Research Article

AI and Intelligent Systems: Engineering, Medicine & Society

Hybrid Method to Enhance Effectiveness and Efficiency for Information Retrieval in Web Search Engines

Zahir Edrees[1]* and Yasin Ortakci[2]

[1] *International Science and Technology University, Warsaw, Poland*

[2] *Department of Computer Engineering, Faculty of Engineering- Karabuk University, Karabuk, Turkey*

*Corresponding Author
Zahir Edrees, International Science and Technology University, Warsaw, Poland.

Submitted: 2025, June 10; **Accepted:** 2025, June 14 ; **Published:** 2025, July 23

Citation: Edrees, Z., Ortakci, Y. (2025). Hybrid Method to Enhance Effectiveness and Efficiency for Information Retrieval in Web Search Engines. *AI Intell Sys Eng Med Society, 1*(2) 01-08

Abstract

Search engines are very important tool today, most of people depend on it to achieve the information they need it. The main issue of search engine is ranking algorithms because it provides top relevant web pages related to the search operation. İn this paper will describe the Page ranking algorithm, as an important part of information retrieval and we modified pagerank algorithm we called our proposed algorithm smart clustering pagerank (SCPR) after that we made hybrid between our proposed algorithm SCPR and BM25 algorithm . Also we compared between them according to time response, efficiency, effectiveness, importance and limitation, the evaluation results show that the our proposed method had better effectiveness and reduced computational time.

Keywords: Page Ranking, Effectiveness, K-means Clustering, Closeness Centrality, Eigenvector Centrality

1. Introduction

Information retrieval (IR) is the process to find material (documents) of an unstructured nature [1]. one of benefits of IR system is that it does not just get documents. It gives the researcher the Uniform Resource Locator of these documents (URL). Each information retrieval systems must handle these issues: Firstly, to give the user related information according to his search this known as the effectiveness of IR system and secondly to minimize the response time to get users requirements we called this as efficiency of IR system, based on these two criteria's the user will decide which search engine can use it. The main difference between information retrieval and data retrieval is that we use artificial query in data retrieval but in information retrieval. we use natural language, also the query may be incomplete in information retrieval but must be complete for data retrieval.We will explain the main components of the web IR system [2]. It is shown in Figure 1:

1. Crawling Process: is the process of browsing documents by using Uniform Resource Locator (URL).
2. Indexing Process: Building the index of the documents.
3. Querying Process: User search for information.
4. Ranking Process: Retrieves for documents that are related to the user's requirements.
5.Users give ffeedback to IR system about satisfaction or not.

In search engine system, user submits a query to the search engine. After that the search engine searches for the relevant results of user's query. The relevance ratio of web page can be obtained by considering the number of in-links and out-links present in a particular web page. If the web page has more number of out-links to a relevant page then that page can be considered as a central page. From this central page, all other web pages are compared for similarity and the most similar pages are grouped together. The grouping of most similar pages together is known as clustering.

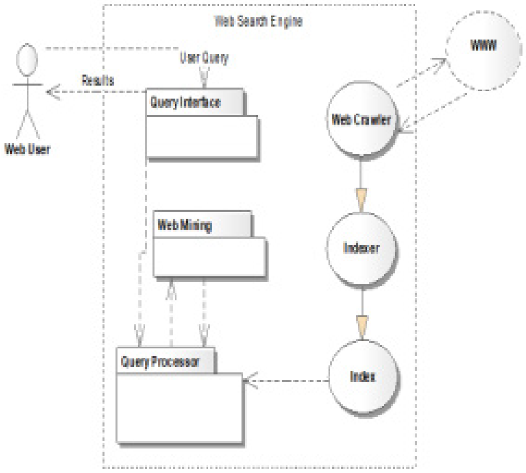

Figure 1: Important Processes of Web Information Retrieval

Clustering can be done based on different algorithms such as hierarchical, k-means, partitioning, etc. The simplest unsupervised learning algorithm that solves clustering problem is K-Means algorithm. It is a simple and easy way to classify a given data set through a certain number of clusters. When documents are clustered using the K-means algorithm [3], the cluster contains more similar documents and increases the relevance of search results. When a user makes a query after this clustering process, they will only get the most relevant cluster that matches their query. They will not get any of the irrelevant pages. Thus, it increases the efficiency of search results and reduces the computation time.

In this paper, we will explore the problem of relevance of the resulting pages to the user query for page ranking algorithm also to improve the efficiency and effectiveness of page ranking algorithm. Our solution by using k means clustering to group web pages into clusters based on closeness, eigenvector and degree centrality to increase the relevance of retrieved web pages. The rest of this paper is organized as follows: in section 2 we introduce related work page ranking algorithm. In section 3, we explain the proposed algorithms. The experiment and discussion are presented in Section 4. Finally, section 5 summarizes the paper.

2. Backgorund

Web page ranking is an optimization technique used by search engines to determine the importance of a web page relative to other pages. There are different criteria used by ranking algorithms, such as link structure or page content [4]. Page ranking algorithms can be classified into two groups: Content-based Page Ranking and Connectivity-based Page Ranking. Content-based Page Ranking ranks pages based on their textual content, taking into account factors such as the number of matched terms with the query string, frequency of terms, and location of terms. Connectivity-based Page Ranking, also known as Link-based ranking, works on the basis of link analysis technique. It views the web as a directed graph where web pages form nodes and hyperlinks between the pages form directed edges [5,15].

2.1. PageRank Algorithm

The PageRank algorithm was developed at Stanford University by Larry Page and Sergey Brin in 1996 [4,16]. The PageRank algorithm is a link-based algorithm that uses the structure of the web to determine the importance of a web page. It works by calculating the probability of a user visiting a page under a random surfer model, where the user randomly clicks on links on a page and switches to another page when they get bored. The algorithm assigns a PageRank value to each web page, which is pre-computed for over 25 billion web pages on the World Wide Web. PageRank is one of commonly algorithms used in ranking, the Google used it in ranking process [3,14]. Formula of Page Ranking algorithm calculation:

$$PR\ (A) = (1-d) + d\ (PR\ (T_1) / C\ (T_1) + \ldots + PR\ (T_n) / C\ (T_n)\ ) \qquad (1)$$

Here, n is the number of pages accounted, d is dampening factor equals 0.85 and it is used to represent pages that have no inlinks to give it some Page Rank value. $C\ (T_1)$, $C\ (T_2)$.. $C\ (T_n)$ are the number of outlinks of pages, T1, T2… T_n are links to the page A.

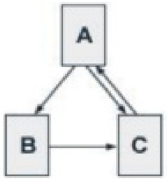

Figure 2: Example of Hyperlinked Structure

Figure 2, is example to shows that how page rank algorithms works, if we have websites that includes A, B and C, page A links to B, C; and B links to the C and C links to the A. To calculate the page rank through these steps

PR (A) = 0.15 + 0.85 PR (C)
PR (B) = 0.15 + 0.85 (PR (A)/2)
PR (C) = 0.15 + 0.85 (PR (A)/2+PR (B))

Iteration	PR (A)	PR (B)	PR (C)
0	1	1	1
1	1.00	0.58	1.06
2	1.05	0.60	1.10
3	1.09	0.61	1.13
4	1.11	0.62	1.15
5	1.13	0.63	1.17
6	1.14	0.63	1.17
7	1.14	0.63	1.17

Table 1: Iteration to Calculate Page Rank for Pages

After the iteration of calculation, we will get the below result of page ranking PR (A) = 1.14, PR (B) = 0.63, PR (C) = 1.17, we can see that PR (C) > PR (A) > PR (B). Page rank uses link structure, this a reason for the result that produced is not relevant to the user's query this problem is called theme drift.

2.1.1.Advantages of PageRank

The strengths of PageRank algorithm are as follows [5]:

- **Scalability:** PageRank is scalable and can handle large web graphs with billions of pages. This makes it suitable for ranking web pages on the entire World Wide Web.
- **Link Analysis:** PageRank algorithm is based on link analysis, which makes it a valuable tool for understanding the structure of the web. It helps to identify important web pages and their relationships with other pages, which is useful for various applications such as search engine optimization, web mining, and social network analysis.
- **Robustness:** PageRank algorithm is robust and can handle noisy and incomplete web graphs. It is less affected by spam and malicious links since it considers the quality and relevance of the linking pages in addition to the quantity of links.

2.1.2. Disadvantages of PageRank

The following are the problems or disadvantages of PageRank [6]:

- **Lack of Consideration for Page Content:** PageRank does not take into account the actual content of a web page, but only its link structure. This means that a web page with relevant and high-quality content may not necessarily have a high PageRank if it lacks backlinks from other high-quality websites.
- **Limited to Static Web Pages:** PageRank was developed to work on static web pages, and may not be as effective for ranking dynamic or personalized web pages.
- **Need for Periodic Updates:** PageRank values need to be periodically updated to reflect changes in the link structure of the web. This process can be time-consuming and resource-intensive for large web search engines
- **Lack of Consideration for Page Content:** PageRank does not take into account the actual content of a web page, but only its link structure. This means that a web page with relevant and high-quality content may not necessarily have a high PageRank if it lacks backlinks from other high-quality websites.
- **Theme Drift:** PageRank can suffer from the problem of "theme drift," where a web page may have a high PageRank due to its link structure, but the content may not be relevant to the user's search query.

3. Proposed Algorithm

Our Proposed Algorithm: Smart clustering PageRank algorithm (SCPR).
Input: Graph G (V; E), number of clusters k.
V: represents pages.
E: represents links between pages.
Output: A partition of the node set V into k clusters (subsets).

3.1. Proposed Algorithm Steps

3.1.1. Dataset Collecting and Visualizing

To check our dataset and how it's connected together we used Gephi tool to visualize our dataset as shows in Figure 3.

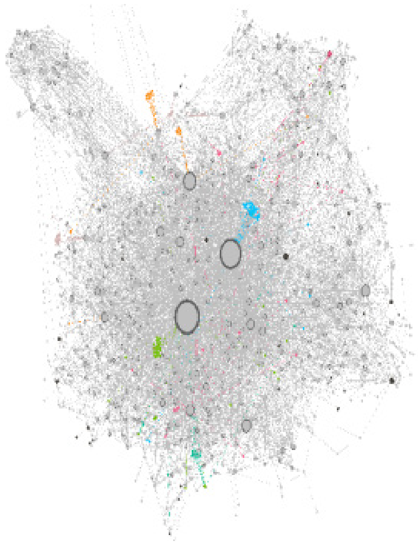

Figure 3: Dataset Visualizing using Gephi tool

3.1.2.Centrality Measures Calculation for Dataset

We calculated degree centrality measures for dataset as shows in Figure 3, Firstly we used Degree centrality of a page is simply its degree and the number of edges it has [7]. Closeness centrality measure indicates how close a node is to all other nodes in the network [13]. It is calculated as the average of the shortest path length from the node to every other node in the network [8]. Also, we used Eigenvector centrality to measure the level of influence of a page within a network, after the centralities has been calculated as shown in Table 2, the dataset becomes ready to apply K means clustering [9,12].

Source	Target	Degree centrality	Closeness centrality	Eigenvector centrality
758	1476	0.00012	0.01028	0.000621
5584	5917	0.00012	0.00011	0.009406

Table 2: Dataset after Calculated Centrality Measures for Web Pages

3.1.3. Applying k Means Algorithm Based on Centrality Measures

We used K means clustering algorithm to group our dataset in to different groups based on centrality measures as shows in Figure 4. K -means algorithm is an eminent clustering technique used to split a set of n items into k clusters with an intention to elevate the resulting intra-cluster resemblance and lower down the inter-cluster resemblance[10]. The resemblance within the cluster is calculated based on the mean value of the items within the cluster [11].

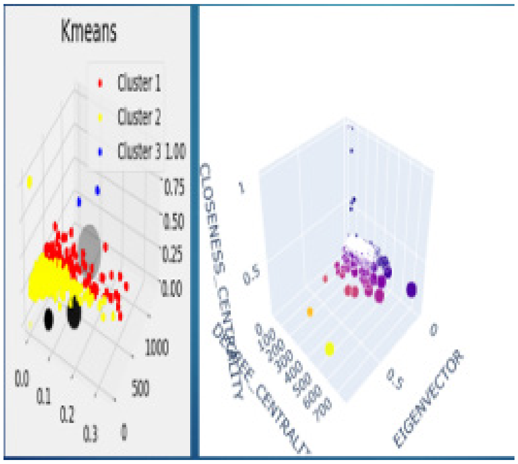

Figure 4: Dataset after Clustered using K means Algorithm

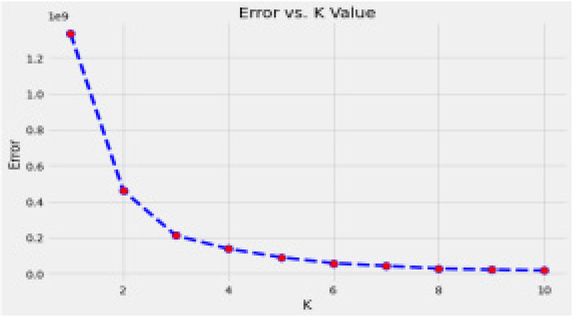

Figure 5: Elbow Method Runs K-Means Clustering on the Dataset

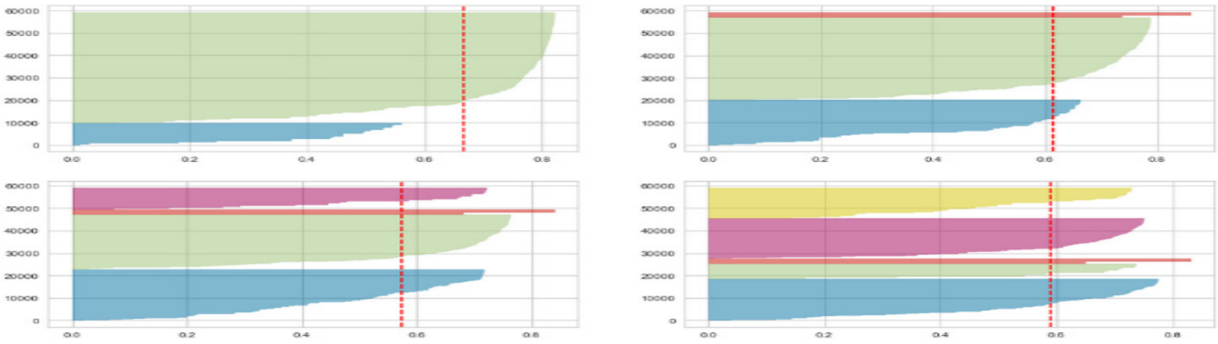

Figure 6: Silhouette Score to Validate Clustering for Dataset

3.1.4. Elbow Method (Clustering)

The elbow method is used in determining the optimal number of clusters in a dataset, Figure 5 shows the elbow method runs k-means clustering on the dataset for a range of values for k (from 1-10) and then for each value of k computes an average score for all clusters.

3.1.5.Accuracy (Quality) of Clusters

We used silhouette score to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster as shows in Figure 6, The value of the silhouette coefficient is between [-1, 1]. A score of 1 denotes the best meaning that the data point is very compact within the cluster to which it belongs and far away from the other clusters. The worst value is -1. Values near 0 denote overlapping clusters.

3.1.6. Calculate PageRank Algorithm for Each Cluster

The last step is to calculate Page rank score after dataset has been clustered in to groups.

3.1.7. Results

After the documents are clustered by using K-Means algorithm, each cluster contains more similar documents Calculate PageRank Algorithm for each cluster. Our proposed algorithm reduced computational time as shows in Figure 7. We used two web document data sets in our experiment. We apply the K means clustering algorithms on our data sets. Clustering is an efficient way of reaching information from raw data and K-means is a basic method for it. Although it is easy to implement and understand.

We have presented an efficient method of combining the restricted filtering algorithm and the greedy global algorithm and use it as a means of

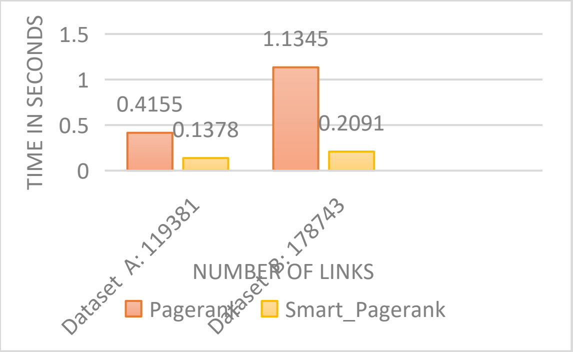

Figure 7: Compare Between Page Rank and Smart PageRank Clustering

Improving user interaction with search outputs in information retrieval systems. We used centrality measure for web graph to identify the web page ranks, also Comparison between the original page rank and smart cluster centrality-based Page rank has been made. We also evaluated the effeciency of our algorithm on two real-world datasets, our evaluation shows that the proposed algorithm performs better than that using only the page ranking algorithm. Our smart cluster page rank (SCPR) limitation is that it favors older pages, because a new page, even a very good one, will not have many links unless it is part of an existing website. To solve these issue in SCPR algorithm we proposed Hybrid algorithm based on content page rank algorithm (BM25) used in our hybrid algorithm and Smart cluster page rank (SCPR).

4. Hybrid Algorithm

We applied the following steps in our proposed Hybrid algorithm

- **Step 1:** A repository database includes our web pages datasets.
- **Step 2:** After creating the database a link structure will be created that will explain how pages are linked to each other. On the basis of links, smart cluster page rank (SCPR) will be calculated for each Page at the beginning.
- **Step 3:** User will add a query and database will be searched for the pages related to user query.
- **Step 4:** Pages will be searched for user query. Web Pages will be selected on the basis of their similarity content and those are similar to user search will be selected for user. Web Page similarity will be calculated using BM25 technique as content based rank.
- **Step 5:** After having the web pages those are matched with user query, their pages will be re-ranked according to BM25 and SCPR ranks as shown in Figure 8. Pages with high page links and content rank will be placed on top of the search result list.

To build our final search list we will consider both web pages content and links. HPR denotes hybrid page rank.

HPR (d) = SCPR (d) * BM25 (d) (2)

By using Eq. (2) a Hybrid Page Rank is computed by combining the percentage of smart cluster page rank (SCPR) and content based page rank (BM25) to get a final score and to generate a final list of pages.

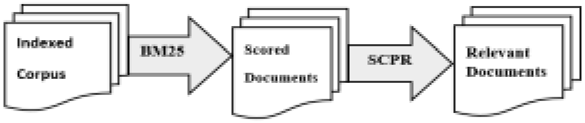

Figure 8: Hybrid Page Rank Algorithm

4.1. BM25 Algorithm

In information retrieval, Best Matching (BM25) is a ranking function used by search engines to estimate the relevance of documents to a given search query BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of their proximity within the document. It is a family of scoring functions with slightly different components and parameters. One of the most prominent instantiations of the function is as follows Eq. (3).

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)}$$

$$L = \sum_i |d_i|/N$$

$$IDF(q_i) = \log_e \frac{N - DF(q_i) + 0.5}{DF(q_i) + 0.5}$$

- The equation gives the BM25 score for query consisting of the words given in document.
- TF (q, d) is the count of number of times word appears in document.
- | d | in the length of document in words.
- L is average length of document in corpus.
- K and b used cross validation k=2, b=0.75.
- IDF is inverse frequency document for word.
- DF is number of documents that contain word.
- N (in BM25 equation) is number of terms in query.
- N (in IDF equation) is total number of documents.

4.2. Assessment Criteria’s

In this evaluation, precision at the position of n (P@n), recall and F1-score were used to evaluate our proposed algorithm. They are defined as follows:

4.2.1. Precision

Precision at n calculates the relevance of n webpages at the top of the list of ranking results with respect.

Precision = TP / (TP + FP)

4.2.2. Recall

Recall is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data. It is sometimes also referred to as Sensitivity. The formula for it is

Recall = TP / (TP + FN)

4.2.3. F1-Score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account, F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall.

F1 = 2 * (Precision * Recall) / (Precision + Recall) (6)

The performance of an IR system can be evaluated with standard precision and recall metrics. These measurements are used to measure accuracy in the ranking and search of documents, respectively. Precision is a measure of exactness, and it estimates how well it eliminates unwanted documents, whereas Recall is a measure of completeness which measures how well an IR system finds what user wants.

4.3. Experiments and Results

Figure 9 shows comparison on the basis of Accuracy between BM25 and Hybrid algorithm for the top first 10 documents retrieved. Our experiment compares the accuracy at the standard points of retrieved documents. These points are 5, 10 and 100 returned documents. Table 3 shows the percentage of P@x improvement for hybrid algorithm over the BM25. The user is usually browsing the top first 10 documents retrieved. So, it is very important to improve retrieval to achieve more relevant documents at the top of the list to ease the user search task.

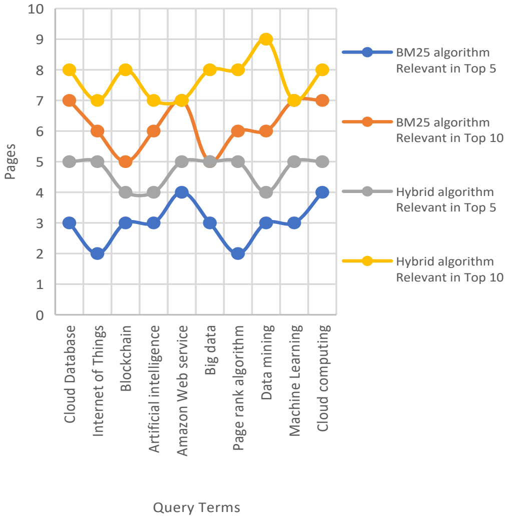

Figure 9: Comparison on The Basis of Accuracy Between Bm25 And Hybrid Algorithm

Table 3, Figure 10. Explain Hybrid algorithm enhancement is ranged from 43 % to 81% in P@5 from 32% to 76% in P@10 and from 21% to 45% in P@100.

Evaluation Measures	BM25 algorithm	Hybrid algorithm
P@5	0.43	0.81
P@10	0.32	0.76
P@100	0.21	0.45
Recall	0.14	0.34
F1- score	0.42	0.65

Table 3: Comparison on Precision, Recall and F1- Score

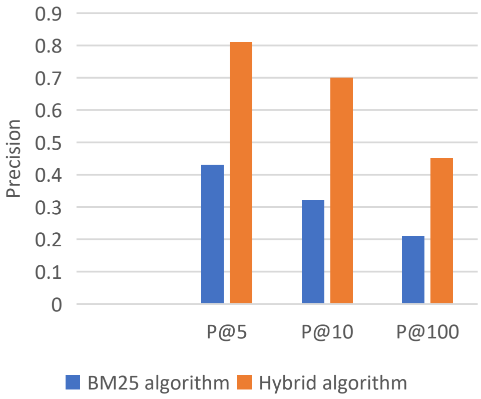

Figure 10: Comparison on the Precision Values P@5, P@10 and P@100

5. Conclusion

Both link and content based algorithms are important to calculate a final score or page rank of a web page, while the researcher always wants to get the best in a short time. In order to rank a lot of web pages accurately and effectively, we proposed a Hybrid Page Rank Algorithm which computes the score on the basis of content as well as link structure of the web pages. A comparison is made between the BM25 as content based structure and our proposed Hybrid page rank algorithm on the basis of accuracy. The results indicated that the our proposed method had better effectiveness in comparison to content-based, connection-based, and hybrid methods such as BM25 with respect to the evaluation criteria of P@n, recall, and F1-score. In the future work we want to use the reinforcement learning methods to achieve more enhancement in effectiveness.

Availability of Data and Material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Manning, P. Raghavan, && H. Schutze. (2009). An introduction to information retrieval. *Cambridge University Press.*
2. Rijsbergen. (1979). Information retrieval. *Butterworths*.
3. Karypis, M. S. G., Kumar, V., & Steinbach, M. (2000, May). A comparison of document clustering techniques. In *TextMining Workshop at KDD2000 (May 2000)* (pp. 428-439).
4. Ridings., && Shishigin, M. (2002). Pagerank Uncovered, Technical report, Aallan borodin, Link Analysis Ranking: Algorithms, Theory,and Experiments, *University of Toronto*.
5. Grover, N., & Wason, R. (2012). Comparative analysis of pagerank and hits algorithms. *International Journal of Engineering Research & Technology (IJERT), 1*(8), 1-15.
6. Selvan, M. P., Chandra Sekar, A., & Priya Dharshini, A. (2012). Ranking Techniques for Social Networking Sites based on Popularity. *Journal of Computer Science and Engineering (IJCSE), 3*(3), 522-526.
7. Maharani, W., & Gozali, A. A. (2014, October). Degree centrality and eigenvector centrality in twitter. In *2014 8th international conference on telecommunication systems services and applications (TSSA)* (pp. 1-5). IEEE.
8. Goldstein, R., & Vitevitch, M. S. (2017). The influence of closeness centrality on lexical processing. *Frontiers in psychology, 8*, 1683.
9. Li, C., Li, Q., Van Mieghem, P., Stanley, H. E., & Wang, H. (2015). Correlation between centrality metrics and their application to the opinion model. *The European Physical Journal B, 88*(3), 65.
10. Sharma, D. K., & Sharma, A. K. (2010). A comparative analysis of web page ranking algorithms. *International Journal on Computer Science and Engineering, 2*(08), 2670-2676.
11. Janßen, A., & Wan, P. (2020). K-means clustering of extremes. *Electronic Journal of Statistics.*
12. Bonacich, P., & Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social networks, 23*(3), 191-201.
13. Roddenberry, T. M., & Segarra, S. (2021). Blind inference of eigenvector centrality rankings. *IEEE Transactions on Signal Processing, 69,* 3935-3946.
14. Zhao, H., Xu, X., Song, Y., Lee, D. L., Chen, Z., & Gao, H. (2019). Ranking users in social networks with motif-based pagerank. *IEEE Transactions on Knowledge and Data Engineering, 33*(5), 2179-2192.
15. Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*(5), 604-632.
16. Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford infolab.