Variants of Term Frequency and Inverse Document Frequency of Vector Space Model for Effective Document Ranking In Information Retrieval

Deepa Yogish, Manjunath T N, Ravindra S Hegadi

Abstract: ***Advances in the world of internet has made information grow exponentially which make people tend to use information retrieval system more often like Google, Ask, Yahoo etc. to extract relevant and contextual information for their query. The task of information retrieval system is to retrieve relevant document from a huge volume of data sets underlying in the internet using appropriate model. Vector space model is an unconventional model in information retrieval for document ranking. VSM adopts similarity measure for matching between documents and user query, and assign scores from the biggest to smallest .The variants of vector space model are used for information retrieval to rank the documents based on similarity values. The proposed model pre-processes the documents and queries using natural language processing techniques like tokenization, stop word removal and stemming to increase the accuracy of the retrieval process and to reduce the search space. The documents and query are assigned with weights using term frequency and inverse document frequency method. To find relevant document to the query term the document ranking function cosine similarity score is applied for every document vector and the query term vector. The documents having high similarity scores will be considered as relevant documents to the query term and they are ranked based on these scores. This paper emphasizes on different approaches of vector space model using variants of term frequency and inverse document frequency to compute similarity values to rank set of documents for a given query. This paper provides comparison analysis of different variants of vector space model for document ranking.***

Index Terms**: *Information Retrieval (IR), Inverse Document Frequent(idf), Natural Language Processing (NLP),Term Frequency(tf), Vector Space Model (VSM).***

I. INTRODUCTION

In the world of internet, the information on the internet is growing extremely and searching play a vital role to retrieve the relevant answers for the user specific queries. Therefore information retrieval becomes a challenging task to understand the meaning of users' questions and extract relevant document to user The information retrieval (IR) came to existence in the 1950. The first information retrieval system developed at Cornell University is the SMART system in 1960 and still used widely [5]. Information retrieval systems are designed in such a way which helps users to find quickly useful relevant information .The two major problems exist in IR systems are fetching some irrelevant information together with the relevant one and not capable of retrieving all relevant documents. Documents are transformed in to suitable representation for efficient information retrieval. Information Retrieval (IR) is the process in which collection of documents are represented and indexed in suitable form, stored, and searched for the purpose of retrieving information as a response to a user query by comparing each document with that of the query using similarity function and listing the results in order of relevancy. Question answering system is an information retrieval system in which the system gives directly the answer for user query rather than set of documents/references which have possibilities as the answer. Some techniques used by question answering system such as template based, N-gram and key word based etc., but not suitable for unstructured data and for long documents[3]. Factors which affect information retrieval system to satisfy user query with respect to documents are topicality, novelty, freshness, authority, topical relevance and User relevance.

Most widely used information retrieval models are Boolean model, Vector Space model and Probabilistic model. The IR models are classified as set-theory model and statistical model. The set-theory model represents documents as sets of words or phrase. Boolean model belongs to set theory model. The statistical models are vector space and probabilistic model which uses the term frequency as statistical information.

The Boolean model is an exact-matching model which depends on set theory and Boolean algebra whether true or false. In Boolean model[14], query is formed with the use of operators like AND, OR, NOT and documents are associated with a set of keywords .Boolean based IR model returns all documents that satisfy the Boolean expression and without ranking documents in order. The Boolean model returns a document as either relevant or irrelevant because all the documents matched with query will be returned. The major advantage and challenging task of Boolean model is users has control to formulate good Boolean query is too specific or too broad.

Revised Manuscript Received on May 06, 2019

Deepa Yogish, Research Scholar, VTU-RC- Department of Information Science and Engineering BMS Institute of Technology and Management , Bangalore -560061,Karnataka, India

Manjunath T N, Professor and Head of Department, ISE, BMSIT&M, Bangalore

Ravindra S Hegadi Professor, School of Computational Sciences, Solapur University, Solapur, Maharashtra, Solapur.

Boolean model is good for users for those who are comfortable at the usage of information retrieval system but difficult novice's users of computers or IR systems.

Probabilistic model was introduced in 1976 by Robertson and Sparck Jones, which later is also known as the binary independent retrieval (BIR) model. The probabilistic retrieval model rank the documents based on their probability of relevance to the query. Documents and queries are represented by binary vectors ~d and ~q, each vector element indicating whether a document attribute or term occurs in the document or query or not[14]. The probabilistic model is a partial matching method to find number of documents better matching with query based on probability of the number of times a term of query appeared in a document. The probabilistic model is an iterative approach which improves and increases search results based on every repetition of the query. Probabilistic model consume more time to get the desired result which makes user to wait for long time waiting for immediate result.

The vector space model gives nearly exact matching results compared to Boolean and probabilistic models. VSM employ TF-IDF weighting scheme for efficient document ranking.

II. RELATED WORK

Ogheneova et.al.,[1] analyzed with his experiments and suggested vector space model technique is best technique used to retrieve relevant data for user query based on measuring similarity . Author illustrates the document and query matching method based on six dimensional vector space using 6 terms and 5 documents and using the threshold value of 0.5 to determine if a document is relevant or not. If a document is greater than the threshold value, then document returned as relevant. Based on cosine similarity function, documents are ranked which is easier to retrieve most relevant document in IR.

Niranjal et.al.,[2] analyzed group of unstructured documents by applying vector space model. Author pre-processed unstructured documents in to a vector space as a dictionary of terms by using tokenization, stop word removal and stemming and applied weighting technique using term frequency and inverse document frequency to find similarity between unstructured data space and query.

Singh et.al.,[4,6,7,8,10],discussed and analysed vector space model using different methods of term frequency and inverse document frequency to measure weights of terms for better evaluation of search engines. In [7,8] discussed term count model, tf*idf model and normalized vector space model. All the three approaches are compared with set of documents and query. All the 3 approaches of vector space model perform well for long documents where the frequency tem in document is high. In [6] compared different variations of inverse document frequency by computing each method with certain queries.

The IDF methods are

$$idft_1 = \log \frac{D}{dft} \qquad idft_2 = \log \frac{D+1}{dft}$$

$$idft_3 = \log \frac{D}{dft} + 1 \qquad idft_4 = log\left(\frac{D}{dft}\right)^2$$

After computing the value of IDF using keyword based search for certain queries and concluded that $idft_3$ gives better weight of terms compared to other methods which helps us to find similarity values using cosine function for ranking the documents.

In [4] presented a new approach for evaluating the performance of search engines on the web by computing relevance scores of hits and ranking.

Author proposed new method of computing similarity value compared to classical method of VSM and conducted experiments on 3 popular search engines Google, yahoo and MSN based on TREC queries. Author proposed new method of IDF as $idft_3 = \log \frac{D}{dft} + 1$ and

$$Sim(Q, D_i) = \frac{\sum_{j=1}^{V} W_{Q,j} \times W_{i,j}}{\sqrt{length\ of\ document_j - number\ of\ stop\ words}}$$

Author compared both manual and other methods for search engines for set of TREC queries ,where proposed method have given better result compared to classical method and Google outperformed other search engines.

III. VECTOR SPACE MODEL

The VSM is a statistical Information Retrieval model uses unconventional method where information is retrieved as partial matching [15]. The vector space model represents natural language document and queries as vectors in a multidimensional space. Generally Vector space model is divided into three stages: Document representation/indexing as terms, weighting the indexed terms to enhance retrieval of document relevant to the user and ranking the documents by similarity measure. Cosine similarity measure is used to determine angle between the document vector and the query vector and ranking according to order of relevance.

The vector space model gives nearly exact matching results compared to Boolean and probabilistic models. VSM employ TF-IDF weighting scheme for efficient document ranking.

The term frequency (tf) is defined as the number of times terms appeared in document or query and an inverse document frequency (idf) factor measuring the rarity of a term in the whole document collection. IDF was first introduced in 1976 by spark jones for improving information retrieval system.

TF-IDF weight solve the problem and tells how important a term in a document and incorporates local and global parameters, because it takes in to consideration not only the isolated terms but also the term within document collection[13].

Advantages of Vector Space Models are considering both local (*tf*) and global (*idf*) word occurrence frequencies, partial matching, efficient implementation for large document collections, term-weighing, Cosine ranking and Document length normalization.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Problems with Vector Space Model are there is no real theoretical basis for the assumption of a term space, Terms are not really orthogonal dimensions, Missing semantic information ,Missing syntactic information and assumption of term independence [9].

The frequency of a term t in the document d is referred as the tf factor is given by

$$tf_{t,d} = \frac{freq_{t,d}}{\max freq_{t,d}}$$

Where $freq_{t,d}$ frequency of term t in the document d, which is normalized by the maximum frequency computed over all terms in the document.

Inverse document frequency idf is the logarithmically scaled inverse fraction of the documents that contain the word. It checks whether the term is common or rare across all documents in the corpus. Inverse document frequency is given by

$$idf_t = \log \frac{N}{df_t}$$

Where N is total number of documents in the collection. df_t is the number of documents in which the term t appears.

R is a ranking function which associates ssimilarity of the query q and document d is calculated by cosine similarity function.

$$sim_{q,d} = \frac{\vec{V}_d \times \vec{V}_q}{\left|\vec{V}_d\right| \times \left|\vec{V}_q\right|}$$

Denominator is a product of Euclidean length of both vectors and it represents their cosine document factor normalization. Normalization is necessary to decrease the advantage of longer documents to smaller ones. Cosine similarity function is also represented as:

$$sim_{q,d} = \frac{\sum_{i=1}^{M} w_{q,i} \times w_{d,i}}{\sqrt{\sum_{i=1}^{M}(w_{q,i})^2} \times \sqrt{\sum_{i=1}^{M}(w_{d,i})^2}}$$

The quality of the results returned by an information retrieval system in a response to user query is measured through two basic evaluation metrics such as precision and recall. Precision is the ratio of the number of relevant documents to the number of documents that have been retrieved and recall is the ratio of the number of documents actually available in the repository to the number of relevant documents that have been retrieved.Combined measure that assesses precision/recall tradeoff is F measure (weighted harmonic mean) which is given as

$$F = \frac{2 \times Precision \times Recall}{Precision + recall}$$

IV. VARIANTS OF TF-IDF WEIGHTING SCHEMES

Formulating a precise query and each document with index terms which enhances the accuracy of the retrieved information. Each document in the vector space model is identified by index terms. Natural language Preprocessing techniques are applied to find index terms in documents. These index terms are represented as vectors in the vector space of dimension M, where M corresponds to maximum number of unique terms in the document.

Weight $w_{d,t}$ represents the importance of the term t in the document d and the document is represented with the vector as $\vec{V}_d = (w_{d,1}, w_{d,2}, w_{d,3}, \ldots\ldots, w_{d,M})$

Query itself can be considered a short document, the importance of the term t in the query q is the weight $w_{q,t}$, where query is represented with the vector as $\vec{V}_q = (w_{q,1}, w_{q,2}, w_{q,3}, \ldots\ldots, w_{q,M})$

Steps for construction /preprocessing document and query:

First step is the tokenization, in tokenization removing the punctuations, converting text to lower case and converting each sentence as tokens. In second step, filtering process is done using stop word removal for reducing search time and final step is stemming process performed.

Method-I

Term frequency Model: In term frequency model weight of terms will be computed only using local parameter that is considering term frequency only.

Weight=$W_{t,d}$=$Freq_{t,d}$

Where $Freq_{t,d}$ =Frequency of term t in document d.

Method –II

Classical TF-IDF model-1:The weight of a term in a document vector is determined by its two components- term frequency tf and inverse document frequency idf that is both local and global information.

$$TF_{t,d} = freq_{t,d}$$

$$IDFt_t = \log \frac{N}{df_t}$$

Method -III

Normalized TF-IDF model: Normalization is a way of control/penalizing the term weights for a document and query.

Normalization is done to prevent unfairness towards longer document which may have higher term count regardless of the actual importance of that term in the document to give a measure of the importance of the term within the particular document.

$$TF_{t,d} = \frac{freq_{t,d}}{\max freq_{t,d}}$$

$$IDF_t = 1 + \log \frac{N}{df_t}$$

Method -IV

Sub-Linear Normalized TF-IDF model: Adding log is to dampen/reduce the importance of term that has a high frequency.

Table 1: Weights calculation for method-1

METHOD-I								
$TF_{t,d} = freq_{t,d}$ Weight=W_t=$Freq_{t,d}$								
QUERY : Beauty Life								
D1 : Peace is the beauty of life								
D2 : Loneliness adds beauty to life and beauty is power , a smile is its sword								
D3 : The future belongs to those who believe in the beauty of their dreams								
	Term Frequency-$TF_{t,d}$					Weights = $TF_{t,d}$			
Terms t	Q_t	D 1	D 2	D 3		$Q_{w,t}$	$D_{w1,t}$	$D_{w2,t}$	$D_{w3,t}$
Peace	0	1	0	0		0	1	0	0
Beauty	1	1	1	1		1	1	1	1
Life	1	1	0	0		1	1	0	0
Loneliness	0	0	1	0		0	0	1	0
Adds	0	0	1	0		0	0	1	0
Power	0	0	1	0		0	0	1	0
Smile	0	0	1	0		0	0	1	0
Sword	0	0	1	0		0	0	1	0
Future	0	0	0	1		0	0	0	1
Belongs	0	0	0	1		0	0	0	1
Believe	0	0	0	1		0	0	0	1
Dreams	0	0	0	1		0	0	0	1

$$TF_{t,d} = 1 + \log\,(freq_{t,d})$$

$$IDF_t = 1 + log\frac{N+1}{df_t+1}$$

Method –V

Classical TF-IDF model-2: The combination of term frequency and inverse document frequency is very effective term weighting scheme and most widely applied in information retrieval for document ranking. Documents have strong attachment with the terms which has high TF-IDF values.

$$TF_{t,d} = freq_{t,d} \quad IDFt_t = \quad log\frac{N+1}{df_t}$$

V. EXPERIMENTS AND OUTCOMES

In our paper, we analyzed different varian ts of vector space model for computing the weights of terms of documents and query for document ranking. The experiment is conducted on 3 short documents and set of queries. Initially preprocessing of document and query is done using tokenization, stop word removal and stemming. Later calculate weights for each terms of document and query using different variants of term frequency and inverse document frequency of vector space model. Calculate similarity between each query with all the documents on corpus and rank according in decreasing order of relevance. The weights for each terms of documents and query is calculated using different methods of vector space model as shown in Table 1-5 as experimental results.

Table 2: Weights calculation for method-2

METHOD-II									
$TF_{t,d} = freq_{t,d}$ $IDFt_t = \square\, log\frac{N}{df_t}$									
QUERY : Beauty Life									
D1 :Peace is the beauty of life									
D2 :Loneliness adds beauty to life and beauty is power , a smile is its sword									
D3 :The future belongs to those who believe in the beauty of their dreams									
	Term Frequency-$TF_{t,d}$						Weights = $TF_{t,d}$ * IDF_t			
Terms t	Q_t	D1	D2	D3	IDF_t		$Q_{w,t}$	$D_{w1,t}$	$D_{w2,t}$	$D_{w3,t}$
Peace	0	1.0	0	0	0.47		0	0.47	0	0
Beauty	1	1.0	2.0	1.0	0.15		0.15	0.15	0.31	0.15
Life	1	1.0	1.0	0	0.23		0.23	0.23	0.23	0
Lonelin ess	0	0	1.0	0	0.47		0	0	0.47	0
Adds	0	0	1.0	0	0.47		0	0	0.47	0
Power	0	0	1.0	0	0.47		0	0	0.47	0
Smile	0	0	1.0	0	0.47		0	0	0.47	0
Sword	0	0	1.0	0	0.47		0	0	0.47	0
Future	0	0	0	1.0	0.47		0	0	0	0.47
Belongs	0	0	0	1.0	0.47		0	0	0	0.47
Believe	0	0	0	1.0	0.47		0	0	0	0.47
Dreams	0	0	0	1.0	0.47		0	0	0	0.47

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 3: Weights calculation for methods-3

METHOD-III

$$TF_{t,d} = \frac{freq_{t,d}}{\max freq_{t,d}}$$

$$IDF_t = 1 + log\frac{N}{df_t}$$

QUERY : Beauty Life

D1 :Peace is the beauty of life

D2 :Loneliness adds beauty to life and beauty is power , a smile is its sword

D3 :The future belongs to those who believe in the beauty of their dreams

	Term Frequency-$TF_{t,d}$						Weights = $TF_{t,d}$ * IDF$_t$			
Terms t	Q_t	D1	D2	D3	IDFt		$Q_{w,t}$	$D_{w1,t}$	$D_{w2,t}$	$D_{w3,t}$
Peace	0	0.33	0	0	1.477		0	0.49	0	0
Beauty	0.5	0.33	0.25	0.2	1.0		0.5	0.33	0.25	0.2
Life	0.5	0.33	0.125	0	1.17		0.5	0.39	0.14	0
Loneliness	0	0	0.125	0	1.47		0	0	0.18	0
Adds	0	0	0.125	0	1.47		0	0	0.18	0
Power	0	0	0.125	0	1.47		0	0	0.18	0
Smile	0	0	0.125	0	1.47		0	0	0.18	0
Sword	0	0	0.125	0	1.47		0	0	0.18	0
Future	0	0	0	0.2	1.47		0	0	0	0.29
Belongs	0	0	0	0.2	1.47		0	0	0	0.29
Believe	0	0	0	0.2	1.47		0	0	0	0.29
Dreams	0	0	0	0.2	1.47		0	0	0	0.29

Table 4: Weights calculation for method-4

METHOD-IV

$$TF_{t,d} = 1 + \log\left(freq_{t,d}\right)$$

$$IDF_t = 1 + log\frac{N+1}{df_t + 1}$$

QUERY : Beauty Life

D1 :Peace is the beauty of life

D2 :Loneliness adds beauty to life and beauty is power , a smile is its sword

D3 :The future belongs to those who believe in the beauty of their dreams

	Term Frequency-$TF_{t,d}$						Weights = $TF_{t,d}$ * IDF$_t$			
Terms t	Q_t	D1	D2	D3	IDF$_t$		$Q_{w,t}$	$D_{w1,t}$	$D_{w2,t}$	$D_{w3,t}$
Peace	0	1.0	0	0	1.30		0	1.30	0	0
Beauty	1.0	1.0	1.30	1.0	1.0		1.0	1.0	1.30	1.0
Life	1.0	1.0	1.0	0	1.12		1.12	1.12	1.12	0
Loneliness	0	0	1.0	0	1.30		0	0	1.30	0
Adds	0	0	1.0	0	1.30		0	0	1.30	0
Power	0	0	1.0	0	1.30		0	0	1.30	0
Smile	0	0	1.0	0	1.30		0	0	1.30	0
Sword	0	0	1.0	0	1.30		0	0	1.30	0
Future	0	0	0	1.0	1.30		0	0	0	1.30
Belongs	0	0	0	1.0	1.30		0	0	0	1.30
Believe	0	0	0	1.0	1.30		0	0	0	1.30
Dreams	0	0	0	1.0	1.30		0	0	0	1.30

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 5: Weights calculation for method-5

METHOD-V									
$TF_{t,d} = freq_{t,d}$ $IDFt_t = \log \frac{N+1}{df_t}$									
QUERY : Beauty Life									
D1 :Peace is the beauty of life									
D2 :Loneliness adds beauty to life and beauty is power , a smile is its sword									
D3 :The future belongs to those who believe in the beauty of their dreams									
	Term Frequency-$TF_{t,d}$					Weights = $TF_{t,d}$ * IDF_t			
Terms t	Q_t	D1	D2	D3	IDF_t	$Q_{w,t}$	$D_{w1,t}$	$D_{w2,t}$	$D_{w3,t}$
Peace	0	1.0	0	0	0.60	0	0.60	0	0
Beauty	1.0	1.0	2.0	0	0.12	0.124	0.12	0.24	0
Life	1.0	1.0	1.0	0	0.30	0.30	0.30	0.30	0
Loneliness	0	0	1.0	0	0.60	0	0	0.60	0
Adds	0	0	1.0	0	0.60	0	0	0.60	0
Power	0	0	1.0	0	0.60	0	0	0.60	0
Smile	0	0	1.0	0	0.60	0	0	0.60	0
Sword	0	0	1.0	0	0.60	0	0	0.60	0
Future	0	0	0	1.0	0.60	0	0	0	0.60
Belongs	0	0	0	1.0	0.60	0	0	0	0.60
Believe	0	0	0	1.0	0.60	0	0	0	0.60
Dreams	0	0	0	1.0	0.60	0	0	0	0.60

Table 6: Similarity values based on different methods of weights

QUERY ID	METHODS	DOCUMENTS		
		D1	D2	D3
BEAUTY LIFE	1	0.81	0.67	0.31
	2	0.51	0.32	0.09
	3	0.72	0.50	0.20
	4	0.75	0.26	0.23
	5	0.47	0.26	0.03
BEAUTY POWER	1	0.40	0.67	0.31
	2	0.09	0.48	0.05
	3	0.26	0.58	0.17
	4	0.30	0.53	0.21
	5	0.03	0.67	0.31
BEAUTY	1	0.57	0.63	0.44
	2	0.28	0.27	0.16
	3	0.46	0.49	0.32
	4	0.50	0.38	0.35
	5	0.18	0.17	0.10
PEACE LONLINES	1	0.40	0.22	0.0
	2	0.60	0.29	0.0
	3	0.48	0.25	0.0
	4	0.46	0.27	0.0
	5	0.40	0.22	0.0
PEACE DREAMS	1	0.40	0.0	0.31
	2	0.60	0.0	0.34
	3	0.48	0.0	0 .33
	4	0.46	0.0	0.33
	5	0.62	0.0	0.35
FUTURE LIFE	1	0.40	0.22	0.31
	2	0.19	0.09	0.44
	3	0.34	0.18	0.37
	4	0.36	0.21	0.35
	5	0.19	0.09	0.44

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 6 shows experimental results for similarity values based on different variants of weights calculation on different query sets on 3 documents. We considered 3 short documents and 6 queries using different methods of weights calculation to find similarity value between document and query.

The query set contains 6 queries such as

1. Beauty life
2. Beauty power
3. Peace loneliness
4. Peace dreams
5. Future life

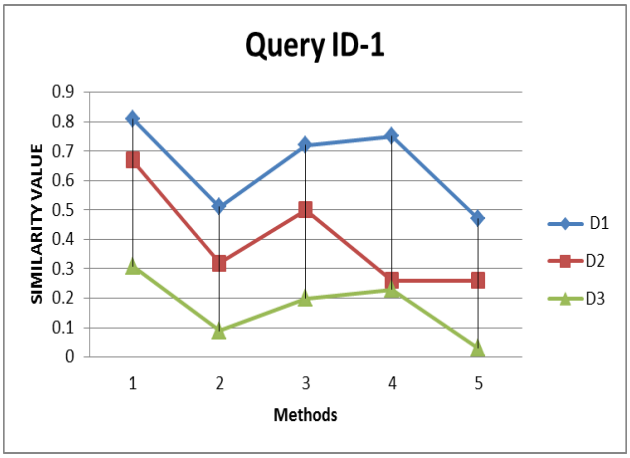

Fig.1.Comparision on 5 methods based on query id-1

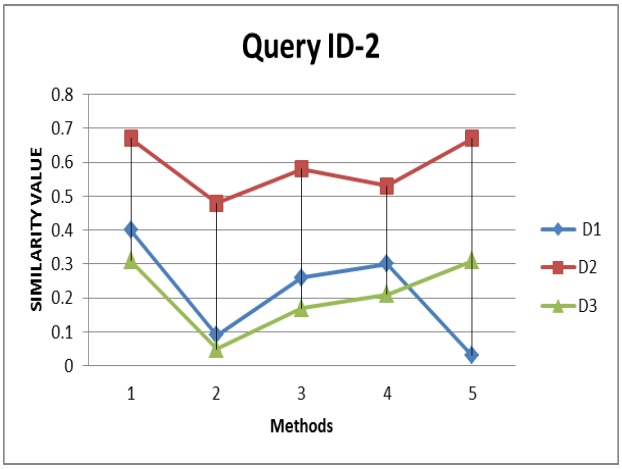

Fig.2.Comparision on 5 methods based on query id-2

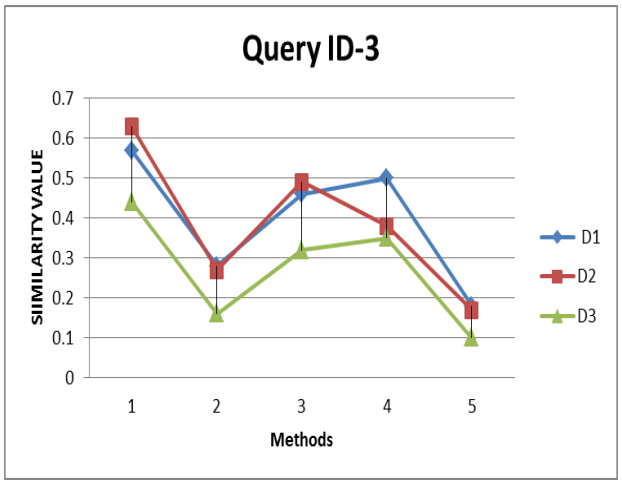

Fig.3.Comparision on 5 methods based on query id-3

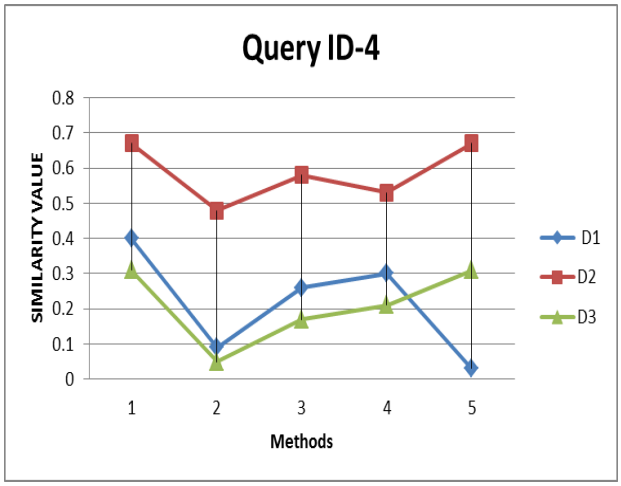

Fig.4.Comparision on 5 methods based on query id-4

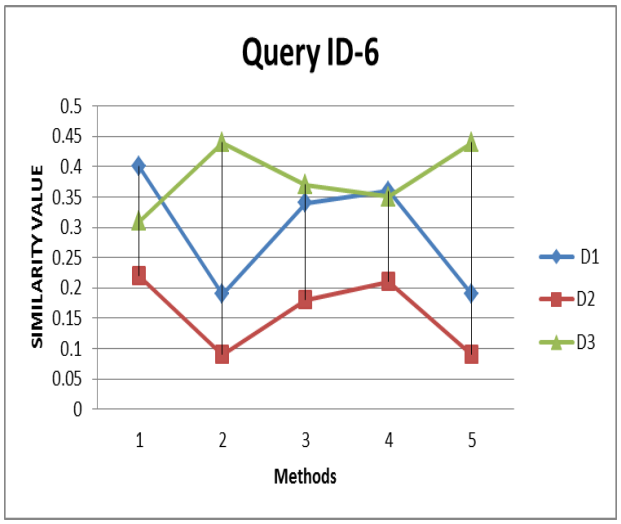

Fig.6.Comparision on 5 methods based on query id-6

Fig 1-6. Shows comparison results on 3 documents for 6 queries on differents methods of weigths caculation for similarity value between documents and queries.Based on the experiments, the certain observations are made. Method-I term frequency model computes weights for terms by considering only local information that is term frequency and produces higher similarity rank to the shorter documents and smaller value for longer document which affects in document ranking. The Tf-Idf weight scheme is a statistical methods which shows the importance of words in the document. Idf is generally used for filtering stop-words which is not much use in documents. That is why method II and V that is classical Tf-Idf method is much better than term frequency model.

Document ranking generally affected by intuition that documents having more repetitive words are more relevant than having fewer words in documents. So Normalisation is required for frequency of terms appear in document which decreases similarity value by improving document rankling. Normalizing term frequency by using $Maxfreq_{t,d}$ that is maximum frequency computed over all terms in the document ,which reduces the problem that longer documents producing higher tf-idf scores. So method III gives better result compared to method I ,II and method V.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

It is not necessarily the case that more the occurrence of a term in a document more is the relevant so the contribution of term frequency to document relevance is essentially a sub-linear function. Hence the log is used to approximate this sub-linear function which is applied in method IV for both term frequency and idf. So method IV gives much better results compared to method I,II and V. Same experiments was conducted on longer documents also. Based on comparison results which are depicted in table 6 and fig 1-6, method III and IV gives better similarity value for both shorter and longer documents for any query.

VI .CONCLUSION

In the world of internet, the information on the internet is growing extremely and searching play a vital role to retrieve the relevant answers for the user specific queries. Therefore information retrieval becomes a challenging task to understand the meaning of users' questions and extract relevant document to user. For efficient information retrieval the documents and query are represented in a suitable form for proper weighing scheme. The concepts behind vector space modeling is to represent documents and queries in a term-document space which allows to compute the similarities between documents and queries and ranked according to the similarity measure between them. The combination of term frequency and inverse document frequency is very effective term weighting scheme and most widely applied in information retrieval for document ranking.

In this paper, we performed extensive analysis of the variants of vector space model. We preprocessed both document and query by using NLP techniques such as tokenization, stop word removal and stemming to increase retrieval efficiency by reducing search space. We computed term frequency and inverse document frequency using different methods of vector space model and computed similarity values for 3 documents with query set. Based on experimental results Normalized and sub-linear normalized TD-IDF methods give best similarity value for document ranking for both short and long documents. For future work, we will focus on applying proposed method of vector space model for better evaluation of intelligent answering system.

REFERENCES

1. E. E. Ogheneovo,R. B. Japheth,[2016] ,"Application of Vector Space Model to Query Ranking and Information Retrieval", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 5, May 2016 .
2. Niranjan Lal, Samimul Qamar,[2015], "Comparison of Ranking Algorithm with Dataspace", International Conference On Advances in Computer Engineering and Application (ICACEA),pp. 565-572, March 2015.
3. Jovita, Linda, Andrei Hartawan, Derwin Suhartono,[2015] "Using Vector Space Model in Question Answering System ", Procedia Computer Science, International Conference on Computer Science and Computational Intelligence (ICCSCI 2015).
4. Singh, J.N. and S.K. Dwivedi,[2015] "Performance Evaluation of Search Engines Using Enhanced Vector Space Model", Journal of Computer Science 2015, DOI: 10.3844/jcssp.2015.692.698
5. D. Manning P. Raghavan S. Hinrich [2013], " Introduction to Information Retrieval", Cambridge University Press.
6. Singh, J.N. and S.K. Dwivedi,[2014] , "Comparative Analysis of IDF Methods to Determine Word Relevance in Web Document", IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 1, No 1, January 2014.
7. Singh, J.N. and S.K. Dwivedi, [2013]," A comparative study on approaches of vector space model in information retrieval", Proceedings on International Conference on Reliability, Infocom Technologies
8. Singh, J.N. and S.K. Dwivedi,[2012], "Analysis of vector space model in information retrieval" In Proceedings of the National Conference on Communication Technologies and its Impact on Next Generation computing (CTNGC"12), International Journal of Computer Applications (IJCA), 2012,pp: 14-18.
9. Jasmina Armenska, Katerina Zdravkova,[2012], "Comparison of Information Retrieval Models for Question Answering ", ConferencePaper,September2012,DOI: 10.1145/2371316.2371348
10. Singh, J.N. and S.K. Dwivedi, and Rajesh Gotam,[2011], "Information Retrieval Evaluative Model", FTICT,2011: Proceedings of the 2011, International conference on Future Trend in Information & Communication Technology,Ghaziabad, India.
11. B. Yates and R. Neto [2012], " Modern information retrieval",Addison Wesley, 2011.
12. R. Baeza-Yates and B. Ribeiro-Neto, [2009],"Modern Information Retrieval", ACM Press, ISBN: 0-201-39829-X.
13. Gerald Salton & Chris Buckley,[1998], "Term weighting approaches in automatic text retrieval", Information Processing and Management, 24(5): No. 5.
14. G. Salton, E. A. Fox. and H. Wu,[1983]," Extended Boolean Information Retrieval," Communications of the ACM,sVol. 26, No. 11, pp. 1022-1036.
15. Salton, G., Wong A., and Yang C. S,[1975], "A vector space model for information retrieval", *Communications of the ACM*,18(11):613–620.

AUTHORS PROFILE

Deepa Yogish, received Bachelor's Degree in Electronics and Communication Engineering Visvesvaraya Technological University, Belgaum, Karnataka, during the year 2004 and M. Tech in Computer Science and Engineering from Visvesvaraya Technological University, Belgaum,Karnataka during the year 2009. Pursuing Ph.D.degree from Visvesvaraya Technogical University since from 2016.Having total 13 years of teaching experience. My areas of interests are Data mining, Natural Language Processing and Information Retrieval Systems. I have published and presented three papers in international conference and journal.

Dr. Manjunath T N. ,Professor & Dean External Relations),BMSIT&M, Bangalore, received his Bachelor's Degree in computer Science and Engineering from Bangalore University, Bangalore, Karnataka, India during the year 2001 and M. Tech in computer Science and Engineering from VTU, Belgaum, Karnataka, India during the year 2004.Completed Ph.D degree from Bharathiar University, Coimbatore. He is having total 18 years of Industry and teaching experience. His areas of interests are Data Warehouse & Business Intelligence, multimedia an d Databases. He has published and presented papers in journals, international and national level conferences.

Dr. Ravindra S Hegadi, Professor and Head Department of Computer Science Director, School of Computational Sciences Solapur University, Solapur, received his Master of Computer Applications (MCA) in 1995& M.Phil in 2003 and Doctorate of Philosophy (Ph.D). in year 2007 in computer science from Gurbarga University, Karnataka; He is having 25years of Experience. He has visited overseas to various universities as SME.His area of interests are Image Mining, Image Processing and Databases and business intelligence. He has published and presented papers in journals, international and national level conferences.