Improved TFIDF weighting techniques in document *Retrieval*

Fadi Yamout
Computer Science
Lebanese International University
Beirut, Lebanon
fadi.yamout@liu.edu.lb

Rachad Lakkis
Computer Science
Lebanese International University
Beirut, Lebanon
rachad.lakis@liu.edu.lb

Abstract—**In information retrieval, documents are usually retrieved using lexical matching which matches where words in a user's query with words found in a set of documents. A significant model used in information retrieval is the vector space model where these words are represented as a vector in space and are assigned weights using a favorite weighting technique called TFIDF (Term Frequency Inverse Document Frequency). In this thesis, we have devised three new weighting techniques to improve the TFIDF weighting technique. The first technique is Dispersed Words Weight Augmentation (DWWA) which gives more weight to the words distributed in most of the document's paragraphs; we consider that those words are more significant than words found in few paragraphs. The second technique is called Title Weight Augmentation (TWA) which gives more weight to the words found in the document's title and first paragraphs. The third technique is called First Ranked Words Weight Augmentation (FRWWA) which increments further the weight of the most frequent words in a document. We tested the three techniques, and we found more relevant documents were retrieved in our system.**

Keywords—Information Retrieval, TFIDF, Precision-recall

I. Introduction

Gerard Salton, a pioneer in information retrieval and one of the leading figures in this area from 1960s to 1990s, proposed the following definition in his classic 1968 textbook [1]: "Information retrieval is a field concerned with the structure, analysis, organization, storage, searching and retrieval of information".

Despite the huge advances in understanding the technology of search engines in the past forty years, Salton's definition is still appropriate and accurate. Information retrieval includes work on a wide range of information and a variety of applications related to search engines [2]. Information retrieval domain is composed of many components: indexing, model applied [3], weighting technique [4], searching for documents, and finally ranking the results. In information retrieval, we call the objects to be retrieved as "documents" even though in actuality they may be web pages, PDFs or even fragments of code [5]. For the model applied, Vector Space Model is a popular model, and TFIDF is one of the weighting techniques used in the vector space model [6][7][8]. This paper aims to improve retrieval of documents by updating the TFIDF weighting technique. We will update based on the distribution of words in paragraphs, the occurrence of words found in the document's title and first paragraph and the most repeated words in a document [9].

Similar techniques exist. For instance, in "Title language model for information retrieval" [10], a new language model called title language model is proposed. The technique is based on the probability of using query Q as the title for document D. In a different technique found in "Document Title Patterns in Information Retrieval" [11], the document's title is frequently used to obtain document keywords. However, both techniques did not alter the weights of the TFIDF based on the number of occurrences.

II. Contribution

Some contributions were made in this paper to improve retrieval of documents. The contributions are:

A. Words Distributed in Many Paragraphs

We consider that words distributed in many paragraphs are more relevant than words found in one or a few paragraphs. For instance, a word repeated five times in five paragraphs should be weighted much more than a word repeated five times in one paragraph. Therefore, we increase the weight of words distributed in many paragraphs.

B. Words Found in Title and First Paragraph

We consider that words appearing in the first paragraphs and the document's title are more significant than other words. Therefore, we have increased the weight of these words to improve the retrieval of documents. Therefore, we increase the weight of words found in the title and first paragraph

C. Words Repeated the Most in Document

We consider that the most repeated words in a document tend to be the most relevant ones. We increase the weight of those considering that they represent the subject of the document accurately. Therefore, we increase the weight of words repeated the most in documents.

III. The three techniques

This section describes the three new techniques devised to improve retrieval of documents. The first technique is called Dispersed Words Weight Augmentation which gives more weight to the words distributed all over the document. The second technique is called Title Weight Augmentation and gives more weight to the words found in the documents' titles and first paragraphs. The third technique is called First Ranked Words Weight Augmentation which increments further the weight of the most frequent words in each document

A. Dispersed Words Weight Augmentation - DWWA

The new technique, DWWA, increments the weight of words spread across a document. If a word is repeated many times in a single paragraph, its weight will remain the same, whereas if repeated in many paragraphs, its weight will be

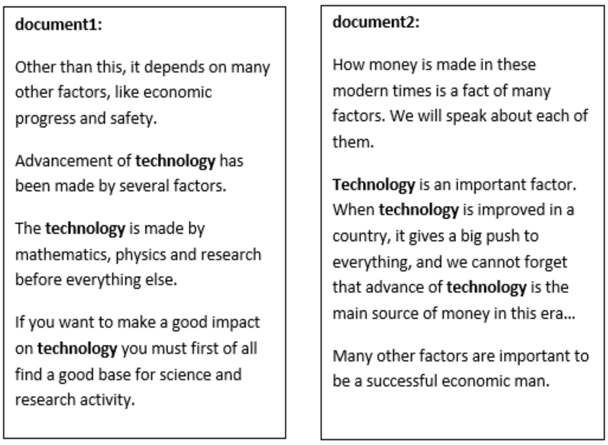
978-1-5386-5244-2/18/$31.00 ©2018 IEEE

augmented proportionally to the number of these paragraphs. For example, the word "life" having TFIDF of 1.5 will remain the same if all of its occurrences are in one paragraph but will become three if all of its occurrences are in two paragraphs. We show the weight in equation 1:

$$\textit{derived weight of the word}_i\textit{=initial weight of word}_i\times P_i \quad (1)$$

In equation 1, P_i is the number of paragraphs containing $word_i$. As a result, the documents that contain the words in the query will be pushed further high in the list of retrieved documents which is the main reason for improving precision at high recall values.

Let's consider the two documents in figure 1:



document1:

Other than this, it depends on many other factors, like economic progress and safety.

Advancement of **technology** has been made by several factors.

The **technology** is made by mathematics, physics and research before everything else.

If you want to make a good impact on **technology** you must first of all find a good base for science and research activity.

document2:

How money is made in these modern times is a fact of many factors. We will speak about each of them.

Technology is an important factor. When **technology** is improved in a country, it gives a big push to everything, and we cannot forget that advance of **technology** is the main source of money in this era...

Many other factors are important to be a successful economic man.

Fig. 1. Example of two documents for DWWA

In document1, we find the word "technology" in three paragraphs, and in one paragraph in document2. We consider that a word found in many paragraphs tends to be more significant than words found in fewer paragraphs since they reflect the main subject of this document. Before applying DWWA, the information retrieval system would give document1 and document2 identical scores (Table I), but with DWWA, we assign to the word "technology" a bigger weight (Table II), and consequently increases the relevancy of document1 if a query contains the word "technology".

TABLE I. TERM TO DOCUMENT BEFORE DWWA

	technology	$word_2$	$word_3$	…	$word_n$
document1	3	..	..	..	..
document2	3	..	..	..	..

TABLE II. TERM TO DOCUMENT AFTER DWWA

	technology	$word_2$	$word_3$	…	$word_n$
document1	**9**	..	..	..	..
document2	3	..	..	..	..

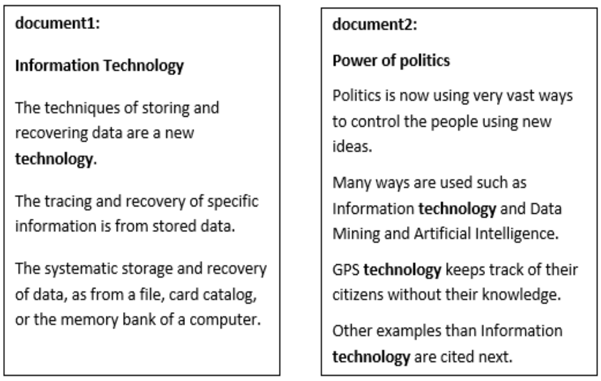
B. Title Weight Augmentation - TWA

The new technique, TWA, adds more weight to words found in the document's title and first paragraphs since we consider that those words describe better the content of the document compared to other words in the body of the document. For example, if the word "life" has a TFIDF weight of 0.5 and is found in the title of the document, the TFIDF of this word will be increased according to a specific criterion, determined by testing as shown in equation 2.

$$\textit{derived weight of the word}_i\textit{=initial weight of word}_i + c \quad (2)$$

In equation 2, we find the $word_i$ in the document's title and first paragraphs and c is the additional weight added to $word_i$. After submitting a query, we give priority to the words in the documents' titles and first paragraphs that are similar to the query. Consequently the documents that contain these words are pushed further higher in the list of retrieved documents, and consequently, it will improve precision as we will show in the experimental results.

Let's consider the two documents in figure 2:



document1:

Information Technology

The techniques of storing and recovering data are a new **technology**.

The tracing and recovery of specific information is from stored data.

The systematic storage and recovery of data, as from a file, card catalog, or the memory bank of a computer.

document2:

Power of politics

Politics is now using very vast ways to control the people using new ideas.

Many ways are used such as Information **technology** and Data Mining and Artificial Intelligence.

GPS **technology** keeps track of their citizens without their knowledge.

Other examples than Information **technology** are cited next.

Fig. 2. Example of two documents for TWA

The word "technology" is found twice in document1 and three times in document2. However, one of the word occurrences in document1 is in the document's title, and another is in the first paragraph. A query that searches for the word "technology", will favor document2 having more occurrences of this word than document1. We consider augmenting the weight of the word "technology" in document1 since it is mainly related to the main topic of the document. Here are some tables representing the calculation of TFIDF, where the table III shows the traditional TFIDF and table IV shows the suggested improvement.

TABLE III. TERM TO DOCUMENT BEFORE TWA

	technology	$word_2$	$word_3$	…	$word_n$
document1	2	..	..	..	..
document2	3	..	..	..	..

TABLE IV. TERM TO DOCUMENT AFTER TWA

	technology	$word_2$	$word_3$	…	$word_n$
document1	**6**	..	..	..	..
document2	3	..	..	..	..

Before the improvement, the information retrieval system would select document2 having more occurrences of the word in the query, but after the improvement, it will select document1.

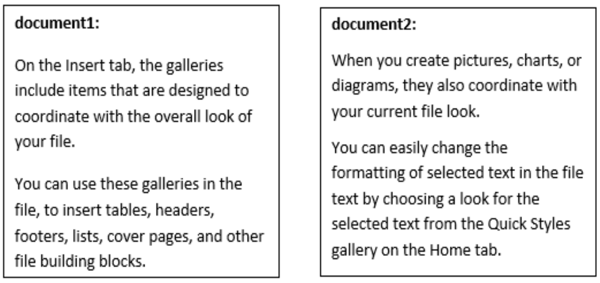
C. First Ranked Words Weight Augmentation - FRWWA

The new technique, FRWW, promotes the most repeated words in each document. We promote the words' weight as shown in equation.3.

$$\forall\, word_t \in set\ N;$$
$$derived\ weight\ of\ the\ word_t = initial\ weight\ of\ word_t + c \quad (3)$$

In equation 3, set N contains the most repeated words, and c is the additional weight added to word$_t$. As a result, documents containing those words will be pushed further high in the list of retrieved documents once found in a query, and consequently, it will improve precision at high recall values. The maximum number of words selected to have weight augmentation varies between one and seven based on the size of the documents. As for the weight to be added, it will be specified next, according to testing experiments.

Let's consider the two documents in figure 3:



document1:

On the Insert tab, the galleries include items that are designed to coordinate with the overall look of your file.

You can use these galleries in the file, to insert tables, headers, footers, lists, cover pages, and other file building blocks.

document2:

When you create pictures, charts, or diagrams, they also coordinate with your current file look.

You can easily change the formatting of selected text in the file text by choosing a look for the selected text from the Quick Styles gallery on the Home tab.

Fig. 3. Example of two documents for FRWWA.

Table V represents the weights of the terms in the documents.

TABLE V. BEFORE FRWWA

	file	gallery	text	look	Insert	. . .
document 1	**3**	2	0	1	2	
document 2	2	1	**3**	2	0	

In document 1, we find three occurrences of the word "file", repeated more than other words. So, we consider that the main topic of this document tends to be "file". In document 2, although we find the word "file two times, however, it is not the most repeated one. Therefore, we consider the word "text" as the most representative of this document. For this reason, we increment its weight as shown in Table VI.

TABLE VI. TERM TO DOCUMENT AFTER FRWWA

	file	gallery	text	look	Insert	. . .
document 1	**6**	2	0	1	2	
document 2	2	1	**6**	2	0	

IV. Experimental Design

This section describes the technical requirements for performing the experiments of the techniques regarding test data and evaluation

A. Data

The data used is the Medline[1], a collection of articles from a medical journal, and NPL [2] (National Physical Laboratory) test collections as shown in table VII. Medline and NPL are held at the University of Glasgow[3]. These test collections include relevance judgments generated manually by experts. Medline test collection is composed of 1,033 documents with 30 queries for testing whereas NPL test collection is composed of 11,429 documents with 93 queries for testing.

TABLE VII. TEST COLLECTIONS USED

Test Collection	No of documents	No of queries
Medline	1,033	30
NPL	11,429	93

B. Evaluation

The evaluation technique used to assess the result is by comparing the results found by the information retrieval system to the relevant judgment list using precision and recall values. For each query, 11 we calculate the values recall coordinates, and for the total number of queries, all the numbers were averaged to have a single graph to represent the system. In many search applications, users tend to look at the top-ranked documents. In these situations, the focus of an adequate measure should be on how well precision is at high recall values [7].

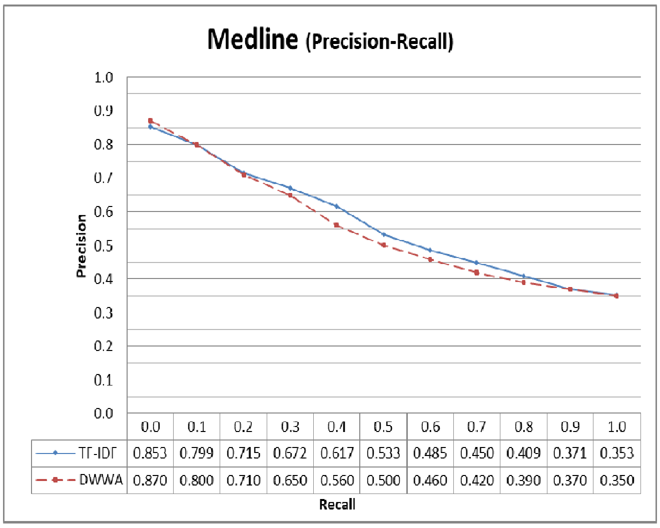
V. Experimental Results

This section describes the experiments done using the newly devised techniques DWWA, TWA, and FRWWA. It then compares their results with TFIDF weighting measure. We test the new techniques on two test collections: Medline and NPL. The intention from the experiments is to get better precision than the baseline at higher recall values [8] since this will be equivalent to search engines getting more relevant documents on the first pages

[1] http://ir.dcs.gla.ac.uk/resources/test_collections/medl/
[2] http://ir.dcs.gla.ac.uk/resources/test_collections/npl/
[3] https://www.gla.ac.uk/

A. *Experiments using DWWA*

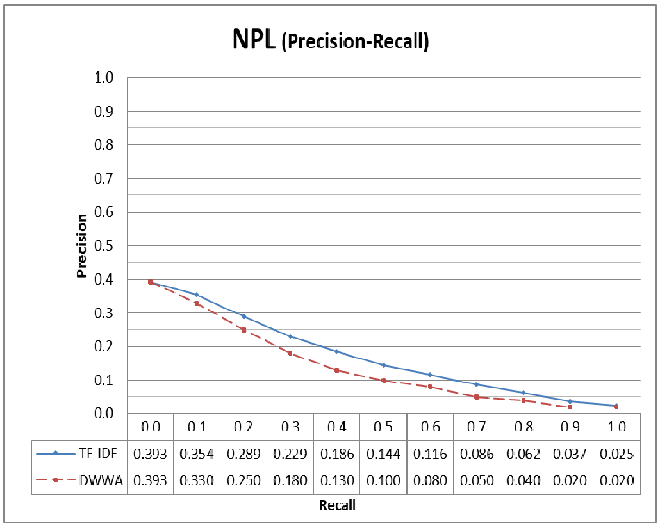
When we tested the new technique DWWA on Medline test collection, the results were better than the baseline at higher recall (figure 4). For instance, results were better until recall 10%. After that and until recall value 20%, the results were the same and then DWWA performed less than TFIDF.



	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
TF-IDF	0.853	0.799	0.715	0.672	0.617	0.533	0.485	0.450	0.409	0.371	0.353
DWWA	0.870	0.800	0.710	0.650	0.560	0.500	0.460	0.420	0.390	0.370	0.350

Fig. 4. DWWA applied to Medline.

For NPL test collection, DWWA performed almost the same as the baseline at a recall value 2% only and then the results deteriorated as shown in figure 5.



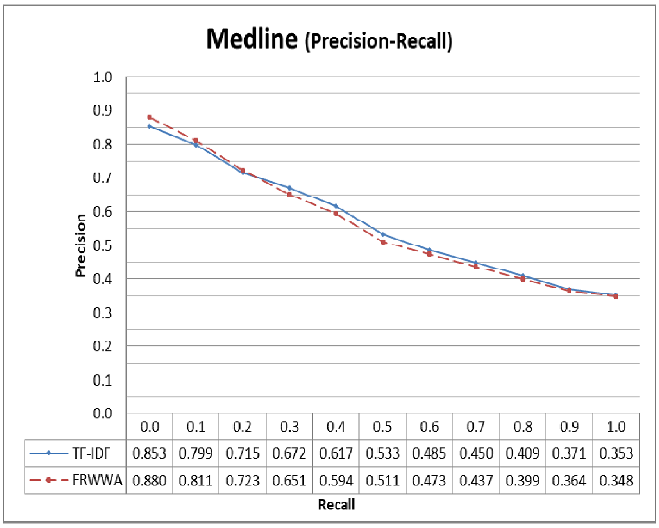
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
TF-IDF	0.393	0.354	0.289	0.229	0.186	0.144	0.116	0.086	0.062	0.037	0.025
DWWA	0.393	0.330	0.250	0.180	0.130	0.100	0.080	0.050	0.040	0.020	0.020

Fig. 5. DWWA applied to NPL

Since the newly devised technique performed better on one test collection, DWWA is a promising technique that requires further research and adjustment to improve on large test collection such as NPL and others.

B. *Experiments using FRWWA*

When testing FRWWA on Medline test collection, it outperformed the baseline at high recall values as shown in figure 6. The optimal result is when the weight of the top three ranked words in each document was augmented by 1.2

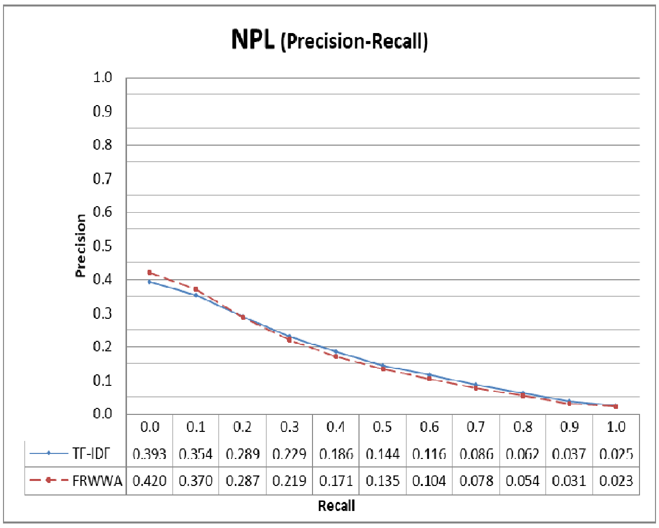


	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
TF-IDF	0.853	0.799	0.715	0.672	0.617	0.533	0.485	0.450	0.409	0.371	0.353
FRWWA	0.880	0.811	0.723	0.651	0.594	0.511	0.473	0.437	0.399	0.364	0.348

Fig. 6. FRWWA applied to Medline.

At recall 30%, FRWWA performed less than the baseline but close to it. Even though it performed further less at low recall values, FRWWA was better at high recall value

When testing FRWWA on NPL test collection, it outperformed the baseline at high recall values as shown in figure 7. To reach an optimal result, we augment the weight of the top five ranked words in each document by 0.9.



	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
TF-IDF	0.393	0.354	0.289	0.229	0.186	0.144	0.116	0.086	0.062	0.037	0.025
FRWWA	0.420	0.370	0.287	0.219	0.171	0.135	0.104	0.078	0.054	0.031	0.023

Fig. 7. FRWWA applied to NPL.

Since the newly devised technique performed better on both test collections, Medline and NPL, at high recall values. FRWWA proved the hypothesis that promoting the most repeated words in each document improves the retrieval of documents

C. Experiments using TWA

TWA outperformed the baseline and gave better precision even to more than 50% recall value as shown in figure 8. We augment the words' weight in the title by one.

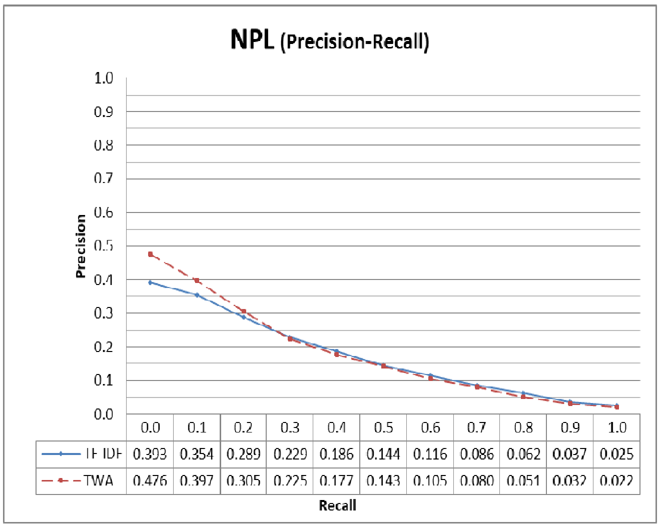


Fig. 8. TWA applied to Medline.

After the 30%, the two techniques performed the same. However, what counts here is that the new technique outperformed the baseline at high recall value. For the large test collection, NPL, TWA gave better results than the baseline (figure 9). Although an increase of precision up to 10% recall would be enough, TWA has increased in precision up to 30% recall value. The additional weight given to the words in titles was 3.5 for the extensive test collection NPL as compared to a weight of 1 given to the small test collection Medline. In further work, we should relate the weight Conclusion

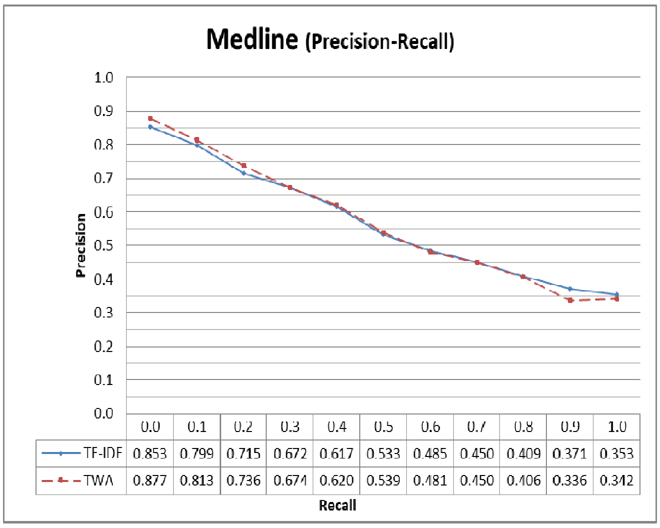


Fig. 9. TWA applied to NPL.

VI. Conclusion

The experiments showed that the newly devised techniques improved the performance of information retrieval according to Precision-Recall. Table VIII shows the parameters used in the three techniques where some of them where determined based on experiments mainly in TWA and FRWWA. However, we add in DWWA based on the number of documents in the collection

TABLE VIII. THE PARAMETERS USED FOR THE THREE TECHNIQUES

			Medline	NPL
DWWA	P_i	number of paragraphs containing $word_i$	additional weight is based on the test collection	additional weight is based on the test collection
FRWWA	N	contains most repeated words	3	5
	c	additional weight added to $word_t$ found in $set\ N$	+1.2	+0.9
TWA	c	additional weight added to $word_i$ found in the document's title	+1	+3.5

In summary, testing on different test collections proved that the new weighting techniques improved the retrieval of documents for high recall values

References

[1] Salton, Gerard. Automatic information organization and retrieval."1968.

[2] Croft WB, Metzler D, Strohman T. Search engines: Information retrieval in practice. Reading: Addison-Wesley; 2010 Feb.

[3] Tala FZ. A study of stemming effects on information retrieval in Bahasa Indonesia. Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands. 2003 Jul.

[4] Verberne S, Sappelli M, Hiemstra D, Kraaij W. Evaluation and analysis of term scoring methods for term extraction. Information Retrieval Journal. 2016 Oct 1;19(5):510-45.

[5] Jain A, Jain A, Chauhan N, Singh V, Thakur N. Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model. International Journal of Computer Applications. 2017 Apr;164(6).

[6] Jain A, Jain A, Chauhan N, Singh V, Thakur N. Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model. International Journal of Computer Applications. 2017 Apr;164(6).

[7] Manning CD, Raghavan P. Hinrich Schütze An Introduction to Information Retrieval Draft. Online edition. Cambridge University Press. -2009.-544 p.

[8] Baeza-Yates R. and Ribeiro-Neto B. Modern Information Retrieval. New York: Addison-Wesley. P 118 1999

[9] Langer, Stefan; Gipp, Bela (2017). "TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections" (PDF). iConference.

[10] Rong Jin, Alex G. Hauptmann, and Cheng Xiang Zhai - "Title language model for information retrieval" - SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - Pages 42-48 2002

[11] Manuel Montes-y-Gomez, Alexander F. Gelbukh, and Aurelio Lopez-Lopez – "Document Title Patterns in Information Retrieval" - TSD '99, LNAI 1692, pp. 372–375, 1999