2019 International Conference on Computing, Power and Communication Technologies (GUCON)
Galgotias University, Greater Noida, UP, India. Sep 27-28, 2019

Ranking Web Pages Using Cosine Similarity Measure

Shadab Irfan
Research Scholar (SCSE)
Galgotias University Greater Noida
Uttar Pradesh, India
shadab710@gmail.com

Subhajit Ghosh
Professor (SCSE)
Galgotias University Greater Noida
Uttar Pradesh, India
subhajit.ghosh@galgotiasuniversity.edu.in

***Abstract*— The work presented in this paper throws light on various models of Information Retrieval and proposed an approach to rank web pages using document clustering on the basis of content similarity. The paper presents a method where the content of the documents are matched on the basis of the query and thereby the most similar documents are clustered together for the ranking process. The given approach overall reduces the computation complexity and help in retrieving and ranking the most relevant pages in less amount of time.**

Keywords—Information Retrieval; Web Mining; Clustering; Similarity Measure

I. INTRODUCTION

The information on the web is increasing in a rapid phase and possesses various challenges for the researchers to mine the relevant information as per their need and requirement. As the amount of information is too massive on the web so in order to extract the relevant information various information retrieval models and ranking algorithm are proposed that help to find the much needed information. The algorithms normally find information on the basis of content of web pages, link among web pages or log information of web pages. The search engine plays a major role in filtering the much needed information in a constrained time period. One of the important criteria that are handled by all search engines is to find high quality pages which are normally based on the conditions set by the users. Diverse methods are used for rating the web pages and in order to perform this task web crawler and indexer play a major role[1]. Ranking is one of the important measures which help in arranging the web pages in order of importance on the basis of the query entered by the users. Various algorithms have been proposed that help in accessing the relevant information [2].

The paper provides an overview of various models of information retrieval and the similarity measures which are used for finding the similarity among the web pages. The mining of the information is done based on the content of the web pages found and thereby clustering the web pages on the basis of similarity and ranking them.

II. INFORMATION RETRIEVAL PROCESS AND MODELS

The process of Information Retrieval focuses on fulfilling the requirement of the user by finding the most appropriate information according to the need from the World Wide Web. The process of information retrieval encompasses query formation, matching function and documentary database and is depicted in "Fig 1" [3].

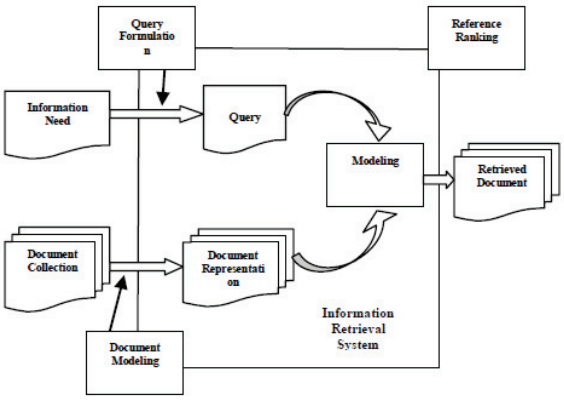

Fig. 1. IRS Structure [4]

Various measures are used to increase the quality of the documents retrieved in the information retrieval process. In order to increase the precision and recall for the information retrieval system document clustering is initiated. It helps in arranging the information into a structured format that eases the process of retrieval. It has been found out that clustering help to organize the documents on the basis of similarity and it can be divided into two stages- first to perform the preprocessing of the documents and second is to analyze them and partition them into clusters [5].

It has been observed that various models for Information Retrieval have been proposed that ease the process of retrieval. The proposed information retrieval models are Boolean Model, Vector Space Model, Region Model, Probabilistic Model, Bayesian Network Model, 2-Poisson Model, Language Model and Google's Page Rank Model. The Boolean model provide flexibility but lack ranking of documents, Vector space model represents query in form of vectors, Probability model is based on probability relevance while Google uses page ranking for ranking the documents. Though these models on one hand provide various measures to solve complex problem but on the other hand lack on various grounds in finding the relevant information [6].

III. WEB MINING PROCESS

The process of mining the information from the web takes various forms and it facilitates the extraction of the information by using various ways. Web Mining help to extract relevant pages from the web by using various forms of measures like content, link and log information and on this concept it is basically divided into Web Content Mining, Web Structure Mining and Web Usage Mining. The mining process helps in extracting interesting patterns that help in

further processing. As the web is huge and semi structured in nature so it possess various challenges for the researchers to retrieve information [7].

IV. Document Clustering and Similarity Measure

In order to manage large group of data and to find relevant information from such large pool of data clustering of documents can boost the retrieval process. The precision and recall of the documents can be enhanced if clustering is employed for the retrieval of the information [8]. The concept of web clustering helps to organize similar web pages into groups removing irrelevant pages which are not important from search point of view [9]. Document clustering helps the user to search for a specific pattern according to their requirement. It helps in improving the method of search and retrieval [10]. Clustering web pages help in organizing similar pages which are tightly coupled with each other. The organization of similar pages into group helps in the various process of analysis [11].It has been found that clustering not only help in finding similar patterns which lie in dataset but also ease the process of finding the information in minimum amount of time [12].

Document similarity can be calculated by using different measures. The similarity of the documents is described by cosine or angle between two vectors [13]. The term frequency is calculated where each term in the document is assigned a weight equals to the number of occurrences. A score is computed between query term 't' and document 'd' and is denoted as $tf_{t,d}$.

Since term frequency give all term equal importance so in this respect inverse document frequency (idf) is calculated for N collection of documents and is defined as in (1)-

$$\text{idf}_t = \log \frac{N}{\text{df}_t}. \tag{1}$$

Now the composite weight for each term in the document is calculated as tf-idf weight and represented as shown in (2)-

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t. \tag{2}$$

After calculating the tf-idf score the similarity of the documents with respect to the query is evaluated to assess the similarity effect among documents and rank them in order of high relevance. One of the standard ways to evaluate the similarity between documents d1 and d2 or between query q and document d is to calculate the cosine similarity of the vectors $\vec{V}(d_1)$ and $\vec{V}(d_2)$ as given in (3)-

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}. \tag{3}$$

Finally cosine similarity between query vector and document vector is calculated which help to rate the top documents for a particular query [14].The final value is calculated as given in (4)-

$$\text{score}(q, d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)||\vec{V}(d)|}. \tag{4}$$

V. Proposed Approach

The proposed approach accepts the query and on the basis of documents retrieved calculates tf-idf score, thereby cluster the documents on basis of different cases and calculate similarity measure. The outline model of the proposed approach can be depicted in Fig 2.

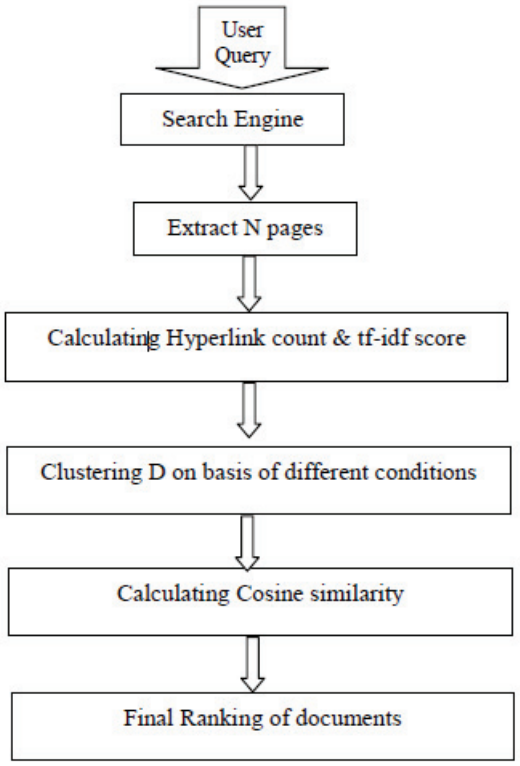

Fig. 2. Proposed Approach Model

In order to implement the proposed approach we will take an example of hyperlink structure comprises of seven web pages/documents as shown in Fig 3.

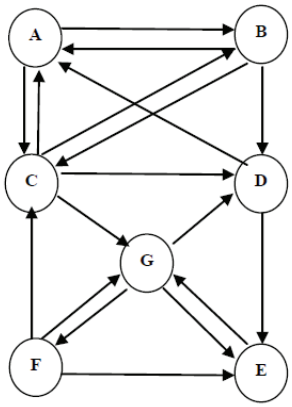

Fig. 3. Hyperlink Structure of web pages

The procedure for the proposed algorithm is:

Input: User Query
Output: Ranked Documents/Web Pages
Steps:

1. Enter the query 'Q' comprises of 'n' terms to extract 'N' pages.
   a. Remove stopwords.
   b. Count Number of terms 'n' of query 'Q'.
2. Calculate Hyperlink count (Inlink + Outlink).
3. Calculate tf-idf score.
4. Group documents/webpages 'D' on the following criteria-
   Case I: If all term 'n' of 'Q' appear in Di

where i=1,2,3,…M group them in Ci, where i=1,2,3.

Case II: If 'n-1> D <n/2' group them in Ci, where i=1,2,3.

Case III: If no term 'n' match group them in cluster Ci, where i=1,2,3.

5. Calculate cosine similarity between Di where i=1,2,3,…M and Q which are then grouped into clusters.
6. Finally the documents are ranked in the clusters.

For a given Query "Information Retrieval Evolutionary Computation, it is assumed that all the seven pages (A, B, C, D, E, F, G) shown in Fig 3 contain the frequency of the terms in the following order given in Table 1-

TABLE I. TERM FREQUENCY IN DOCUMENTS

	A	B	C	D	E	F	G
Information	20	8	8	10	20	0	25
Retrieval	10	12	13	15	10	0	0
Evolutionary	12	0	6	8	5	0	0
Computation	8	0	2	6	2	6	0

The hyperlink count for the graph given in Fig. 3 is calculated and shown in Table II.

TABLE II. HYPERLINK COUNT

Node	Inlink	Outlink	Hyperlink Count
A	3	2	5
B	2	3	5
C	3	4	7
D	3	2	5
E	3	1	4
F	1	3	4
G	3	3	6

It is assumed that page A contains 500 words, B contains 300 words, C contains 200 words, D contains 400 words, E contains 300 words, F contains 300 words and G contains 200 words. On the basis of this the tf-idf score is calculated and finally the cosine similarity is calculated as shown in Table III and Table IV.

TABLE III. TF-IDF SCORE FOR TERM

Document \ Term	tf-idf Score			
	Information	**Retrieval**	**Evolutionary**	**Computation**
A(500)	0.008564	0.009708	0.0193752	0.0077664
B(300)	0.0057093	0.019416	0	0
C(200)	0.008564	0.031551	0.024219	0.004854
D(400)	0.0053525	0.0182025	0.016146	0.007281
E(300)	0.0142733	0.01618	0.013455	0.003236
F(300)	0	0	0	0.009708
G(200)	0.0267625	0	0	0

After calculating the tf-idf score for respective term the tf-idf score for the query is calculated as shown in Table IV.

TABLE IV. TF-IDF SCORE FOR QUERRY

Query	tf-idf
Information	0.053525
Retrieval	0.12135
Evolutionary	0.201825
Computation	0.12135

After calculating the tf-idf score finally cosine similarity of all the documents with respect to query is calculated as shown in Table V

TABLE V. COSINE SIMILARITY

Cosine Similarity	
cos(Q,A)	0.97755
cos(Q,D)	0.925222
cos(Q,C)	0.88167
cos(Q,E)	0.842342
cos(Q,B)	0.48662
cos(Q,F)	0.448983
cos(Q,G)	0.198037

The graph depicted in Fig 4 displays the cosine similarity of all the documents.

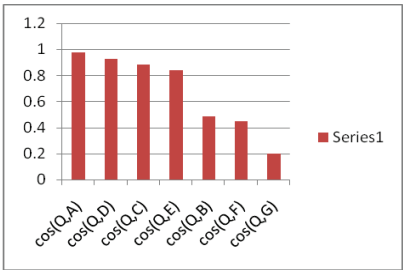

Fig. 4. Cosine Similarity

Now applying the proposed approach and inspite of calculating the cosine similarity of all the documents it has been found that only A,D,C and E contain all the words of the query and hence applying case I are clustered in one group and on document B case II is applied which is placed in second cluster while the two documents F and G are left out. Finally cosine similarity of only A,D,C,E and B is calculated inspite of all and documets and they are ranked on basis of similarity measure. The Table VI and Table VII show the cosine similarity.

TABLE VI. COSINE SIMILARITY FOR CLUSTER 1

Cluster 1	
cos(Q,A)	0.97755
cos(Q,D)	0.925222
cos(Q,C)	0.88167
cos(Q,E)	0.842342

TABLE VII. COSINE SIMILARITY FOR CLUSTER 2

Cluster 2	
cos(Q,B)	0.48662

The Fig 5 shows the cosine similarity of the final clustered documents.

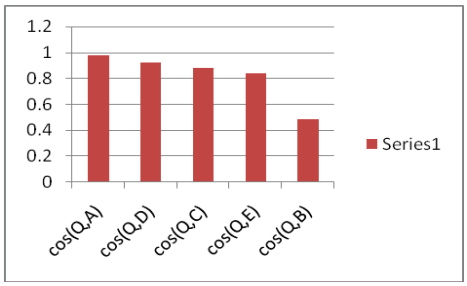

Fig. 5. Cosine Similarity of Clustered Documents

The Table VIII show the final ranking of the documents A,D,C,E and B.

TABLE VIII. RANKED DOCUMENTS

Ranked Documents
A
D
C
E
B

The importance of the proposed approach is that in spite of calculating the cosine similarity of all the documents for ranking, only few calculations is done on the documents which are clustered together on the basis of the proposed algorithm thus overall reducing the complexity of the calculation and also minimize the time for retrieval. It has also been found that though the hyperlink count for node G is high but still due to lack of query words in it, it is left out in the processing task.

VI. CONCLUSION

The paper proposed a new approach for ranking the web pages. After investigation of various models for information retrieval and ranking algorithm it has been found that content mining play a major role and in order to rank the documents content similarity has been proposed by various researchers. By applying the document clustering with similarity measure we can retrieve the desired information on the basis of query by overall reducing the complexity of the computation process.

The experimental implementation is done by using the basic approach and proposed approach. The results shows that the given approach help to cluster the documents on the basis of terms present in it and calculate cosine similarity for checking the similarity measure of only those web pages which satisfy the given condition and clustered together.

Thus it has been found that the time is reduced overall by calculating the similarity measure of only clustered documents and leaving the rest documents which are of less relevance. As the process of information retrieval is a typical task which require abundance of time in finding the required information so this process also helps in reducing the computation complexity of the process and help in getting the required information easily thus overall reducing the time and cost .

REFERENCES

[1] Jaskirat Singh and Mukesh Kumar, "A Meta Search Approach to Find Similarity between Web Pages Using Different Similarity Measures", ICAC3, CCIS 125, pp. 150–160, Springer-Verlag Berlin Heidelberg 2011.

[2] Shadab Irfan, Subhajit Ghosh,"A Review on Different Ranking Algorithms",IEEE International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018.

[3] Anubha Jain, Swati V. Chande, Preeti Tiwari ,*"* Relevance of Genetic Algorithm Strategies in Query Optimization in Information Retrieval*",* (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (No.4), 2014.

[4] Anubha Jain and Swati V Chande, "Impact of Different Selection Strategies On Performance Of GA Based Information Retrieval " , International Journal on Computational Science & Applications (IJCSA) Vol.5,No.6, December 2015.

[5] Julian Sedding,Dimitar Kazakov,"WordNet-based Text Document Clustering".

[6] Shadab Irfan, Subhajit Ghosh,"Optimization of Information Retrieval UsingEvolutionary Computation: A Survey",International Conference on Computing, Communication and Automation (ICCCA2017).

[7] Amar Singh, Navjot Kaur,"A Survey on k-Means Clustering Algorithm Using Different Ranking Methods in Data Mining", International Journal of Computer Science and Mobile Computing,IJCSMC, Vol. 2, Issue. 4, April 2013, pg.111 – 115.

[8] Nadella Sandhya and A. Govardhan, "Analysis of Similarity Measures with WordNet Based Text Document Clustering ", Proceedings of the InConINDIA 2012, AISC 132, pp. 703–714, Springer-Verlag Berlin Heidelberg

[9] Jeevan H E , Prashanth P P , Punith Kumar S N , Vinay Hegde ,"Web Pages Clustering: A New Approach",International Journal Of Innovative Technology & Creative Engineering (ISSN:2045-8711) Vol.1 No.4 APRIL 2011.

[10] Oren Zamir and Oren Etzioni and Omid Madani and Richard M. Karp, "Fast and Intuitive Clustering of Web Documents",KDD-97 Proceedings. Copyright © 1997, AAAI.

[11] Antonio LaTorre, Jose M. Pena, Victor Robles, Maria S. Perez,"A Survey in Web Page Clustering Techniques".

[12] Shadab Irfan, Gaurav Dwivedi, Subhajit Ghosh," Optimization of K-Means clustering Using Genetic Algorithm", IEEE International Conference on Computing And Communication Technologies For Smart Nation(IC3TSN),2017.

[13] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney,"Impact of Similarity Measures on Web-page Clustering",AAAI Technical Report WS-00-01. Compilation copyright © 2000, AAAI.

[14] Manning, C.D., Raghavan, P.," An introduction to Information Retrieval", Preliminary draft© 2008 Cambridge UP (2008).