf), the number of times a term t appears in a document d as the basis for calculating the score [2]. *tf.idf* is the most commonly used term weighting scheme where the inverse document frequency (idf), introduced by Karen Sparck Jones, computes the term specificity. The formula for idf is given by,

$$idf_t = \log(N/df_t) \quad (1)$$

- N is the total number of documents in Collection C,
- df_t is the document frequency for term t.
- tf combined with idf give high weights to rare terms and low weights to more frequent ones.

The statistical models usually use different weighting models for ranking the documents. These term weighting methods on documents are based on the query input by the user. Most of the time, the Collection that is to be searched contain relevant and irrelevant information. IR faces the two-sided problem of the searchers not being able to frame the best suitable query for their information need and also lack of information regarding the collection used for retrieval [3]. This led to the development of the concept called Relevance Feedback where [4];

1. User submits a query
2. System retrieves an initial set of results
3. User marks this set as relevant or irrelevant
4. Based on the user's feedback, system retrieves a better-set of results.

Manually skimming through the initial set of documents and marking them as relevant or irrelevant is a tedious task. Pseudo Relevance Feedback or blind Relevance Feedback automates this marking system and it assumes that the top *k* ranked documents of the initially retrieved results are relevant. Terms related to the search query are selected from these documents to improve the query representation with the help of query Expansion [5]. The process of adding more significant and contextually similar words to the original query is called *query expansion*. Most often, queries contain terms that may not match the indexed terms leading to lesser accuracy in retrieval process [15]. This problem is addressed by relevance feedback, an automatic process of query reformulation, where important words are chosen from previously retrieved documents that are relevant to the query [16]. Thus the basic idea behind query expansion is to augment the query with related terms like synonyms, plurals, modifiers, category keywords etc. for improving the retrieval accuracy [6]. Many Techniques have been proposed by researchers for query expansion. For our study, we select three QE models namely Bo1, Bo2 and KL. A comparative study on the retrieval effectiveness of state-of-the-art retrieval models on two different languages is introduced in this paper. Also, we investigate the effectiveness of applying query expansion to improve the retrieval accuracy. The study is mainly done to understand which baseline works most effectively on multiple languages of the FIRE Adhoc 2011 Test collection. We also try to understand the effectiveness of the optimal QE model and how it improves the retrieval accuracy. Terrier™ is used as Information Retrieval framework for all our experiments [17].

The paper is organized as follows:
The weighting models selected for retrieval and QE in our experiments are discussed in Section 2 and 3 in detail. Section 4 showcases the Experimental Results of different Retrieval models and the effect of Query expansion on them. Section 5 contains concluding remarks.

2 WEIGHTING MODELS

Several weighting models have been proposed by many researchers in IR. In this paper we will discuss a few parametric and non-parametric models. In the parametric models, there involves a hyper parameter tuning for length normalization. Since each query behaves differently the optimal parameter setting is different for each of them. Through our experiments, we find the optimal values for normalization parameters by selecting the one that gives the highest MAP. Discussed are BM25, I (n) L2 and PL2 Parametric Models and DPH, DLH parameter free models. These models are based on Terrier's Divergence from Randomness DFR Models [7].

2.1 Parametric Models

2.1.1 OKAPI's BM25

BM25 is one of the best known term-weighting schemes derived from the probabilistic model. BM25 is a family of scoring functions and BM stands for Best Match. It takes into account the three components namely, the term frequency, inverse document frequency and the length of the document [8]. In this method, each document D is scored against a Query q given by the formula:

$$w_t = tf_d.(A/B) \quad (2)$$

where,

$$A = (\log[(N - n + 0.5)/(n + 0.5)] \quad (3)$$

$$B = (k_1.((1 - b) + b.(dl/avdl) + tf_d) \quad (4)$$

- w_t is the relevance weight assigned to a document due to query term t,
- tf_d is the number of times t occurs in document
- N is the total number of documents, n is the number of documents containing at least one occurrence of t; dl is the length of the document and avdl is the average document length.
- k_1 is the term-frequency influence parameter $1.0 \leq k_1 \leq 2.0$
- b is the normalization parameter $0.0 \leq b \leq 1.0$, for document length. b can be set to zero of the document length need not be considered.

2.1.2 I (n)L2

An Inverse document Frequency model with LaPlace after-effect normalization 2. The scoring function is given by:

$$w_t = (1/tf_n + 1)\,.(tf_n.log_2\,[N + 1/N_t + 0.5] \quad (5)$$

where, tf_n is the normalized term frequency given by the formula:

$$tf_n = tf\,.log_2[1 + c.(avgl/l)] \quad (6)$$

- c is the term frequency normalization parameter
- l is the document length which corresponds to number of tokens in a document and
- *avg_l* is the average document length in the collection.

2.1.3 PL2:

A Posisson model with Lapalce after effect and normalization 2. PL2 is one of the Divergence from Randomness weighting models [9]. Scoring function PL2 is given by:

$$w_t = (1/(tf_n + 1))\,(A + B + C) \quad (7)$$

where,

$$A = tf_n.log_2[tf_n/\lambda] \quad (8)$$

$$B = (\lambda + (1/12.tfn) - tfn).log_2 e \quad (9)$$

$$C = 0.5.log_2(2\pi.tfn) \quad (10)$$

- tfn is the normalized term frequency as explained in (4).
- λ is the mean and variance of a Poisson distribution.

2.2 PARAMETER FREE MODELS

2.2.1 DLH: DLH hyper geometric DFR Model

This is a DFR model based on the hyper geometric distribution of tf. For a workable weighting function, the hyper geometric function is reduced to binomial distribution with non-uniform term priors [10]. It is a parameter free model and there is no need for expensive training. This model assumes that the occurrences of a query term in a document are samples from the whole collection instead of from the document [11]. The scoring function is given by:

$$Score(d,Q)_t = \sum_{t \epsilon Q} qtw\,.((1/(tf + 0.5))\,.(log_2[(tf.avg\_l)/l\,.(N/F)]) + 0.5\,log_2(2\pi tf(1 - (tf/l)) \quad (11)$$

where,

- F is given by tf/l is within document frequency
- l is the document length in tokens.
- avg_l is the average document length in collection
- tf is the term frequency in the collection.

2.2.2 DPH

DPH is a parameter free scoring technique which is derived from the Divergence from Randomness model [12]. The scoring function is given by[19]:

$$Score(d,Q) = \sum_{t \epsilon Q} ((qtw\,(1 - F)^2/(tf + 1))\,.(tf.log_2(tf.(avg\_l/l).(N/TF))) + 0.5\,.log_2(2\pi.tf.(1 - F)) \quad (12)$$

DPH, like DLH is a parameter free model.

- qtw = qtf/qtf_{max},
- where, qtf is the query term frequency and qtf_{max} is the maximum query term frequency among all query terms.
- N is the total number of documents
- avg_l is the average document length in collection
- tf is the term frequency in the collection.

3 QUERY EXPANSION MODELS

For our experiments, we use Terrier's DFR-based Term weighting models namely, Bo1, Bo2 and KL. Terrier employs a Divergence from Randomness based QE mechanism which is a generalization of Rocchio's method [18]. In the first step, the term weights of the terms from top ranked documents are calculated. The DFR model calculates the informativeness of a term by the divergence of its distribution in top ranked documents from random distribution [10]. The top most informative terms are then extracted ad merged with the original query to form an expanded one. Weighting schemes for the three models mentioned are as given in the sections 3.1, 3.2 and 3.3 [13].

3.1 Kullback Leibler

The Scoring function is given by,

$$W(t)=P_x . log_2[{}^{x}/_{P_c}] \quad (6)$$

- $P_x = tf_x / l_x$;
- tf_x is the frequency of the query term in the top-ranked documents.
- l_x is the sum of the length of the exp_doc top ranked documents where exp_doc is a parameter of the query expansion methodology.

3.2 Bo1

This model is based on the based on the Bose Einstein Statistic and the weight of the term t in the top ranked documents (rank ranging from 3 to 10) is given by [14].

$$w(t) = tf_x . log_2 \frac{\quad}{P_n} + log_2(1 + P_n) \quad (7)$$

- tfx is the frequency of the query term in the top-ranked documents
- Pn is given by F/N, where F is the term frequency in the collection and N is the number of documents in the collection.

3.2 Bo2

The scoring function of Bo2 is given by :

$$w(t) = tf_x . log_2 \frac{\quad}{P_f} + log_2(1 + P_f) \quad (8)$$

- tf_x is the frequency of the query term in the top-ranked documents
- P_n is given by F/N, where F is the term frequency in the collection and N is the number of documents in the collection.
- $Pf_x = (tf_x . l_x)/token_c$; where lx is the sum of the length of the exp_doc top ranked documents where exp_doc is a parameter of the query expansion methodology.
- F, is the term frequency of the query term in the whole collection.
- $token_c$, is the total number of tokens in the whole collection.

4 EXPERIMENTS AND RESULTS

Our experiments are performed with Terrier Information Retrieval framework. It provides indexing, retrieval and evaluation for English and non-english documents. For the evaluation of various retrieval models and performance of QE models on them, we use both English and Hindi collections. This is provided by Forum for Information Retrieval Evaluation (FIRE) and the dataset conforms to the TREC style Format. 100 topics were chosen with 50 each for one language. They are numbered from 126-175 for Adhoc English and Hindi 2011dataset. The corpus in encoded in UTF-8 format and the tags are as follows:

```
<topics> <top>
<num>126</num>
<title>Swine flu vaccine</title>
<desc>Indigenous vaccine made in India for swine flu prevention</desc>
<narr>Relevant documents should contain information related to making indigenous swine flu vaccines in India, the vaccine's use on humans and animals, arrangements that are in place to prevent scarcity / unavailability of the vaccine, and the vaccine's role in saving lives.</narr>
</top></topics>
```

Each of the FIRE Topic consists of three fields: title, description and narration. All the three types of queries were experimented to understand the impact of query length [2]. We evaluate the performance of this dataset on (i) different retrieval models (parametric and parameter-free) and (ii) enhancing the retrieval models using query expansion. We experiment with Short Queries (title field only), Normal Queries (Description field) and Long Queries (title + description + narration)

Evaluation is done for TF_IDF, PL2, I(n)L2,DLH, DPH and BM25 to study the impact of short, normal and long queries. The optimum values for the parameters b and c have been set manually. The value that gives the highest MAP is considered as optimum.

4.1 Experiments with Short Queries

We tested with the 6 models out of which DLH and DPH are parameter free. The MAP, R Precision (R is the relevant retrieved documents), Precision at 10 and 20 documents are reported in Table 1.

- Results from Hindi_Short_Query (Table 1): The highest MAP was obtained for PL2 model where term frequency normalization parameter c is set to be 4.0 for optimum result. This was set manually. R Precision also was highest for PL2. But Precision at 10 and 20 documents was better for DPH model .Thus, if we consider the MAP, PL2 model with c set to 4.0 gave the best results for Hindi short queries followed by I(n)L2 and DPH.

TABLE 1
MAP FOR HINDI SHORT QUERIES

	TF_IDF	DLH	DPH	PL2 (C=4.0)	BM25 (b=0.25)	InL2 (c=1.0)
MAP	0.2241	0.2295	0.2433	**0.2467**	0.2214	**0.2454**
R Precision	0.2504	0.2613	0.2772	**0.2848**	0.2576	0.2775
P@10	0.3580	0.3640	**0.4040**	0.3980	0.3820	0.3920
P@20	0.3140	0.3290	**0.3670**	0.3600	0.3450	0.3550

- Results from English_Short_Query (Table 2): The Highest MAP was obtained for PL2 at c=5.0 and BM25 at b=0.25. R Precision is marginally better for DPH compared to PL2 and BM25. P@10 was best obtained for PL2 and P@20 for DPH. Thus we got a distributed result but overall, we can conclude that both PL2 and BM25 retrieval worked well for English short queries.

TABLE 2
MAP FOR ENGLISH SHORT QUERIES

	TF_IDF	DLH	DPH	PL2 (C=5.0)	BM25 (b=0.25)	InL2 (c=2.0)
MAP	0.2965	0.2881	0.3102	**0.3138**	**0.3139**	0.2977
R Precision	0.3208	0.3148	**0.3303**	0.3273	0.3258	0.3136
P@10	0.4200	0.3980	0.4560	**0.4640**	0.4440	0.4440
P@20	0.3820	0.3740	**0.4060**	0.4030	0.4050	0.3900

4.2 Experiments with Normal Queries

For Normal queries, only the description field was considered with number of words ranging from 7-10.

- Results for Hindi_Normal_Query (Table 3): The highest MAP is obtained for I(n)L2 model with c value at 0.75. Rest of the values: R Precision, P@10 and P@20 are also highest for this model.

TABLE 3
MAP FOR HINDI NORMAL QUERIES

	TF_IDF	DLH	DPH	PL2 (C=2.0)	BM25 (b=0.25)	InL2 (c=0.75)
MAP	0.2498	0.2529	0.24	0.25	0.2072	**0.2651**
R Precision	0.2805	0.2803	0.2691	0.2797	0.2505	**0.2951**
P@10	0.4040	0.3860	0.4040	0.4100	0.3740	**0.4300**
P@20	0.3480	0.3420	0.3500	0.3550	0.3210	**0.3630**

- Results for English_Normal_Query (Table 4): Highest MAP is obtained for BM25 at b=0.5. Precision at 10 documents was best obtained for DPH. Precision at 20 was best at PL2 with c=2.0.

TABLE 4
MAP FOR ENGLSIH NORMAL QUERIES

	TF_IDF	DLH	DPH	PL2 (c=2.0)	BM25 (b=0.5)	InL2 (c=0.75)
MAP	0.3694	0.3586	0.3765	0.377	**0.3816**	0.3696
R Precision	0.3888	0.379	0.3893	0.3912	**0.4008**	0.388
P@10	0.5160	0.5120	**0.5360**	0.5260	0.5200	0.5060
P@20	0.4640	0.4580	0.4740	**0.4780**	0.4740	0.4640

4.3 Experiments with Long Queries

Long queries comprise of the (i) title field (ii) description field and (iii) narration field. The average query length is 20-30 words.

- Results for Hindi Long Queries (Table 5): The highest MAP is obtained for BM25 with b =0.75. Rest of the values for Recall and P@10 and 20 are also higher with this model.

TABLE 5
MAP FOR HINDI LONG QUERIES

	TF_IDF	DLH	DPH	PL2 (C=2.0)	BM25 (b=0.75)	InL2 (c=0.5)
MAP	0.2267	0.2047	0.2158	0.2126	**0.3506**	0.2402
R Precision	0.26	0.2384	0.2534	0.2487	**0.3695**	0.2711
P@10	0.3600	0.3300	0.3500	0.3420	**0.4820**	0.3740
P@20	0.3360	0.2960	0.3110	0.3010	**0.4590**	0.3450

- Results for English Long Queries (Table 6): The highest MAP is obtained for InL2 model with c value at 0.5. However, PL2 at c=1.0 fared well for P@10 and 20 documents.

TABLE 6
MAP FOR ENGLISH LONG QUERIES

	TF_IDF	DLH	DPH	PL2 (C=1.0)	BM25 (b=0.25)	InL2 (c=0.5)
MAP	0.3463	0.3313	0.33	0.3488	0.1554	**0.3539**
R Precision	0.3667	0.3561	0.3546	0.3722	0.1906	**0.3741**
P@10	0.4900	0.4800	0.4900	**0.5040**	0.2960	0.4980
P@20	0.4550	0.4370	0.4460	**0.4740**	0.2540	0.4690

4.3 Experiments for Query Expansion with Bo1, Bo2 and KL models on short queries

We perform enhancement of the retrieval models explained here using Query expansion on short queries and test whether the results are significantly different. 1000 documents each are retrieved initially using each of the retrieval models explained in Table 7&8. Each of the models is enhanced with Bo1, Bo2 and KL models and is checked for the improvement in MAP. For QE, top 10 documents are used and 10 words are used for expansion with "title" query which is short. Table 7 & 8 shows the results obtained after query expansion with 10 terms from top 10 documents.

QE on Hindi Short Queries: Results indicate that the highest MAP was obtained for I(n)L2 enhanced with Bo1 model with c value set to 1.0. The delta obtained from baseline for this is 24%. Among the baseline, PL2 with c=4 and InL2 with c=1 gave the best MAP for initial retrieval with 0.2467 and 0.2454 respectively. The highest delta +34% was obtained for BM25 (b-0.25), enhanced with Bo1. We can also observe that Query expansion using Bo2 weighting hurt the MAP drastically and deteriorated the results in DLH (-18%), PL2(-4.8%), BM25 (-80%).

TABLE 7
QE ON HINDI SHORT QUERIES

	MAP (%improvement)	RPrecision	P@10	P@20
TF_IDF	0.2241	0.2504	0.3580	0.3140
TF_IDF_Bo1	0.2785 (+24%)	0.2924	0.418	0.372
TF_IDF_Bo2	0.2415(+7%)	0.267	0.354	0.332
TF_IDF_KL	0.2768(+23%)	0.2921	0.404	0.368
DLH	0.2295	0.2613	0.3640	0.3290
DLH_Bo1	0.2919(+27%)	0.3097	0.4240	0.3860
DLH_Bo2	0.1861(-18%)	0.2145	0.3240	0.2750
DLH_KL	0.2912(+27%)	0.3039	0.4260	0.3850
DPH	0.2433	0.2772	0.4040	0.3670
DPH_Bo1	0.3004(+23%)	0.3181	0.434	0.367
DPH_Bo2	0.2467(+1.39%)	0.2706	0.392	0.349
DPH_KL	0.3(+23%)	0.3155	0.436	0.391
PL2_C4.0	0.2467	0.2848	0.3980	0.3600
PL2_Bo1	0.3033(+22%)	0.3253	0.43	0.398
PL2_Bo2	0.2348(-4.8%)	0.2746	0.382	0.342
PL2_KL	0.3021(+22.45%)	0.3218	0.436	0.399
BM25_b0.25	0.2214	0.2576	0.3820	0.3450
BM25_b0.25_Bo1	**0.2976(+34%)**	0.3045	0.432	0.398
BM25_b0.25_Bo2	0.0426(-80%)	0.0599	0.0918	0.0776
BM25_b0.25_KL	0.2973(34%)	0.3094	0.436	0.396
InL2_c1.0	0.2454	0.2775	0.392	0.355
InL2_c1.0_Bo1	**0.3056**(+24%)	0.3169	0.432	0.418
InL2_c1.0_Bo2	0.2609(+6.3%)	0.2887	0.392	0.362
InL2_c1.0_KL	0.3034(23.6%)	0.3191	0.43	0.417

QE on English Short Queries: Results on English data show that the highest MAP was obtained for PL2 (c=5.0) enhanced with Bo1 with an 18% improvement over baseline. The highest MAP among baseline was given by PL2 and BM25 (b=0.25). The highest delta was obtained for TF_IDF (+23%) with Bo2 and InL2 (23.9%) with c=2.0. The KL and Bo1 models performed almost similarly in all the cases.

TABLE 8
QE ON ENGLISH SHORT QUERIES

	MAP (%improvement)	RPrecision	P@10	P@20
TF_IDF	0.2965	0.3208	0.4200	0.3820
TF_IDF_Bo1	0.3548(+19.66%)	0.3661	0.4580	0.4090
TF_IDF_Bo2	0.3652(+23%)	0.3721	0.4880	0.4350
TF_IDF_KL	0.3543(+19%)	0.3694	0.4500	0.4040
DLH	0.2882	0.3148	0.3980	0.3740
DLH_Bo1	0.3473(+20.5%)	0.3666	0.4460	0.3990
DLH_Bo2	0.3423(+18.77 %)	0.3553	0.4420	0.4070
DLH_KL	0.3452(+19.77%)	0.3641	0.4480	0.4000
DPH	0.3102	0.3303	0.4560	0.4060
DPH_Bo1	**0.3723(+20%)**	0.3778	0.4900	0.4360
DPH_Bo2	0.3712(+19.66%)	0.3771	0.5020	0.4460
DPH_KL	0.3703(+19.37%)	0.3764	0.4960	0.4360
PL2_C5.0	0.3138	0.3273	0.4640	0.4030
PL2_Bo1	**0.3724(+18.6%)**	0.3766	0.4880	0.4430
PL2_Bo2	0.3686(+17.46%)	0.3699	0.4920	0.4260
PL2_KL	0.3712(+18.29%)	0.3771	0.5040	0.4340
BM25_b0.25	0.3139	0.3258	0.4440	0.4050
BM25_b0.25_Bo1	0.3651(16.3%)	0.3715	0.4860	0.4330
BM25_b0.25_Bo2	0.3633(+15.7%)	0.371	0.4900	0.4240
BM25_b0.25_KL	0.3629(+15.6%)	0.3682	0.4700	0.4220
InL2_c2.0	0.2977	0.3136	0.4440	0.3900
InL2_c1.0_Bo1	0.3629(+21%)	0.3626	0.4760	0.4190
InL2_c1.0_Bo2	0.369(+23.9%)	0.3684	0.4860	0.4300
InL2_c1.0_KL	0.3627(+21.8%)	0.3625	0.4760	0.4270

5 CONCLUSION

In our experiment we conducted a study on effectiveness shown by QE. This is an initial step towards identifying a baseline for our future experiments that involves finding a term weighting strategy for Query Expansion using PRF. We also investigated the improvement shown by state-of-the-art QE models on FIRE Collection. The results of our study show that there is a relation between the retrieval effectiveness and query expansion as mentioned by previous researchers. Also, Query Expansion has improved the MAP of the retrieval by 18-20% for the FIRE 2011 Collection.

For Hindi dataset, Bo1 model gave the best results where as all three models performed similarly for English dataset. It is

also observed that the b parameter in BM25 was optimum at 0.25 for both the Collections, in case of short queries.

For Hindi dataset, InL2 performed better with different c values for short and normal queries. Though PL2 gave the best result for short queries, the result is not very significantly different from InL2. Applying QE on PL2 did not show much improvement and incase of Bo2, it hurt the MAP by 4.8%. Infact, Bo2 consistently did not improve the MAP in any significant manner for Hindi dataset.Hence we can support the suggestion that it is not only the quality of the top ranked documents but also the quality of the reweighting fo the query terms that improves the retrieval effectiveness [11]. In case of English dataset, BM25 at b=0.25 and PL2 gave the highest MAP during initial retrieval for short and normal queries. MAP was improved by 17-18% using Bo1, Bo2 and KL for English dataset.

It is observed that the drawback with the parametric models is that they require the parameter tuning and in case of automatic query expansion, setting the parameter automatically would in itself be a research problem. Our future study aims at formulating a QE model that will find the optimal values for parameters, if any, automatically and yield better results compared to the state-of-the-art models. We would also like to consider the length of the query while reformulating it as this can reduce the iterations during retrieval. Thus our future work aims at integrating both these aspects effectively and giving improved results for retrieval.

ACKNOWLEDGMENT

The authors sincerely thank FIRE for providing test collections especially on Indian Languages and for the relevance judgement files. The authors would also like to thank the Terrier team for the Terrier Retrieval Engine which was used for indexing, retrieval and evaluation of the experiments on different languages.

REFERENCES

[1] Turtle, H. R., & Croft, W. B. (1992). A comparison of text retrieval models. The computer journal, 35(3), 279-290.

[2] He, Ben, and Iadh Ounis. "Term frequency normalisation tuning for BM25 and DFR models." Advances in Information Retrieval. Springer Berlin Heidelberg, 2005. 200-214.

[3] Ruthven, Ian, and Mounia Lalmas. "A survey on the use of relevance feedback for information access systems." The Knowledge Engineering Review 18.02 (2003): 95-145.

[4] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Vol. 1. Cambridge: Cambridge university press, 2008.

[5] Lv, Yuanhua, and ChengXiang Zhai. "Positional relevance model for pseudo-relevance feedback." Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010.

[6] Vaidyanathan, Rekha, Sujoy Das, and Namita Srivastava. "Query Expansion Based on Equi-Width and Equi-Frequency Partition." Multilingual Information Access in South Asian Languages. Springer Berlin Heidelberg, 2013. 13-22.

[7] Plachouras, Vassilis, Ben He, and Iadh Ounis. "University of Glasgow at TREC 2004: Experiments in Web, Robust, and Terabyte Tracks with Terrier." TREC. 2004.

[8] Hawking, David, Trystan Upstill, and Nick Craswell. "Toward better weighting of anchors." Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004.

[9] G. Amati and C. J van Rijsbergen. Probabilistic models of Information Retrieval based on measuring the divergence from randomness. In ACM Transactions on Informaton Systems (TOIS), volumen20 (4), pages 357-389, 2002.

[10] Lu, Sha, Ben He, and Jungang Xu. "Hyper-geometric Model for Information Retrieval Revisited." Information Retrieval Technology. Springer Berlin Heidelberg, 2013. 62-73.

[11] He, Ben, and Iadh Ounis. "Combining fields for query expansion and adaptive query expansion." Information processing & management 43.5 (2007): 1294-1307.

[12] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G.Gambosi. FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track. In Proceedings of TREC 2007.

[13] Plachouras, Vassilis, Ben He, and Iadh Ounis. "University of Glasgow at TREC 2004: Experiments in Web, Robust, and Terabyte Tracks with Terrier." TREC. 2004.

[14] Macdonald, C., He, B., Plachouras, V., & Ounis, I. (2005). University of Glasgow at TREC 2005: Experiments in terabyte and enterprise tracks with terrier. In Proceedings of the 14th text retrieval conference (TREC 2005). Gaithersburg, MD.

[15] Harman, Donna. "Relevance Feedback and Other Query Modification Techniques." (1992): 241-263.

[16] Salton, Gerard, and Chris Buckley. "Improving retrieval performance by relevance feedback." Readings in information retrieval 24.5 (1997).

[17] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006, August). "Terrier: A high performance and scalable information retrieval platform". In Proceedings of the OSIR Workshop (pp. 18-25).

[18] Rocchio, J. (1971). Relevance feedback in Information Retrieval. In G. Salton (Ed.), The SMART retrieval system: Experiments in automatic document processing (pp. 313–323). Prentice-Hall Englewood Cliffs.

[19] McCreadie, R., Macdonald, C., Ounis, I., Peng, J., & Santos, R. L. (2009). University of glassgow at trec 2009: Experiments with terrier. GLASGOW UNIV (UNITED KINGDOM).