IMPROVING SEARCH ENGINES BY DEMOTING NON-RELEVANT DOCUMENTS

Fadi Yamout and Mireille Makary

The International University of Beirut, Department of Computer Science, Beirut

ABSTRACT

A good search engine aims to have more relevant documents on the top of the list. This paper describes a new technique called "Improving search engines by demoting non-relevant documents" (DNR) that improves the precision by detecting and demoting non-relevant documents. DNR generates a new set of queries that are composed of the terms of the original query combined in different ways. The documents retrieved from those new queries are evaluated using a heuristic algorithm to detect the non-relevant ones. These non-relevant documents are moved down the list which will consequently improve the precision. The new technique is tested on WT2g test collection. The testing of the new technique is done using variant retrieval models, which are the vector model based on the TFIDF weighing measure, the probabilistic models based on the BM25, and DFR-BM25 weighing measures. The recall and precision ratios are used to compare the performance of the new technique against the performance of the original query.

KEYWORDS

Information retrieval, TFIDF, BM25, DFR-BM25.

1. INTRODUCTION

Search engines extract user-specified information from documents and files, ranging from books to online blogs, journals, and academic articles [1]. The primary objective of search engines is to quickly and precisely retrieving relevant documents related to the user's request [2]. Search engines cannot be 100% accurate because the document relevance is subjective and depends on the user's judgment, which depends on many factors such as his knowledge about the topic, the reason for searching, and his satisfaction with the returned result [3].There are many challenges involved in making a search engine successful [2,4]. These challenges include acquiring lots of relevant documents from many sources, extracting useful representations of the documents to facilitate search, ranking documents in response to a user request, and presenting the search results effectively by posting the most relevant document on the top of the list [5,6,7].

This paper describes a new technique to improve search engines performance. The new technique locates non-relevant documents among the documents retrieved and moves them down the list. As a result, more relevant documents are lifted up the list, and consequently, the performance of the search engine [8] improves. This is done by generating new queries from the original query, retrieve a set of document for each new generated query, combine them into one set and use a heuristic to determine the most non-relevant documents. The new technique is tested on WT2g[1] test collection using the vector model [9,10,11,12] based on the TFIDF weighing measure[13,14], the probabilistic models [15] based on the BM25, and DFR-BM25 weighing measures[16,17,18]. The recall and precision ratios are used to compare the performance of the new technique against the performance of the original query.

[1] http://ir.dcs.gla.ac.uk/wiki/Terrier/WT2G

DOI: 10.5121/ijnlc.2019.8401

2. MODELS

The Information Retrieval (IR) model defines a way to represent the documents and queries to compare them [19, 20]. The most common IR models are vector [9, 10] and probabilistic models[15].

2.1. Vector Model

In a vector model, both the documents and the queries are represented as vectors in multi-dimensional space, where the terms become the dimensions' vector [9, 10, 11, 12]. Therefore, document d is represented as a vector of terms, as shown in Equation [11]:

$$d = (t_1, t_2 \dots t_m) \quad (1)$$

In this equation, m represents the number of unique terms in the collection and t_i denotes the presence or absence of term i in document d. Vector model is based on linear algebra allowing documents to be ranked based on their possible relevance to the query [11].

2.2. Probabilistic

In a probabilistic model, the documents and queries are viewed as vectors. However, the weight of a term in a document is given by a probability [15]. Probabilistic models have been extended to different models; Best Match 25 (BM25) [16, 17, 18] Okapi [21], Statistical Language Modelling (SLM) [22], and Divergence From Randomness (DFR-BM25) [17].

3. WEIGHTING TERMS

Information retrieval system has various methods for weighting terms [23, 24, 25]. The primary weights are TFIDF [14, 13], BM25 [18] and DFR_BM25 [16, 17, 18], weighting measure. Assigning a weight for each term in each document in the collection has shown great potential for improving retrieval effectiveness [24].

3.1. TFIDF

TFIDF weighting measure is a combination of local and global weights [13, 14]. The term frequency (TF_{ij}) is based on the notion that terms that frequently occur in the text of documents are essential in that text. Therefore, it represents the occurrences of a term i in a document j. The global weight is the document frequency ($DocFreq_i$), which represents in how many documents the term i occurs. Inverse Document Frequency (IDF_i) of term i has $DocFreq_i$ scaled to the total number of documents in the collection (N) as shown in Equation 2:

$$IDF_i = \log_{10}\left(\frac{N}{DocFreq_i}\right) \quad (2)$$

In equation (2), N is the total number of documents, $DocFreq_i$ is the total number of documents containing term i, and the log is based 10. The logarithm reduces the large value obtained due to N since IDF_i is to be multiplied by the small value TF_{ij} to derive the TFIDF as shown in Equation 3:

$$TFIDF = TF_{ij} \times IDF_i \quad (3)$$

3.2. BM25

BM25 [18] also known as "Best Match 25", is the main weighting measures for probabilistic models. BM improved from the traditional probabilistic weighting scheme to BM25 through BM11 and BM15. It is the best of the known probabilistic weighting schemes by recent TREC tests.

The weights assigned to the documents' terms are given by a probability shown in Equation 4:

$$BM25 = \frac{\mathrm{TF_{ij}} \times (\mathrm{k_3}+1) \times \mathrm{QF_{iq}}}{\left((\mathrm{k_3}+\mathrm{QF_{iq}}) \times \mathrm{K}\right)} \times \log_2\left(\frac{\mathrm{N}-\mathrm{DocFreq_i}+0.5}{\mathrm{DocFreq_i}+0.5}\right) \quad (4)$$

Where QF_{iq} represents the occurrences of a term i in a query q, k3 is set to 1000, as proposed in [18], and K is shown in Equation 5:

$$\mathrm{K} = \mathrm{k_1} \times \left((1-\mathrm{b}) + \mathrm{b} \times \frac{\mathrm{DocL_j}}{\mathrm{averageDocL}}\right) + \mathrm{TF_{ij}} \quad (5)$$

Where DocLj is the length of document j, averageDocL is the average length of all documents in the test collection, k1 is set to 1.2, and b is set to 0.75 as proposed in [18].

3.3. DFR-BM25

DFR_BM25 [16, 17], is derived by measuring the divergence of the actual term distribution from that obtained under a random process. The weight of the term in a document is computed as a function of two probabilities $Prob_1$ and $Prob_2$. Equation 6 shows the weight of a term as a product of two components.

$$w = (1 - Prob_2) \times (-\log_2(Prob_1)) \quad (6)$$

$Prob_2$ measures the information gain of the term concerning the set of all documents in which the term occurs. It is measured by the counter-probability (1- $Prob_2$), where the less the term is expected in a document concerning its frequency in the set of all documents in which the term occurs, the more the amount of information is gained with this term. The counter-probability 1 – $Prob_2$ is computed, as shown in Equation 7:

$$(1 - \mathrm{Prob_2}) = \left(\frac{\mathrm{TF_{ij}}+1}{\mathrm{DocFreq_i} \times (\mathrm{TF_i}+1)}\right) \quad (7)$$

$Prob_1$ measures the information content of the term in a document. The component ($-\log_2(Prob_1)$) provides the equivalent amount of information and is computed, as shown in Equation 8:

$$(-\log_2 Prob_1) = \mathrm{TF_i} \times \log_2\left(\frac{\mathrm{TF_i}}{\lambda}\right) + \left(\lambda + \frac{1}{12 \times \mathrm{TF_i}} - \mathrm{TF_i}\right) \times \log_2 e + 0.5 \times \log_2(2 \times \pi \times \mathrm{TF_i}) \quad (8)$$

Where TF_i is the term frequency of the term i in the collection, and $\lambda = TF_i/N$.

4. TEST COLLECTION

A test collection consists of a large collection of documents, a set of queries, and a relevance judgment list which matches each query to its relevant documents [7, 26, 27]. In this paper, WT2g[2] test collection is used for the experiments. It has a size of 2 GB and consists of 247491

[2] http://ir.dcs.gla.ac.uk/wiki/Terrier/WT2G

documents. WT2g has 50 topics with a variant number of terms in each query. In our technique, we use the queries that are composed of three terms, as shown in Table 1, since we can generate more queries out of three terms.

Table 1. Queries of WT2g Composed of Three Terms

Query number	Query content	Query number	Query content
401	foreign minorities, Germany	423	Milosevic, Mirjana Markovic
404	Ireland, peace talks	426	law enforcement, dogs
407	poaching, wildlife preserves	427	UV damage, eyes
409	legal, Pan Am, 103	428	declining birth rates
411	salvaging, shipwreck, treasure	430	killer bee attacks
414	Cuba, sugar, exports	432	profiling, motorists, police
415	drugs, Golden Triangle	433	Greek, philosophy, stoicism
416	Three Gorges Project	435	curbing population growth
419	recycle, automobile tires	437	deregulation, gas, electric
420	carbon monoxide poisoning	439	inventions, scientific discoveries
421	industrial waste disposal	443	U.S., investment, Africa
422	art, stolen, forged	450	King Hussein, peace

Query 428, for example, has the following three terms: “declining”, “birth”, and “rates”. WT2g also has a Relevance Judgment List (RJL) that indicates the relevant documents for each query [7, 27]. For example, RJL indicates that query 401 “foreign minorities, Germany“, has 2739 relevant documents.

An example of a few relevant documents to query 401 is shown in Table 2.

Table 2. A Few Relevant Documents to Query 401

Relevant documents
FBIS3-100090, FBIS3-106290, FBIS3-115210, FBIS3-100590, FBIS3-107660, FBIS3-127970, FBIS3-131970, FBIS3-148320, FBIS3-153870, FBIS3-133220, FBIS3-149260, FBIS3-155350

When a query is submitted, many documents are retrieved with the most relevant ones on the top of the list. Table 3 shows a few documents retrieved by query 401. The two notations “R” and “NR” are used to indicate a relevant document and non-relevant one respectively. Document FBIS4-18372, for example was not found in the RJL whereas FBIS3-20090 was found in the RJL.

Table 3. A Few Documents Retrieved by Query 401

Rank	Documents	RJL	Rank	Documents	RJL
1	FBIS4-18372	NR	4	LA022790-0091	NR
2	FBIS3-20090	R	5	FT922-14939	R
3	FT941-1403	NR			

5. ASSESSMENT

The evaluation measures used to assess the effectiveness of Information Retrieval (IR) are precision and recall [5].

The precision is the number of relevant documents retrieved over the retrieved documents, as shown in Equation 9:

$$Precision = \frac{Relevant\ documents\ Retrieved}{Retrieved\ documents} \quad (9)$$

The recall is the number of relevant documents retrieved over the total number of relevant documents, as shown in Equation 10:

$$Recall = \frac{Relevant\ documents\ Retrieved}{Relevant\ documents} \quad (10)$$

The precision in this paper is represented using the precision-recall curve with pre-established recall levels (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0) where all queries are put together and averaged at these levels.

6. THE NEW TECHNIQUE: DNR

In this section, an example is used to describe the new technique proposed in this paper, "improving search by demoting non-relevant documents (DNR). The example has the sample query 555 that contains three terms "improved", "search", and "engines". Table 4 lists the documents retrieved by query 555.

Table 4. List of Documents Retrieved by Query 555

Rank	Document	RJL	Rank	Document	RJL	Rank	Document	RJL	Rank	Document	RJL
1	1872	N	4	0091	R	7	1796	N	10	1883	R
2	2090	R	5	1439	N	8	1882	R			
3	1403	N	6	5536	N	9	6528	N			
3	1403	N	6	5536	N	9	6528	N			

Note : R: Relevant – N: Non-relevant

The documents 2090, 0091, 1882, and 1883 are considered relevant with the notation 'R' in column RJL whereas non-relevant documents have the notation 'N'.

The first step is to generate queries from the terms of the initial query 555. The first set of queries generated contains only one term, and are called them query-size-one. Table 5 shows three query-size-ones generated: The first query generated is query 5551 and contains the term "improved". The second query, 5552, contains the term "search". The third query, 5553, contains the term "engines".

Table 5. Query-Size-One Generated from Query 555

Query generated	Content
Query 5551	improved
Query 5552	search
Query 5553	engines

The second set of queries generated contains two terms and are generated by combing the terms in all possible ways. These set of queries are called query-size-two. Table 6 shows the three query-size-two generated. Query 5554 contains the terms "improved", and "search", query 5555 contains the terms "improved", and "engines", query 5556 contains the terms "search", and "engines".

Table 6. Query-Size-One Generated from Query 555

Query generated	Content
Query 5554	improved, search
Query 5555	improved, engines
Query 5556	search, engines

Each of the generated queries will retrieve its unique list of documents; a few or a lot of those retrieved documents are also retrieved by the original query 555. Table 7 compares the retrieved documents by generating queries against the retrieved documents by the original query.

Table 7. Comparison between the Original Query and Generated Queries

Rank	Original Query 555	RJL	Query-size one			Query-size two		
	Documents		5551	5552	5553	5554	5555	5556
1	1872	N	0	1	1	1	0	0
2	2090	R	1	1	1	1	1	1
3	1403	N	0	0	0	1	0	0
4	0091	R	0	1	1	1	1	1
5	1439	N	0	0	0	1	0	1
6	5536	N	0	0	0	0	0	1
7	1796	N	0	1	1	1	1	1
8	1882	R	0	0	0	0	0	1
9	6528	N	0	0	0	1	1	1
10	1883	R	0	1	1	1	1	1

Note: R: Relevant – N: Non-relevant

For example query-size-one 5552 retrieved the document “1272”, as indicated by the digit ‘1’, which is also retrieved by the original query 555 as indicated by the notation ‘R’. Next, the list of retrieved documents by the query-size-one and query-size-two generated queries are examined. In Table 7, some of the documents were not retrieved by the query-size-one queries such as “1439”, and “6528”. A few other documents were retrieved by query-size-one and query-size-two queries such as “1872”, “2090”, “0091”, 1796”, and “1883”. DNR considers a document to be non-relevant when it is retrieved by none of the query-size-one generated queries and at most by one of the query-size-two queries. This condition detects the non-relevant documents as illustrated in Table 8.

Table 8. Selection of the Non-Relevant Documents

Rank	Original Query 555			Query-size one					Query-size two					Status
	Documents	RJL		5551	5552	5553	S1		5554	5555	5556	S2		
1	1872	N		0	1	1			1	0	0	T		
2	2090	R		1	1	1			1	1	1			
3	1403	N		0	0	0	T		1	0	0	T		S
4	0091	R		0	1	1			1	1	1			
5	1439	N		0	0	0	T		1	0	1			
6	5536	N		0	0	0	T		0	0	1	T		S
7	1796	N		0	1	1			1	1	1			
8	1882	R		0	0	0	T		0	0	1	T		S
9	6528	N		0	0	0	T		1	1	1			
10	1883	R		0	1	1			1	1	1			

Note: R: Relevant – N: Non-relevant

In Table 8, the column S1 has a "true" value when a document is retrieved by at most one query-size-one query. Column S2 has a "true" value when at most two query-size-two queries retrieve a document. Finally, a document is considered non-relevant when S1 and S2 have a "true" value, and this is indicated by "S" in the "Status" column. Therefore, three out of the ten documents are selected as non-relevant documents. The documents are "1403", "5536", and "1882". These documents are labeled with "S" in column "Status". The results are classified as "False Alarm", "Relevant Rejected", "Missed", and "Not-relevant Rejected" as shown in Table 9.

Table 9. Summary of the Results

	Selected	Not selected
R (Relevant)	False Alarm	Relevant Rejected
NR (Non-relevant)	Non-relevant Selected	Missed

For example, the non-relevant documents "1403" and "5536" in Table 8 are classified as "Non-relevant Selected" since our technique selected them; the third document, "1882", is classified as "False Alarm" since it is relevant in the RJL and was classified by our technique as non-relevant. The non-relevant documents "1872", "1439", "1796", and "6528" are classified as "Missed" since they are classified as non-relevant in the RJL but were not selected as non-relevant by our technique. The relevant documents, "2090", "0091", and "1883", are classified as "Relevant Rejected" since our technique did not select them.
Table 10 lists the classifications of the documents

Table 10. Classifications of the Documents

Rank	Original query 555	RJL	Status	Classification
1	1872	N		Missed
2	2090	R		Relevant Rejected
3	1403	N	S	Non-relevant Selected
4	0091	R		Relevant Rejected
5	1439	N		Missed
6	5536	N	S	Non-relevant Selected
7	1796	N		Missed
8	1882	R	S	False Alarm
9	6528	N		Missed
10	1883	R		Relevant Rejected

Note: R: Relevant – N: Non-relevant

The best categories are the "Non-relevant Selected" and "Relevant Rejected"; the "Non-relevant Selected" detects the non-relevant documents based on our condition whereas the "Relevant Rejected" detects the relevant documents that should not be selected. The "False Alarm" documents affect the precision of the retrieved documents severely since relevant documents in RJL were considered to be non-relevant by our technique. Although the "Missed" category missed the non-relevant documents, however, it has no significant effect on precision as will be shown in the experiments. Finally, the non-relevant documents are moved down the list to improve the precision of the retrieved documents, as shown in Table 11.

Table 11. Relevant Documents Moved Higher

Rank	Initial results	RJL	Status	Rank	New results	RJL	Status
1	1872	N		1	1872	N	
2	2090	R		2	2090	R	
3	1403	N	S	3	0091	R	
4	0091	R		4	1439	N	
5	1439	N		5	1796	N	
6	5536	N	S	6	6528	N	
7	1796	N		7	1883	R	
8	1882	R	S	8	1403	N	S
9	6528	N		9	5536	N	S
10	1883	R		10	1882	R	S

Note: S: Selected - R: Relevant – N: Not-relevant

The new technique moves a few of the selected non-relevant documents to the bottom of the list to improve the precision. For example, documents “1403“, “5536”, and “1882” are moved to the bottom of the list as shown in Table 10. Consequently, documents “0091” and “1883”, which are relevant (R), are lifted. As a result, more relevant documents are moved to the top.

7. EXPERIMENTS AND RESULTS

The following sections describe the experimental results of “Improving search engines by demoting non-relevant documents” (DNR) against the baseline in the vector and probabilistic models. Although the experiments were done on all WT2g’s documents, this paper shows the results of the top twenty documents on one of the queries.

7.1 Using the vector model based on TFIDF

When DNR is tested in the vector model based on TFIDF weighting measure [14, 13], 3781 documents were found non-relevant in the RJL. These documents were pushed down the list and consequently precision improved. The technique also classified 116 relevant documents as non-relevant. Also, 506 documents, that are relevant in the RJL, were detected to be as relevant and therefore, were not selected. Finally, 15419 documents that are non-relevant in the RJL were missed. Table 12 summarizes the results, and Table 13 shows only the top twenty documents retrieved for query 451.

Table 12. Statistics on the Vector Model, based On TFIDF

	Selected		Not selected	
R (Relevant)	False Alarm :	116	Relevant Rejected :	506
NR (Non-relevant)	Non-Relevant Selected:	3781	Missed :	15419

Table 13. Selection of the Non-Relevant Documents for Query 451

Rank	Original Query 451			Query-size one					Query-size two				Status	Classified
	Documents	RJL		4511	4512	4513	S1		4514	4515	4516	S2		
100	FT932-4802	R		0	0	1			0	1	1			RR
101	FT933-496	R		0	0	1			0	1	1			RR
102	FT924-13548	N		0	0	1			0	1	1			M
103	LA090190-0081	N		1	0	0			1	1	0			M
104	LA092990-0090	N		0	0	1			0	1	1			M
105	LA102190-0152	N		0	0	1			0	1	1			M
106	FT921-16122	R		0	0	0	T		1	0	0	T	S	FA
107	FT943-11390	R		0	0	1			0	1	1			RR
108	FT911-4070	R		0	0	1			0	1	1			RR
109	FT943-12373	N		0	0	1			0	1	1			M
110	FT944-6889	N		0	0	0	T		1	0	0	T	S	NS
111	LA090690-0256	N		0	0	1			0	1	1			M
112	LA051989-0015	R		0	0	1			0	1	1			RR
113	LA092290-0079	N		0	0	1			0	1	1			M
114	LA111290-0065	N		0	0	1			0	1	1			M
115	LA082490-0156	N		0	1	0			1	0	1			M
116	FT941-15071	N		0	0	1			0	1	1			M
117	FT942-4603	R		0	0	1			0	1	1			RR
118	LA053190-0100	R		0	0	1			0	1	1			RR
119	LA042289-0128	N		1	0	1			1	1	0			M

Note: FA: False Alarm – M: Missed – RR: Relevant Rejected NS: Non-relevant Selected –
T: True – S: Selected – R: Relevant – N: Not-relevant

In Table 13, two non-relevant documents “FT921-16122”, and “FT944-6889” were detected by our technique to be non-relevant and therefore were shifted down the list. Consequently, the relevant documents “FT943-11390”, “FT911-4070”, “LA051989-0015”, “FT942-4603” and “LA053190-0100” will be lifted two ranks to be positioned higher in the list. Therefore, precision will improve. It should be noted here that the relevant document “FT921-16122” is classified as a non-relevant and is a “False Alarm”.

7.2 Using the probabilistic model based on BM25

When DNR is tested in the probabilistic model based on BM25 weighting measure [18] it classified 3631 non-relevant documents as non-relevant. These documents were pushed down the list and consequently precision improved. The technique also classified 97 relevant documents as non-relevant. Also, 526 relevant documents were classified to be as relevant and therefore, were not selected.

Finally, 15568 documents that are non-relevant in the RJL were missed as shown in Table 14, and Table 15 shows only the top twenty documents retrieved for query 451.

Table 14. Statistics on the Probabilistic Model, based On BM25

	Selected		Not selected	
R (Relevant)	False Alarm :	97	Relevant Rejected :	526
NR (Non-relevant)	Non-Relevant Selected	3631	Missed :	15568

Table 15. Selection of the Non-Relevant Documents for Query 451

Rank	Original Query 451			Query-size one					Query-size two					
	Documents	RJL		4511	4512	4513	S1		4514	4515	4516	S2	Status	Classified
250	LA121590-0052	R		0	1	0			1	0	1			RR
251	LA110690-0017	N		0	0	0	T		1	0	0	T	S	NS
252	LA092290-0008	N		0	0	1			0	1	1			M
253	FT942-11348	N		0	1	0			1	0	0	T		M
254	LA082290-0085	N		0	0	0	T		1	0	0	T	S	NS
255	LA120290-0218	N		0	0	0	T		1	0	0	T	S	NS
256	FT934-10273	N		0	1	0			1	0	1			M
257	LA040490-0001	R		0	1	0			1	0	1			RR
258	LA090290-0168	N		0	1	0			1	0	1			M
259	LA111990-0044	N		0	1	0			1	0	1			M
260	FT911-4579	N		0	1	0			1	0	1			M
261	LA110190-0016	N		0	0	1			0	1	1			M
262	LA100690-0121	N		0	1	0			1	0	1			M
263	LA120690-0183	N		0	0	1			0	1	1			M
264	FT943-16550	N		0	1	0			1	0	1			M
265	LA120190-0115	N		0	1	0			1	0	1			M
266	FT934-10071	N		0	0	1			0	1	1			M
267	LA092190-0053	R		0	1	0			1	0	1			RR
268	LA101690-0044	N		0	1	0			1	0	1			M
269	FT923-12577	N		0	1	0			1	0	1			M

Note: FA: False Alarm – M: Missed – RR: Relevant Rejected NS: Non-relevant Selected –
T: True – S: Selected – R: Relevant – N: Not-relevant

Three documents, “LA110690-0017”, “LA082290-0085”, and “LA120290-0218” were detected by our technique to be non-relevant and therefore, are shifted down the list. Consequently, the relevant documents “LA040490-0001” and “LA092190-0053” are lifted three ranks to be positioned higher in the list.

7.3 Using the probabilistic model based on DFR_BM251

When DNR is tested in the probabilistic model based on DFR_BM25 weighting measure [16, 17, 18] DNR classified 3632 non-relevant documents as non-relevant. These documents were pushed down the list and consequently precision improved. The technique also classified 93 relevant documents as non-relevant. Also, 533 documents, that are relevant in the RJL, were detected to be as relevant and therefore, were not selected. Finally, 15564 documents that are non-relevant in the RJL were missed as shown in Table 16, and Table 17 shows only the top 20 documents retrieved for query 451.

Table 16. Statistics on the Probabilistic Model, based On DFR-BM25

	Selected		Not selected	
R (Relevant)	False Alarm :	93	Relevant Rejected :	533
NR (Non-relevant)	Non-Relevant Selected	3632	Missed :	15564

Table 17. The Selection of the Non-Relevant Documents for Query 451

Rank	Original Query 451			Query-size one					Query-size two					
	Documents	RJL		4511	4512	4513	S1		4514	4515	4516	S2	Status	Classified
100	FBIS4-6284	N		0	0	0	T		1	0	0	T	S	NS
101	LA090190-0081	N		0	1	0			1	0	1			M
102	LA092990-0090	N		0	1	0			1	0	1			M
103	LA090690-0256	N		0	0	0	T		1	0	0	T	S	NS
104	FT933-499	R		0	1	0			1	0	1			RR
105	FT943-11390	R		0	1	0			1	0	1			RR
106	LA082490-0156	N		0	1	0			1	0	1			M
107	FT933-496	R		0	1	0			1	0	1			RR
108	LA092290-0079	N		0	1	0			1	0	1			M
109	FT923-7518	N		0	0	0	T		1	0	0	T	S	NS
110	LA081090-0021	N		0	1	0			1	0	1			M
111	FT932-4802	R		0	0	1			0	1	1			RR
112	FT941-15071	N		0	0	1			0	1	1			M
113	LA051989-0015	R		0	1	0			1	0	1			RR
114	LA081690-0050	N		0	1	0			1	0	1			M
115	FT921-16122	R		0	0	0	T		1	0	0	T	S	F
116	LA042289-0128	N		0	0	1			0	1	1			M
117	FT911-4070	R		0	1	0			1	0	1			RR
118	FT923-6463	N		0	1	0			1	0	1			M
119	LA053190-0100	R		0	1	0			1	0	1			RR

Note: FA: False Alarm – M: Missed – RR: Relevant Rejected NS: Non-relevant Selected –
T: True – S: Selected – R: Relevant – N: Not-relevant

Four documents, “FBIS4-6284”, “LA090690-0256”, “FT923-7518”, and “FT921-16122” were classified as non-relevant and therefore are shifted down the list. Consequently, the relevant documents “FT933-499”, “FT943-11390”, “FT933-496”, “FT932-4802“, “LA051989-0015”, “FT911-4070”, and “LA053190-0100” are lifted four ranks to be positioned higher in the list. Therefore, precision will improve. It should be noted here that the document “FT921-16122” detected by our technique as non-relevant is actually relevant in the RJL and therefore, is classified as “False Alarm”.

7.4 Analysis of the precision and recall

Table 18 compares the results of the baseline and the new technique for all models using precision and recall at pre-established recall levels.

Table 18. Precision Values at Different Recall Levels

Rank	Vector model using TFIDF			Probabilistic model using BM25			Probabilistic model using DFR-BM25		
	Precision		%	Precision		%	Precision		%
	Base	DNR	Improve	Base	DNR	Improve	Base	DNR	Improve
0.0	0.025	0.029	15	0.047	0.045	-5	0.032	0.033	3
0.1	0.021	0.022	4	0.021	0.026	27	0.021	0.024	15
0.2	0.019	0.020	4	0.019	0.023	20	0.019	0.021	12
0.3	0.018	0.018	1	0.018	0.021	22	0.018	0.018	0
0.4	0.017	0.017	0	0.016	0.020	23	0.017	0.017	0
0.5	0.015	0.015	0	0.015	0.016	6	0.015	0.015	0

0.6	0.013	0.013	0	0.013	0.013	2	0.013	0.013	0
0.7	0.011	0.011	0	0.007	0.007	0	0.011	0.011	0
0.8	0.004	0.004	0	0.004	0.004	0	0.004	0.004	0
0.9	0.001	0.001	0	0.001	0.001	0	0.001	0.001	0
1.0	0.000	0.000	0	0.000	0.000	0	0.000	0.000	0

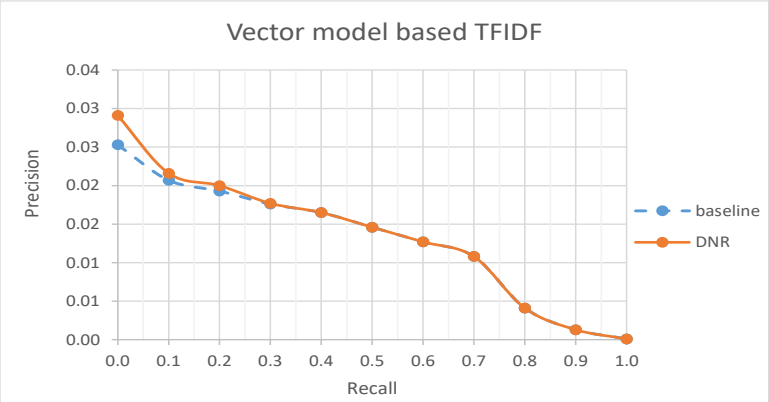

Figure 1. Comparing the baseline to DNR based on TFIDF

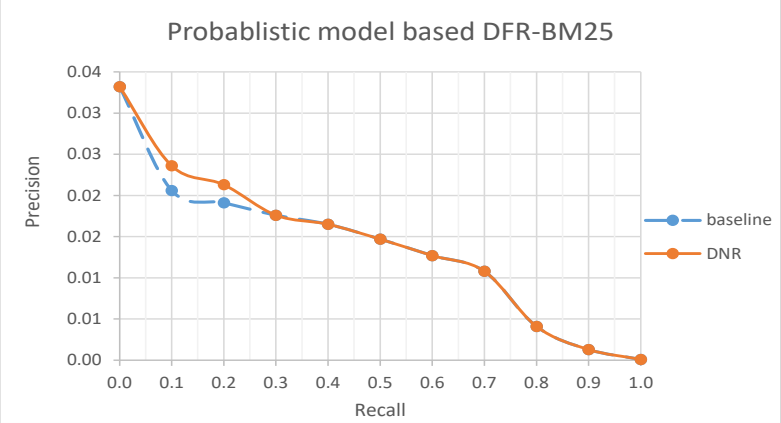

Figure 2. Comparing the baseline to DNR based on BM25

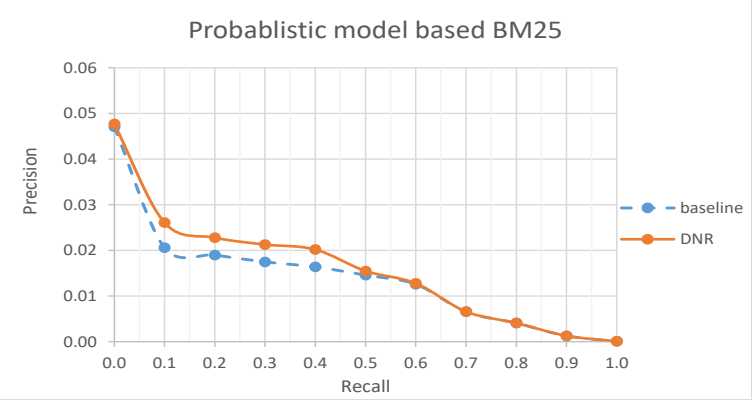

Figure 3. Comparing the baseline to DNR based on DFR-BM25

8. CONCLUSIONS

In this paper, we have introduced a new technique, Demoting Non-Relevant documents (DNR) that improves the precision of search engines by detecting and demoting non-relevant documents. The new technique is tested on WT2g test collection using variant retrieval models. The results show that the new technique outperformed the baseline on low recall level when tested using the vector model based on the TFIDF weighing measure, the probabilistic models based on the BM25, and DFR-BM25 weighing measures.

When DNR is tested in the vector model based on TFIDF weighting measure, 3781 documents were found non-relevant and were pushed down the list. When DNR is tested in the probabilistic model based on BM25 weighting measure, 3631 documents were found non-relevant and were pushed down the list. When DNR is tested in the probabilistic model based on DFR_BM25 weighting measure 3632 documents were found non-relevant and were pushed down the list.

The main limitation of the DNR technique is the time required to generate the new queries and to apply the heuristic to the documents retrieved from each query. Further research should be done on larger test collections to determine if the precision and recall values found in this paper can also be applied to different collections. The experiments will be done on GOV2[3] which consist of 25,205,179 documents.

ACKNOWLEDGMENTS

This work was fully supported by the National Council for Scientific Research (CNRS) in Beirut, Lebanon

REFERENCES

[1] R. P. S. H. Manning CD, An Introduction to Information Retrieval DRAFT, vol. 1, Cambridge: Cambridge University Press, 2008.

[2] R.-N. B. Baeza-Yates R., Modern Information Retrieval, vol. Vol. 463. , New York: New York: ACM Press , 1999.

[3] D. M. R. R. G. B. B. Carpineto C., ""An Information Theoretic Approach to Automatic Query Expansion"," ACM Transactions on Information Systems (TOIS), vol. 19, no. 1, pp. Pages 1-27, 2001.

[4] R. P. S. H. Manning C., ""An Introduction to Information Retrieval"," Natural Language Engineering, vol. 16, no. 1, pp. 100-103, 2010.

[5] B. W. Croft, D. Metzler and T. Strohman, ""Search engines"," Information retrieval in practice, vol. 2, no. 2, pp. 13-28, 2010.

[6] B. Croft, D. Metzler and T. Strohman, Search Engines: Information Retrieval in Practice, Pearson Education, Inc., 2015.

[7] G. G. Chowdhurry , Introduction to Modern Information Retrieval, Neal-Schuman Publishers, 2010.

[8] B. W. Croft, D. Metzler and T. Strohman, ""Search engines"," Information retrieval in practice, vol. 2, no. 2, pp. 13-28, 2010.

[9] Q. Ai, L. Yang, J. Guo and W. B. Croft, ""Analysis of the Paragraph Vector Model for Information Retrieval"," in Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2016), Italy, 2016.

[10] C. N. S. V. T. N. Jain A, "Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model," vol. 164, no. 6, 2017.

[3] http://ir.dcs.gla.ac.uk/wiki/Terrier

[11] C. W. B. Zamani H., ""Estimating Embedding Vectors for Queries"," in Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR), Newark, DE, USA, 2016.

[12] D. Z. J. E. Berry M. W., ""Matrices, Vector Spaces, and Information Retrieval"," SIAM Review., vol. 41, no. 2, p. Vol. 41, 1992.

[13] J. . H. Paik, "A novel TF-IDF weighting scheme for effective ranking," in SIGIR '13 Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, Dublin, 2013.

[14] J. A. B. A. Kumari M, "Synonyms Based Term Weighting Scheme: An Extension to TF.IDF," vol. 89, no. 1, 2016.

[15] S. Robertson and H. Zaragoza, The Probabilistic Relevance Framework, Hanover, MA, USA: Now Publishers Inc, 2009.

[16] G. Amati and C. J. Van Rijsbergen, "Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness," ACM Transactions on Information Systems, p. 57–389, 2002.

[17] S. Clinchant and E. Gaussier, "Bridging Language Modeling and Divergence from Randomness Models: A Log-Logistic Model for IR," in Conference on the Theory of Information Retrieval - ICTIR 2009: Advances in Information Retrieval Theory, Berlin, 2009.

[18] S. Robertson, "The Probabilistic Relevance Framework - BM25 and Beyond," Foundations and Trends in Information Retrieval, pp. 333-389, 17 December 2009.

[19] Hiemstra D., "Information Retrieval Models," in Information Retrieval: Searching in the 21st Century, Vols. pp 2-19, UK, A John Wiley and Sons, Ltd., Publication, 2009.

[20] T. Roelleke, Information Retrieval Models- Foundations and Relationships, London: Morgan & Claypool Publishers, 2013.

[21] S. Robertson, S. Walker, M. Hancock-Beaulieu, M. Gatford and A. Payne, "OKAPI at TREC-4," in In the Proceedings og the 7th Text retrieval Conference (TREC7), Gaithersburg, 1995.

[22] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?," in Proceedings of the IEEE, 2000.

[23] L. S. B. G. Beel J., "A Novel Term-Weighting Scheme for User Modeling based on Users Personal Document Collections," in Proceedings of the iConference 2017, China, 2016.

[24] S. M. H. D. K. W. Verberne S, "Evaluation and analysis of term scoring methods for term extraction," vol. 19, no. 5, 2016.

[25] Sparck-Jones K., ""Experiments in Relevance Weighting of Search Terms"," Information Processing & Management., vol. 15, no. 3, pp. 133-144. ISSN: 0306-4573, 1979.

[26] M. Sanderson, "Test Collection Based Evaluation of Information Retrieval Systems," The essence of knowledge, Sheffield, 2010.

[27] F. Yamout and M. Makary, "Building Relevant Judgment List with Minimal Human Intervention," International Journal of Advanced Computer Technology (IJACT), 2015.

[28] H. Harankhedkar, "Techspirited," 01 01 2019. [Online]. Available: https://techspirited.com/internet-its-uses-in-our-daily-life.

[29] H. C. M. M. Baeza-Yates R., "Query Recommendation Using Query Logs in Search Engines", vol. 16, Berlin, Heidelberg: Springer, 2005, pp. 588-596.

AUTHORS

Fadi Yamout is a PhD holder in Computer Sciences from the University Of Sunderland, United Kingdom, in the area of Information Retrieval and Search Engines. He worked as Chairman of Computer Science and for Middle East Airlines as Head of planning and control department. He worked in the capacity of System Analyst and Programmer for Societe de Service d'Informatique in Lebanon and Somapro in Canada.

Mireille Makary is an MS holder in Computer Sciences from the University Of Balamand, Lebanon. Worked as Campus coordinator of Computer Science and have extensive teaching experience in Computer Science courses at all different levels.