JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

IRS-WEB: A Unified Framework for Information Retrieval over the Web

[1]Shabina, [2]Sonal Chawla

[1]Research Scholar, [2]Professor,
Department of Computer Science & Applications,
Panjab University, Chandigarh, India

Abstract: World Wide Web (WWW) is a huge repository of hyperlinked documents containing useful information. Extraction of web information is quite challenging particularly due to some of its inherent characteristics such as non-interlinked documents that's why it is generally inaccessible to most of the general search engine crawler. Hence, there is a requirement for development of a novel framework for facilitating extraction of relevant web information. There are certain issues and challenge related to information retrieval such as relevant documents returned with minimal number of accesses and also there is big challenge to identify the relevant domain among the large number of domains available on World Wide Web. This research paper focused on the proposed methodology of the unified framework to address the aforesaid issues.

***Index Terms*: Information Retrieval, Indexing, Clustering, Ranking, Relevancy.**

I. INTRODUCTION

In the recent years, the exponential growth of the information technology has led to large amount of information available through the WWW [25]. The searching of WWW for the useful and relevant information has become more challenging as the size of the Web continues to grow. The contents of the deep web are dynamically generated by web server and return to the user during online query. Web crawler is a program that is specialized in downloading web contents. Conventional web crawler can easily index, search and analyze the surface web having interlinked html pages but they have limitations in fetching the data from deep web. To access deep web, a user must request for information from a particular database through a search interface [26].

An Information Retrieval (IR) system retrieves information about a subject from a collection of data objects. This is different from Data Retrieval, which in the context of documents consists mainly of determining which documents of a collection contain the keywords of a user query. IR deals with satisfying specific need of a user. The IR system must somehow 'interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query. This 'interpretation' of document content involves extracting syntactic and semantic information from the document text [1].

II. BACKGROUND WORK

A detailed study of the literature review has been carried out to study the information retrieval and its relevant techniques to retrieve desired information. Shah [1] has reviewed human (manual) indexing and automatic indexing methods, and observed that human indexing is efficient way to index data if the data size is not too big. Automatic indexing based on various hashing techniques i.e. fuzzy-finger printing and locality-sensitive hashing has also discussed. Malki [2] has focused on three indexing techniques i.e. inverted files, suffix trees and signature files. The comparison of indexing techniques has performed on the basis of 3 parameters viz. performance, stability and limitations. The performance of considered indexing algorithms is calculated on the basis of processing time or response time. The factors considered for the calculation of stability parameter is measure of rewards and risks associated with every technique. Boukhari and Omri [3] proposed a hybrid approach for information retrieval of bio-medical documents. This indexing approach is based on Vector Space Model (VSM) and Medical Sub Headings (MESH) terms. Numerous approaches like statistical approaches, semantic methods, free indexing and controlled indexing has been applied during this study. This approach follows the sequence of preparation, extraction of concepts and filtering to execute. Djenouri Y [4] proposed a novel cluster-based approach for information retrieval that extracts useful patterns from the object collection, named Cluster-based Retrieval using Pattern Mining (CRPM). This approach followed by a pre-processing step that finds frequent and high-utility patterns in each cluster of objects. A variety of algorithms k-means, DBSCAN, and Spectral are proposed to split the database (considered as objects) into clusters. Further, clusters are ranked as per the user's request by implementing the proposed techniques: Weighted Terms in Cluster (WTC) and Score Pattern Computing (SPC). In this approach various data sources viz. CACM, TREC, Webdocs, and Wikilinks have used. Choubey V [5] has performed a theoretical literature review in this paper to study and analyze the various document clustering algorithms. During the study, total 95 research papers were identified and out of these 30 papers were selected. A variety of algorithms and modifications to earlier algorithms proposed for document clustering by different

researchers has also presented in the study to find out the future scope of research in the domain of information retrieval. From the literature study, it has concluded that K-Means clustering technique is most appropriate technique of the document clustering. Sheetrit E [6] proposed a learning-to-rank approach that rank clusters of similar passages and transform the cluster ranking to passage ranking. The dataset used during this study is composed of English Wikipedia articles that are transformed into flat representation by removing all XML markups. The dataset of INEX and AQUAINT has taken into consideration for experimental study. Saini A et al. [7] studied Fuzzy Based Approach to Develop Hybrid Ranking Function for Efficient Information Retrieval. Ranking function is used to compute the relevance score of all the documents in document collection against the query in Information Retrieval system. A new fuzzy based approach is proposed and implemented to construct hybrid ranking functions called FHSM1 and FHSM2 in their paper. The performance of proposed approach is evaluated and compared with other widely used ranking functions such as Cosine, Jaccard and Okapi-BM25. Thomas JR et al. [8] proposed a fusion approach for automatic keywords extraction from e-newspaper articles. The comparative analysis of the proposed approach has done with already existing keyword extraction techniques viz. TF-IDF (term frequency-inverse document frequency), TF-AIDF (term frequency-adaptive inverse document frequency), and NFA (a number of false alarm). Author has tested the proposed algorithm against the 10 different e-newspaper articles considered from diverse newspapers. It has been observed that keyword detection algorithm worked efficiently in comparison to other algorithms with respect to precision, recall, and f-measure.

III. Research Findings

From the literature review of information retrieval the fact comes to surface that there are several issues such as undesired information retrieval results, huge amount of data, quality of results, heterogeneous data over the web and duplicity of data that need to be considered while designing an adept IR mechanism. These issues can be resolved by proposing a novel systematic approach and its relevant techniques i.e. Clustering, Indexing and Ranking. The Clustering approach will help to handle high volume of text documents. Indexing technique will assist in finding the desired data with minimal number of accesses. Ranking algorithm will help to attain the performance parameters by retrieving the relevant documents.

IV. Research Objectives

The overall objective of this research work is to retrieve most relevant documents from the web content in a fast and facile manner. The specific objectives of the proposed work will be:

- To propose an indexing algorithm by exploiting the existing algorithms to improve the speed of the search for effective retrieval of documents.
- To propose a clustering algorithm for efficient retrieval of textual data from heterogeneous web resources based upon similarity vector.
- To propose enhanced ranking algorithm based on relevancy analysis of web documents as per user's inputs and historical data for extracting desired documents on the basis of input query term.
- To design a unified framework for information retrieval over Web which encompasses query processing based on indexing, clustering and ranking.

There are many methods and algorithms proposed in literature but each of them suffers from some limitations. The proposed work in this regard is relevant as it combines the features of context, content and relevancy analysis.

V. Methodology

This section presents the system methodology of our proposed information retrieval system that has been followed. Firstly, we extract the content using the web scrapper from Times of India archive and store in the MongoDB database. Then data cleaning has performed to remove the numbers, punctuations and white spaces between the words and convert the text in lower-case and store the output as 'content_lower'. Furthermore, data pre-processing has performed for removal of stop words, tokenization and lemmatization and store the output as 'content_words'. To the pre-processed data, indexing is applied for sorting out and ordering of the data in more efficient way. Further to the indexed documents, clustering is applied to made clusters of each word that is present in an article on basis of synonyms. In the end, ranking of the articles done based on score value assigned to each article on the basis of maximum occurrence of keywords.

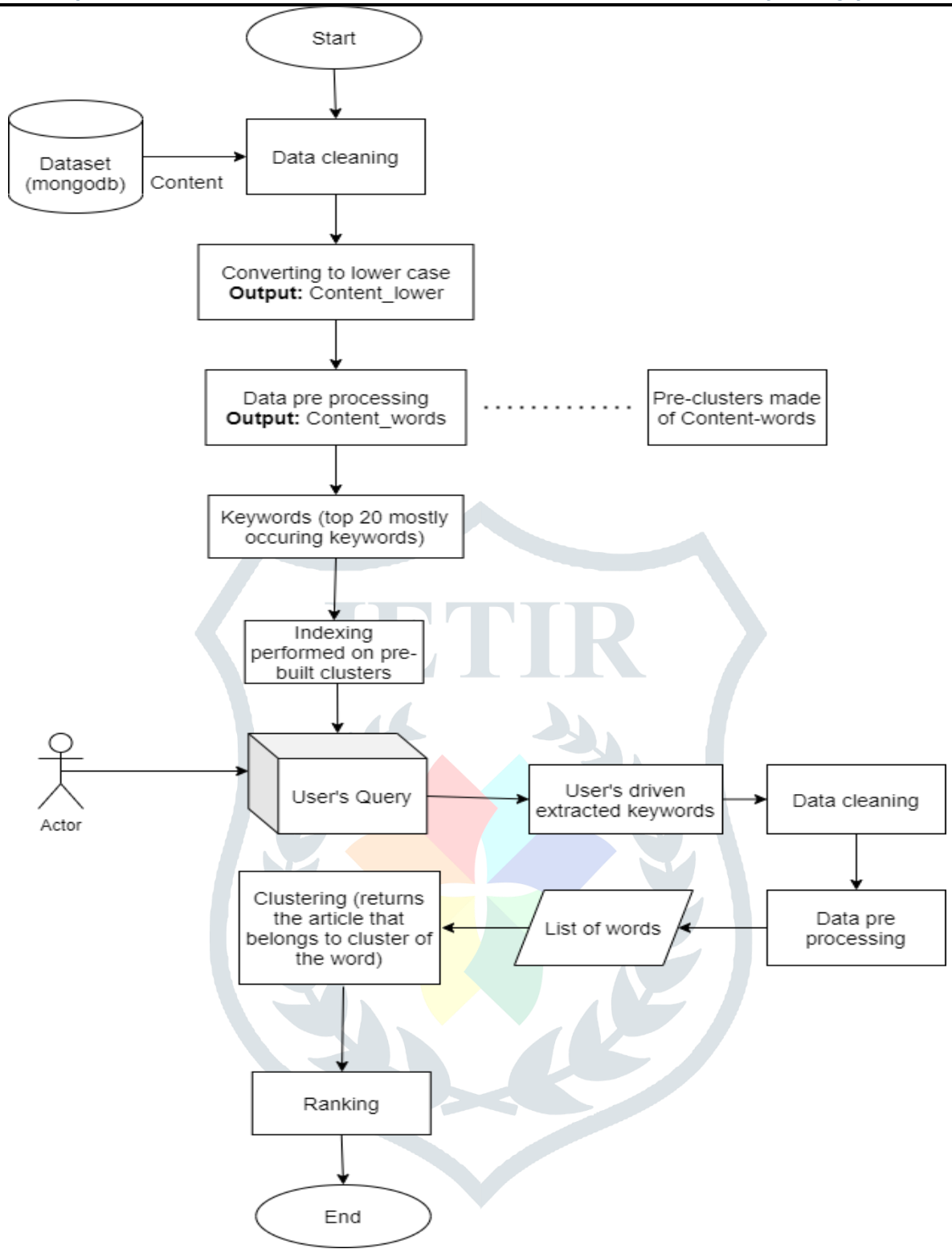

Fig 1: Proposed Methodology of the Retrieval System

5.1. Information Retrieval System

The working of our IRS-web system starts with indexing process. Here, the various keyword based indexes has generated based on the words listed under the user's query. This process is followed by Multi-level Clustering. Here, multi-level array based clusters has been generated up to the level of 4 to deeply understand the user's query based on keyword based indexes [20]. In continuation with this, the last step is to compute the scoring of each document retrieved through indexed based clustering. This scoring is used to grant the rank of each document to sort it into relevant order. The subsequent sub-sections describe the detail of each phase along with their flowchart.

5.1.1. Keyword based Indexing

Indexes are used to reduce the size of search space and to trace the desired data without inefficient and expensive table scans. By applying indexing technique, we need to search only through the ordered index first, rather than searching through each and every document. The index scan works in the form of key-value pair, where the key is the field's value that we have indexed on and key's value is the actual document itself. If we don't use indexing, then database will have to look up every document in the collection. It results in more searching time that affects the query performance. To improve the query processing time, different indexing techniques has been used. Table 1 depicts the comparative analysis of various indexing techniques: B-tree, B+ -tree, Hashing, Bitmap Indexing, and Function based Indexing has done corresponding to different parameters: basic data structure, advantages, limitations, space and time complexity.

Type of Index	**Basic Data Structure**	**Deliverables**	**Complexity**
B-Tree [9]	B-tree	• It performs inefficiently with low cardinality data • It does not support ad hoc queries. • More I/O operations are needed for a wide range of queries. • The indexes cannot be combined before fetching the data.	O(log n)
Hashing [12]	Hash Table	• Hashing is generally better at retrieving records having a specified value of the key. • A hash index organizes the search keys, with their associated pointers, into a hash file structure.	O(n) O(1)
Bitmap Indexing [11]	2-D array	• To use bitmap indexes, records must be ordered sequentially, starting from zero • Designed for easy and efficient querying on multiple keys/ • Useful for querying on multiple attributes	O(n)
B^+tree [10]	B^+tree	• Performance degrades as file size increases • Alternative to indexed-sequential files	O(t logt n) O(log n)

A comparative study of different indexing techniques has done based on some defined parameters that will focus on usefulness and limitations of various techniques. We use B-trees for indexing rather than other indexing strategies, as it helps for supporting range queries and exact lookups that result in flexible and efficient query performance.

- The B-tree index has a hierarchical tree structure where top is header block.
- This block has pointers associated to the appropriate block for any range of given values.
- The branch block will point to the leaf block which is appropriate for more specific range of values, for a larger index which points to another branch block.
- Leaf block contains a list of key values and specific pointers pointing to the location of particular documents on the disk.

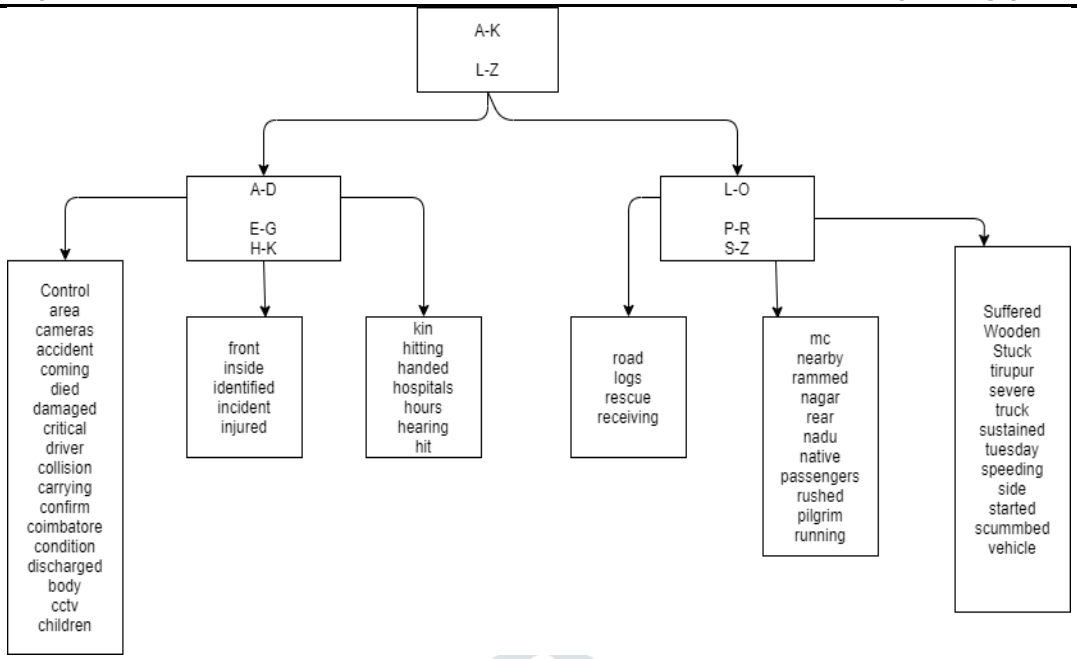

Fig 2: B-tree Index structure

Examining the figure 2 above, Top block is header block. If any word comes, suppose we want to access the record for 'control', first of all we would access the header block. Header block will tell that key values starting from A to K are stored in left most branch block and after accessing this block, key values from A to D are stored in leftmost leaf block. By accessing this leaf block, we will find the record 'control' and its associated disk location that helps to get the concerned document. Leaf blocks maintain the links to both previous and next leaf block that helps us to scan the index in either ascending or descending order.

In B-tree, as each leaf node is at same depth, which makes the performance very predictable. In fact, as the header block will almost loaded in the memory and branch blocks that are usually loaded in memory, the actual number of physical disk reads is usually one or two.

5.1.2. Synonyms based Clustering

Clustering technique is used to partition set of data items in such a manner that similar item will appear together. Various clustering methods have developed for diverse applications and have applied for information retrieval over many decades. Existing clustering methods used for set of data items are primarily based on K-means. Author [13] use keywords clustering followed by the scoring process, while Becker et al. create a feature space based on elements viz. tags and location to learn similarity metric for document clustering.

A variety of clustering methods have been proposed for collection of data items with huge dimensionality, but the most suitable one is K-means to handle large document corpora, while Hierarchical clustering is suitable for smaller collection of documents relevant to a query. The comparative analysis of different clustering algorithms [19]: K-means, BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), STING (Statistical Information Grid Clustering Algorithm), PAM (Partition around Medoids), OPTICS (Ordering Point to Identify Clustering Structure Clustering Algorithm), DBSCAN (Density-based spatial clustering of applications with noise) has done corresponding to the parameters: type of dataset, Scalability and efficiency, shape of cluster, advantages and disadvantages of each technique as discussed in Table 2.

Table 2: Comparison of Clustering Algorithms

Algorithm	Dataset	Scalability and Efficiency	Shape of Cluster	Deliverables
K-means [14]	Numerical data	Scalable in processing large datasets	Works well only with clusters of convex shape	• Easy to implement and interpret • Depends on initial value of k, sensitive to noise and outliers
BIRCH [17]	Numerical data	Shows linear scalability with respect to number of objects	Performs clustering well only with spherical data	• Handles noise easily • Order sensitive
STING [16]	Spatial dataset	Handles noise efficiently		• Arbitrary shaped clusters • Works efficiently with numeric values
OPTICS [15]	Heterogeneous dataset	Produces a visualization of reachability distances	Uses visualization to cluster the data.	• Communication overhead is less • Requires more computational power
DBSCAN [18]	Heterogeneous dataset	Handles noise efficiently	Good at finding clusters of arbitrary shape	• Number of clusters need not to be specified • Cannot handle high dimension data efficiently

In this paper, our major contribution is to examine use of extrinsic techniques that are pertinent to information retrieval: how the relevant documents corresponding to a user's query are distributed across clusters and like-wise the distribution across clusters of the documents retrieved in response to a query by our proposed Information Retrieval System. Clustering has applied as technique to increase the effectiveness of our proposed system, by finding the similarity between queries and documents on the basis of similarity vector. For each query, a cluster is selected according to search strategy, it is a technique used for grouping the unlabeled data. After user enters the query, data cleaning, lower case and data preprocessing is performed. An API call is made to find the synonyms of each word and we store them in a 4 level undirected graph where only unique words are considered as mentioned in figure 3. As an output we get 4 lists of synonyms each of every level and searching for each word in a tree is done and checks if that word/synonym is present in the 'content_words' of any article, that article is returned and implies that article belongs to that specific word cluster and merge the cluster which consist of list of articles.

5.1.3. Score based Ranking

Scoring is a methodology which helps to evaluate the complex type of data related to some criteria. In our project, we did scoring of the articles based on mostly occurring keywords [20] [22] and keywords that are present in an article and present the total score which will help us to identify that in which article that words present in the user's query are mostly occurring and as much as highest score of an article, higher will be the priority of the words that are searched by the user [23] [24]. To implement the scoring see if the word is present in keywords, add 10 to the score and else if it is present in Content_words, add 2 to it, then see if the word is present in first list, divide the score by 2, if in 2nd list divide by 4 and else if in 3rd list then divide by 8. After this, we get out final score based on which articles are ranked [21].

This type of scoring gives more priority to the more relatable and content driven article based on the similarity of our query. Thus making searching process more optimized as ranking the articles on the basis of their score value.

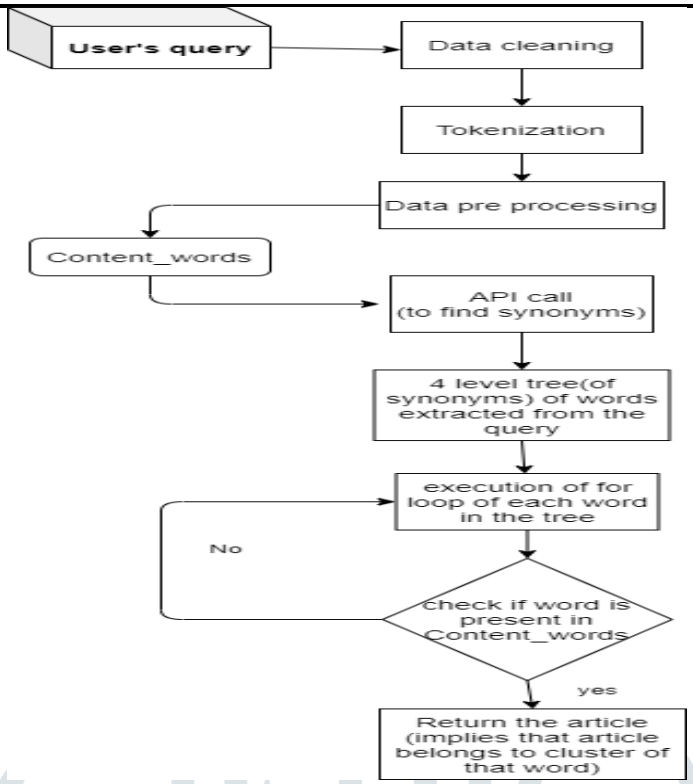

Fig 3: Flowchart of Clustering

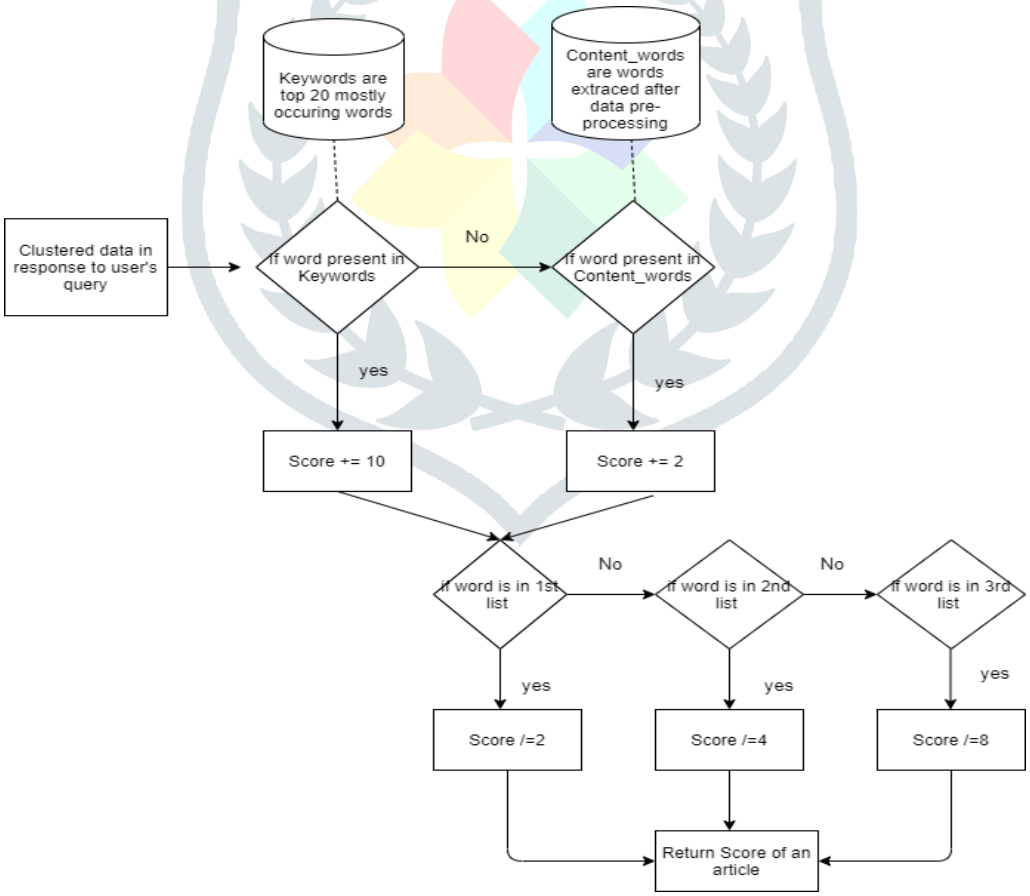

Fig 4: Flowchart of Ranking

VI. CONCLUSION

The proposed unified framework for information retrieval over the web will help in answering diverse research questions by applying various information retrieval algorithms. The evaluation of queries will be effectively pursued by applying varied indexing structures. Clustering helps to extract the information based on similarity vector resulting effective information retrieval and management of available data over the web for further information extraction. Ranking will help to find out the most retrieved results for desired information. In all, the proposed framework will retrieve the relevant query results after passing through unified framework of clustering, indexing and ranking. The information need in the web environment is associated with a given task that is not known in advance and varies from user to user, even if the query specification is same. This research work will help in resolving the challenges related to undesired information retrieval over web i.e. irrelevant information and information overload. The proposed framework will help to attain the desired results by exploiting various algorithms and techniques thereby providing effective methods of content retrieval, indexing, clustering based on similarity vector and ranking of the web contents based on relevancy.

VII. REFERENCES

[1] Shah, N. S. (2016). Review of indexing techniques applied in information retrieval. *Pakistan Journal of Engineering, Technology & Science*, *5*(1).

[2] Malki, Z. (2016). Comprehensive study and comparison of information retrieval indexing techniques. *International Journal of Advanced Computer Science and Applications*, *7*(1).

[3] Boukhari, K., & Omri, M. N. (2017, July). Information retrieval approach based on indexing text documents: Application to biomedical domain. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 2213-2220). IEEE.

[4] Djenouri, Y., Belhadi, A., Djenouri, D., & Lin, J. C. W. (2021). Cluster-based information retrieval using pattern mining. *Applied Intelligence*, *51*(4), 1888-1903.

[5] Choubey, V., & Dubey, S. K. (2020). An Analytical Approach to Document Clustering Techniques. In *ICT Systems and Sustainability* (pp. 35-42). Springer, Singapore.

[6] Sheetrit, E., & Kurland, O. (2019, November). Cluster-based focused retrieval. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2305-2308).

[7] Saini, A., Gupta, Y., & Saxena, A. K. (2015). Fuzzy based approach to develop hybrid ranking function for efficient information retrieval. In *Advances in Intelligent Informatics* (pp. 471-479). Springer, Cham.

[8] Thomas, J. R., Bharti, S. K., & Babu, K. S. (2016, August). Automatic keyword extraction for text summarization in e-newspapers. In *Proceedings of the international conference on informatics and analytics* (pp. 1-8).

[9] Koruga, P., & Baca, M. (2010). Analysis of B-tree data structure and its usage in computer forensics. In *Central European Conference on Information and Intelligent Systems* (p. 423). Faculty of Organization and Informatics Varazdin.

[10] Rosnan, S., Abd Rahman, N., Hatim, S. M., & Ghul, Z. H. (2019, November). Performance evaluation of inverted files, B-Tree and B+ Tree indexing algorithm on Malay text. In *2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)* (pp. 1-6). IEEE.

[11] Goyal, N., Zaveri, S. K., & Sharma, Y. (2006, December). Improved bitmap indexing strategy for data warehouses. In *9th International Conference on Information Technology (ICIT'06)* (pp. 213-216). IEEE.

[12] Knuth, D. (1996). Comparison of indexing techniques. *Term Indexing*, 201-231.

[13] Chen, J., Shankar, S., Kelly, A., Gningue, S., & Rajaravivarma, R. (2009, May). An adaptive bottom up clustering approach for Web news extraction. In *2009 18th Annual Wireless and Optical Communications Conference* (pp. 1-5). IEEE.

[14] Choubey, V., & Dubey, S. K. (2020). An Analytical Approach to Document Clustering Techniques. In *ICT Systems and Sustainability* (pp. 35-42). Springer, Singapore.

[15] Kadhim, A., Abdul, G. H., & Ali, R. S. (2016). Dynamic Clustering for Information Retrieval from Big Data Depending on Compressed Files. *scanning*, *7*(1).

[16] Handa, R., Krishna, C. R., & Aggarwal, N. (2019). Document clustering for efficient and secure information retrieval from cloud. *Concurrency and Computation: Practice and Experience*, *31*(15), e5127.

[17] Sheetrit, E., & Kurland, O. (2019, November). Cluster-based focused retrieval. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2305-2308).

[18] Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 86-97.

[19] Giakoumi, I., Makris, C., & Plegas, Y. (2015). Language Model and Clustering based Information Retrieval. In *WEBIST* (pp. 479-486).

[20] Khan, H. U., Nasir, S., Nasim, K., Shabbir, D., & Mahmood, A. (2021). Twitter trends: A ranking algorithm analysis on real time data. *Expert Systems with Applications*, *164*, 113990.

[21] Dash, A., Zhang, D., & Zhou, L. (2021). Personalized ranking of online reviews based on consumer preferences in product features. *International Journal of Electronic Commerce*, *25*(1), 29-50.

[22] Azarbonyad, H., Dehghani, M., Marx, M., & Kamps, J. (2021). Learning to rank for multi-label text classification: combining different sources of information. *Natural Language Engineering*, *27*(1), 89-111.

[23] Gupta, Y., Saini, A., & Saxena, A. K. (2015). A new fuzzy logic based ranking function for efficient information retrieval system. *Expert Systems with Applications*, *42*(3), 1223-1234.

[24] Saini, A., Gupta, Y., & Saxena, A. K. (2015). Fuzzy based approach to develop hybrid ranking function for efficient information retrieval. In *Advances in Intelligent Informatics* (pp. 471-479). Springer, Cham.

[25] Ceri, S., Bozzon, A., Brambilla, M., Valle, E. D., Fraternali, P., & Quarteroni, S. (2013). The information retrieval process. In *Web Information Retrieval* (pp. 13-26). Springer, Berlin, Heidelberg.

[26] Sharma, M., & Patel, R. (2013). A survey on information retrieval models, techniques and applications. *International Journal of Emerging Technology and Advanced Engineering*, *3*(11), 542-545.