State of the art document clustering algorithms based on semantic similarity

Niyaz Mohammed Salih [a], Karwan Jacksi [b]

[a] *Information System Engineering Department, Erbil Polytechnic University, 44001 Erbil, Iraq*
[b] *Computer Science Department, University of Zakho, 42002 Zakho, Iraq*

ABSTRACT

The constant success of the Internet made the number of text documents in electronic forms increases hugely. The techniques to group these documents into meaningful clusters are becoming critical missions. The traditional clustering method was based on statistical features, and the clustering was done using a syntactic notion rather than semantically. However, these techniques resulted in un-similar data gathered in the same group due to polysemy and synonymy problems. The important solution to this issue is to document clustering based on semantic similarity, in which the documents are grouped according to the meaning and not keywords. In this research, eighty papers that use semantic similarity in different fields have been reviewed; forty of them that are using semantic similarity based on document clustering in seven recent years have been selected for a deep study, published between the years 2014 to 2020. A comprehensive literature review for all the selected papers is stated. Detailed research and comparison regarding their clustering algorithms, utilized tools, and methods of evaluation are given. This helps in the implementation and evaluation of the clustering of documents. The exposed research is used in the same direction when preparing the proposed research. Finally, an intensive discussion comparing the works is presented, and the result of our research is shown in figures.

Keywords:
Clustering Documents
Semantic Similarity
Clustering Algorithms
Traditional Method

I. Introduction

Nowadays, the search engine is a very useful and fast method to solve any query [1]. The quickest learning method is the Internet by understanding and solving the problems or acquiring information from a global knowledge base [2], [3]. However, sometimes when looking for some queries, we also get a lot of unrelated information about our query, with less related information [4], [5]. Therefore, document clustering is being used in all search engines to show the queries' results in an ordered and efficient way [6], [7]. Clustering documents aim to assist humans in searching and understanding information [8]. The clustering of documents is an unsupervised technique that gathers related documents in the same category. Clustering related documents into a more relevant group to the document in a group than a document belonged to some other group [9]. This technique is called the clustering of documents. However, most of the traditional algorithms of clustering are based primarily on Back-of-Words (BOW); the semantic similarity between clusters and documents is incomplete [10]. Clustering documents handle the non-structured document, which poses several problems. Text documents are usually comprised of complex ideas that are hard to describe by using traditional text mining techniques. Because of this lack, traditional document clustering algorithms cannot describe semantic relations into a less output quality between words and sentences [6]. Clustering could be used to group the search results into useful groups, making it easier for web users to scan a few structured groups than several individual documents [11]. There are many ways to resolve this problem resulting in using the traditional approach. The use of Ontology involves different ways to solve the problem. Ontology can be used as context information that can help in detecting the meanings relevant to the words that appear in documents [12]–[14]. This article's remainder is arranged as follows: a comprehensive literature review is summarized in section II. Section III is the methodology. Results and Discussions of the reviewed papers are presented in section IV. Conclusions and suggestions are presented in section V.

II. Literature Review

This section presents a tabulated exhibition of all related research of semantic similarity based on document clustering. In their research, Patil and Thakur determine every document's participation score for every particular cluster; semantic relation between the text documents and the cluster is found. Many authors are focusing on semantic features of clustering the document to re-develop output in clusters. For this reason, the authors are using many external information bases, such as WordNet, Wikipedia, Lucene, etc. The proposed method in their research provides the use of WordNet to enhance the functionality of cluster memberships. The research result reveals that clustering efficiency has significantly improved with the use of the proposed semantic method system [6].

Gao et al., suggested a modern method to measure the WordNet-based similarity semantically among a pair of words. The approach is sub-categorized into three strategies, including to the various methods of the shortest path length weighting: method 1 measures all the edges along the shortest path among the words related; method 2 uses word width to measure the shorter path length and; method 3 uses word Information Content (IC) to measure the shorter path length. The outputs of the experiment display that when $1 \geq \alpha \geq 0.25$, shortening the shortest path length (method 2) among pairs of words before nonlinear translation increases the precision of the estimation of semantic similarity, and when $0.25 \geq \alpha \geq 0.05$, expanding the length (method 3) obtains a higher success rate. The system suggested has a good correlation with standardized benchmarks. This received the highest success rate of 0.883 with the test collection given by Miller and Charles, and 0.885 with respect to the experiment by Rubenstein and Good enough. In addition, the system is less influenced by the parameters because for most parameter settings of α and β one of the strategies still has adequately high success rate values with human ratings. Simplified ways to calculate the shortest path length eventually make the process fast and simple to calculate. From the suggested way measurements, it may be established that the system would obtain more reliable results if the background analysis is considered before the similarity test [8].

Wang and Koopman performed several cluster methods on the basis of the relationship between documents. However, the semantic relationship applied to cluster the papers. The method provided in this paper is to construct during the initial step, the semantics of papers basis on the involvement of the objects in those papers, and three vectors of representations for each paper. One measures the average vector for all entities, one with all objects but without quotations, and another only with quotation objects. Another step is to recognize vector-based clusters of papers by using clustering (k-mean) and group detection by applying the Louvain (methods of clustering Network-based) once the paper vectors are generated [11].

Zafar et al., proposed a Bisecting K-Means strategy, indicating that their variance results better in the clustering of documents than traditional K-Means. Firstly, the algorithm of traditional K-means is used to generate two initial dataset sub-clusters. The sub-cluster that displays the highest similarity is chosen as the dataset for inputs. The above process is correctly repeated until it exceeds the desired number of clusters [15].

Sumathy and Chidambaram proposed a hybrid approach for determining semantic similarity between documents. Semantic similarity plays an important role in text processing and information retrieval. This work gave a summary of semantic similarities and their current strategies. The semantic measure of similarity measures the similarity between words, sentences, and documents. The proposed methods are separated into two folds: The proposed method uses the ontology-based similarity model in a first fold and counts the dependent similarity approach in order to determine the similarity of documents. The proposed method uses ontology and corpus in a second fold to evaluate the similarity of the document. The proposed method achieves high accuracy in the similarity to the document evaluation due to the hybrid approach [16].

Wei et al., introduced a new approach for clustering documents using WordNet and the lexical chain. In their research, ontology hierarchical structure is used to recognize the measure of similarity and to retrieve semantic characteristics of text documents [17].

Zandieh and Shakibapoor proposed an algorithm to cluster text documents automatically and their clustering efficiency is higher than traditional hierarchical clustering algorithms. Therefore, its efficiency and accuracy of clustering are improved by using the proposed Semantic TF-IDF matrix since the meanings and definitions of the terms are taken into consideration. Since clustering is done based on concept and meaning, problems like marking nodes, consistency, the unpredictability of quality and results, and the irrationality of clusters in the hierarchy in traditional clustering methods are solved [18].

Melasagare and Thombre suggested a method that is useful for summing up and classifying the data collection. The system offers a fine description of the data set in the form of clustering and retrieval of text data. This method produces flat clusters, preprocessing (text feature extraction), and representation of data points (normalization, filtering, tokenization, etc.). This method achieves a limitation in noisy data which helps to classify data sets more accurately as flat clusters. The use of hierarchical clustering in this process provides high precision and data classification [19].

Bafna et al., suggested merging K-means with hierarchical algorithms. It initially begins from the small dataset and progresses to an expanded one when building different clusters. In fact, the entire classification process requires 10,000 papers. The result shows that the combined algorithm is capable of classifying two positive-negative classes and three positive-neutral groups [20].

Blokh and Alexandrov presented a strategy to news data clustering using evaluation of ontology-based similarity preprocessing on specific Facebook news data of mass media. This method combination allowed us to obtain the dissemination of news clusters over time. As a data source, they took official accounts from Facebook media news and gathered 415 000 messages from Jan. 2014 to May 2017. Findings display that messages may be clustered into thematic clusters where each cluster represents a theme. They have written the theme distribution by the sum of messages of news. Furthermore, they might determine how strongly selected mass media examined a subject described by a cluster during the period observed [21].

Afreen and Srinivasu provided a clustering approach using the dis-ambiguous definitions and lexical chains. For word sense disambiguation, a modified term-based semantic similarity measure is proposed, and lexical chains are used to extract core semantic features that express the topic of documents, determining the number of clusters, and assigning appropriate explanations for the clusters produced. More significantly, with a reduced number of features in the document clustering phase, they show that the lexical chain features (core semantics) can dramatically improve the quality. Although lexical chains widely have been used in many fields of application, this study is one of the few that attempts to investigate the potential impact of lexical chains on text clustering [22].

Awajan performed a K-means bisecting via the MapReduce method. The aim behind it is to suggest a framework that resolves the issue of clustering of concentrated data documents. In addition, the bisecting K-means clustering algorithm is combined with WordNet to obtain the semantic relation between words to enhance the process of clustering. The Elastic MapReduce is being used for testing to implementing the bisecting k-means algorithm. WordNet's use of semantic relationships decreases the size of features and the clustering of big data became apparent due to the decrease in the number of dimensions. WordNet lexical categories are also implemented for nouns that improve the measure of internal results [23].

Mousavi et al., proposed a systematic approach for evaluating and planning the license agreements for end-users. This is the first systematic approach to EULA comprehension and interpretation, to the best of our knowledge. The method included a systematic ontology describing relevant words, an ontology-based extraction of information, an excerpt clustering process, and a web-based user interface for analysis of self-service licenses. Our assessment found that the clustering is successful and greatly decreases the number of related words for users to initially concentrate on. Furthermore, EULAide is more visual and easier to digest EULAs according to the usability study report and saves about 75 percent of the time. Nonetheless, they are aware that this comes at a marginal price of 10.5 percent loss of useful knowledge, which is an appropriate trade-off, given the amount of time saved by users-particularly because a lot of users do not read the full

EULAs. They find this work to be a major step forward in making it more user-friendly to define the rules and regulations regulating online services, software tools, portals and applications [24].

Kamath S and V S presented a semantic-based Web service retrieval system using natural language processing (NLP) techniques to retrieve usable knowledge from the services has been suggested. They used this extracted information to measure the phonetic similarity between service pairs and to create service tags. In the service classification process, similarity and tags were used through Weka, 128 Semantic Similarity Based Context-Aware Web Service Discovery Using NLP strategies after which created classes are also tagged, taking into account the tags of all member services. The Artificial Neural Network (ANN) classifier obtained a classification accuracy of 91 percent, and class tags were created for each class based on the tags of their member services. In addition, the system also provided a natural language interface for user needs and context capture. The retrieval of related resources was based on the cosine similarity between the query vector and the class/service vectors. Experimental findings found that a query size of 2-4 terms resulted in the highest accuracy as the user context relations could be adequately identified and thus the most relevant resources could be retrieved. Furthermore, during service exploration when the Semantic Query method was used a major improvement of 16 percent growth in precision and more than a 40 percent increase in recall was observed [25].

Kavitha et al., proposed a comparison of the Artificial bee colony (ABC) algorithm and Support Vector Machine (SVM) training using the Reuters dataset alone and with a combination of Reuters and web documents data dataset. The overall F1 classification estimate based on the methodology suggested using the ABC algorithm is 88%. The F1 estimate for the Reuters dataset classification-based ABC algorithm using the Reuters data set alone is 83%. The experiment results show that because of the dynamic updating of web content and a thorough review of concepts, the proposed approach based on web documents yields better performance of precision and recall on unstructured documents. The excerpts also include the semantically linked documents, which are used to enhance the consistency of the classification. The F1 for SVM-based classification for a web document combined with the Reuters dataset is 85%. The F1 classification metric based on SVM is 81 percent for features from the Reuters dataset only. The F1 calculation on the SVM classifier with features from 20newsgroup and web documents is 78 percent for the 20Newsgroup data set and 84 percent for the ABC classifier. The results demonstrate that the suggested approach based on the ABC algorithm offers better performance compared to SVM as it requires fewer control parameters and a more detailed analysis of the concepts. Future work may include the parallelization of the clustering phase to reduce the processing time [26].

Avanija and Ramar proposed a C-Means Fuzzy technique called Semantic Hybrid Ontology Document Clustering (SEMHYBODC). The clustering technique would go like this: Initially, the documents are remarked before clustering by the "KIM" module. The documents are then clustered utilizing Fuzzy particle swarm optimization with Fuzzy C-Means (FPSOFCM), and the weight of the concept is measured. The SEMHYBODC algorithm recalculates the term weight through the steps defined in it. Additionally, clustering accuracy is calculated using cluster purity and comparison to different other approaches to hybrid. Usage of F-measurement and purity during the evaluation. The output shows with the use of swarm clustering intelligence is not acceptable for huge amounts of data in order to the processing time has been more accurate. However, the hybrid model combines with algorithms like FCM and PSO to solve this problem.

Nevertheless, the PSO is merged with K-mean it still contributes to the best solution; however, the computational duration is much higher, so in the end, the FPSO+FCM demonstrates the progress of another algorithm like Fuzzy C-means, K-mean, and Hybrid. This is mixed with the capacity to search from its globally connected and quick PSO aggregation algorithm FCM. Has been used as an outcome of the preliminary crop algorithm FPSOFCM, which has been implemented with the final result of the purifying. Web pages are compiled from the Internet for evaluation of proposals [27].

Agirre et al. discussed the results of a joint challenge for STS 2016. The STS 2016 attendance has risen significantly. The English STS subtask contains 119 submissions from 43 participating

teams. This is an improvement in active teams by 45 percent over 2015. The Spanish-English STS pilot cross-lingual subtask has 26 submissions from 10 teams, which is remarkable considering that this is the first year such a demanding subtask has been attempted. Ironically, on pairs drawn from the same sources, the cross-lingual STS systems tend to be performing competitively to monolingual systems. Which means that a more clear contrast between cross-lingual and monolingual systems would be important [28].

Ali and Melton provided a novel approach to the clustering of semantic-related papers, focused on the graph's mental science and theory. For human semantic memory, they applied the statistical mental system of semantic interaction, the ICAN system. The ICAN system was used to produce ICAN semantic-graphs at the document level. An emotional and graph-based strategy for the semantic elimination of ICAN graphs was applied. Then a corpus-level graph production algorithm proposed to generate the clustered documents for the generation of corpus-graphs from ICAN graphs, which are then clustered by applying the Louvain technique of community-detection. Their findings displayed a significant performance of their strategy over the LSA-based approach using the purity and entropy parameters to assess the clustering's consistency. Their research also highlighted the use of graph theory of emotional science for the clustering of semantic-based documents. In fact, using WordNet's lexical ontology, the ICAN-based method has the disadvantage of not being completely data-driven [29].

Romeo et al., proposed a framework for document clustering, called SeMDocT (segment-based Multilingual Document Clustering the Tensor Modeling). All documents into segments decomposed by SeMDocT; each section defines a subtopic and is depicted by a vector of BabelNet synsets or word occurrences. Parts are then clustered in clusters k. For each cluster a document is represented in the feature document matrix by a feature vector, the vector is the sum of all the vectors of document segment expose in the cluster. The system employs the document-feature matrices built to build a third-order tensor, where a tensor is a multi-dimensional sequence. And finally, the records are organized into clusters K [30]. Elsayed et al., provided a Bisecting K-Means method in [31], arguing that their variance provides better results in clustering documents than the standard K-Means. Next, the standard K-means algorithm is used to render the initial dataset two sub-clusters. The subcluster with the lowest similarity is chosen as the dataset for inputs. The above process is properly repeated until it exceeds the desired number of clusters.

Conrad and Bender published a research on clustering news stories around an event's sub-themes. The seed clusters for an event are defined from the subject labels provided editorially, called "sluglines." Their approach involves fuzzy deduplication of papers to define extremely related documents and then cluster them using language, persons and subject-based similarities. As it is difficult to collect ground-truth data from ideal clusters, the authors opted for a qualitative evaluation of this method with human evaluations of the clusters' cohesiveness and accuracy using a 5-point Likert scale [32]. Kolhe and Sawarkar proposed a clustering of idea oriented documents using WordNet. The established approach defines the prevalent notion and produces clusters automatically based on the notation of the matrix. A higher number of transformation matrices contribute to challenging memory requirements. The semantic algorithm can offer a particular application such as the product of web search clustering [33].

Glänzel and Thijs in their research have pointed out two concerns; firstly, to the clustering results, and the second is the position of core documents. In the 13-cluster solution, cluster hierarchy and concordance were the cluster's high exceptions on 'Pulsars of Radio. In the seven-cluster solution, this cluster was divided nearly equally among the clusters on ' Energy of Dark' and ' Explosion of Gamma-Ray.' Nevertheless, the hierarchical classification of 'Turbulence Atmospheric' in theory 2 was still quite 'fuzzy' but in the first scenario had a key concurrence of over 60 percent of the documents with 'Looping of Coronal. For all other cases, concordances were similar documents near or even above 90 percent. The second category of prominent remarks applies to key papers. These documents reflect the cluster-wide connections, along with the clusters 'the structure of an internal topic. In this case, they will reiterate that core-document identification is theoretically independent of clustering and therefore do not entail any cluster processing or detecting of the group, however, it may be incorporated smoothly into clustering activities, offered

the same group of links, i.e. linking of bibliographic, co-citation, the similarity of text or hybrid, are used. The core of documents affirms observation about the hybrid clustering's essential performance. Key cluster papers on 'Neutrino' and 'Energy of Dark' directly from the structure's core. The choosing of the two solution points occurred in a structure hierarchically that verified the suitability of the method applied [34].

Renukadevi and Sumathi presented research about the methodology of clustering and discuss their findings as to the advance of information technology, and the growing accessibility of the Internet radically changes all fields of operation in modern times. Consequently, it will allow a very large number of people to communicate with computer systems more regularly. It is important to provide systems capable of handling inputs in a variety of ways, such as printed / handwritten paper papers, of making the man-machine interaction more effective in these circumstances. The computer has to process the scanned images of printed documents effectively, and the methods have to be more sophisticated. The text documents are preprocessed, Term Frequency and Inverse Text Frequency (TF-IDF) are used to rate the paper. Different knowledge is then grouped using Fuzzy C – Means Clustering Algorithm [35]. Lin et al. presented methods in paper [36], first identifying similarity among two documents then expanding to two groups of documents, then applying clustering and grouping to a group of documents on the bases of similarity measures. The suggested measure of similarity called a measure of similarity for processing text (SMTP) and the algorithm classification applied is a classification single-label based on KNN (SL-KNN), classification multi-label based on KNN (MLKNN), and algorithm of clustering employing Clustering K-mean and Hierarchical Agglomerative Clustering. The output confirms the effectiveness of the measure of similarity SMTP which applies in-text applications (SL-KNN, HAC, MLKNN, and K-mean). Comparing the SMTP results with the other five measurements, Euclidean, Cosine, IT-Sim, Pairwise-adaptive (Pairwise) and Extended Jaccard (EJ), using specific k values and measurement measures such as (AC (Accuracy), EN (Entropy). This shows that the utility measure of similarity will based on 1) feature format, e.g., tf-idf or word count; 2) domain use, e.g. text or picture; and 3) clustering algorithm or classification.

Desai and Laxminarayana proposed a clustering model that integrates the resolution of Coreference and abuses semantic relationships among words by addressing synonymy and polysemy, employing semantic similarity and WordNet. The suggested method is composed of five elements (Preprocessing, Resolution of Coreference, Recognition of Synonymy, Features Selection, Disambiguation of Meaning, and Bisecting of K-means). The aim of all these five models is to classify the keywords in the document at first Coreference Resolution, This is achieved in three stages (POS marking, Elimination of stop words and stemming), after which Disambiguation of Meaning and Recognition of Synonym tackle the question of synonymy and polysemy in the text and use the WordNet model for this proposal, Instead the selection of the features is achieved by the words weighing in a document applying tf-idf, and finally by using Bisecting K-mean to cluster the documents. For testing purposes, the Article uses four popular Datasets (CISI is forty-four Documents, CACM is sixty Documents, MED consists of fifty-five Documents, and CRAN is forty-four Documents,) and is used to test the purity of the cluster. The result shows a contrast among origin and suggested system, thus the origin is the system without (the resolution of Coreference, disambiguation of meaning, and term pruning), after the contrast, the result term purity shows that the suggested system is observed to obtain 30 percent improvement in purity of the cluster, the purity percentage of origin configuration is 0.55 thus the suggested system is 0.8 [37].

Stanchev in his research paper [38], discussed how WordNet knowledge can be used to create a probabilistic graph. By adding part of the speech tag to each word and meaning, they expanded existing algorithms. They showed then how to extend the graph with DBPedia knowledge. They experimentally validated the algorithm by comparing it to an algorithm that uses the metric of cosine similarity and an algorithm that uses only WordNet knowledge. The results show that adding DBPedia information makes the Reuters-21578 benchmark both more reliable and more reminiscent of the algorithm. Nanayakkara and Ranathunga introduced a Sinhala news article clustering method consisting of two modules, namely the module for data collection of Article and

the module for grouping of Article. The classification of objects was implemented using a similarity analysis algorithm based on a corpus. This algorithm was able to achieve an accuracy of 77 percent in a news article set of 9 news providers, given its simplicity [39].

Zheng et al., proposed an evolutionary neural network-based model to classify report-pairs of imaging exams and pathological tests involving overlapping body sites by detecting semantic similarity. Compared to other traditional models such as keyword mapping, LSA, LDA, Doc2Vec, Siamese LSTM, our model showed superior performance and a system based on NER. They also leveraged the embedding graph methods for using external medical ontology knowledge and obtained further development. Furthermore, they implemented the LIME algorithm to evaluate our model behavior visually. The results suggested that our model could extract semantic features from texts automatically and fairly and make precise judgments. It may help collect patients or records more effectively to assess the quality of the imaging diagnosis [40].

Kenter and Rijke proposed a methodology for calculating the semantic similarity of short texts by combining word embedding with methods based on external sources of information. They used different features of text to train a supervised learning algorithm. They used a modification of the Okapi BM25 feature to rate documents in information retrieval and modify it to measure short texts' semantic similarity. They demonstrated that their approach outperforms other baseline methods to test the semantic similarity of short texts [41]. In another work, the query's semantic was obtained as an expanded collection of keywords with the use of synonyms. Each keyword and its corresponding synonyms are organized into different clusters that help to derive the user query's semantic meaning. The ranking was achieved by applying a membership rating to every keyword, which serves as the cluster's head. The presented model based on a fuzzy-cluster is compared to the model of classical Boolean. The findings obtained illustrate the model's efficacy as compared with the traditional approach. Nevertheless, the method can be further improved by considering the semantic relationships of a word to word. Several powerful ranking algorithms can also be used to obtain better results [42].

Radu et al., estimated the accuracy of the Doc2Vec model by integrating it with four distinct algorithms of clustering: DBSCAN, LDA, K-Means, Spherical K-Means, and, by using the model of TF-IDF, they compared the findings with those achieved. To verify that the embeddings of Doc2Vec are independent of the preprocessing process, they preprocessed the input corpus with a lemmatizer and two stemmers. The experiments empirically show that decreasing the dimensions increases the quality of the clusters, thus preserving the structures of the internal cluster and the semantic similarity space of terms by maintaining the relationships between the document [43]. Fatimi et al. show how web semantic techniques could be used to cluster and analyze textual documents. It highlighted some of the current works of modifying RDF data and was motivated by them to introduce a linked pipeline of semantic operations for the RDF-based semantic text clustering. The important component is to present an overall structure for semantic text clustering based on data modeling from RDF. This approach integrates several strategies to create an accurate and effective system such that textual documents can be analyzed using machine learning techniques integrated with semantic web concepts. The method lets RDF representation, topic modeling, clustering, and summarizing clusters of documents as well as information retrieval applying both reasoning tools and RDF querying. The objective is to use the semantic web to enhance document exploration and improve the use of semantics in the entire process [44].

This research aims to perform patient data clustering by listening carefully to the similarities of the patient's illness. ICD code for determining a patient's illness is also used as a guide. The K-means approach is integrated with semantic similarity to determine the proximity of the patient's ICD code. In this research, the approach used to determine the semantic similarity within data is the semantic similarity of Chodorow & Girardi, Leacock, Jaccard, and Rada. Measuring the quality of the clusters uses the method of the silhouette coefficient. The way of determining semantic similarity data is able to create higher quality results of clustering than without semantic similarity, depending on the experimental findings. The highest accuracy rate for all semantic similarity approaches is 91.78 percent, while the highest accuracy rate without semantic similarity is 84.93 percent [45].

Curiskis et al., illustrated the experimental findings of the proposed approach applying the domain Ontology which applies the semantic relationship in the extraction phase of the function and the algorithm Swarm Optimization (PSO). The performance of intracluster similarity is highly acceptable according to the result. In addition, it is discovered that the similarity between the clusters also achieves an accurate result. The suggested method provides a satisfactory outcome of the F-measure value (100 documents) on the News Group dataset [46]. Curiskis et al., tested four feature representation techniques derived through embedding matrices and TF-IDF (the matrix of TF-IDF, the matrix of mean Word2Vec, the matrix of mean Word2Vec weighted by the scores of TF-IDF and the matrix of Doc2Vec to every document) in combination with four strategies of clustering. (Algorithms for K-Means, Hierarchical Agglomerative, K-Medoids, and Matrix of Positive Factorization) on comments of Reddit and Twitter data. The Doc2Vec method, combined with the K-Means algorithm, gains the best Optimized Rand Index scores [47].

In this research work, using word embedding representations and Machine Translation (MT), they have provided two ways of measuring the semantic relationships among Arabic-English cross-language sentences. The key concept is based on the use of word semantic properties that have been used in the word embedding method. They have used a mixture of IDF, word alignment, and POS weighting to further verify the most informative words from every sentence to analyze the semantic sentence similarity further. In addition, they estimated their suggestions on the four STS shared task datasets SemEval-2017. They have shown in the experiments how the Bag-of-words approach significantly enhanced the results of the correlation. Their proposed methods' performance was established through the Pearson correlation among both their allocated semantic similarity measures and human judgments. They actually achieved the highest correlation value compared with all of the participating systems in SemEval-2017's cross-language STS Arabic-English subtask [48].

Cao et al., generated named entities through the clustering of documents that have not been properly examined, to their knowledge. As the first step in the scope of this Article, they are not worried about the best algorithm for clustering NE-based documents. Alternatively, they apply the traditional keyword-based VSM to describe a document through respect to the named entities' various characteristics that exist in the document, including their types, identifiers, and names. Then they will use k-means, a standard algorithm of clustering, to explain the benefits of clustering documents through NE knowledge [49].

Several steps have been used in the technique. The method is called WMDC (Wikipedia matching document classification). Firstly, choose the Wikipedia knowledge and concept. Secondly, to select the relevant concepts by applying heuristic selection. Third, determining similarities between both text using an integration of (Wikipedia-based semantic similarity and Textual similarity depend on matching keywords) and using the algorithm of K-mean for classification in this method due to its accuracy and efficiency. Additionally, in this paper, assessment research is consisting of two sections. The first section focuses on the effectiveness of the rules of heuristic selection; thus those rules are utilized to capture the relevant concept of the word (Rule 1: all the titles, Rule 2: any keywords and Rule 3: all the keywords) in order to determine the effectiveness of these three rules through selectivity measurement and quality evaluation using relevance. The second section, assessing the approach's effectiveness [50].

The proposed approach brings in an association between both the disciplines of text mining and natural language processing. The "semantic-based mining method" integrated into the proposed approach significantly shows a favorable increase in cluster performance. Advancement in the output of clustering is due to better recognition of the sentence's semantic structure in an article; this determination activity is apparent and plays a significant role in the performance of the tests. Compared to that of the traditional single term-based method, the reliability of the output clusters produced by this process may have significant improvements [51].

III. Methodology

In this section, we have planned to survey on document clustering based on semantic similarity. By our plan, we searched on the Internet for the related papers. First, eighty articles are selected by our plan when we are planned from our topic of semantic similarity based on document clustering, In a quick review, ten papers were eliminated due to they are duplicated papers and

seventy papers remain. After that, twenty articles are eliminated because they are not related to our objective. The remaining papers were reviewed in detail based on three groups of features that are classified to similarity measures, methods and clustering algorithms, and evaluation measures. Finally, after reviewing all papers, ten papers are omitted thus their quality of issues is not good which means forty papers to be used in our research. We have used several excel sheets to collect the remaining papers. The searching process steps are indicated in Fig. 1.

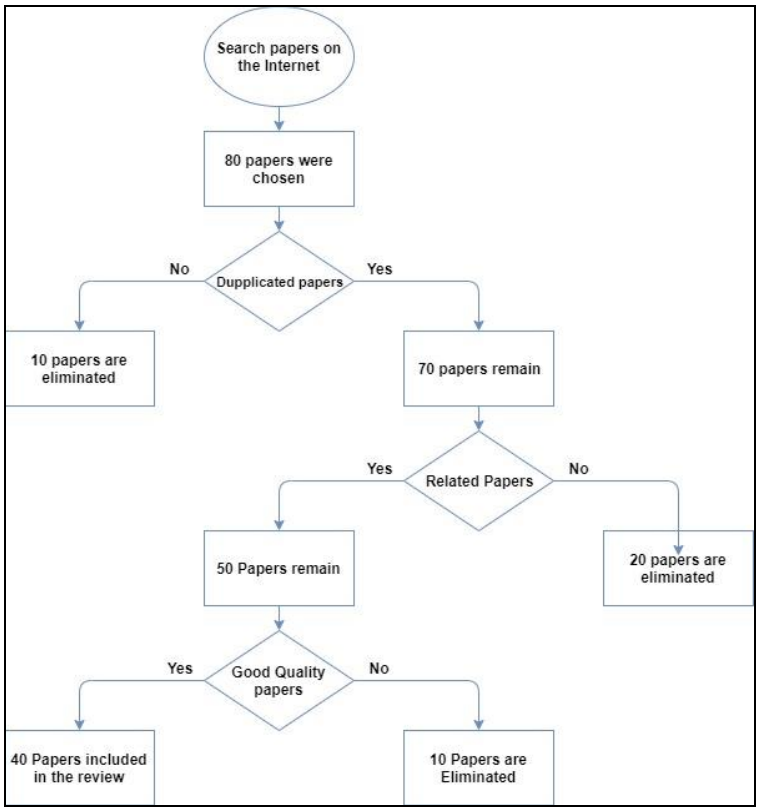

Fig. 1. Searching Steps Process

IV. Results and Discussion

In this section, a comprehensive discussion of the reviewed papers is presented in Table 1. The features of the thirty papers, such as similarity approaches, used datasets, similarity measures, clustering algorithms, and evaluation measures, are summarized in a table. Furthermore, our significant findings of the studies have also been explained. The term frequency-inverse document frequency (TFIDF) matrix is the most common and popular similarity measure the authors use in their work. When using TF-IDF, the accuracy and quality of the clustering are increased. The researchers proposed a new technique for detecting similarities and semantic associations between words using the TF-IDF matrix and WordNet ontology. It is used to increase quality and quality [18]. It uses the TF-IDF technique, which removes the most common terms and derives from the corpus only the most important terms [20]. Furthermore, it is used to calculate the importance of a specific tag participant for a given service compared to its significance in other data set services [25]. It is also the easiest option for a paper to use the basic frequency of a term [35].

The cosine similarity measure is used fewer than TF-IDF by authors to measure the similarity between documents in their works. Cosine similarity parameters had good performance than Complete -Linkage, Average-Linkage hierarchical algorithms [18]. Therefore, the class/service is unrelated to the query. In this way, a ranked list of classes/services for a given query can be generated using the computed cosine similarity values [25]. The measure of cosine similarity is employed to describe the power of the bibliographic relation among the documents [34].

The Jaccard and Euclidean distance similarity measures are nearly used the same by some authors in their works. Jaccard's coefficient of similarity determines the similarity between the

documents and ontology shows an important role in the retrieval of document similarity [15]. Euclidean similarity parameters had better performance than the Complete –Linkage, Average-Linkage hierarchical algorithms [18]. However, researchers rarely use the dice similarity measure to determine the similarity between documents in their work. The document's similarity can be calculated using Dice, based on the keywords collection [15].

The researchers in their work apply the most popular and common technique is WordNet. WordNet ontology integration provides more semantics to the clusters and thus allows it easier for users to search for more versatility [15]. WordNet is used to determine the document context, as it has been used widely and consistently to facilitate a better understanding of documents [16]. With its semantic associations of terms, it has been used widely to increase the quality of the cluster of the text [3]. WordNet may improve the efficiency of the calculation of similarities. WordNet has been used widely in improving document clustering quality [6]. WordNet's use of clustering identifies the relationships among the words and helps to define the exact document clustering. A new technique for finding similarity and semantic relationships between terms using the TF-IDF matrix and WordNet ontology is introduced. Then, in order to improve reliability and accuracy, a new approach is proposed using WordNet ontology based on hierarchical clustering algorithms, which categorize data more accurately and qualitatively [18]. For determining the similarity of the synset based on the distance between two terms in WordNet taxonomy [25]. Although using WordNet to decrease document dimensionality, document clustering increases Efficiency [31]. WordNet uses the word form for vocabulary to refer to both words and phrases. Note that a word form is not exact in its meaning. For instance, the word "spring" may mean, among other meanings, the season after winter, a metal elastic device or natural groundwater flow. That is why WordNet uses a meaning concept [38].

The K-means and Hierarchical Clustering algorithms are used the same by authors in their work to clustering documents. Because it's a simple, highly scalable clustering method that operates directly on the articles' vector representations [11]. This correlates well to a large number of samples and has been used in many different fields including scientometrics across a wide variety of application areas [11]. And k means clustering is one of the non-supervised learning strategies used to comprehend the underlying dataset structure [19]. By taking into account the hyponyms and synonyms of prevalent notions from commonly happening words in the text corpus, WordNet can achieve a stronger clustering [33]. A new approach is proposed using WordNet ontology based on hierarchical clustering algorithms, which categorize data more accurately and qualitatively [18]. Within this method, the use of hierarchical clustering provides high-level reliability and data classification. The benefits of similarity of document and hierarchical clustering are helpful in increasing the speed of processing and information retrieval [19]. Throughout the wide dataset [20], Hierarchical Agglomerative Clustering (HAC) is used. They adopt an approach in which they preprocess the corpus in order to extract noisy, less valuable information. They apply TF-IDF followed by the HAC and fuzzy K-means algorithm's well-known technique, which has been shown to be an effective method for clustering text and documents. Moreover, the average linkage has been proved to be the most suitable one for text categorization among various HAC approaches. In addition, the average linkage has been shown to be the most suitable one for text categorization among various HAC methods [24].

The Fuzzy C-means clustering and bisecting k-means algorithms are used the same by authors in their work but fewer than Hierarchical and K-means clustering algorithms. Using Fuzzy C-means on the large dataset [20]. They adopt an approach in which they preprocess the corpus in order to extract noisy, less valuable information. They apply TF-IDF, followed by HAC, and the fuzzy C-means algorithm and K-Means are well suited for large datasets. It has been suggested that an optimum number of clusters be produced and that record recovery is better accurate since both effective and easy to enforce [27]. Finding the best set of characteristics and metrics to assess whether a pair of singletons or local clusters warrant merging into larger clusters while remaining sufficiently coherent and, second, finding the optimal sequence to compare such clusters when considering merging[52]. The fuzzy C-mean algorithm offers good efficiency for record clustering compared to the Purity and Entropy accuracy clustering [35]. The authors used the Bisecting k-

means algorithm to test if our approach is capable of producing better classification labels for derived clusters [17]. Bisecting the k-means algorithm was also used to decrease the dimensionality problem by a few authors along with the map reduces programming model [31]. Some researchers have used the combination of bisecting k-means and tensor-based model, which produced strong results for multi-lingual documents with no particular disadvantage [30].

Some measures such as precision, recall, purity, F-score, Adjust Rand Index, and entropy are commonly used for evaluation of results by different authors. The precision and recall evaluation measures are common, and popular tools are used to estimate the clusters' quality and performance. Many small clusters may increase the precision since they comprise only the articles considered to be in the same cluster [11]. Experimental findings showed that a query size of 2-4 terms resulted in the highest precision since the user context relationships could be correctly identified and thus the most relevant services could be retrieved [25]. Precision is determined as the ratio of the amount of related records obtained to the total amount of significant and important records collected [33]. A 100 percent recall can be accomplished by increasing all the items in one cluster [11]. The recall values are marginally increasing because more resources will suit the words in the query [25]. The recall is calculated as the ratio of the amount of related records obtained in the database to the total amount of related records [33].

Entropy, purity, and F-Score evaluation measures are used the same by authors in their work to estimate the experimental results. Entropy is the number of individual entropies measured by the size of the cluster [17]. Measurement of Entropy and F-measure was determined for each dataset. The Best algorithm is indicated by the lowest entropy and highest F-measure value [20]. The highest scoring collection of synonymy concepts has the lowest and nearly similar [15]. Purity suggests that the texts they received from a cluster are part of that cluster's actual class [17]. A cluster's integrity is the proportion of a number of most frequent class occurrences to the cluster size [33]. They used F-Score to determine the consistency of the clusters. The F-Score values are within the range [0..1], and the largest F-score value indicates higher cluster output. They compared our algorithm's F-Score value to other algorithms [6]. This was used to evaluate program effectiveness and implemented modifications [39]. The authors never use the Adjusted Rand Index assessment test to assess the experimental outcome. In [18], the Change Rand Index unmonitored evaluation criterion is used to determine the clustering's efficiency.

Dataset Reuters-21578 is commonly used for clustering document. This dataset is separated into four partitions. The first partition includes 193 scripts and 10 groups where the distribution of groups is standardized. The second partition has 1726 scripts and 20 groups, of which the length of the groups is of great difference and the distribution of the classes is not either. The third partition has 1240 scripts and 9 groups, where each group has huge documents and there is a uniform distribution of groups. There are 6714 scripts and 3 groups in partition Four.

From this survey and discoveries, we discover which combining techniques such as Ontology-based and Lexical Chain-based approaches could also provide higher performances for the clustering of semantic documents. Lexical chains could resolve issues with clustering documents, like cluster naming, synonymy, high dimensionality, and polysemy.

Table 2. Includes a list of the different algorithms of clustering that could be applied for document clustering. Included in Table 2. are different algorithms of clustering like the Hierarchical Agglomerative, Bisecting k-means, Fuzzy C means, and K-means clustering algorithms.

Fig. 2 shows the similarity measurements which are used in the previous papers. The data on this chart has been taken from the table of the survey. Various similarity measures such as Cosine, Jaccard, Dice, Euclidean, and the term frequency-inverse document frequency (TF-IDF) are included in Table 1. Fig. 3 displays the methods and clustering algorithms that are applied in the previous papers. Various clustering algorithms such as K-means, Bisecting k-means, Fuzzy C-means, Hierarchical Clustering algorithms, and WordNet are included in Table 1. Fig. 4 indicates the evaluation measures that determine the clusters' performance that has been used in previous papers, such as purity, entropy, precision, recall, f-score, and Adjust Rand Index.

Table 1. Survey on Document Clustering based on Semantic Similarity.

References	Date	Semantic Approach	Dataset	Similarity Measures					Method and Clustering Algorithms					Evaluation Measures					
				Cosine	Jaccard	Euclidean	Dice	TD-IDF	fuzzy (C-Means)	K-means	Hierarchical Clustering	WordNet	Bisecting K-Means	Entropy	Precision	Recall	Purity	F-Score	Adjust Rand Index
[26]	2014	web document classification	20Newsgroup					x				x			x	x			
[30]	2014	Semantic-Based Multilingual Document Clustering	Reuters Corpus Volume 2 (RCV2).										x		x	x			x
[52]	2014	Semi-Supervised Events Clustering	The news repository.						x		x					x			
[35]	2014	Text Classification and Clustering	21578 news documents	x	x	x	x	x	x										
[36]	2014	Text Classification and Clustering	several real-world data sets	x	x	x	x	x		x	x			x					
[17]	2015	text clustering using WordNet and lexical chains	reuters-21578 corpus									x	x	x	x	x	x		
[8]	2015	combining edge-counting and information content theory	MC Dataset			x			x			x							
[23]	2015	reducing Arabic texts dimensionality	BBC Arabic news									x							
[31]	2015	Ontology based document clustering	21578 news documents	x				x			x	x	x	x	x	x	x		
[41]	2015	Short Text Similarity with Word Embeddings	Microsoft Research Paraphrase Corpus data set	x		x		x				x			x	x		x	
[16]	2016	Measuring Semantic Similarity between Documents	A collection of 50 documents from the Australian news mail service.		x		x	x				x			x	x			
[19]	2016	Document Classification and Clustering	several real world data sets							x	x			x					
[20]	2016	Document Clustering	NEWS 20, Reuters, and emails	x				x	x	x	x			x	x	x			
[25]	2016	Using nlp techniques	WSDL document	x				x				x			x	x			
[28]	2016	Semantic Textual Similarity	news and multi-source	x								x			x	x			
[37]	2016	Document Clustering	classic4	x				x				x	x				x		
[11]	2017	Clustering articles based on semantic similarity	the astromy dataset										x		x	x		x	
[18]	2017	Clustering Data Text Based on Semantic	The 20Newsgroups dataset	x		x		x			x	x							x
[21]	2017	News Clustering	Several news mass media official page in facebook									x							

[22]	2017	Semantic based document clustering	20Newsgroup									x							
[24]	2017	Improved End-User Interpretation	comprising 1.000 EULAs								x				x	x			
[27]	2017	Web Document Clustering	Reuters			x		x	x	x					x	x			
[33]	2017	A Concept Driven Document Clustering	Open Directory Project (ODP) and Ambient datasets	x				x				x			x	x	x		
[34]	2017	the representation of clusters and topics	the astromy dataset	x	x			x											
[50]	2017	An efficient Wikipedia semantic matching	Wikipedia and Reuters-21578[6]	x	x			x		x									
[15]	2018	Ontology Based Document Data Analysis	Annual reports of banks					x		x		x					x		
[6]	2018	text document clustering using frequent itemsets and WordNet	Classic4, Reuters WAP, Hitech and Re0					x				x						x	
[29]	2018	Semantic-Based Text Document Clustering	BBC, BBC Sport and classic300		x	x		x		x		x		x			x		
[38]	2018	Semantic Document Clustering	the Reuters-21578 benchmark	x				x		x		x			x	x			
[39]	2018	ClusterigSinhala News Articles	Sinhala news articles	x				x				x						x	
[48]	2018	Measuring Semantic Similarity of Arabic-English Sentences	STS Cross-lingual Arabic-English	x	x		x				x								
[49]	2018	Introduce name entities into document	960 documents collected mainly in the politics domain from BBC News	x				x		x	x			x					
[51]	2018	Web Document Clustering	Real document	x	x			x				x				x			
[40]	2019	Detection of medical text semantic similarity	16,354 imaging and pathology report-pairs from 1926 patients	x	x										x	x		x	
[45]	2019	Estimate The Cost of Hospitalization	244 patient data		x		x			x									
[46]	2019	globalized search in the entire solution space	News Group dataset (100 documents)								x				x	x			
[47]	2019	An evaluation of document clustering and topic modeling in two online social networks	Twitter data and Reddit comments	x				x		x	x								x

[42]	2020	Semantic information retrieval	document repository						x			x			x	x			
[43]	2020	reduce the documents representations dimensionality	Sexualities of Mexico.	x				x		x									x
[44]	2020	Semantic text clustering	RDF data of DBPedia	x				x		x	x	x			x				

Table 2. Comparison of Algorithms of Clustering

Algorithms	Pros	Cons
Hierarchical Clustering	Implementation is simple. Deductive data is not about cluster no. provided. The quality of the cluster is adequate.	The algorithm isn't working backward. Computing time is valuable. Hard to define by dendrogram the exact cluster no.
Bisecting K-Means	It also has a low expense, and its clustering performance is very adequate. No need to consider the amount of k.	No cons points.
Fuzzy C-Means	Provides the highest overlapping data set outcome and compared batter than the K-means.	Apriori cluster number specification. Euclidean distance measurements may cause underlying factors to weight unequally.
K-Means	Specifically generates strong clusters if the clusters are through it in shape.	Hard to guess the amount of k. Clusters are non-hierarchy and therefore do not Conflicting.

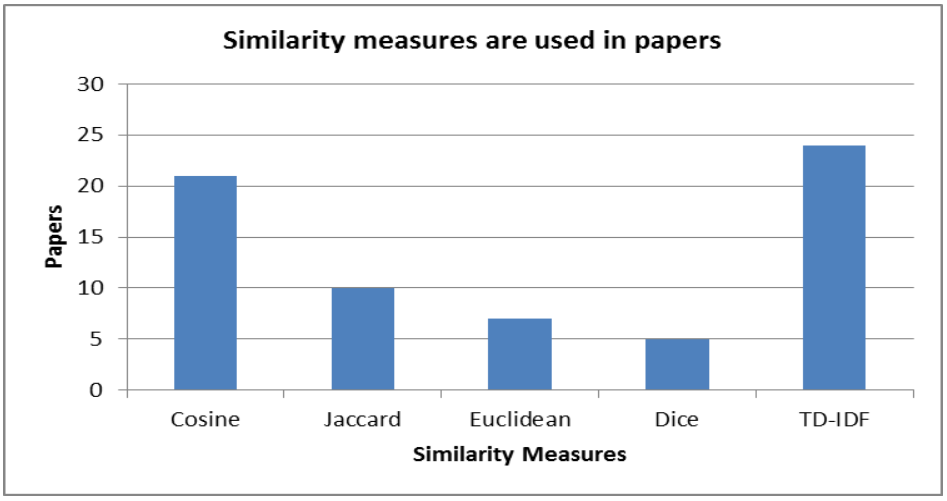

Fig. 2. Similarity Measures on Survey

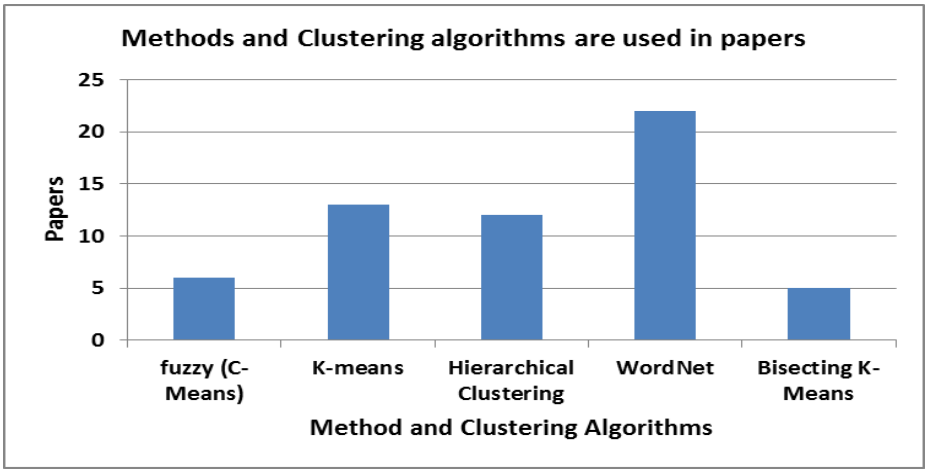

Fig. 3. Methods and Clustering Algorithms on Survey

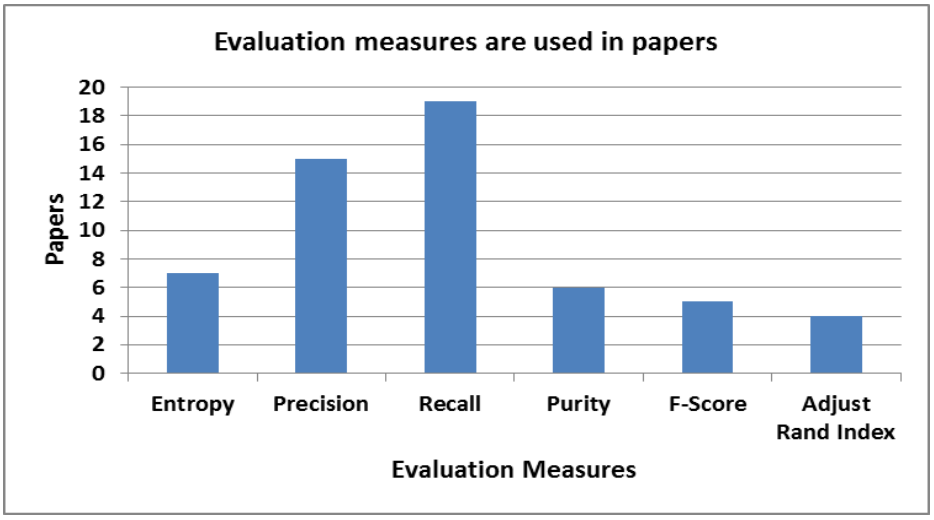

Fig. 4. Evaluation Measures on Survey

V. Conclusions

In this paper, a comprehensive review of the semantic similarity based on the clustering of documents has been presented in depth. The research has been conducted on various techniques, text processing methods, ontologies, and various clustering algorithms. We have shown datasets, measures of the similarity, and the methods and clustering algorithms that are used in each paper. We have also presented in detail the evaluation measures that are used to evaluate the performance of the result. Nevertheless, one of the most common methods used is the English WordNet dataset for accurate clustering. Thus clustering is performed using algorithms of clustering. Also, the most popular one would be the algorithm of K-mean in order to its ease of use. It can provide closer clustering as well as there is another such as Fuzzy C-mean, bisecting K-mean, and Hierarchical agglomerative clustering. At last, seeing as we have seen that the main objective of all techniques is to improve the performance, ensure the accuracy of clustering has been used the theoretical measure to evaluate proposes including (Precision, Recall, Entropy, and Purity, etc.) demonstrate that the semantic clustering provides a better outcome in all circumstances. Finally, the results of our survey have been shown by using several figures. The survey findings revealed that in terms of the quality and accuracy of clusters, the semantic technique is better than the traditional techniques. This helps in the implementation and evaluation of the clustering of documents. The exposed research is used in the same direction when preparing the proposed research.

References

[1] K. Jacksi, S. R. M. Zeebaree, and N. Dimililer, "LOD Explorer: Presenting the Web of Data," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 9, no. 1, 2018, doi: 10.14569/IJACSA.2018.090107.

[2] K. Jacksi and S. Abass, "Development History Of The World Wide Web," *Int. J. Sci. Technol. Res.*, vol. 8, pp. 75–79, 2019.

[3] K. J. A Zeebaree SRM Zeebaree, "Designing an Ontology of E-learning system for Duhok Polytechnic University Using Protégé OWL Tool," *J Adv Res Dyn Control Syst Vol*, vol. 11, no. 5, pp. 24–37, 2019.

[4] K. Jacksi, N. Dimililer, and S. R. M. Zeebaree, "A Survey of Exploratory Search Systems Based on LOD Resources," in *PROCEEDINGS OF THE 5TH INTERNATIONAL CONFERENCE ON COMPUTING & INFORMATICS*, COLL ARTS & SCI, INFOR TECHNOL BLDG, SINTOK, KEDAH 06010, MALAYSIA, 2015, pp. 501–509.

[5] K. Jacksi, N. Dimililer, and S. R. Zeebaree, "State of the Art Exploration Systems for Linked Data: A Review," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 7, no. 11, pp. 155–164, 2016, doi: dx.doi.org/10.14569/IJACSA.2016.071120.

[6] H. Patil and R. Thakur, "A semantic approach for text document clustering using frequent itemsets and WordNet," *Int. J. Eng. Technol.*, vol. 7, p. 102, Jun. 2018, doi: 10.14419/ijet.v7i2.9.10220.

[7] R. Ibrahim, S. Zeebaree, and K. Jacksi, "Survey on Semantic Similarity Based on Document Clustering," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 4, no. 5, pp. 115–122, 2019, doi: 10.25046/aj040515.

[8] J.-B. Gao, B.-W. Zhang, and X. H. Chen, "A WordNet-based semantic similarity measurement combining edge-counting and information content theory," *Eng Appl AI*, vol. 39, pp. 80–88, 2015, doi: 10.1016/j.engappai.2014.11.009.

[9] K. Jacksi and S. Badiozamany, "General method for data indexing using clustering methods," *Int. J. Sci. Eng.*, vol. 6, no. 3, pp. 641–644, Mar. 2015.

[10] K. Jacksi, "Toward the Semantic Web and Linked Data Exploration," 2019, pp. 227–227.

[11] S. Wang and R. Koopman, "Clustering articles based on semantic similarity," *Scientometrics*, vol. 111, pp. 1017–1031, 2017, doi: 10.1007/s11192-017-2298-x.

[12] A.-Z. Adel, S. Zebari, and K. Jacksi, "Football Ontology Construction using Oriented Programming," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 1, pp. 24–30, 2020.

[13] K. Jacksi, "Design and Implementation of E-Campus Ontology with a Hybrid Software Engineering Methodology," *Sci. J. Univ. Zakho*, vol. 7, no. 3, pp. 95–100, 2019.

[14] S. R. M. Z. Adel AL-Zebari Karwan Jacksi and Ali Selamat, "ELMS–DPU Ontology Visualization with Protégé VOWL and Web VOWL," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 1, pp. 478–485, 2019.

[15] A. Zafar, M. Awais, and M. A. Aftab, "Ontology Based Document Data Analysis," p. 7, 2018.

[16] M. K. L. Sumathy and D. Chidambaram, "A Hybrid Approach for Measuring Semantic Similarity between Documents and its Application in Mining the Knowledge Repositories," *Int. J. Adv. Comput. Sci. Appl. Ijacsa*, vol. 7, no. 8, 2016, doi: 10.14569/IJACSA.2016.070831.

[17] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 2264–2275, Mar. 2015, doi: 10.1016/j.eswa.2014.10.023.

[18] P. Zandieh and E. Shakibapoor, "Clustering Data Text Based on Semantic," 2017.

[19] S. Melasagare and V. Thombre, "Document Classification and Clustering using Feature Extraction for Similarity Measure," 2016.

[20] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Mar. 2016, pp. 61–66, doi: 10.1109/ICEEOT.2016.7754750.

[21] I. Blokh and V. Alexandrov, "News clustering based on similarity analysis," *Procedia Comput. Sci.*, vol. 122, pp. 715–719, Jan. 2017, doi: 10.1016/j.procs.2017.11.428.

[22] S. AFREEN and D. B. SRINIVASU, "SEMANTIC BASED DOCUMENT CLUSTERING USING LEXICAL CHAINS," 2017.

[23] A. Awajan, "Semantic Similarity Based Approach for Reducing Arabic Texts Dimensionality," *Int. J.*

Speech Technol., Jun. 2015, doi: 10.1007/s10772-015-9284-6.

[24] N. Mousavi, S. Scerri, and S. Auer, "Semantic Similarity based Clustering of License Excerpts for Improved End-User Interpretation," Sep. 2017, doi: 10.1145/3132218.3132224.

[25] S. Kamath S and A. V S, "Semantic similarity based context-aware web service discovery using NLP techniques," *J. Web Eng.*, vol. 15, Mar. 2016.

[26] C. Kavitha, S. Sadhasivam, and S. Kiruthika, "Semantic similarity based web document classification using Artificial Bee Colony (ABC) algorithm," *WSEAS Trans. Comput.*, vol. 13, pp. 476–484, Jan. 2014.

[27] J. Avanija and K. Ramar, "Semantic Similarity-Based Clustering of Web Documents Using Fuzzy C-Means," *Int. J. Comput. Intell. Appl.*, vol. 14, Sep. 2015, doi: 10.1142/S1469026815500157.

[28] E. Agirre *et al.*, "SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, Jun. 2016, pp. 497–511, doi: 10.18653/v1/S16-1081.

[29] I. Ali and A. Melton, "Semantic-Based Text Document Clustering Using Cognitive Semantic Learning and Graph Theory," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, Jan. 2018, pp. 243–247, doi: 10.1109/ICSC.2018.00042.

[30] S. Romeo, A. Tagarelli, and D. Ienco, "Semantic-Based Multilingual Document Clustering via Tensor Modeling," Oct. 2014, doi: 10.13140/2.1.2947.7765.

[31] A. Elsayed, H. Mokhtar, and O. Ismael, "Ontology Based Document Clustering Using MapReduce," *Int. J. Database Manag. Syst.*, vol. 7, May 2015, doi: 10.5121/ijdms.2015.7201.

[32] J. G. Conrad and M. Bender, "Semi-supervised events clustering in news retrieval.," 2016, pp. 21–26.

[33] S. R. Kolhe and S. D. Sawarkar, "A concept driven document clustering using WordNet," in *2017 International Conference on Nascent Technologies in Engineering (ICNTE)*, Jan. 2017, pp. 1–5, doi: 10.1109/ICNTE.2017.7947888.

[34] W. Glänzel and B. Thijs, "Using hybrid methods and 'core documents' for the representation of clusters and topics: the astronomy dataset," *Scientometrics*, vol. 111, Feb. 2017, doi: 10.1007/s11192-017-2301-6.

[35] D. Renukadevi and S. Sumathi, "TERM BASED SIMILARITY MEASURE FOR TEXT CLASSIFICATION AND CLUSTERING USING FUZZY C-MEANS ALGORITHM," 2014.

[36] Y.-S. Lin, Y. Jiang, and S.-J. Lee, "A Similarity Measure for Text Classification and Clustering," *Knowl. Data Eng. IEEE Trans. On*, vol. 26, pp. 1575–1590, Jul. 2014, doi: 10.1109/TKDE.2013.19.

[37] S. S. Desai and J. A. Laxminarayana, "WordNet and Semantic similarity based approach for document clustering," in *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Oct. 2016, pp. 312–317, doi: 10.1109/CSITSS.2016.7779377.

[38] L. Stanchev, "Semantic Document Clustering Using Information from WordNet and DBPedia," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, Jan. 2018, pp. 100–107, doi: 10.1109/ICSC.2018.00023.

[39] P. Nanayakkara and S. Ranathunga, "Clustering Sinhala News Articles Using Corpus-Based Similarity Measures," in *2018 Moratuwa Engineering Research Conference (MERCon)*, May 2018, pp. 437–442, doi: 10.1109/MERCon.2018.8421890.

[40] T. Zheng *et al.*, "Detection of medical text semantic similarity based on convolutional neural network," 2019, doi: 10.1186/s12911-019-0880-2.

[41] T. Kenter and M. de Rijke, "Short Text Similarity with Word Embeddings," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, Melbourne, Australia, 2015, pp. 1411–1420, doi: 10.1145/2806416.2806475.

[42] D. Mahapatra, C. Maharana, S. P. Panda, J. P. Mohanty, A. Talib, and A. Mangaraj, "A Fuzzy-Cluster based Semantic Information Retrieval System," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, Mar. 2020, pp. 675–678, doi: 10.1109/ICCMC48092.2020.ICCMC-000125.

[43] R.-G. Radu, I.-M. Radulescu, C.-O. Truica, E.-S. Apostol, and M. Mocanu, "Clustering Documents using the Document to Vector Model for Dimensionality Reduction," in *2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*, Cluj-Napoca, Romania, May 2020, pp. 1–6, doi: 10.1109/AQTR49680.2020.9129967.

[44] S. Fatimi, C. El, and L. Alaoui, "A Framework for Semantic Text Clustering," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, 2020, doi: 10.14569/IJACSA.2020.0110657.

[45] I. B. G. Sarasvananda, R. Wardoyo, and A. K. Sari, "The K-Means Clustering Algorithm With Semantic Similarity To Estimate The Cost of Hospitalization," *IJCCS Indones. J. Comput. Cybern. Syst.*, vol. 13, no. 4, p. 313, Oct. 2019, doi: 10.22146/ijccs.45093.

[46] Wai Wai Lwin, "Impressive Approach for Documents Clustering Using Semantics Relations in Feature Extraction," presented at the 2019 the 9th International Workshop on Computer Science and Engineering, 2019, doi: 10.18178/wcse.2019.03.007.

[47] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," *Inf. Process. Manag.*, vol. 57, no. 2, p. 102034, Mar. 2020, doi: 10.1016/j.ipm.2019.04.002.

[48] E. M. B. Nagoudi, J. Ferrero, D. Schwab, and H. Cherroun, "Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences," in *Arabic Language Processing: From Theory to Practice*, vol. 782, Cham: Springer International Publishing, 2018, pp. 19–33.

[49] T. H. Cao, V. M. Ngo, D. T. Hong, and T. T. Quan, "Semantic Document Clustering on Named Entity Features," *ArXiv180707777 Cs*, Jul. 2018, Accessed: Jul. 04, 2020. [Online]. Available: http://arxiv.org/abs/1807.07777.

[50] Z. Wu *et al.*, "An efficient Wikipedia semantic matching approach to text document classification," *Inf. Sci.*, vol. 393, pp. 15–28, Jul. 2017, doi: 10.1016/j.ins.2017.02.009.

[51] Dr. N. Krishnaraj, D. P. Kumar, and S. K. Bhagavan, "Conceptual Semantic Model for Web Document Clustering Using Term Frequency," *EAI Endorsed Trans. Energy Web*, vol. 5, no. 20, p. 155744, Sep. 2018, doi: 10.4108/eai.12-9-2018.155744.

[52] J. G. Conrad and M. Bender, "Semi-supervised events clustering in news retrieval.," in *NewsIR@ ECIR*, 2016, pp. 21–26.