Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 5, 115-122 (2019)

www.astesj.com

Special Issue on Advancement in Engineering and Computer Science

ASTESJ

ISSN: 2415-6698

Survey on Semantic Similarity Based on Document Clustering

Rowaida Khalil Ibrahim[*,1], Subhi Rafeeq Mohammed Zeebaree[2], Karwan Fahmi Sami Jacksi[1]

[1]Computer Science Department, University of Zakho, 420011, Zakho, Iraq

[2]Department of IT, Akre Technical College, Duhok Polytechnic University, 42004, Akre, Iraq

ARTICLEINFO

Article history:
Received: 15 June, 2019
Accepted: 16 July, 2019
Online: 03 September, 2019

Keywords:
Clustering
Text Clustering
Semantic Similarity
Ontology
Evaluation Measures

ABSTRACT

Clustering is a branch of data mining which involves grouping similar data in a collection known as cluster. Clustering can be used in many fields, one of the important applications is the intelligent text clustering. Text clustering in traditional algorithms was collecting documents based on keyword matching, this means that the documents were clustered without having any descriptive notions. Hence, non-similar documents were collected in the same cluster. The key solution for this problem is to cluster documents based on semantic similarity, where the documents are clustered based on the meaning and not keywords. In this research, fifty papers which use semantic similarity in different fields have been reviewed, thirteen of them that are using semantic similarity based on document clustering in five recent years have been selected for a deep study. A comprehensive literature review for all the selected papers is stated. A comparison regarding their algorithms, used tools, and evaluation methods is given. Finally, an intensive discussion comparing the works is presented.

1. Introduction

Clustering is a problem of unsupervised learning; the main task is to group a set of objects in a manner that the group of objects in the same collection called the cluster are more similar (in meaning) to each other than those in the other cluster [1][2]. Document clustering or (text grouping) is a block analysis application on text documents, has applications in automatic organization of document, filtering or fast information retrieval and subject extraction [3][4]. The text group divides a set of text documents into different categories so that the documents in the same category group describe the same subject. Text clustering is an essential function in the text mining process [5]. Problem of Description clustering is one of the key issues for the traditional algorithm of document clustering. The objects can cluster in traditional algorithm, but the result of clustering cannot give description concept. It is therefore necessary to find a way to resolve the problem of grouping the document description in such a way to meet the specific needs of document clustering [6]–[8].

For conceptual and meaningful collecting of documents into groups (clusters), it is necessary to exploit the semantic relationships between words and concept in the documents [9]. For this case semantic similarity is needed. The term semantic similarity is a specific measure for several pamphlets/concepts in contrast with the similarity that capable of roughly- calculations with respect to represent grammar (such as the format of the string). For this reason, the mathematical instruments are depended for asset estimation semantic association among language units and ideas or examples, by means of a numerical description obtained by comparing information that supports its meaning or description of its nature [10]–[12].

It is often confused with the term semantic relatedness and semantic similarity. The semantic relationship involves any relationship between two periods, whereas the semantic similarity only includes a relationship “is a” [13]. For example, "car" and “bus" are similar, but is related to “driving" and “road”. By determining topological similarity, Semantic similarity can be estimated using the ontology to determine the distance between concepts/term [14][15].

For semantic similarity purpose between concepts many tools are used, the most common one is the WordNet. The term of WordNet Is the lexicons of the English language database. It’s collecting the words into a group of synonyms named syn-sets, provides general, short info and register a different semantic relationship between these sets of synonyms. Especially WordNet is suitable for similarity procedures, because it regulates nouns and

*Rowaida Khalil Ibrahim, Email: rowayda.gravi@gmail.com

verbs in the hierarchy of "is a" relation [16]. Several tools for finding semantic similarity between concepts and words have been proposed in our literature but most of them have been used WordNet. For this paper we reviewed many papers that used semantic similarity but we choose only thirteen systems in these five recent years that using semantic similarity for clustering propose and we give some important information of each of them. Our paper organized as follow: in section 2 it's a literature review of papers, in section 3 we give table of discussion of all that systems we reviewed before, then we conclude all work in section 4.

2. Literature Review

The main framework of the approach addressed by [17] consists of three modules, first one is Page Counts based measures, second Snippets and third is taxonomy-based approach. for page count four co-occurrence standard measures are used namely (WebOverlap, WebJaccard, WebPMI and WebDice), after the page count of two words are calculating then checking the context of those two words locally by using snippets based on probability measure to find out the pattern exist between both words. Also, they have used taxonomy-based measure (LEACOCK and CHODOROW (LCH) [18], WU & PALMER (WUP) [19], and RESNIK [20]) that using WordNet for finding and computing similarity between words (X and Y). then for combination the result using tow class of Support vector Machin (SVM and libSVM). The term of dataset used for evaluation, for training propose using MEN dataset [21] Selecting synonymous word pair at random from dataset and create non-synonymous for each one. However, for testing propose using MC approach [22], and RG [23], WordSimilarity-353 (WS) [24]). And using WEKA [25] for testing and training of the SVM with suitable demonstrating approach LibSVM and the qualified prototypical via determining the division ratio is about 70.51%. The results of accuracy were obtained near by 72%.

As explained in [26], the approach in this paper at first finding similarity between tow documents then extend to between two document sets, after that using clustering and classification to group the documents based on similarity measures .the proposed measure of similarity called (similarity measure for text processing) SMTP and the classification algorithm used are KNN based (SL-KNN) single label classification, KNN based (ML-KNN) multi label classification and for clustering algorithm using K-mean Clustering and (HAC) Hierarchical Agglomerative Clustering. The result checks the SMTP similarity measure's effectiveness that applying in text application (SL-KNN, ML-KNN, K-mean and HAC). Also, Compare SMTP performance with other five metrics, Cosine, Euclidean, IT-Sim, Extended Jaccard (EJ) and Pairwise-adaptive (Pairwise) by using different k value and using evaluating measures like (AC (Accuracy), EN (Entropy)). It shows that the similarity measure usefulness may rely on: 1) the format of feature, e.g., tf-idf or word count; 2) application of domains, e.g., image or text; and 3) classification or clustering algorithm. In addition, the term of datasets which are used for evaluating are three different datasets WebKB download from internet, Reuters-8 [27], and RCV1 [28]. Finally, in each case the result show the proposed measure SMTP have better performance.

In [29], the system approach for clustering the text document is an ensemble approach by using concept from Wikipedia. Two kind of clustering used, First approach is Partitional Clustering (lexical document clustering (LDC)) have three main phases are (Clustering term, finding documents lexical seed and Clustering text documents), in phase1 for the clustering purpose the fuzzy c-means algorithm is used to collecting the columns of matrix the (document-term matrix) hence the Fuzzy c-means collect the document-terms into k term clusters then in phase2 extract the documents representative, that are using as seeds in order for clustering all documents later, finally in phase3 based on those document seeds of each cluster term the centroid of document will be computed. Then the distances among the centroids and documents are used for clustering the documents. Second approach is Ensemble clustering ELSDC (Ensemble Lexical-Semantic Document Clustering) the main phases are (clustering term and topic key term selection, lexical seed documents extraction, find documents seeds, tagged and by using the consensus method collecting the documents). In phases 1 and 2, is same as of (LDC) but in phases 3 and 4, BOC and BOW are used for documents representation. The relevant concepts extract from Wikipedia by wikify module. The Naive Bayes classifier is the component of consensus method, by using documents in the same cluster in both clustering they are trained. The final cluster is generated after the remaining document are classified. However, the feature selection Var-TFIDF method are used for both approaches.

The experimental result has two rounds, in first round find comparison between LDA model and lexical document clustering (LDC) algorithm, the LDA's main idea is that the document can be rate as a distribution probabilistic on the underlying themes seen each subject as the terms of probability distribution. The C++ implementation are used for the LDA model; in this round It has been shown that the LDC can generate results similar to the LDA model on some real text data sets based on the representation of the duration document. In second round of experiments, find comparison between LDC and proposed ensemble algorithm (ELSDC). Show the benefits of incorporating Wikipedia concepts into document clustering. For evaluation proposed using Normalized Mutual Information (NMI) hence It measures the amount of information obtained from the categories given to a group of clusters, and has a maximum value of one clustering process when re-create the layers perfectly and the minimum it is zero. Also, for experiment propose eight datasets are used (20Newsgroups (News-sim3, News-rel3 and 20ng-whole), Reuters-21578, Classic4 is created from SMART data repository, WebKB, SMS Spam Collection and Cade is gathered from the content of Brazilian web pages).

As show in [30], extracting document feature and document vectorization to find similarity between Arabic webpages and showing the semantic annotation. The main steps in proposed approach are, in the first step Extraction of main class features using (Arabic VerbNet). The second one is document vectorization according to semantic feature by two solutions (semantic class probability distribution or Semantic class density) and in third step doing clustering and annotation. the model is work as follow the annotated document is take to be classified and all opinion word will be recognized by using semantic feature extraction and all the word will be aggregate to give semantic annotation to the document. The result gathered tow experiments, at the first one is

showing comparison between clustering only using K-mean (standard) and clustering using K-mean with semantic class density however, k-mean with semantic probability distribution and this is done by using evaluation measures (purity, MICD and DBI) hence the purity show that the clustering with vectorization signifies with high score, MICD tend that proposed model is out perform than standard k-mean and DBI show that proposed model appear to have smaller value than standard is mean that is better performance. The Runtime of standard k-mean is longer than the time consumed for tow solutions k-mean with vectorization. The second experiment showing annotation by using Mean score among cluster with both vectorization solution and Mean ratio among corpus (MRAC) also with two solution of vectorization. However, the datasets used for evaluation propose are collected corpus from the archives of online Arabic newspaper and VerbNet download from internet.

In [31], the approach of semantic clustering the text documents by using lexical chain and WordNet. The proposed approach in this paper are work as follow, at first the polysemy and synonymy are two main problem that effect the representation of text, this problem is solve by using WordNet based on Word sense disambiguation, after that introducing the lexical chain to capture the main theme of text and find the relationship between the word sense, then is show that the method can find a true number of clustering which is useful for finding the number of k that use in K-mean clustering algorithm. Finally, for experimental propose the paper show the comparison between all three (Base, DC, DCS) and with another system like ASG03 [32], LMJ10 [33] and CSF11 [34], hence Base is mean (all nouns) the system use all basic preprocessing techniques, i.e. term set extraction, stop words removing, stemming word and identification of nouns from term set but without performing WSD. However, DC is (disambiguated concepts) corresponding to the Base, but with preforming WSD and finally DCS (disambiguated core semantics) also corresponding to the DC, but adds the core semantics extraction process and is the proposed method.

In addition, in this paper for evaluation propose using Reuters-21578 corpus dataset and the evaluation measure using purity, entropy and F1-measure. Finally, the result show that in all experimental scheme the LMJ10 is the worst and the term of dimensionality reduction the feature account that derived from ASGO3 is lower than the number of base and the reduction of DSC is between 10% and 40% hence that the CSF11 greater than 74%, as well as the cluster quality obtained using the core semantic feature is better than using all nouns and using disambiguated concept (or at least comparable to) however the performance of using the disambiguation concepts is better than using all nouns, This suggests that the proposed disambiguation standard can solve the problem in a highly commendable and volatile manner, improve quality to a certain extent, and that the features of the semantic strings produced by the DCS approach do not only reduce the number of semantic concepts without losing much information, also adequately the main topic of the document that helps the clustering.

As [9], the model of clustering proposed in this paper incorporate Coreference resolution and exploit semantic relationship among the words by tackling polysemy and synonymy problem using WordNet and semantic similarity. The proposed approach consists of five modules are (Coreference Resolution, Preprocessing, Synonymy Identification and Sense disambiguation, Feature Selection and Bisecting k-means) modules. the purpose of those five model are, at first Coreference Resolution is the process of identifying Coreference words that occur in the document, then the propose of preprocessing model to transform the document in more suitable form and this is done in three steps (POS tagging ,Stop word Elimination and stemming),after that Sense Disambiguation and Synonym Identification deal with polysemy and synonymy problem in document and the model use WordNet for this propose, then the feature selection done by weighing the words in document using tf-idf, finally the documents are clustered by using Bisecting K-mean. For experimental propose the paper using four classic Datasets (CACM are 60 Documents, CISI are 44 Documents, CRAN are 44 Documents, and MED consists of 52 Documents) and for evaluating the quality of cluster using purity. The result show comparison between Base and proposed model hence the Base is the model without (the Coreference resolution, sense disambiguation and pruning term), after comparison the term of purity result show that the proposed model is observed that achieve 30% of improvement in clustering purity, that the purity rate of Base configuration is 0.55 hence the proposed model is 0.8.

As [35], Based on style Similarity and structure Clustering Web Pages, the model which used in this paper is DOM (Document Object Model) tree. The approach using two main measures, one of them for structural similarity to find similarity on DOM using TED (Tree Edit Distance) for HTML pages and another one using (Jaccard similarity measure) for CSS which called Stylistic similarity on style sheet information, after that based on those similarity measures using (Near neighbor clustering technique) for clustering propose, however to combine this information using Jarvis and Patrick method (shared nearest neighbors) [36]. It is a respectable feature aimed at aggregating Internet pages built for pairs similarity metrics. For evaluating propose using different threshold value of near neighbor at first using 90% edge for resemblance toward luxury near neighbors' pamphlets and 90% threshold for communal close towards combine clustering then altered the threshold to 95% until found the optimal threshold. The result show that when the threshold was 90% total number cluster was 12, and when it was 95% the total was 24 cluster as well as when it was 85 the total number of clusters was only 5, so after few trails they found that for test dataset the optimal threshold for both resemblances and intersection for communal close parties was near by 85%. As it declares that the result is positive as early evaluation.

As [37], the approach based on Hierarchical agglomerative clustering method (HAC) that is based of conceptual annotation of documents. The key steps of approach are: (1) pairwise calculation of the initial similarity matrix (2) Explanation of newly cluster and updating the matrix of similarity (3) post processing of HAC tree result. The approach chooses to rely on Lin Measure for pairwise similarity and on Base Match Average (BMA) for GroupWise semantic similarity, the proposed method uses GroupWise semantic similarity to compute the pairwise similarities between document for creating the label-base similarities matrix. When two clusters are agglomerative a new cluster are creating then automatically compute the annotation for this new cluster and then the label-based matrix are iteratively update, finally using

postprocessing of the tree hierarchy to take the branch lengths advantage in the tree to only keep the most meaningful cluster. For evaluation propose using (HSC heavy semantic clustering and LSC light semantic clustering) approaches for comparison with baseline (classical HAC with naive cluster annotation) by using Normalized Robinson-Foulds Distance measure hence the lower the distance the higher resemblance. The result after comparison between approaches (HSC, LSC and baseline) the HSC and LSC clearly outperform the baseline but for runtime the baseline was faster. then it shows the comparison of approaches (LSC, Baseline) with postprocessing and without post processing (pp and nopp) show that both are bad without postprocessing, finally it evaluating cluster labeling between our clustering labeling (CL), naive merge annotation (baseline) and expert also, the result show CL is better. And the dataset which used for evaluating propose was derived from WordNet-based disambiguation version of the aforementioned bookmark annotation [38] however, all resources and result of this paper are available at [39].

As [40], the approach for semantic document clustering based on the graph similarity. Graph is mainly used information from WordNet to the degree of semantic similarity between 150,000 of the most common in English terminology. The proposed approach at first Adds all the documents in to the similarity graph, then the distance between a pair of documents is measured by evaluating the paths between them, where a path in the graph can go through several terms that are semantically similar. The idea is to creating a node for each document and connecting this node to the graph, then to find the semantic similarity between the two documents will be measured by calculating the distance between the two nodes, and using three choices for the distance metric: the cosine, linear or logarithmic. Finally, for clustering propose using K-mean algorithm to cluster a set of documents and the algorithm relies on a way for computing the distance between two documents. The term of experimental result, the dataset of documents from Reuters-21578 benchmark and using F-measure for evaluating. The end of evaluating after applying the K-mean algorithm in three different distance metrics (cosine, linear and logarithmic) at first the result show that Using the similarity graph can lead to both higher precision and recall. And with Noting that using the linear or logarithmic similarity metric did not make a difference. The reason is that the two metrics apply different monotonic functions on the average of the sum of the forward and backward paths. Applying these monotonic functions has no effect on the ordering of the distances between nodes and on the clustering result.

In [41] paper the approach named (An Ontology-based and Domain Specific Clustering Methodology for Financial Documents). The main steps in approach are (preprocessing and feature extraction, sense word disambiguation, representation document and clustering). The approach works as follow: at step1 the preprocessing and data extraction is responsible to extract text from document in order to transform word in to more meaningful form and this is done in three steps (Stop Word Removal, Noun Extraction and Lemmatization). In step2 word sense disambiguation the correct sense for the noun are identified and this is done by using (WordNet ontology/database and an information content file) and in this study two disambiguation techniques are used which used different external information to remove ambiguity, used one technique (Brown Content File) which is the default for measuring Resnik and other technical used (file content financial information) that is proposed in this study. In step3 the document representation the features are represented as term frequency (TF or tf) vector to be prepare for clustering. Then, in step4 clustering is the final step which clusters the document vector by using algorithm and in this approach two kind of algorithms are studies (K-mean and sequential Information Bottleneck).

The dataset which used in this study was downloaded from EMMA [42] online repository using 446 documents, EMMA documents are classified under the US jurisdiction and their sectors or purposes and have achieved three types of labeling settings according to the categories provided by EMMA (setting1: purpose at the same time called sector as the class label, setting2: class label as state and setting3: class label as mixing of both state and sector). Then for evaluating propose in this study using purity for external evaluation to evaluate the cluster performance. The result showing the purity rate of both algorithm K-mean and sIB based on three settings also showing comparison as follow: 1) comparison between all term and noun only, Although the names are selected, the number of features is reduced significantly, the purity values of both algorithms do not have a considerable difference; 2) comparison between tf-idf of nouns and without tf-idf, it shows that with tf-idf better purity; 3) comparison between no disambiguation and with disambiguation, both methods of disambiguation have yielded good results, but the finest result produced with one the used financial information content file. Moreover, based on experimental results, the sIB algorithm can be determined as the most appropriate algorithm for aggregation.

As explained in [43], The system called WMDC (Wikipedia matching document classification) several steps are used in this approach. First, select knowledge and concept from Wikipedia. Second, using heuristic selection to pick up related concept. Third, finding similarity between document using combination of (Semantic similarity based on Wikipedia machine and Textual similarity based on keywords matching), and choosing K-mean algorithm for classification in this approach because of its efficiency and accuracy. Also, the evaluation experiment in this paper are divided in to two part. First part, focus on effectiveness of heuristic selection Rules, hence three rules are used to pick up the related concept of word (Rule1: all title, Rule2: all keyword and Rule3: any keyword) for evaluating efficiency of those three rules using selectivity measure and evaluating quality by using relevance. Second part, evaluating the effectiveness of approach.

Using Purity for measuring the accuracy of document classification and for evaluating the efficiency of approach using (vector construction time). The datasets used in this paper are Wikipedia dataset which download directly from internet and published in 2011-10-7, and document dataset from Reuters-21578 divided in to 82 cluster by removing those cluster with documents less than 15 and more than 200 only 30 cluster reminding to enhance the experimental effect. The result in this paper show that after using 4 different thresholds 0,0.5,0.10 and 0.15 for evaluating all rules of heuristic selection given that the smaller rule of selectivity is better efficiency and higher rule of relevance is better quality. then as the overall given that the proposed approach can accurately find out the related concept for a given document. Then the same threshold used for purity and (vector construction time) for evaluating classification accuracy and efficiency it concluded

that the proposed approach can improve the efficiency of language and for the corresponding Wikipedia under the precondition of not compromising the accuracy classification of the document.

As in [44] the system called SEMHYBODC related to (Fuzzy C-Manners). The methodology of clustering works as follow: At first the documents are annotated by "KIM" plugin before clustering. Then the documents are clustered using FPSOFCM (FPSO+FCM) and concept weight is calculated. The term of weight is recalculated through the steps specified in SEMHYBODC algorithm. And the Accuracy of clustering is computed using measures such as cluster purity and compared with various other hybrid approaches. after evaluating by using F-measure and purity the result show that, using swarm intelligence for clustering is not suitable for large dataset because the computational time is more but it give better accuracy in order to solve this problem a hybrid approach combine with algorithms such as PSO and FCM, however the PSO is combine with K-mean it's also leads to optimal solution but the time for computational is still higher, then as finally result the FPSO+FCM it show improvement over another algorithms like K-mean, Fuzzy C-means and Hybrid, It is combined with the ability to search from the globalized and fast algorithm PSO convergence algorithm FCM. It is used as a result of the initial seed FPSO algorithm FCM, which is applied to the purifying with generate concluding outcome. For evaluating propose webpages are collected from internet.

As explain in [45] many methods of clusters are operate based on similarity between documents as well as we explained before in our literature but in this paper the semantic similarity used for clustering the articles. The approach which proposed in this paper are at the first step the semantics of articles based on participating the entities in these articles, build three vectors of representations for each article: One calculates the average vector for all entities, one also with all entities but excluding citations and the other with only citation entities. Once the article vectors are generated, the next step is to identify clusters of articles based on vectorization using (k-mean) Clustering and using the Louvain (Network-based clustering methods) for community detection. For experiments the both clustering methods are using the Astro dataset [46]. At first four solutions of clustering are collected, namely CWTSC5, STS-RG, UMSI0 and ECOOM-BC13 for the pseudo-ground-truth based on adjustF1 that the best K are choosing which gives a highest score of adjustF1 also using adjustedF1 score to evaluate the results of clustering as well from the Louvain method.

The result shows the quality scores of three cluster based on pseudo-ground-truth (no citation, only citation, all entities) and OCLC Louvain then gives the average Adjusted Mutual Information scores (AMI) between current solution and the all four other solutions named, STS-RG, UMSI0, CWTS-C5 and ECOOM-BC13. It indicates that, if citations are used only, the resulting combinations correspond to other solutions of clustering than those if not using citation whose adjusted F1 score is also the lowest. Not surprisingly, the other clustering solutions depend heavily on citation information. So, even if they use different methods of citations, the citation information is still bringing enough agreement between them. Use all entities to represent articles has highest score adjustedF1 and agree with most others. Also, find AMI results among these three solutions and groups based on the Louvain method. Again, the cluster based on citation only agree with the results of Louvain almost as much as using cluster that use all entities. In accordance with these measures, they decided to use all entities as a major choice of characteristics, and to maintain 31 sets as key results for K-Means, which is named OCLC-31.

3. Discussion

The table below show the survey of approaches that using semantic similarity based on clustering. The table give a number of papers each paper has specific method and different tools are used for each of this method as well as provide information of each paper given different options like Measures for similarity and evaluation propose and algorithms for clustering or classification also, provide datasets that used by each paper.

As show in the survey, most of the method using WordNet, the term of WordNet as we mentioned before in introduction is the lexicons of the English language database, as well as in information system the WordNet was used for a number of proposals, for example (information retrieval, word sense disambiguation, text classification automatically, summarization of text automatically, machine translation and also can used for crossword puzzle generation automatically) [47] but one of the WordNet most common application is using to determine the similarity between words as we seen in all that method which are using WordNet the main propose was used for (word sense disambiguation) that deal with polysemy and synonymy problem. However, rather than WordNet there are another tool are used like Wikipedia or Kim Plugging also, there are method based on clustering, similarity graph or even network. Also, most of people using K-mean algorithm for clustering hence it's a must popular clustering algorithm that grouping the documents using a similarity metric that is based on keywords matching, and its useful for huge variables for this reason most of time computationally the k-mean is faster than hierarchical clustering but if keeping (K) small, and another advantage is it can make a tighter cluster [48]. However, many of research using bisecting K-mean because it can deal with large dataset also its very satisfactory quality of clustering with low cost. And another algorithm used as shown in surveys like fuzzy c-mean, HAC, Naïve base classifier, near neighbor, sIB, Louvian method and using PSO with clustering algorithm for optimization propose.

As it can be seen that other measures are used like tf-idf its mean (term frequency-inverse document frequency) it's using as weighting factor the main job is determine how much important the word is to the document in corpus or in group [49], as we seen in survey most of paper used this measure for feature selection. Also, other measures are used like Var-TFIDF, heuristic selection etc. for evaluation and experimental result many types of dataset are used and there are specific measures that using for evaluation purpose most popular one is Purity hence its very simple and primary evaluation measure that using for validation to determine the quality of cluster [50], however there are another one like F-measure, Entropy etc. the main propose are to determine the performance of cluster.

Table 1: Summery of Approaches Using Semantic Similarity Based On Clustering

No.	papers	Date	Method	classification and clustering Algorithm	Evaluation measures	Other Measures	Dataset
1.	L. Sahni, et al. ,[17]	2014	Semantic similarity between words using WordNet	SVM	Training the model by determining a split ratio of about 70.51%.	using 4 co- occurrence measures, probability measure and measures-based approach (LCH, WUP and Resnik)	using MEN dataset for training and 3 Benchmark datasets for testing propose
2.	Y. S. Lin et al. ,[26]	2014	Text classification and clustering	SL-KNN, ML-KNN, k-means and HAC	Accuracy and Entropy in (tf-idf, word count, Z-score)	SMTP proposed measure	Three data sets namely RCV1, WebKB and Reuters-8
3.	S. Nourashrafeddin et al. ,[29]	2014	Document clustering using Wikipedia	fuzzy c-means and Naive Bayes classifier	Normalized Mutual Information (NMI)	Var-TFIDF	using 8 datasets 1-20Newsgroups (News-sim3, News-rel3 and20ng-whole). 2-Reuters-21578. 3- from data repository SMART creating Classic4. 4-WebKaB. 5-SMS Spam Collection 6- Cade from Brazilian web pages content is gathered
4.	H. M. Alghamdi et al ,[30]	2014	Arabic VerbNet	K-Mean	Purity, Mean intra cluster distance (MICD) and Davis Bouldin Index (DBI)	using Mean score among cluster, Mean ratio among corpus (MRAC) for Annotation	Corpus collected from online Arabic newspaper archive
5.	T. Wei et al. ,[31]	2014	Text clustering Using WordNet version 2.0	Bisecting k-mean	Purity, Entropy and F1-measure	Using Lexical chains	Reuters-21578 corpus
6.	S. S. Desai and J. A. Laxminarayana ,[9]	2016	Document clustering Using WordNet	Bisecting K-mean	Purity	Tf-idf for feature selection	using 4 classic Datasets 1-60 Documents from CACM, 2- 44 Documents from CISI, 3-44 Documents from CRAN, 4- 52 Documents from MED
7.	T. Gowda and C. Mattmann ,[35]	2016	Webpage clustering using (DOM)tree	Near Neighbor Clustering technique	using different value of threshold	TED and JS	Dataset from a popular weapons classifieds site
8.	N. Fiorini et al. ,[37]	2016	Semantic clustering using WordNet	HAC	Normalized Robinson-Foulds Distance	lin Measure, BMA, post processing	Derived from WordNet-based disambiguation version
9.	L. Stanchev ,[40]	2016	Document clustering Based on similarity Graph	K-mean	F-measure	cosine, linear, and logarithmic	Reuters-21578 benchmark
10.	C. Kulathunga and D. D. Karunara,[41]	2017	Document Clustering Using WordNet and information content file	K-mean and sIB	Purity	Tf-idf	EMMA
11.	Z. Wu et al. ,[43]	2017	Document clustering Using Wikipedia	K-mean	Selectivity, relevance, purity and vector construction time	Heuristic selection	Wikipedia Dataset published from internet and document dataset from Reuters-21578
12.	J. Avanija et al. ,[44]	2017	Document Clustering using KIM plugging tool	Fuzzy C-mean with PSO and K-mean with PSO	Purity and F-Measure	Tf-idf	Web pages collect from internet
13.	S. Wang and R. Koopman ,[45]	2017	Article clustering based on Network	K-means and Louvain method	F1-measure and adjustedF1	vectoring	Astro dataset

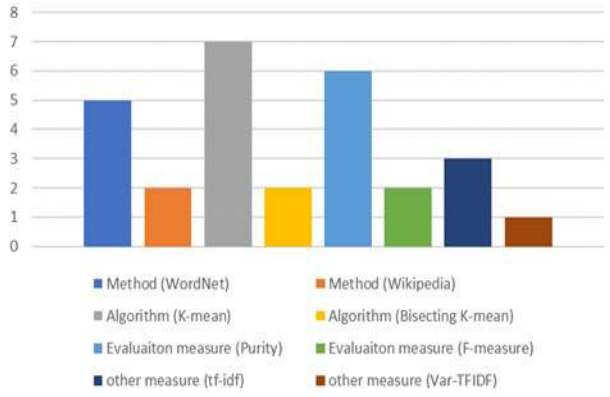

Figure 1: Showing the usage number of main features (Methods, Algorithms and measures).

4. Conclusion

In this paper, after we reviewed all that papers, we conclude that each paper has specific approaches and various tools and measures are available for each approach, and as it can be seen that most popular steps are used in approaches were preprocessing to transform the document in better format and different steps are used in preprocessing like removing stop words, stemming, tokenization etc., word sense disambiguation are used for solving the synonymy and polysemy problems and feature selection another important step that many of approaches was using tf-idf for this case and there was another like Var-TFIDF, heuristic selection etc. however one of the most popular tools used is WordNet the English dataset for meaningful clustering, then clustering done by using clustering algorithms also the most usage one was K-mean algorithm because of its simplicity to use and can give tighter cluster as well as there is another like bisecting K-mean, fuzzy c-mean, hierarchical agglomerative clustering etc. finally, as we see that the main goal of all approaches trying to a chive a better efficiency, accuracy and quality of clustering. after using the experimental measure for evaluating propose like (Purity, F-measure, Entropy etc.) in all case show that the semantic clustering giving a better performance.

References

[1] B. Everitt and Wiley InterScience (Online service), Cluster Analysis. Chichester, West Sussex, U.K: Wiley, 2011.
[2] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and S. R. M. Zeebaree, "Combination of k-means clustering with genetic algorithm: A review," Int. J. Appl. Eng. Res., vol. 12, no. 24, pp. 14238–14245, 2017.
[3] W. L. Chang, K. M. Tay, and C. P. Lim, "A New Evolving Tree-Based Model with Local Re-learning for Document Clustering and Visualization," Neural Process. Lett., vol. 46, no. 2, pp. 379–409, 2017.
[4] K. Jacksi, N. Dimililer, and S. R. M. Zeebaree, "a Survey of Exploratory Search Systems Based on Lod Resources," Turkey. Univ. Utara Malaysia, no. 112, pp. 501–509, 2015.
[5] H. H. Tar and T. Nyunt, "Enhancing Traditional Text Documents Clustering based on Ontology," Int. J. Comput. Appl., vol. 33, no. 10, pp. 38–42, 2011.
[6] Z. Chengzhi and X. Hongjiao, "Clustering description extraction based on statistical machine learning," Proc. - 2008 2nd Int. Symp. Intell. Inf. Technol. Appl. IITA 2008, vol. 2, pp. 22–26, 2008.
[7] K. Jacksi, S. R., and N. Dimililer, "LOD Explorer: Presenting the Web of Data," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, pp. 45–51, 2018.
[8] K. Jacksi, N. Dimililer, and S. R., "State of the Art Exploration Systems for Linked Data: A Review," Int. J. Adv. Comput. Sci. Appl., 2016.
[9] S. S. Desai and J. A. Laxminarayana, "WordNet and Semantic similarity-based approach for document clustering," 2016 Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2016, pp. 312–317, 2016.
[10] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic Similarity from Natural Language and Ontology Analysis," Synth. Lect. Hum. Lang. Technol., vol. 8, no. 1, pp. 1–254, May 2015.
[11] Y. Feng, E. Bagheri, F. Ensan, and J. Jovanovic, "The state of the art in semantic relatedness: a framework for comparison," Knowl. Eng. Rev., pp. 1–30, Mar. 2017.
[12] F. M. Couto and A. Lamurias, "Semantic Similarity Definition," in Encyclopedia of Bioinformatics and Computational Biology, Elsevier, 2019, pp. 870–876.
[13] A. Ballatore, M. Bertolotto, and D. C. Wilson, "An evaluative baseline for geo-semantic relatedness and similarity," Geoinformatica, vol. 18:4, pp. 747–767, 2014.
[14] "Semantic similarity." [Online]. Available: https://en.wikipedia.org/wiki/Semantic_similarity. [Accessed: 27-Nov-2018].
[15] A. A.-Z. K. J. A. S. Subhi R. M. Zeebaree, "Designing an Ontology of E-learning system for Duhok Polytechnic University Using Protégé OWL Tool," J. Adv. Res. Dyn. Control Syst., vol. Volume 11, no. 05-Special Issue, pp. 24–37, 2019.
[16] M. A. Hadj Taieb, M. Ben Aouicha, and A. Ben Hamadou, "Ontology-based approach for measuring semantic similarity," Eng. Appl. Artif. Intell., vol. 36, pp. 238–261, 2014.
[17] L. Sahni, A. Sehgal, S. Kochar, F. Ahmad, and T. Ahmad, "A Novel Approach to Find Semantic Similarity Measure between Words," Proc. - 2014 2nd Int. Symp. Comput. Bus. Intell. ISCBI 2014, pp. 89–92, 2015.
[18] C. Leacock and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification," WordNet An Electron. Lex. database., no. January 1998, pp. 265–283, 1998.
[19] W. Zhibiao and M. Palmer, "VERB SEMANTICS AND LEXICAL Zhibiao W u," Proc. 32nd Annu. Meet. Assoc. Comput. Linguist., pp. 133–138, 1994.
[20] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy."
[21] K. Sparck Jones and P. Willett, Readings in information retrieval. Morgan Kaufman, 1997.
[22] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," Lang. Cogn. Process., vol. 6, no. 1, pp. 1–28, Jan. 1991.
[23] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," Commun. ACM, vol. 8, no. 10, pp. 627–633, Oct. 1965.
[24] L. Finkelstein et al., "Placing search in context: the concept revisited," ACM Trans. Inf. Syst., vol. 20, no. 1, pp. 116–131, 2002.
[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update."
[26] Y. S. Lin, J. Y. Jiang, and S. J. Lee, "A similarity measure for text classification and clustering," IEEE Trans. Knowl. Data Eng., vol. 26, no. 7, pp. 1575–1590, 2014.
[27] "Datasets for single-label text categorization - Ana Cardoso Cachopo's Homepage." [Online]. Available: http://ana.cachopo.org/datasets-for-single-label-text-categorization. [Accessed: 05-Dec-2018].
[28] D. D. Lewis, Y. M. Yang, T. G. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research," J. Mach. Learn. Res., vol. 5, pp. 361–397, 2004.
[29] S. Nourashrafeddin, E. Milios, and D. V. Arnold, "An ensemble approach for text document clustering using Wikipedia concepts," Proc. 2014 ACM Symp. Doc. Eng. - DocEng '14, pp. 107–116, 2014.
[30] H. M. Alghamdi, A. Selamat, and N. S. Abdul Karim, "Arabic web pages clustering and annotation using semantic class features," J. King Saud Univ. - Comput. Inf. Sci., vol. 26, no. 4, pp. 388–397, 2014.
[31] 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, 2015.
[32] H. Andreas, S. Staab, and G. Stumme, "Wordnet improves Text Document Clustering," 2003.
[33] L. Jing, M. K. Ng, and J. Z. Huang, "Knowledge-based vector space model for text clustering," Knowl. Inf. Syst., vol. 25, no. 1, pp. 35–55, Oct. 2010.
[34] S. Fodeh, B. Punch, and P.-N. Tan, "On ontology-driven document clustering using core semantic features," Knowl. Inf. Syst., vol. 28, no. 2, pp. 395–421, Aug. 2011.
[35] T. Gowda and C. Mattmann, "Clusteringweb pages based on structure and style similarity," Proc. - 2016 IEEE 17th Int. Conf. Inf. Reuse Integr. IRI 2016, pp. 175–180, 2016.
[36] R. A. Jarvis and E. A. Patrick, "Clustering Using a Similarity Measure Based on Shared Near Neighbors," IEEE Trans. Comput., vol. C–22, no. 11, pp. 1025–1034, Nov. 1973.
[37] N. Fiorini, S. Harispe, S. Ranwez, J. Montmain, and V. Ranwez, "Fast and reliable inference of semantic clusters," Knowledge-Based Syst., vol. 111, pp. 133–143, 2016.
[38] R. Bernardi and I. AT4DL 2009 (2009: Trento, Advanced language technologies for digital libraries: international workshops on NLP4DL 2009,

Viareggio, Italy, June 15, 2009 and AT4DL 2009, Trento, Italy, September 8, 2009. Springer, 2011.
[39] "SC - Semantic clustering." [Online]. Available: http://sc.nicolasfiorini.info/. [Accessed: 13-Dec-2018].
[40] L. Stanchev, "Semantic Document Clustering Using a Similarity Graph," Proc. - 2016 IEEE 10th Int. Conf. Semant. Comput. ICSC 2016, pp. 1–8, 2016.
[41] C. Kulathunga and D. D. Karunaratne, "An -ontology-based and domain specific clustering methodology for financial documents," 17th Int. Conf. Adv. ICT Emerg. Reg. ICTer 2017 - Proc., vol. 2018-Janua, pp. 209–216, 2018.
[42] "Municipal Securities Rulemaking Board:EMMA." [Online]. Available: https://emma.msrb.org/. [Accessed: 13-Dec-2018].
[43] Z. Wu et al., "An efficient Wikipedia semantic matching approach to text document classification," Inf. Sci. (Ny)., vol. 393, pp. 15–28, 2017.
[44] J. Avanija, G. Sunitha, and K. R. Madhavi, "Semantic Similarity based Web Document Clustering Using Hybrid Swarm Intelligence and FuzzyC-Means," vol. 7, no. 5, pp. 2007–2012, 2017.
[45] S. Wang and R. Koopman, "Clustering articles based on semantic similarity," Scientometrics, vol. 111, no. 2, pp. 1017–1031, 2017.
[46] J. Gläser, W. Glänzel, and A. Scharnhorst, "Same data—different results? Towards a comparative approach to the identification of thematic structures in science," Scientometrics, vol. 111, no. 2, pp. 981–998, May 2017.
[47] "WordNet Application."
[48] Santini, "Advantages & Disadvantages of k- - - Means and Hierarchical clustering (Unsupervised Learning)," pp. 1–5, 2016.
[49] A. Rajaraman and J. D. Ullman, "Data Mining," in Mining of Massive Datasets, Cambridge: Cambridge University Press, 2011, pp. 1–17.
[50] S. C. Sripada, "Comparison of Purity and Entropy of K-Means Clustering and Fuzzy C Means Clustering," Indian J. Comput. Sci. Eng., vol. 2, no. 3, pp. 343–346, 2011.