\Engineering and Technology Journal e-ISSN: 2456-3358
Volume 05 Issue 10 October-2020, Page No.-746-750
DOI:10.47191/etj/v5i10.01, I.F. – 6.127

Improving Results of TF-IDF based Retrieval System using Co-reference Resolution and Pronoun Substitution

S. Srihari[1], Dr. M Premalatha[2]

[1]SCOPE, Vellore Institute of Technology,Chennai Chennai, Tamilnadu, India
[2]Associate Professor, SCOPE, Vellore Institute of Technology Chennai Chennai, Tamilnadu, India

ABSTRACT: Information Retrieval systems involve the process of retrieving relevant information based on user queries. TF-IDF is one of the most popular techniques of Information Retrieval. It is widely used and been successful in retrieving relevant information. But still it has some disadvantages. In this paper we propose a method to improve the performance of TF/IDF based systems using Co-reference Resolution and Pronoun Substitution. The system is found to be effective as there has been significant changes in the order of rankings of documents retrieved due to the relative increase in the amount
of content that have taken into consideration during the retrieval process. Graphical analysis of the observed improvement is given by visualizations of TF-IDF, Cosine Similarity and Effective improvement in rank for various documents before and after the change of algorithm.

I. INTRODUCTION

Everyone has an obsession with searching and surfing things online. Googling something is one of the most frequent activities we engage ourselves in daily life. But what goes behind that is nothing short of fascination. There are a lot of algorithms that work behind the scene for a single Google search.

Google is just one example of a Search Engine. There are a lot of Search Engine Systems. These systems form of part of a larger category called Information Retrieval Systems. Information Retrieval systems deal with info organization, storage, revival and evaluation form document repositories/databases. Various algorithms are used to retrieve relevant information for user queries.

TF-IDF method is one of most popular methods of information retrieval. This method, even though popular, is not one the most accurate and reliable information retrieval system .This method has some disadvantages. One of them as we have observed is .3that all the content that is relevant to a particular query may not be considered with traditional TF – IDF systems. We in this paper propose an algorithm based on co-reference resolution and pronoun substitution to overcome the same problem faced in traditional TF IDF algorithm.

II. LITERATURE SURVEY

A. What is Information Retrieval and a Search Engine

Information Retrieval systems are systems that deals with information organization, storage, retrieval and evaluation from document database/repositories typically wherein data is semi-structured or unstructured[1]. IR systems are prevalent especially with text based document repositories. Search Engine Systems are the most popular example of IR systems. Google particularly is used especially on a day to day basis.

Search Engine System is a type of system in which users search for variety of information and the system collects related information and documents based on organized information.[2] The retrieved documents are ordered based on their relevancy and they are retrieved and displayed to the user who searches it.

B.TF/IDF Based Information Retrieval

TF/IDF method is one of the most popular strategies of Information Retrieval. Before going into the strategy we should know about some basic terms like TF,IDF.

Consider we have a set of n documents. Lets say we want TF i.e Term frequency of a term t in document d can be written as tf(t,d) = N(t,d) which is the number of times the term appeared in the document. But there is a need to normalize for unbiased results . So the normalized TF would be N(t,d)/||D|| where ||D|| is the number of terms in the document D .

IDF is called Inverse Document Frequency.
IDF(t) = log(N/df(t)). N here stands for the number of documents and df(t) is the number of documents. This is used to get the data on the relative importance of the term.

$$\textbf{tf-idf(t,d) = tf(t,d) * idf(t,d)} \qquad (1)$$

Based on the preprocessed text of documents a table is maintained for each term and documents and based on a similar processing for the query (q). So based on cosine

similarity we can order the documents which are most similar to least similar and then display to the user the documents based on relevancy.

C. Disadvantages of the above method

Although the method proves to be pretty effective in retrieving relevant information and documents there are few significant disadvantages to this method. The TF IDF feature does not consider the distribution of information among word class[3]. Also in this method few terms which could be more important/meaningful which could not be given apt or sufficient weight[4]. In the method of TF-IDF factors like position, semantics and co-occurrence of terms are not taken into consideration. TF-IDF being a unsupervised feature selection technique does not give an idea on a term being relevant to a particular class or not. TF/IDF method has the assumption which is not always true that the counting of different terms gives independent evidence of similarity[5].

D. Various Improvements Suggested

Xu R et.al [4] suggests an improvement of the TF-IDF algorithm using the concept of Parts of Speech. It suggests that verbs and nouns are more meaningful and important than adverbs and adjectives which in turn are more important than other POS tags. So based on this theory the author suggests a POS Weighted TF-IDF Algorithm. This can also be modified further so that user preference on importance of different terms .

Roul et.al [5] suggests 4 types of Modified TF-IDF. One of them is TF-IDF based on modified IDF. Here IDF is modified as follows

$$\textbf{Modified}_{IDF}\textbf{(i) = log10(no of documents in P + 1/document frequency of term i)} \quad (2)$$

Here one is added to the number of documents to make sure that discriminatory power is not reduced for both unique and non unique terms. At the same time if a term is important but prevalent in all documents this modification would allow not to lose its significance. The modified TF-IDF value is given below:

$$\textbf{Weight(i,j) = TF(i,j) * Modified}_{IDF}\textbf{(i)} \quad (3)$$

Some other basis for modifying TF-IDF to overcome traditional TF-IDF as given by [5] is as follows:

Inter-class dispersion: Inter-class dispersion can be defined as a term's contribution within a class that helps in correct classifying/retrieval and hence correct decision making.

Class frequency: This helps identifying extent of relevancy of a term to a class. Classification and Retrieval Accuracy could be enhanced based on this statistic.

Normalized Length: Increases the importance of TF and reduces the weights of terms that are not that frequent but that have relatively high TF-weighting in the document.

Based on the above three factors Modified TF-IDF values can be used as an improvement over traditional TF-IDF.

Guo,A et.al [3] also suggests a method wherein the modified IDF formula increases the weight of those items which appear frequently in the class.

Wang, N et.al [6] suggested using the function of distribution degree of a term in a class and using that subsequently to determine their classification ability. For eg – if a term is prevalent in a lone class then distribution degree = 1 and has the weakest classification capability and if the case is the opposite wherein the term is present in every class then the distribution degree is subsequently 0 and has the strongest classification ability.

E. Performance Metrics

Various performance metrics can be used for determining and comparing the performance of various TF-IDF based Information Retrieval Techniques.[5] suggests the following performance techniques.

1) **Precision:** Precision can be defined as the ratio of the number of relevant records retrieved and total number of records retrieved. The more the precision the more the relevancy is and this method has proved to be effective in retrieving more relevant documents.

2) **Recall:** Recall can be defined as the ratio of the number of relevant records retrieved and total number of records in the database that are relevant . Recall is the fraction of relevant documents that were retrieved.

3) **Mean Average Precision :** Mean Average Precision for all queries put together.

III. PROPOSED METHODOLOGY

As discussed before TF-IDF is one of the most popular methods of Information Retrieval. It is based on the Term Frequency and the Inverse Document Frequency of various terms across various documents. Documents are retrieved and ordered based on cosine similarity.

While noting the success of TF-IDF in retrieving relevant documents , it is also to be noted that there are some disadvantages in using this method. One of the disadvantages which we have identified in the method of TF-IDF is that given a query the retrieval system may not potentially identify all the content that is related to that particular query. To explain in detail assume the following document.

"MS Dhoni has been in great form. His hard work has been paying dividends lately. He and Yuvraj Singh have steered India to the final of ICC World Cup 2011".

Consider two queries q1 = "MS Dhoni" and q2 = "Yuvraj Singh". As we can see there are three sentences in the document that is related to q1 i.e MS Dhoni but only one sentence related to q2 i.e Yuvraj Singh. But if we traditionally implement TF-IDF both queries will give more or less the same result. But this result is not acceptable because more the content, more the document

is related and it should be ranked accordingly.

In order to make this possible we have identified possibly one of the possible solutions. This could be done with the help of co-reference resolution and pronoun substitution. Co-reference resolution refers to the process of finding the expressions (especially pronouns) that relate to a particular entity. Then pronoun substitution is substituting appropriate pronouns with corresponding proper noun. To explain this in detail a comparative study of the TF-IDF algorithm before and after this process must be done.

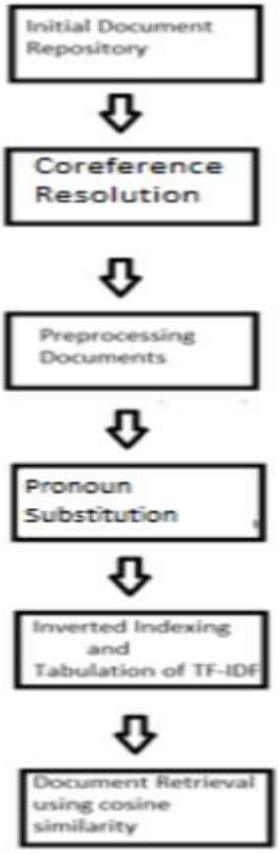

Figure 1 : Sequence of events in the proposed methodology

Given N documents in the repository for information retrieval, for each document we tabulate the TF values(no of times term appears in a document) of preprocessed documents
(documents after stop words removal etc). It is normalized. IDF values is also calculated for each term which is nothing but $\log(N/\log(n_i))$ where n_i refers to number of documents where term i exists and N refers to total number of documents in the repository. Then we create a table with TF-IDF values which is the product of TF values of each term in the document with IDF of the term. Generally TF-IDF value is calculated as follows:

$$\textbf{TF-IDF(i,d) = TF(i,d) * IDF(i)} \quad \textbf{(4)}$$

We create another repository wherein we store documents after co-reference resolution and pronoun substitution. This is done using neural coref package of python which is used for co-reference resolution. Based on that pronoun substitution was correspondingly done. For both the repositories TF-IDF values are tabulated.

We can note that IDF values does not change for both the repositories. But there is a noticeable change in TF for some terms in both the repositories.
So,

$$\textbf{TF-IDF1(i,d) = TF1(i,d) * IDF(i)} \quad \textbf{(5)}$$

$$\textbf{TF-IDF2(i,d) = TF2(i,d) * IDF(i)} \quad \textbf{(6)}$$

Here TF1 is the Term frequency before pronoun substitution and TF2 is the term frequency after pronoun substitution..

Also one can notice the length of the documents would have also changed after pronoun substitution. So, corresponding documents in repositories R1 and R2 have different lengths.

$$\textbf{Length of a document} = \sqrt{(\sum(\textbf{TF-IDF(i,d)})\text{^}2} \quad \textbf{(7)}$$

Now for the query a similar process is followed wherein TF-IDF values are tabulated and length values are calculated.
Now based on the queries, tabulated values of TF-IDF for both the repositories and the query and calculated lengths for each document and the query, based on the cosine similarity ,the relevant articles are retrieved and are displayed to the user based on the order of relevance.

IV. IV. RESULTS AND DISCUSSION

Our results show significant change in the order of relevancy of both the methods. We took a document repository of 284 documents collected from Deccan Chronicle.

For example we took "Faf Du Plessis" as the query for the two repository. 22 documents in both repositories were found to be relevant to the given query. Different results of documents were observed for the same query for different repositories. Following graphs and screenshots illustrate that.

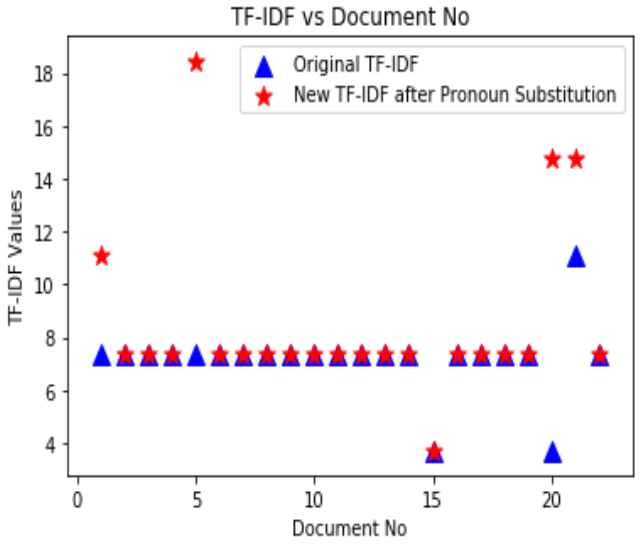

Figure 2 : TF-IDF vs Document No

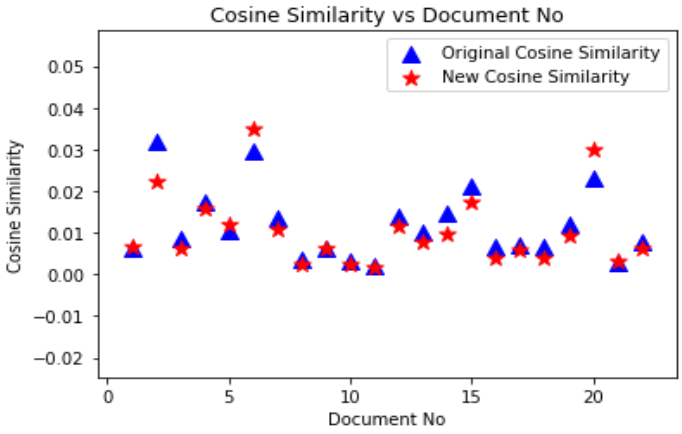

Figure 3 : Cosine Similarity vs Document No

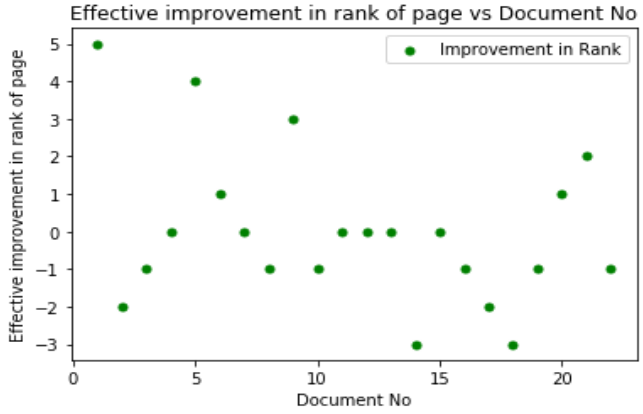

Figure 4 : Effective improvement in rank of page vs Document No

In Figure 2 for a part of the query 'Faf' TF-IDF change in TF-IDF values were displayed in the graph wherein for some of the documents there was an observable difference in the values because of the change due to pronoun substitution.

Correspondingly this change (along with the observed change in length) has an effect on the cosine similarity as one can see in Figure 3. So when we take this into account and plot the observed change in rank or rather the effective improvement in rank many documents have been impacted and effective improvement up to 5 places in the repository of 22 documents were seen (See Figure 4). This shows that there has been a significant impact due to pronoun substitution.

An evidence of how more related content was taken into account by using the proposed procedure can be provided with the following example. Of the 22 documents retrieved for the above observation document no 6 initially had an rank of 2 which subsequently became 1. This can be explained with the help of the below images.

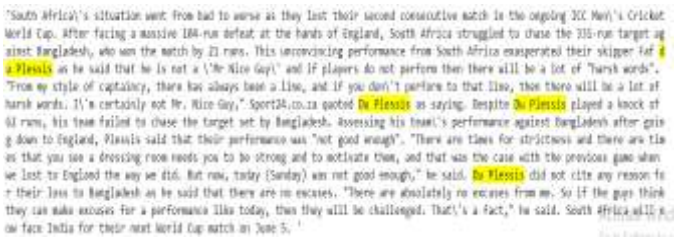

Figure 5 : Before Pronoun Substitution

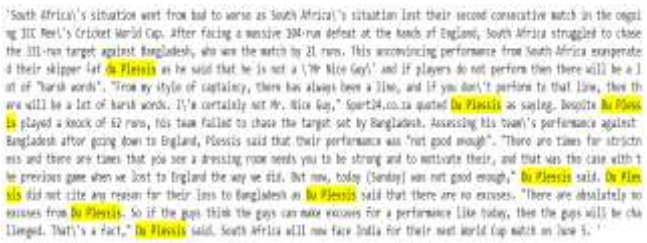

Figure 6 : After Pronoun Substitution

As you can see in Figure 5 , there were only 4 sentences related to Du Plessis was discovered and correspondingly the TF-IDF algorithm worked to give a rank of 2 for the document. But after Co-reference Resolution and subsequently pronoun substitution 8 instances of Du Plessis was found in sentences as one can see in Figure 6 highlighting that more content was actually related to the search query and because of which after modification TF-IDF gives a higher ranking of 1 for the same document.

V. CONCLUSION

Through this paper we successfully conclude that using co-reference resolution and pronoun substitution a higher degree of relevant content can be taken into account and make sure that the results are a more accurate description and ranking of relevant documents. Although this improvement is significant this can be improved further. One way is , we take popular queries and for all those queries we create a database of alias names and this could be taken into account to produce more accurate results. For eg – MS Dhoni and Mahi represent the same person but if the query is "MS Dhoni" the content referring to Mahi is not considered. So this can be done using the above mentioned process. Like the suggested method many other ways can be suggested to improve the present work.

REFERENCES

1. Sanderson, M., & Croft, W. B. (2012). The history of information retrieval research. *Proceedings of the IEEE*, *100*(Special Centennial Issue), 1444-1451.
2. Jin, Y., Lin, Z., & Lin, H. (2008, December). The research of search engine based on semantic web. In *2008 International Symposium on Intelligent Information Technology Application Workshops* (pp. 360-363). IEEE
3. Guo, A., & Yang, T. (2016, May). Research and improvement of feature words weight based on TFIDF algorithm. In *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference* (pp. 415-419). IEEE.
4. Xu, R. (2014, July). POS weighted TF-IDF algorithm and its application for an MOOC search engine. In *2014 International Conference on Audio, Language and Image Processing* (pp. 868-873). IEEE.

5. Roul, R. K., Sahoo, J. K., & Arora, K. (2017, December). Modified TF-IDF term weighting strategies for text categorization. In *2017 14th IEEE India Council International Conference (INDICON)* (pp. 1-6). IEEE.
6. Wang, N., Wang, P., & Zhang, B. (2010, June). An improved TF-IDF weights function based on information theory. In *2010 International Conference on Computer and Communication Technologies in Agriculture Engineering* (Vol. 3, pp. 439-441). IEEE.
7. Mishra, A., & Vishwakarma, S. (2015, December). Analysis of tf-idf model and its variant for document retrieval. In *2015 international conference on computational intelligence and communication networks (cicn)* (pp. 772-776). IEEE.
8. Liu, Q., Wang, J., Zhang, D., Yang, Y., & Wang, N. (2018, December). Text features extraction based on TF-IDF associating semantic. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)* (pp. 2338-2343). IEEE.
9. Sanderson, M., & Croft, W. B. (2012). The history of information retrieval research. *Proceedings of the IEEE*, *100*(Special Centennial Issue), 1444-1451.
10. Jin, Y., Lin, Z., & Lin, H. (2008, December). The research of search engine based on semantic web. In *2008 International Symposium on Intelligent Information Technology Application Workshops* (pp. 360-363). IEEE