Semantic Conceptual Relational Similarity Based Web Document Clustering for Efficient Information Retrieval Using Semantic Ontology

Selvalakshmi B[1], Subramaniam M[2*], and Sathiyasekar K[3]
[1] Assistant Professor, Dept. of CSE, Tagore Engineering College, Chennai
[e-mail: brslakshmi@gmail.com]
[2] Professor, Dept. of CSE, SRMIST-VDP, Chennai- 600026
[e-mail: subbu.21074@gmail.com]
[3] Professor, Dept. Of ECE, Prathyusha Engineering College, Tiruvallur
[e-mail: ksathiyasekar@gmail.com]
[*]Corresponding author: Subramaniam M

Received December 11, 2020; revised June 12, 2021; accepted August 30, 2021;
published September 30, 2021

Abstract

In the modern rapid growing web era, the scope of web publication is about accessing the web resources. Due to the increased size of web, the search engines face many challenges, in indexing the web pages as well as producing result to the user query. Methodologies discussed in literatures towards clustering web documents suffer in producing higher clustering accuracy. Problem is mitigated using, the proposed scheme, Semantic Conceptual Relational Similarity (SCRS) based clustering algorithm which, considers the relationship of any document in two ways, to measure the similarity. One is with the number of semantic relations of any document class covered by the input document and the second is the number of conceptual relation the input document covers towards any document class. With a given data set Ds, the method estimates the SCRS measure for each document Di towards available class of documents. As a result, a class with maximum SCRS is identified and the document is indexed on the selected class. The SCRS measure is measured according to the semantic relevancy of input document towards each document of any class. Similarly, the input query has been measured for Query Relational Semantic Score (QRSS) towards each class of documents. Based on the value of QRSS measure, the document class is identified, retrieved and ranked based on the QRSS measure to produce final population. In both the way, the semantic measures are estimated based on the concepts available in semantic ontology. The proposed method had risen efficient result in indexing as well as search efficiency also has been improved.

Keywords: Document Clustering, Web Search, Social Networks, Semantic Ontology, Information Retrieval, Semantic Conceptual Relational Similarity, Query Relational Semantic Score.

http://doi.org/10.3837/tiis.2021.09.001 **ISSN : 1976-7277**

1. Introduction

The use of web has been growing every day and the number of web documents is increasing at every movement. The people spend most of their time in the web and they approach the web for everything to learn about. For example, to learn about a topic, the students approach the web to see some informative web pages. As the size of web increases, maintaining or indexing the web becomes more complicated for the popular search engines. The search engines like Google, indexes the web pages based on the terms or key words present in the Meta data. They do not look on the concept present in the webpage.

This is not necessary that the page should contain the detail relate to the Meta data. So this introduces higher irrelevancy in the search result. For example, if you submit a query "Networking," to the Google Search engine, it will produce lot of results. The page links present in the first few pages would contain information related to the query but when you move on to the links after five pages, you can see more irrelevant results. The reason for this is the way of indexing and the search engine does not bother about the topic being discussed in the web page. This increases the requirement of efficient clustering of web pages or tweets or any other document. The clustering of web documents is performed with frequent item sets and the feature selection performed rider moth search scheme towards web document clustering [1]. Earlier the similarity between the documents is measured as Euclidean distance, cosine similarity, Pearson Correlation Coefficient, Jaccard Coefficient towards document clustering [2]. Similarly Komal Maher and Madhuri S. Joshi [3] compared the effectiveness of different similarity measures. The above mentioned similarity measures are used in different document clustering algorithms, still the performance of clustering is questionable and needs attention. The most similarity measures suffer with higher dimensionality in text processing.

On the other side, the impact of social network has become higher in many domains. In the same web search, the tweets of the various users would have been shared and the people would try to search the similar tweets posted by various users. To identify such similar tweets, it is necessary to index them in efficient manner. Similarly, the user would expect different form of results and it is necessary to consider the semantic meaning which is represented by the query submitted. For example, the user would submit a query "Hotel," which means "Hotel, Boarding, and Lodging". These terms are highly related in semantics. So the semantic meanings of the query terms have to be identified and the documents have to be indexed based on their semantic properties. The keyword based search scheme and indexing misses the semantic features. By considering this, an ontology based information retrieval is discussed [4].

The semantic ontology is the concept of maintaining different concepts and relations and terms in one root. By maintaining the related terms or concepts and their properties with relation, it's more supportive to identify the class of document. In this paper such clustering approach has been presented. The most information retrieval algorithms consider only the features of the document but do not consider their semantic relation with the query. It is necessary to measure their semantic relationship between the query and documents of each class. Different similarity measures on information retrieval have been evaluated in [5]. Similarly, query sensitive similarity measure is discussed in [6], to support information retrieval which considers the semantic relationship. However, the proposed scheme takes both conceptual and semantic relations for evaluating the similarity of document. The evaluation of SCRS has been measured based on the level of frequency the concept has been related with the documents of the class. The semantic ontology can be used in measuring the value. Similarly the QRSS is the score being measured based on the relationship between the query terms and semantic class. The detailed approach is discussed below.

2. Literature Review

The web document clustering issue is handled by different methods. Such methods are analyzed in detail in this part.

Vajenti Mala and D. K. Lobiyal [7] discussed an information retrieval scheme which uses keywords and semantic terms present in the query. The method is evaluated with popular search engines like Google and key word based search engines such as Hakia. Their performance is measured on different input submitted and their results are classified according to their relevancy. Similarly, Weiguang Fang et al [8] presented an indexing scheme to support information retrieval according to the ontology set, which index the documents using the semantic information. The method indexes the documents of engineering concepts which combines lucene method with semantic information to extend the indexing performance. Avani Chandurkar and Ajay Bansal [9] presented an information retrieval scheme, which is capable of retrieving the required information from structured knowledge base. The method receives the query from the user and generates answers according to the result of information retrieval. The method combines the key words and applies the natural language processing techniques in producing answers to the user. Sanjib Kumar Sahu et al [10] performed a performance study on different information retrieval schemes which uses semantic information. Due to the rapid increase in size of the web, the process of searching the information becomes more complex. The relevancy among the information should be considered in information retrieval. However quantifying the retrieval is important in case of web search. Thaer SamarMyriam C et al [11] discuss a bias function to quantify the result of information retrieval. Their method indexes different version of documents independently and generates results according to aggregate versions. The content similarity is measured to retrieve the documents and further the documents versions are collapsed according to their URL. The retrievability is measured according to them.

The performance of information retrieval is depending on how much the result is related to the query. According to this, Bashir S and Rauber A [12] introduced a scheme which does not think about the query characteristics but uses the bias function to measure the depth of retrieve ability. Klein M and Nelson M.L [13] performs a study on the performance of various schemes of information retrieval on the basis of individual and combined manner. Traub M.C et al [14] presented an information retrieval scheme to support retrieval from large newspaper corpus and investigate the efficiency of retrieve-ability measure using a large digitized newspaper corpus. Azadeh Mohebi [15] discussed a retrieval scheme for scientific documents using subject features. The method generates a probability model according to the key phrase and retrieval the article from all Iranian papers. The classification is performed using naïve bayes algorithm. Hany M Harb et al [16] approached the web document retrieval with semantic concepts to support jaundice disease document retrieval. The ontology is generated based on the crawled information and semantic concepts measured according to the query. The wordnet dictionary is used in generating semantic ontology. Yanti Idaya Aspura M.K. and Shahrul Azman Mohd Noah [17] presented a semantic text-based image retrieval scheme which uses the high level concept and low level features from the images. The semantic concepts and features also used to support information retrieval to support retrieval of sports documents (SIRSD). Similarly, a co-occurrence base information retrieval scheme is presented. The model combines the result of semantic information retrieval with the co-occurrence analysis [18].

Shengtao Sun [19] discussed an uncertain model using semantic ontology towards spatial data retrieval, which considers ontology consisting incomplete data. According to these features, the semantic relationship quantitative (SRQ) assess with possibility and probability as SRQPP is computed towards data retrieval. Mohamed Marouf Z et al [20] applied the idea of semantic based data retrieval over medical system which works over gene ontology (SOR). The method classifies the genes towards various groups. Similarly, the classification of sports document retrieval is handled with semantic ontology with word net [21]. Kara, Soner, et al [22] discussed the indexing of documents according to the value of semantic measures. Razieh Rahimi [23] presented an axiomatic approach towards cross language information retrieval according to the corpus available. The CLIR approach is targeted in translating the knowledge on document to rank them towards any cross language retrieval. Eilon Sheetrit [24] discussed a passage based approach to rank the documents towards information retrieval. The method ranks the documents according to the query and based on that information retrieval is performed. Haotian Zhang [25] discussed a sentence level feedback in relevancy in document to measure the performance of information retrieval and improve the performance of information retrieval. The method receives the feedback about the relevancy of documents being produced as result and based on that the performance is improved.

The methods analyzed in the above section identified as they endure to bring forth higher performance in clustering which introduces pathetic performance in retrieving the information.

3. Problem Statement

The clustering web document and retrieving information problem is analyzed in detail. Earlier the document clustering and information retrieval is carried out with the Term Frequency / Inverse Document Frequency (TF/IDF) based approach which considered only the terms present in the document and measure their frequency in different documents towards clustering and retrieval. Further to improve the performance of clustering and information retrieval, different methods like K means, SVM, Rule based approaches are presented in several articles. The problem with those approaches is that they considered only the text features of the documents. But the technological growth has allowed the content developer to present content in web documents in several ways like pure text, semantic features, and web links and so on. So, clustering the web documents just based on the text features is not enough and they endure to attain the performance. The existing techniques consider only the topical and conceptual features in clustering and they measure the similarity between the documents based on the texts or terms present in the document. All these challenges the methods in achieving higher performance in web document clustering and information retrieval.

3.1 Contribution made in this Article

Towards improving the performance of web document clustering different approaches are discussed in literature. This article considers different features like semantic, text, conceptual features in document clustering and information retrieval. By considering the semantic features and their relevancy gains the input query, the performance of clustering as well as information retrieval can be improved. Towards this, a Semantic Conceptual Relational Similarity (SCRS) based web document clustering and information retrieval technique is sketched in this article. By adapting the SCRS based approach in web document clustering, the performance of clustering as well as retrieval can be improved. On the other side, towards information retrieval, a QRSS (Query Relational Semantic Score) based technique is presented. The QRSS measure is computed according to the terms of query and the semantic features towards the taxonomy of different document class and their semantic ontology. This supports the improvement of information retrieval.

3.2 SCRS Web Document Clustering and Information Retrieval

The proposed document clustering algorithm utilizes the Open Directory Project (ODP Taxonomy) and WorldNet for the development of semantic ontology. Each class has N number of semantic class and each has different properties and relations. The relations are identified from the synset pointers identified from the WorldNet taxonomy. With the developed semantic ontology, the input data set has been measured for semantic conceptual relational similarity (SCRS) towards each class. Finally a single class has been selected and indexed. The information retrieval is performed by measuring the QRSS value towards each class. Finally based on the QRSS measure a subset of documents has been retrieved as result. The detailed approach is discussed in this section.

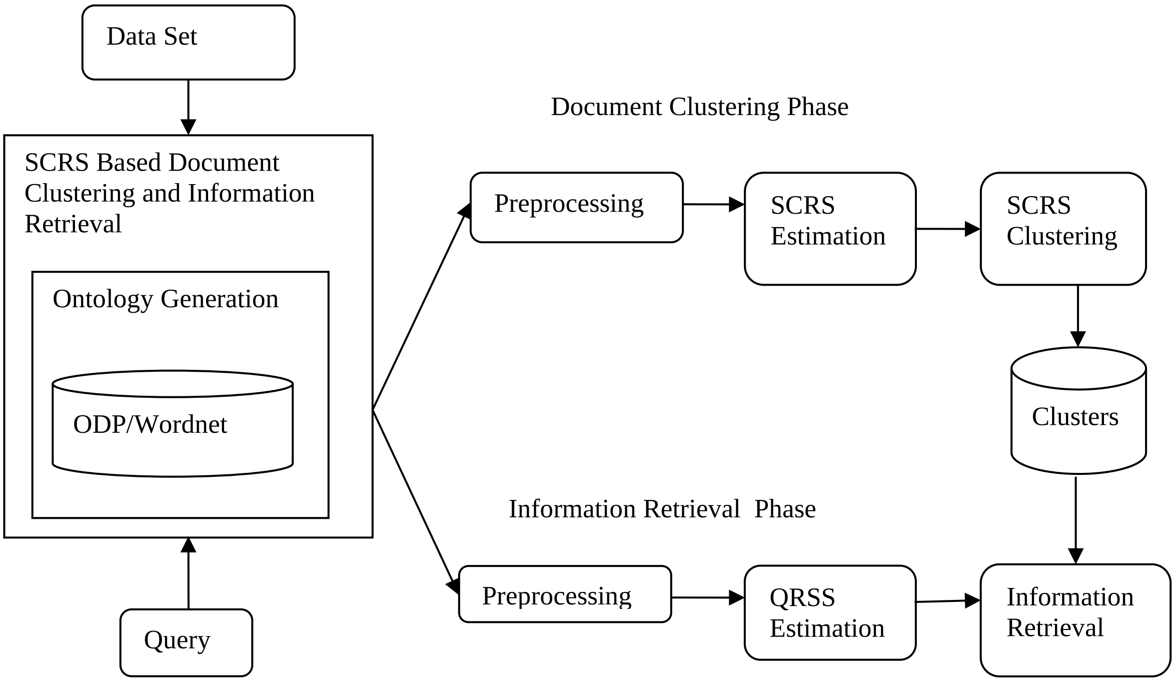

Fig. 1. Architecture of SCRS Clustering and Information Retrieval

The Fig. 1 shows the architecture of SCRS clustering and information retrieval system and shows the components of the system. At the document clustering phase, the documents are preprocessed and the features of the document are extracted. According to the features extracted, the SCRS clustering approach estimates the SCRS measure towards various document clusters. Based on the SCRS value, the method performs SCRS clustering. Similarly, at the information retrieval phase, the query has been preprocessed and the features from the query are extracted to measure the value of QRSS value to identify the class of query and to perform information retrieval.

3.2.1 Preprocessing

The preprocessing algorithm works in two ways according to the input. If the input is a web document then it removes the presentation tags. With the presentation removed text, the list of terms has been generated as Ts. For each term present in the term set Ts, the stopword removal is performed to obtain the key terms. With the key terms, for each key term, the method applies Part of Speech Tagger (POS Tagger) to identify the root words or nouns. Similarly, with the input query or tweet, this performs all the operation other than presentation tag removal. The term set generated by the approach is used in estimating different measures.

Preprocessing Algorithm:

```
Intake: Document/Tweet/Query T, Stop word list Swl
Output: Term Set Ts.
Start
    Read Text T.
    If T.Type==Html then
        T=Remove presentation Tags by applying Html Parser.
    End
    Term set Ts = Split the text T with space character.
    For each term Ti
        If Swl$\in Ti$ Then
            Ts = Ts$\cap Ti$
            Continue;
        Else
            Term Tx = PoS (Ti) //PoS is a Part of Speech tagger provided by
Stanford                                    //university is applied on
each term Ti.
            If Tx.Type!=Noun
                Ts = Ts$\cap Ti$
            End
        End
    End
Stop
```

Table 1. Document Text for preprocessing

“Cloud computing is the paradigm which allows the access of shared resource with least cost. The organizations are not have the capability to afford huge cost to purchase the high end resources. However, by moving the costly resources to the cloud and by providing set of services to access them, the organizations can perform their task. The service requests are scheduled according to different factors like resource utilization, completion ratio, cost and so on. Also, the data security plays vital role in the communication and access of different resources. Different access restriction schemes are available to secure the data present in the cloud as privacy preservation is the most important factor, because the cloud data contains different private information of users. The resource sharing is the most dominant factor in cloud and the access restriction is performed to ensure the privacy of user data.”

The preprocessing algorithm extract the text features and finds the tags, stop words and punctuation marks. Identified noisy features are eliminated from the term set generated and find the root words to produce final term set TS.

The **Table 1** shows the document text considered for preprocessing which has number of terms and the terms marked in red shows the list of stop words identified by the preprocessing algorithm, which is being used for stemming and tagging process. The **Table 2** shows the text content which is eliminated from stop words and the text has been used for tagging. The **Table 3** shows the list of nouns being identified from the stop word removed and stemming content. The identified nouns are used for semantic conceptual relation score measurement.

Table 2. Stop word removed content

“Cloud computing paradigm allows access shared resource least cost. Organizations capability affords huge cost purchase high end resources. Moving costly resources cloud providing services access organizations perform task. Service requests scheduled factors resource utilization, completion ratio, cost data security plays vital role communication access different resources. Different access restriction schemes available secure data present cloud privacy preservation important factor cloud data contains private information users. Resource sharing dominant factor cloud access restriction performed ensures privacy user data.”

Table 3. Stop word removed content

Cloud Computing – Noun
Organization – Noun
Shared Resource – Noun
Service Access – Verb
Scheduling – Noun
Resource Utilization – Noun
Completion Ratio – Noun
Data Security – Noun
Access Restriction – Noun
Privacy Preservation - Noun

3.2.2 SCRS Estimation

The semantic conceptual relational score is the measure which represents the conceptual relation the document has with the documents of the cluster C considered. IF there exist N number of documents in class C, then the similarity of document given towards the set of documents of cluster C has been measured based on the SCRS measure. To measure the SCRS value, the domain ontology has been taken. For each class or concept of the domain ontology (DO), the number of properties gets matched with the term set Ts and the total number of relations the term set has with the concept. Based on these two measures, the SCRS measure has been estimated.

SCRS Estimation Algorithm:

```
Input: Term Set Ts, Domain Ontology DO
Output: SCRS.
Begin
        Read  Ts and   DO.
        For each concept Con
                Compute concept frequency measure CFm.
                CFm = = $\sum_{i=1}^{size(Ts)} Ts(i) \in \forall Properties(Con) \Big/ Size(Con)$
                Compute Total Number of relations of Ts.
            Tnr =
$\sum_{i=1}^{size(Ts)} Ts(i) \in \forall Properties(Con)\ \&\&\ Ts(i).Relations \rightarrow Con \Big/ Size(Con)$
        $\rightarrow$ denotes the relation present in the concept Con
            Compute SCRS = $CFm / Tnr$
        End
Stop
```

The SCRS estimation approach measures the semantic conceptual relational similarity of the term set with the different class or concept available in the domain ontology. Estimated similarity measure is used in clustering. The method estimates the Concept Frequency Measure (CFM) which represents the frequency of concept being discussed in the document and identifies the number of relations covered by the document towards any class. By using both of them, the method computes the value of SCRS.

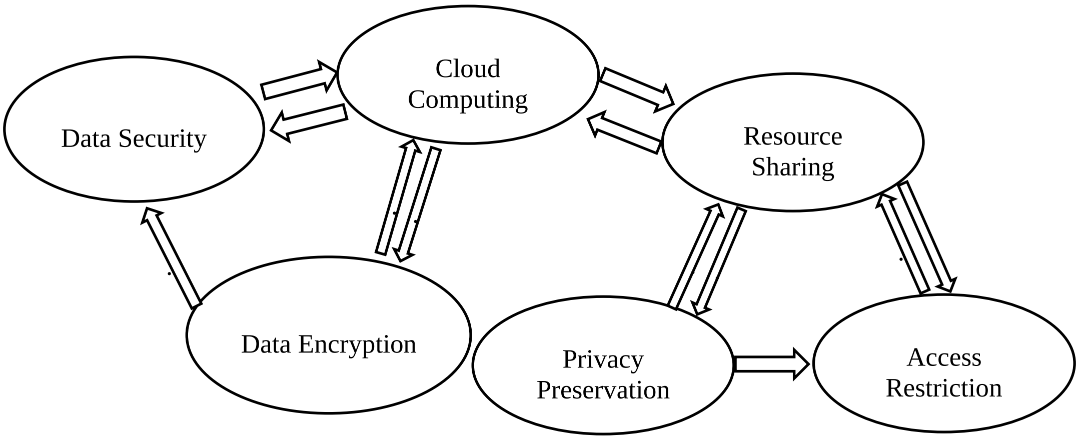

Fig. 2. Semantic Relational Diagram of Data Mining

The Fig. 2 shows the semantic relational diagram of category "Cloud Computing," which shows number of concepts and their relations with others. The arrows just represent there is a relation between two concepts. The size of arrows are not misleading because they have been used to represent there exist a relation between the concepts. The Table 4 shows the total number of terms present in each category of documents considered for retrieval and clustering. Table 5, shows the list of categories considered and the number of documents present in each. Using the documents available, a set of terms has been selected and used for the measurement of semantic conceptual relational measure. Estimated semantic conceptual relational score has been presented in **Table 5**.

Table 4. Terms Count

Category	Total Terms
Cloud Computing	22
Sentiment Analysis	15
Network Security	14
Image Analysis	16

Table 5. Category and Measures

Category	Total Terms Selected	Number of Terms in each category	Number of relations identified	CRS
Cloud Computing	9	22	12	12/22 = 0.54
Sentiment Analysis		15	1	1/15=0.06
Network Security		14	2	2/14=0.142
Image Analysis		16	0	0/16=0.00

3.2.3 SCRS Clustering

The SCRS clustering is performed based on the semantic conceptual relational similarity being estimated between the given document and the set of documents present in each class. For each class, the method estimates the SCRS measure and finally a single class which has higher SCRS value has been selected. The document has been indexed to the class selected.

Algorithm:

```
Input: Document D, Cluster C
Output: Null
Start
        Read document D.
        Term Set Ts = preprocessing (D)
        For each class C
                SCRSc = SCRS-Estimation (Ts, Domain Ontology)
        End
        Class Cs = select the class C with higher SCRS
                Index D to selected class.
Stop
```

The SCRS clustering algorithm compute the value of SCRS towards various class of documents. According to the value of SCRS, a single class has been selected and indexed. According to the values of **Table 5**, the class “Cloud Computing” will be selected and the documents of the class have been retrieved as result.

3.2.4 QRSS Estimation

The query relational semantic score represent the relevancy of the query to the semantic class. It has been measured based on concept linkage the query has with different semantic class. The method measures the number of concept terms present in the query. Second, for each key term from the query, the method identifies the synset terms from the wordnet taxonomy. Based on the synset terms, the method computes synset coverage measure Scm. Using these two, the QRSS measure has been measured.

Algorithm:

```
Input: Query Q, Domain Ontology Do, Wordnet W
Output: QRSS.
Start
        Read Q, Do, W.
        Term Set Ts = Preprocessing (Q)
        Identify synset set SynS.
        For each term Ti
                Syns = $\sum$(synsets $\in$ Syns ) $\cup$ (Synset(Ti) $\in$ W)
        End
        Compute Total Concepts Tc = $\sum_{i=1}^{size(Ts)} Ts(i) \in Do(c)$
Compute Synset Coverage Measure Scm = $size(Syns) \big/ Concepts \in Do(c)$
        Compute QRSS = $\frac{Tc}{TS} \times \frac{Tc}{Scm}$
Stop
```

The QRSS estimation algorithm computes the query relational semantic score for the class being given. The estimated QRSS measure has been used to perform information retrieval. The method computes the total number of concepts present in the query and number of synonyms or synsets of any concept covered by the query. By using both of them, the value of QRSS is measured. According to the value of QRSS a particular document class can be identified to perform information retrieval. Consider the query given is “privacy preservation technique in cloud computing paradigm towards access restriction,” then the list of terms selected for relational score measurement is as follows, which Table 6, shows the list of terms identified for QRSS measurement.

Table 6. Terms identified for QRSS Measurement

Terms Selected for QRSS Measurement
Cloud Computing
Privacy Preservation
Access Restriction

The Table 7 shows the value of SCM and QRSS values measured for different category of documents for the input query. The category “Cloud computing” has least distance in relational semantic similarity and has been selected as the class. According to this, the documents have been retrieved.

Table 7. Estimated Measures

Category	Number of terms identified from Query	Number of Synsets Identified	Total Concepts	SCM	QRSS
Cloud Computing	3	13	22	13/22=0.59	22/(3×0.59)=12.4
Sentiment Analysis			15	13/15=0.86	15/(3×0.86)=5.81
Network Security			14	13/14=0.92	14/(3×0.92)=5.07
Image Analysis			16	13/16=0.81	16/(3×0.81)=6.58

3.2.5 Information Retrieval

In this stage, the method receives the query and performs preprocessing to identify the query term set. Second, for each class of semantic ontology, the method computes QRSS measure. According to QRSS value, the method selects a specific class. Then for each document of the class, the method computes conceptual strength measure CSM. The documents are ranked based on CSM and top documents have been returned as result.

```
Algorithm:
Input: Input_Query Q, Domain_Ontology Do
Output: Information_Result R
Begin
    Read Q, Do
    Set of Terms Ts = Preprocessing (Q)
    For each class C of Do
        QRSS = Estimate-QRSS (Ts, C)
    End
    Class c = Choose the class with higher QRSS.
    For each document Di of C
        Compute conceptual Strength Measure CSM.
        CSM = $\sum Terms(Do(c)) \in Ts \big/ size(Ts)$
    End
    Rank Documents Based on CSM.
    R=Choose Top ranked document.
Stop
```

Afore discussed algorithm Fig.s QRSS measures for the different class and selects a single one. Again for each document of the selected class, the method computes CSM measure to rank and populate the result.

4. Results and Discussion

The proposed SCRS measure based clustering approach is implemented and its performance is evaluated under different constraints. The proposed algorithm has produced efficient results on clustering and information retrieval. The documents are classified into number of classes and for each class domain ontology has been generated.

Table 8. Experimental Details

Parameter	**Value**
Data Set	Clue-Web 12-B13
Number of Classes	10
Number of Documents	503,903,810 pages
Tool used	Advanced Java

Table 9. Relevancy Score Analysis

Test Case	**Case**	**Size of Web Set**	**Size of Tweet Set**	**Relevancy Score (%)**	
				Web Docs.	**Tweets**
1	Group1	1500	1500	97.684	99.31
2	Group2	3000	3000	98.58	99.51
3	Group3	5000	5000	99.44	99.60
4	Group4	7000	7000	98.32	99.30
5	Group5	10000	10000	98.55	98.50

The experimental data and features considered towards performance measurement of proposed approach are listed in **Table 8**. The clue-web data set is used for evaluation which contains 10 language documents and English documents are considered. It contains 503,903,810 pages with 10 different categories. The evaluation is performed with advanced java and the performance is measured on different parameters.

The clustering performance has been measured with varying number of groups and documents. The relevancy of documents of each cluster has been measured and compared. The **Table 9** present the relevancy of clustering being generated by the SCRS scheme at varying size of documents in each class. The SCRS scheme has produced higher relevancy in all the conditions. The results are compared with semantic relationship quantitative (SRQ) assess with possibility and probability as SRQPP [19], Semantic Information Retrieval on Sports Document (SIRSD) [18], and Semantics based Ontology Retrieval (SOR) [20]

Table 10. Comparative analysis on Time Complexity (web documents)

Size of record set	SRQPP		SOR		SIRSD		SCRS	
	Entire Features	Selective Features	Entire Features	Selective Features	Entire Features	Selective Features	Entire Features	Selective Features
1500	0.28	0.26	0.25	0.23	0.20	0.19	0.19	0.17
3000	0.37	0.33	0.32	0.26	0.24	0.21	0.22	0.19
5000	0.48	0.35	0.35	0.29	0.28	0.25	0.23	0.20
7000	0.54	0.41	0.42	0.33	0.32	0.28	0.26	0.24
10000	0.61	0.43	0.46	0.37	0.36	0.28	0.21	0.19

The Table 10, details the analysis result of time complexity introduced by the proposed method towards varying number of documents in the class considered. However, the proposed SCRS scheme achieved lesser time in all the cases considered. Here the entire features represent the consideration of all the features present in the document where the selective features represent the analysis with subset of features selected from the document.

Table 11. Comparative Performance analysis (Tweets)

No. of records considered	SRQPP		SOR		SIRSD		SCRS	
	Entire Features	Selective Features	Entire Features	Selective Features	Entire Features	Selective Features	Entire Features	Selective Features
1500	89.38	91.21	90.13	92.24	92.44	99.22	99. 26	99.28
3000	89.52	91.16	90.16	92.39	92.38	99.33	99. 40	99.41
5000	89.57	91.32	90.13	92.36	92.37	99.41	99.41	99.42
7000	89.61	91.34	90.17	92.41	92.41	99.42	99.44	99.46
10000	89.63	91.32	90.19	92.46	92.46	99.48	99.48	99.49

Table 11 and Table 12; present the comparative analysis of clustering performance obtained by various methods by varying the number of tweets and documents. For all the cases, the proposed method has resulted higher efficiency compare to other methods. Relevance states how well a retrieved document or collection of documents meets the information requirement of the user. Semantic information retrieval technique contains the benefits of the semantic web for retrieving the appropriate data.

Table 12. Comparative Performance analysis

No. of records considered	SRQPP		SOR		SIRSD		SCRS	
	Selective Features	Entire Features	Selective Features	Entire Features	Selective Features	Entire Features	Selective Features	Entire Features
1500	91.25	91.27	92.33	92.37	99.14	99.15	99.17	99.19
3000	91.27	91.29	92.38	92.40	99.16	99.17	99.18	99.22
5000	91.30	91.32	92.38	92.42	99.19	99.20	99.27	99.28
7000	91.32	91.36	92.41	92.43	99.22	99.23	99.31	99.33
10000	91.35	91.37	92.43	92.45	99.25	99.26	99.36	99.38

4.1 Relevancy Analysis

The performance of different schemes on relevancy has been measured in various test cases and compared with other schemes. The analysis is conducted by varying the number of documents in each class. In all the cases, the SCRS approach has achieved higher relevancy than other schemes. The result of analysis is presented in **Fig. 3**, where the proposed SCRS model has produced higher relevancy in all the cases than other schemes.

The performance of various methods on relevancy has been measured in different test cases and compared with other methods. In all the case, the SCRS approach has achieved higher relevancy than other methods. The performance on relevancy based on different tweets is measured and plotted in **Fig. 4**. The SCRS model has produced higher relevancy in all the cases than other methods.

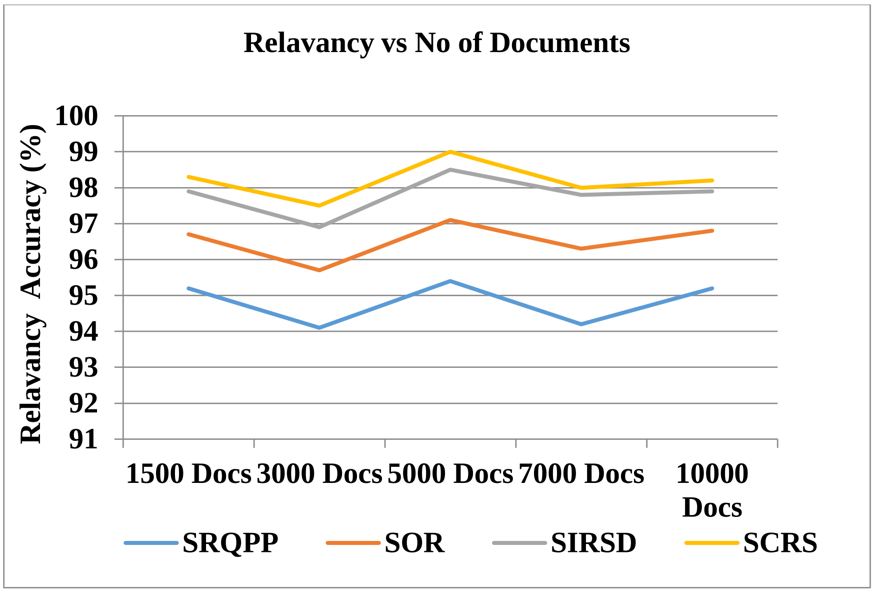

Fig. 3. Comparative on Relevancy vs. No of Documents

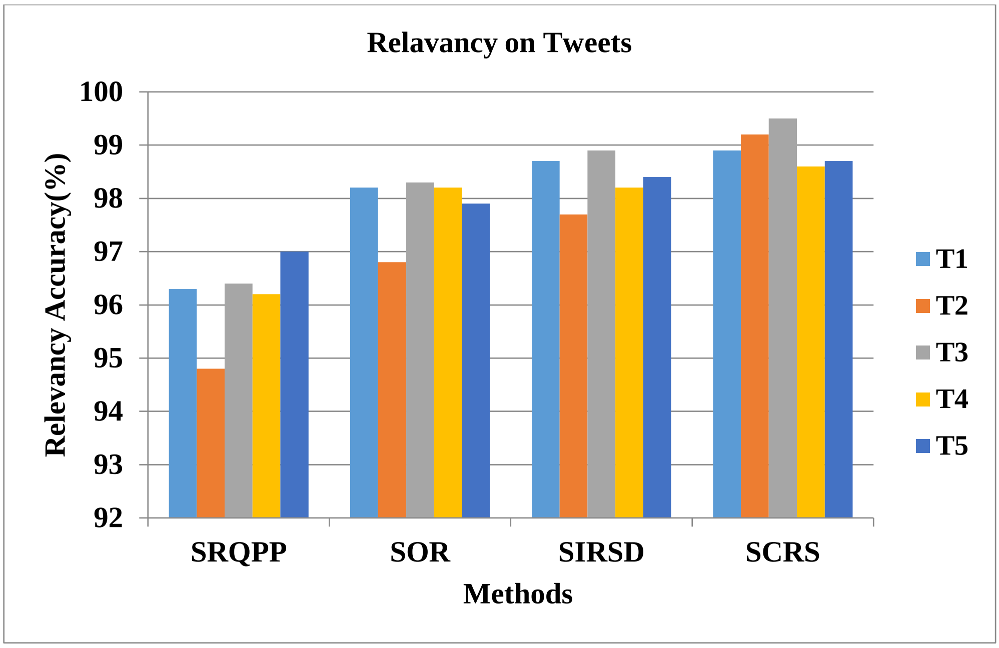

Fig. 4. Relevancy Score Analysis (Tweets)

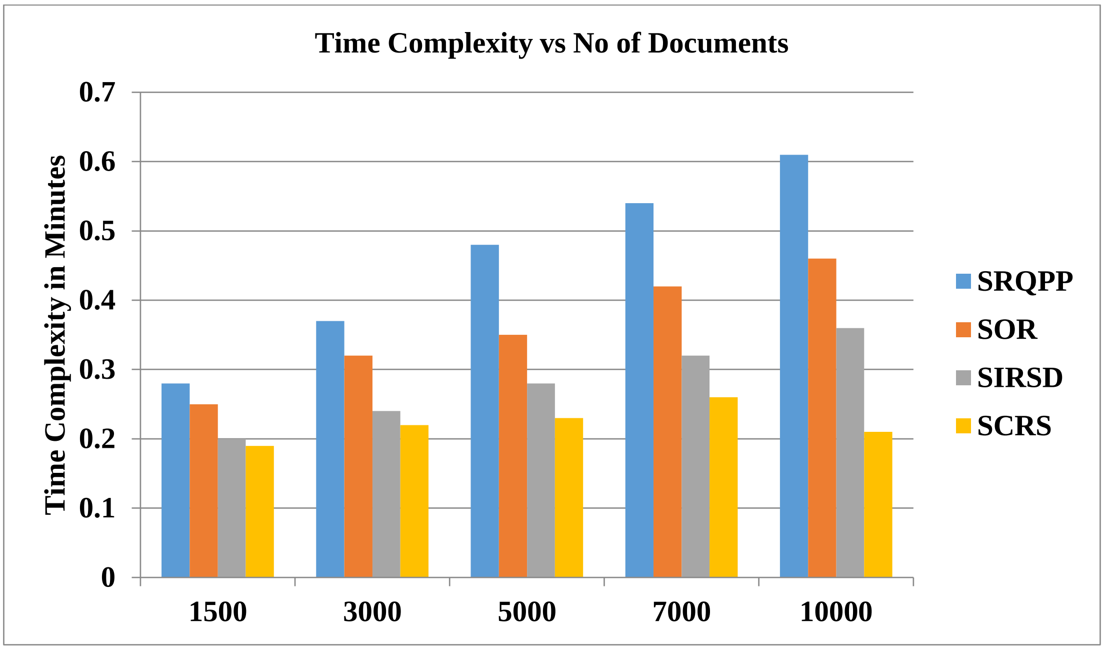

Fig. 5. Performance on Time Complexity

The **Fig. 5**, details the time complexity achieved by various approaches according to number of documents. The SCRS method has resulted in lesser time in all other number of documents than other methods.

5. Conclusion

This paper presented an efficient semantic concept relational similarity based document clustering and information retrieval scheme. The method first develops the semantic ontology based on the open directory project taxonomy and word net. Then with the input document, the term sets are identified using preprocessing technique and for each class an SCRS measure is computed. Based on SCRS measure, a single class has been selected. Similarly when the query has been received, the method estimates QRSS measure to identify the class of query. According to the class identified, concept coverage measure is computed. The documents are ranked based on CCM measure to populate the result. The proposed method has considerably improved the performance of clustering and information retrieval than other methods. The performance in information retrieval is achieved up to 99.3% where the time complexity is reduced up to 170 milli seconds. The same on tweet is achieved up to 99.49% of relevancy and time complexity is reduced up to 14 milli seconds.

References

[1] MadhulikaYarlagadda, K Gangadhara Rao, A Srikrishna, "Frequent itemset-based feature selection and Rider Moth Search Algorithm for document clustering," *Journal of King Saud University-Computer and Information Sciences*, Sep.2019. Article (CrossRef Link).

[2] Nadella Sandhya, A. Govardhan, "Analysis of Similarity Measures with WordNet Based Text Document Clustering," in *Proc. of International Conference on Information Systems Design and Intelligent Applications*, pp. 703-714, 2012. Article (CrossRef Link)

[3] Komal Maher, Madhuri S. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering," *International Journal of Computer Science and Information Technologies*, Vol. 7 (4), pp. 1715-1720, 2016. Article (CrossRef Link)

[4] Binbin Yu, "Research on information retrieval model based on ontology," *EURASIP journal of wireless communication and networking*, vol. 2019, pp. 1-8, Feb. 2019. Article (CrossRef Link)

[5] K. Pradeep Reddy, T. Raghunadha Reddy, G. Apparao Naidu, B, vishnu Vardhan, "Impact of Similarity Measures in Information Retrieval," *International Journal of Computational Engineering Research(IJCER)*, vol. 8, no. 6, pp. 54-59, 2018. Article (CrossRef Link).

[6] Anastasios Tombros, C.J, van Rijsbergen, "Query-Sensitive Similarity Measures for Information Retrieval," *Know. Inf. Sys*, vol. 6, no. 5, pp. 617-642, 2014. Article (CrossRef Link).

[7] Vajenti Mala, D. K. Lobiyal, "Semantic and keyword based web techniques in information retrieval," in *Proc. of ICCCA*, 2016. Article (CrossRef Link)

[8] Weiguang Fang, Yu Guo, Wenhe Liao, "Ontology-based indexing method for engineering documents retrieval," in *Proc. of ICKEA*, 2016. Article (CrossRef Link)

[9] Avani Chandurkar, Ajay Bansal, "Information Retrieval from a Structured Knowledge Base," in *Proc. of ICSC*, 2017. Article (CrossRef Link)

[10] Sanjib Kumar Sahu, D. P. Mahapatra, R. C. Balabantaray, "Analytical study on intelligent information retrieval system using semantic network," in *Proc. of ICCCA*, 2016. Article (CrossRef Link)

[11] Thaer Samar, Myriam C. Traub, Jacco van Ossenbruggen, Lynda Hardman, Arjen P. de Vries, "Quantifying retrieval bias in Web archive search," *Int. J Digit Library*, vol. 19, no. 1, pp. 57–75, 2018. Article (CrossRef Link)

[12] Shariq Bashir, Andreas Rauber, "On the relationship between query characteristics and IR functions retrieval bias," *J. Am. Soc. Inf. Sci. Technology*, vol. 62, no. 8, pp. 1515–1532, August 2011. Article (CrossRef Link)

[13] Martin Klein, Michael L. Nelson "Moved but not gone: an evaluation of real-time methods for discovering replacement web pages," *Int. J. Digit. Library*, vol. 14, pp. 17–38, Feb. 2014. Article (CrossRef Link)

[14] Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries, Lynda Hardman, "Query log-based assessment of retrieve-ability bias in a large newspaper corpus," in *Proc. of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pp. 7–16, 2016. Article (CrossRef Link)

[15] Azadeh Mohebi, Mehri Sedighi, Zahra Zargaran, "Subject-based retrieval of scientific documents, case study: Retrieval of Information Technology scientific articles," *Library Review*, vol. 66, no. 6, pp. 549-569, 2017. Article (CrossRef Link).

[16] Hany M Harb, Khaled M. Fouad and Nagdy M. Nagdy, "Semantic Retrieval Approach for Web Documents," *IJACSA*, vol. 2, no. 9, 2011. Article (CrossRef Link)

[17] Yanti Idaya Aspura M.K., Yanti Idaya Aspura M.K., "Semantic text-based image retrieval with multi-modality ontology and DBpedia," *The Electronic Library*, vol. 35, no. 6, pp. 1191-1214, Nov. 2017. Article (CrossRef Link)

[18] Wen Lou, Junping Qiu, "Semantic information retrieval research based on co-occurrence analysis," *Online Information Review*, vol. 38, no. 1, pp. 4-23, Jan. 2014. Article (CrossRef Link)

[19] Shengtao Sun, Lizhe Wang, Rajiv Ranjan & Aizhi Wu, "Semantic analysis and retrieval of spatial data based on the uncertain ontology model in Digital Earth," *International Journal of Digital Earth*, vol. 8, no. 1, pp. 3-16, June 2014. Article (CrossRef Link)

[20] Mohamed Marouf Z. Oshaiba, Enas M. F. El Houby, and Akram Salah, "Semantic Annotation for Biological Information Retrieval System," *Advances in Bioinformatics*, Hindawi, vol. 2015, Feb. 2015. Article (CrossRef Link)

[21] M. Uma Devi and G. Meera Gandhi, "Wordnet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports Domain," *Journal of Computer Science*, vol. 11, no. 2, pp. 361-371, Feb. 2015. Article (CrossRef Link)

[22] S. Kara, Ö. Alan, O. Sabuncu, S. Akpinar, N. K. Çiçekli and F. N. Alpaslan, "An ontology-based retrieval system using semantic indexing," *Information Systems*, vol. 37, no. 4, 294-305, 2012. Article (CrossRef Link)

[23] Rahimi R., Montazeralghaem A., Shakery A., "An axiomatic approach to corpus-based cross-language information retrieval," *Inf. Retrieval J.*, vol. 23, pp. 191–215, Apr. 2020. Article (CrossRef Link)

[24] Eilon Sheetrit, Anna Shtok, Oren Kurland, "A passage based approach to learning to rank documents," *Inf. Retrieval J.*, vol. 23, pp. 153-186, March 2020. Article (CrossRef Link)

[25] Haotian Zhang, Gordon V. Cormack, Maura R. Grossman & Mark D. Smucker, "Evaluating sentence level reference feedback for high recall information retrieval," *Inf. Retrieval J.*, vol. 23, pp. 1-26, Aug. 2020. Article (CrossRef Link)

B. Selvalakshmi is an Assistant Professor in Tagore Engineering College, Chennai. She received her B.E. Degree in Computer Science and Engineering from Madras University in 1998, M.B.A. Degree from Periyar University, Salem in 2001 and M.E. Degree in Computer Science and Engineering from Anna University, Chennai in 2013. She once worked as Sr. Lecturer in Vinayaga Mission Kirupananda Variyar Engineering College, Salem during 2001 to 2006 and System Analyst in L3 Info Solution during 2007 to 2010. She Joined Tagore Engineering College in 2013. Her research interest includes big data, cloud computing and Networking. She has published around 5 academic papers.

M. Subramaniam (1974) is a Professor, DCSE in SRM Institute of Science and Technology- Vadapalani (Campus), Chennai, (INDIA). He obtained his Bachelor's degree (B.E) in Computer Science and Engineering from University of Madras (1998), Master degree (M.E) in Software Engineering and Ph.D from College of Engineering Guindy (CEG), Anna University Main Campus, Chennai -25 in the year 2003 and 2013 respectively. His research focuses are Computer & Mobile Networks, Cloud, Big-data and Software Engineering, AI & ML. He is an active life member of the Computer Society of India (CSI) and the Indian Society for Technical Education (ISTE). He has eight Research scholars perusing Ph.D. under his guidance. He published many research papers in reputed journals. He is also reviewer in IEEE- International Journal of Communication Systems.

Sathiyasekar K was born in Erode, TamilNadu, India in 1971. He received the B.E. degree in Electrical and Electronics Engineering from Madras University in 1999 and M.Tech. Degree in High Voltage Engineering from the SASTRA University in 2002. He received Ph.D. in High Voltage Engineering from Anna University, Chennai in 2010. He is currently a Professor and Head in the department of Electrical and Electronics Engineering, Prathyusha Engineering College, Chennai, India. He has been an Expert Member for Ph.D, viva-voce examination (University Nominee) and received fund from Ministry of MSME, Government of India for Business Incubation Cell. He had been Centre Head for SPARK Automation Centre in SA Engineering College. He received an award of 'Certificate of Outstanding Contribution in Reviewing' from International Journal of Electrical Power and Energy Systems, Elsevier, Amsterdam, The Netherlands. He is reviewer for reputed journals like IEEE, Elsevier, Technical Gazette and Australian journals etc. He is editorial Board Member in International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering. He was awarded Travel Grant by the Department and Science and Technology, Government of India, to present my research paper in the International Conference INDUCTICA-2010, at Messe Berlin, Germany in 2010 and Best Paper Award in the International Conference on Digital Factory – 2008, held at CIT, Coimbatore.