Semantic-based Web Page Clustering System using K-means Algorithm

Ei Ei Moe[#1], Hnin Hnin Htun[*2], Aye Mon Yi[#3]
[1,2,3]*Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City) PyinOoLwin, Myanmar*

***Abstract*—Web page clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing large sets of web pages into a small number of meaningful clusters. Clustering is a very powerful data mining technique for topic discovery from web pages. Partition clustering algorithm, K-means, is reported performing well on web page clustering. They treat the clustering problem as an optimization process of grouping web pages into k clusters so that a particular criterion function is minimized or maximized. The bag of words representation used for these clustering is often unsatisfactory as it ignores relationships between important terms that do not co-occur literally. So, this system proposes the semantic based web page clustering system by using both the K-means clustering algorithm and word sense disambiguation method. For semantic analysis, this system uses the WordNet as lexicon and K-nearest neighbor (KNN) classifier as the word sense disambiguation method. Based on the semantic logic, this system supports the more accurate about the web page clustering process.**

***Keywords*— Semantic, Word Sense Disambiguation, WordNet, K-means, Clustering.**

I. INTRODUCTION

Clustering is one of the most important unsupervised learning problem. In the clustering process, objects are organized into groups of similar members. Hence, a cluster is a collection of objects which are similar to each other but dissimilar to the objects of other c1usters. A text clustering divides a collection of web pages into different category groups so that web pages in the same category group describe the same topic. A text clustering is the elemental function in the process of text mining. Automated web page processing can include operations such as web page comparison, web page categorization, and web page selection.

Web page clustering has very significant uses in many areas of data mining and information retrieval. Clusters of web pages are generated automatically from the collection of web pages. In traditional method of web page clustering, single, unique, or compound words of the document set are used as features. But, the traditional method doesn't consider semantic relationships into account. The problems such as the synonym problem and the polysemous problem exist in the traditional method; therefore, a bag of original words can't represent the exact content of a document and can't produce meaningful clusters. Therefore, to improve web page clustering, there is a need of clustering techniques that also consider meaning of words into clustering process.

So, the proposed system considers the semantic relationships to enhance the performance of web page clustering. For semantic analysis, this system uses the word sense disambiguation (WSD) method and WordNet. WSD method resolves the ambiguity by pointing which concept is represented by a word or a phrase in a context. Use of WordNet makes it easier to identify related concepts and their linguistic representatives given a key concept, whereas WSD tries to uncover the hidden conceptual relationships among the words. WordNet contains many set of synonym words of same concept and their relationships with different synsets. For clustering, this system uses the K-means clustering algorithm. By using both WSD and K-means methods, this system improves the performance of web page clustering.

The rest of the paper is organized as follows: related work is described in section II. Pre-processing for web page and term weighting are shown in section III and IV. Word sense disambiguation is presented in section V. Semantic web page clustering and clustering methods are expressed in section VI and VII. Proposed system design is presented in section VIII. Explanation of the system is described in section IX. Experimental results are shown in section X. Finally, conclusion is given in section XI.

II. RELATED WORK

In 2012, B. Drakshayani and E. V. Prasad[12] proposed a new model for text document representation. The proposed model followed parsing, preprocessing and assignment of semantic weights to Document phrases to reflect the semantic similarity between phrases and k-means clustering algorithm. They evaluated the proposed model using five different datasets in terms of F-Measure, Entropy, and Purity for K-means clustering algorithm. The results demonstrated a performance improvement compared to the traditional vector space model and latent semantic indexing model. More natural language processing (NLP) techniques may be included to enhance the performance of the text document clustering.

In 2012, J. S. Priya and S. Priyadharshini[13] proposed the

semantic clustering and feature selection method to improve the clustering and feature selection mechanism with semantic relations of the text documents. They also proposed a new text clustering algorithm TCFS, which stands for Text Clustering with Feature Selection. TCFS can incorporate the supervised feature selection method to identify relevant features (i.e., terms) iteratively, and the clustering becomes a learning process. This system is designed to identify the semantic relations using the ontology.

In 2017, P. Gurung and R. Wagh[11] presented the application of k-means algorithm for document clustering that was experimented with two different but related data sets. Experimental results showed that document clustered in abstract corpus and full text corpus differs in many ways. Topic identification and text clustering are two very important tasks in information retrieval domain. Availability of big documents on web impacts the result of these processes as demonstrated in this paper. The results thus highlighted the need for more robust methods for documents with more number of terms. The results emphasized on more robust methods like semantic based methods for big document collections.

Based on the previous researches, this system proposes the semantic based web page clustering system by using both the k-means clustering algorithm and WordNet that is used for semantic analysis.

III. PREPROCESSING FOR WEB PAGE

Preprocessing for web page includes the tokenization, stop words elimination and stemming. These are as follows:

- Tokenization: The data must be processed in the three operations: the first operation is to convert document to word counts which is equal to bag of word (BOW). The second operation is removing empty sequence i.e. this step comprises cleansing and filtering (e.g., whitespace collapsing, stripping extraneous control characters). Finally, each input text document is segmented into a list of features which are also called (tokens, words, terms or attributes)[4].
- Stop words elimination: A stop words list is a list of commonly repeated features which emerge in every text document. The common features such as conjunctions such as or, and, but and pronouns he, she, it etc. need to be removed due to it does not have effect and these words add a very little or no value on the categorization process. For the same reason, if the feature is a special character or a number then that feature should be removed[4].
- Stemming: Stemming is the process of removing affixes (prefixes and suffixes) from features i.e. the process derived for reducing inflected (or sometimes derived) words to their stem. The stem need not to be identified to the original morphological root of the word and it is usually sufficiently related through words map to the similar stem. This process is used to reduce the number of features in the feature space and improve the performance of the clustering when the different forms of features are stemmed into a single feature [4].

IV. TERM WEIGHTING

Term weighting is used for enhanced text document presentation as feature vector. Term weighting helps to locate important terms in a document collection for ranking purpose [5]. TF-IDF is the most widely known and used weighting method [4].TF-IDF term weight is given as follows:

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{|v|j}\}} \quad (1)$$

$$idf_i = \log \frac{N}{df_i} \quad (2)$$

$$w_{ij} = tf_{ij} \times idf_i \quad (3)$$

In this scheme, N is total number of web pages in the system. The $df_i$is number of web pages in which term $t_i$appears at least once. The $f_{ij}$is the raw frequency count of term $t_i$in web paged_j. The $tf_{ij}$is the normalized term frequency. The $idf_i$is the inverse web page frequency of termt_i.

V. WORD SENSE DISAMBIGUATION

Word Sense Disambiguation (WSD) is the task of mappingan ambiguous word in a given context to its correct meaning. WSD also defines as the task of automatically assigning the most appropriate meaning to a polysemous word within a given context. WSD is an important problem in natural language processing (NLP), both in its own right and as a stepping stoneto more advanced tasks such as machine translation, information extraction and retrieval and question answering.

WSD can be divided into supervised approaches and knowledge-basedunsupervisedapproaches.In thesupervised approach, machine learning modelsare trainedby a corpus, in which the correct senses of ambiguous words are already annotated byhuman annotator. On the other hand, the knowledge-based unsupervised approachesutilizelexical knowledge bases (LKBs) such as a WordNet. These approaches have performed WSD by combining contextual Information and semantic knowledge on the LKBs [6].

A. KNN Classifier

K-nearest neighbor (KNN) classifier is very easy to implement and give fairly good results. Also KNN algorithm does not require any prior knowledge regarding data setfor classification. It performs classification purely on similarity (distance) basis [14]. KNN classifier algorithm is shown in Fig.1.

Algorithm: KNN Classifier
Input Parameters: Data set, k
Output: Classified test tuples
Step 1: Store all the training tuples.
Step 2: For each unseen tuple which is to be classified
A. Compute distance of it with all the training tuples using Euclidean distance method.
B. Find the k nearest training tuples to the unseen tuple.
C. Assign the class which is most common in the k nearest training tuples to the unseen tuple.

Fig. 1: KNN Classifier Algorithm

KNN algorithm is applicable in a number of fields such as pattern recognition, text categorization, finance, agriculture, medicine etc.

B. WordNet

Most WSD systems use a sense repository to obtain a set of possible senses for each word. WordNet is a comprehensive lexical database for the English language, and is commonly used as the sense repository inWSD systems. It provides a set of possible senses for eachcontent word (nouns, verbs, adjectives and adverbs) in thelanguage and classifies this set of senses by the POS tags.For example, the word "cricket" can have 2 possible nounsenses: 'cricket#n#1: leaping insect' and 'cricket#n#2: game played with a ball and bat', and a single possible verbsense, 'cricket#v#1: (play cricket)'. WordNet also contains information about different typesof semantic relationships between synsets. These relationsinclude hypernymy, meronymy, hyponymy, holonymy,etc. [7].

VI. SEMANTIC WEB PAGE CLUSTERING

Clustering is considered as one of the most important unsupervised learning problem. In the clustering process, objects are organized into groups of similar members. Hence, a cluster is a collection of objects which are similar to each other but dissimilar to the objects of other c1usters. A text clustering divides a collection of text documents into different category groups so that documents in the same category group describe the same topic. A text clustering is the elemental function in the process of text mining [1]. Web page clustering has very significant uses in many areas of data mining and information retrieval. Clusters of webpages are generated automatically from the collection of documents [2].

A. Semantic and Semantic Analysis

Semantics is concerned with the study of meaning. It focuses on the relation between signifiers like words, phrases, signs, and symbols. The meaning of semantic is related with the meaning in language or logic. It tries to recognize the meaning as an element of language and how it is constructed by language. Semantics looks at meaning in language in isolation, and in the language itself. Semantics checks the different ways in which meanings of words can relate to each other to understand the relationships between them. Sentences can semantically relate to one-another in different ways.

Semantic analysis is the process of relating syntactic structures from phrases, clauses, sentences, and paragraphs to their language-independent meanings. It involves removal of the features specific to particular linguistic and cultural contexts [3].

B. Semantic Clustering

Semantic clustering is a technique to develop relevant keywords by concentrating majorly on keywords and keyword phrases that are closely related and associative. Semantic clustering concerns with partitioning points of a data set into distinct groups (clusters) in a way that two points from one cluster are semantically similar to each other but two points from distinct clusters are dissimilar to each other [3].

C. Advantages of Semantic Clustering

Semantic document clustering has an important benefit of being able to remove irrelevant documents by recognizing conceptual mismatches. Advantages of semantic clustering are as follows:

- Latent semantic indexing (LSI) technique can achieve dynamic clustering on the basis of conceptual contents of documents.
- Clustering is the method to make groups of documents on the basis of their conceptual similarity therefore, it makes the task easier while working with unknown collection of unstructured text.
- LSI can carry out example based categorization as well as cross linguistic concept searching.
- LSI can also process random character strings. This technique is not limited to work only with words.
- It is proven that LSI is good solution for a number of conceptual matching problems. This technique can capture key relationship information containing casual information, goal oriented and taxonomic information.
- Information and Relationship discovery.
- Semantic information retrieval method has exploited the advantages of the semantic web to retrieve relevant data.
- Semantic information is used for improving evaluation measures like precision or recall in information retrieval system and clustering process.
- Common text clustering methods have poor capabilities in explaining its users why certain result is achieved. Because of this, the disadvantage is that these methods cannot relate semantically nearby terms as well as they cannot explain how the result clusters are related to one another [3].

VII. CLUSTERING METHODS

Clustering sets a set of objects and discovers whether there is some correlation between the objects. Clustering methods can be broadly divided as follows:

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Model-based methods.

Among them, this system uses the partitioning method for web page clustering. With partitional clustering, the algorithm generates a group of data non-overlapping subsets (clusters) such that each data object is in correctly one subset. These advances need choosing a value for the preferred number of clusters to be produced [4].

A. K-means Clustering

K-means clustering algorithm is one of the popular and simple clustering algorithms and used in various fields like medical, science finance, engineering etc. This algorithm is also a refined algorithm in which a collection of n objects is partitioned into k clusters, which are updated recursively until they concentrate into a regular partition [4]. K-means clustering algorithm [9] is shown in Fig. 2.

Algorithm: K-means clustering
Step 1: It partitions the data into k groups where k is predefined.
Step 2: Select k points by random as cluster centers (centroids).
Step 3: Assign a objects to their closest cluster centre according to the Euclidean distance function.
Step 4: Calculate the mean for all objects in each cluster.
Repeat step 2, 3 and 4 until the same points are assigned to each cluster.

Fig. 2: K-means Clustering Algorithm

In K-means, distance measure is used to group the data objects into group based on minimum distance [8].

B. Euclidean Distance

Distance or similarity measures are essential to solve many pattern recognition problems such as classification, clustering, and retrieval problems. The Euclidean distance or Euclidean metric is the "ordinary" straight-line distance between two points in Euclidean space.

$$d(q,p)=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\ldots+(q_n-p_n)^2} \quad (4)$$

where, $p = (p_1, p_2,\ldots,p_n)$ and $q = (q_1, q_2,\ldots,q_n)$ are two points in the Euclidean n-space [10].

VIII. PROPOSED SYSTEM DESIGN

This system is implemented as the semantic based web page clustering system to enhance the performance of traditional clustering system. This system first extracts the web pages from the web pages collection (database). This system consists of three parts that are semantic analysis process, terms weight calculation process and clustering process. Proposed system design is shown in Fig. 3.

In the first process, this system performs tokenization, stopword removal and stemming processes. Then, each token (words) are checked to distinguish this word that is ambiguous word or not. For this semantic analysis, this system uses the K-nearest neighbor (KNN) classifier and WordNet. After finishing disambiguation process, this system calculates the weight for each term (word and sense). For weight calculation, this system uses the term frequency - inverse document frequency (TF-IDF) method.

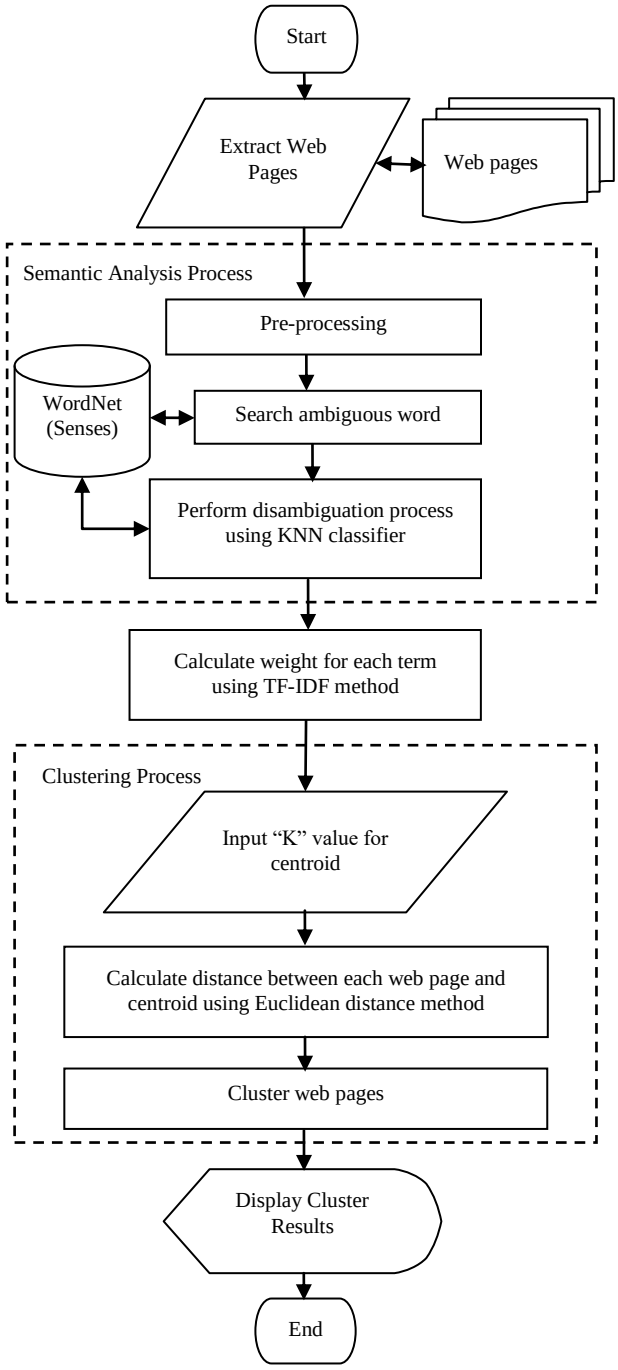

Fig. 3: Proposed System Design

Finally, this system performs the clustering process based on the K-means clustering logic. Using user desired "K" value, this system must produce the "K" clusters that is the user desired cluster numbers. For cluster, this system calculates the distance between each web page and centroid by using Euclidean distance method. This system clusters the same web pages until the cluster number is equal to the "K" value. Then, this system produces these cluster results to the user.

IX. Explanation Of The System

In the explanation of the system, five web pages are tested to show the effectiveness of semantic-based web page clustering. These five web pages are shown in Table I.

TABLE I: Ambiguous Web Pages

Web Page Name	Content
Web page 1	Classification is the act of distributing things into classes.
Web page 2	In categorization, a group of people is arranged by category.
Web page 3	Prediction is a statement about the future.
Web page 4	It is a forecast about how the weather will develop.
Web page 5	Classification is the basic cognitive process of distributing into classes.

After extracting web pages, this system performs the tokenization, stopwords removal and stemming processes. Then, this system disambiguates each ambiguous word in each web page by using KNN classifier and WordNet. Disambiguated web pages are shown in Table II.

TABLE II: Disambiguated Web Pages

Web Page Name	Content
Web page 1	classification, categorization, categorisation, compartmentalization, compartmentalisation, assortment, act, distribut, thing, category, family, class
Web page 2	categorization, classification, categorisation, group, people, arrang, category, class, family
Web page 3	prediction, foretell, forecast, prognostication, statement, future
Web page 4	forecast, prognosis, weather, develop
Web page 5	classification, categorization, categorisation, compartmentalization, compartmentalisation, assortment, basic, cognitive, process, cognitive, process, mental, process, operation, cognitive, operation, distribut, class, category, family

To calculate the distance between centroid and each web page, this system first calculates the weight for each term in each web page. The sample weight results are shown in Table III.

TABLE III: Weight Results

Web Page Name	classification	categorization	assortment	act	prediction	forecast	family
Web page 1	0.2218	0.2218	0.3979	0.6989	0	0	0.2218
Web page 2	0.2218	0.2218	0	0	0	0	0.2218
Web page 3	0	0	0	0	0.6989	0.3979	0
Web page 4	0	0	0	0	0	0.3979	0
Web page 5	0.0738	0.0738	0.1325	0	0	0	0.0738

For clustering, the user must define the “K” value. In this sample, user define K = 2. So, two centroids are randomly chosen and used for clustering.

- Centroid 1 =(0.2218, 0.2218, 0.2218, 0, 0, 0, 0, 0, 0, 0.2218, 0.2218, 0.2218, 0.6989, 0.6989, 0.6989, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
- Centroid 2= (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.6989, 0. 6989, 0.3979, 0.6989, 0.6989, 0.6989, 0, 0, 0, 0, 0, 0, 0, 0)

By using Euclidean distance measure method, this system calculates the distance between each centroid and each web page. The most similar web pages are grouped into each cluster. Cluster results are shown in Table IV.

TABLE IV: Cluster Results

Step	Cluster 1	Cluster 2
1	Web page 2	Web page 3
2	Web page 2, Web page 5	Web page 3, Web page 4
3	Web page 2, Web page 5	Web page 1, Web page 3, Web page 4

X. Experimental Results Of The System

To measure the performance of the semantic-based web page clustering system, this system is tested by using 300 web pages from the technology domain, sport domain and hazard domain. To evaluate clustering results, precision method is used.

$$\text{Precision } (i, j) = n_{ij} / n_i \quad (5)$$

where, n_{ij} is the number of members of class i in cluster j and n_i is the number of members of class i. Experimental and precision results are shown in Table V and Fig. 4.

TABLE V: Experimental Results of the System

Domain Name	Tested Web Pages	Correct Rate (Precision)	Error Rate
Sport Domain	100	88%	12%
Hazard Domain	100	85%	15%
Technology Domain	100	89%	11%

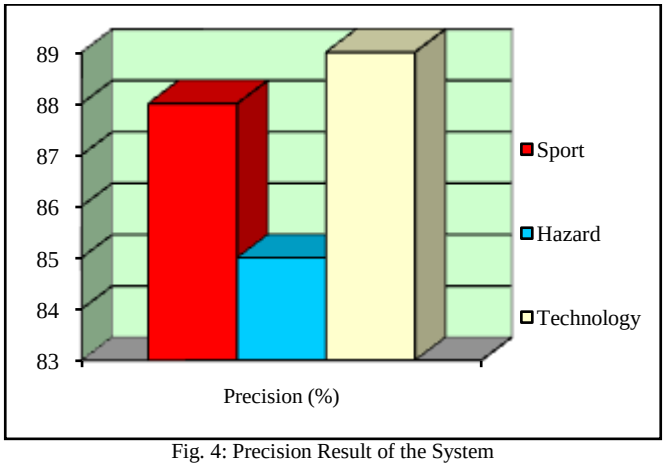

Fig. 4: Precision Result of the System

XI. CONCLUSION

In conclusion, this system described the semantic-based web page clustering system. This system allows the user to choose the desired cluster number according to the K-means clustering algorithm. This system also points out the effectiveness of WSD method for semantic analysis. WordNet is also used for getting meaningful clusters. This system disambiguated the ambiguous word in each web page for the effectiveness of clustering system. By disambiguating ambiguous words from different domains, this system points out the semantic that is effective for web page clustering system.

ACKNOWLEDGMENT

Firstly, the author would like to express her profound gratitude to Dr. Aung Win, for his invaluable directions and managements. The author would like to acknowledge the effective support of her honorable supervisor Dr. HninHninHtun and Dr. Aye Mon Yi as co-supervisor who have supported closely with her at every stage in her research work and valuable suggestions, true line guidance, supervision and editing this research.

REFERENCES

[1]. H. H. Tar and T. T. S. Nyunt, "Enhancing Traditional Text Documents Clustering based on Ontology", International Journal of Computer Applications (IJCA), vol. 33, no. 10, pp. 38-42, 2011.

[2]. B.S. Krishna, "Comparative Study of K-means and Bisecting K-means Techniques in Wordnet Based Document Clustering", International Journal of Engineering and Advanced Technology (IJEAT), pp. 229-234, 2012.

[3]. M. P. Naik, H. B. Prajapati and V. K. Dabhi, "A Survey on Semantic Document Clustering", IEEE, 2015.

[4]. A. I. Kadhim, Y. N. Cheah and N. H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering", 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, IEEE, pp.69-73, 2014.

[5]. M. Alhanjouri, "Pre-processing Techniques for Arabic Documents Clustering", International Journal of Engineering and Management Reserch, vol. 7, no. 2, pp. 70-79, April, 2017.

[6]. O. Dongsuk, K. Sunjae and K. Kyungsun, "Word Sense Disambiguation based on Word Similarity Calculation using Word Vector Representation from a Knowledge based Graph", Proceedings of the 27th International Conference on Computational Linguistics, pp. 2704-2714, August, 2018.

[7]. D. S. Chaplot and R. Salakhutdinov, "Knowledge-based Word Sense Disambiguation using Topic Models", Machine Learning Department, Carnegei Mellon University, 2018.

[8]. A. D. Khandare, "Modified K-means Algorithm for Emotional Intelligence Mining", International Conference on Computer Communication and Informatics, IEEE, January, 2015.

[9]. K. Nithya, R. Shanmugasundaram and N. Santhiyakumari, "Study of Salem City Resource Management Using K-Means Clustering", IEEE Conference on Emerging Devices and Smart Systems, pp. 79-83, March, 2017.

[10]. C. Sunghyuk, "Comprehensive Survey on Disstance/ Similarity Measures between Probability Density Functions", International Journal of Mathematical Models and Methods in Applied Sciences, no. 4, vol. 1, pp. 300-307, 2007.

[11]. P. Gurung and R. Wagh, "A Study on Topic Identification using K-means Clustering Algorithm: Big vs. Small Documents", Advances in Computational Sciences and Technology, vol. 10, no. 2, pp. 221-233, 2017.

[12]. B. Drakshayani and E. V. Prasad, "Text Document Clustering based on Semantics", International Journal of Computer Applications, vol. 45,no. 4, pp. 7-12, May, 2012.

[13]. J. S. Priya and S. Priyadharshini, "Clustering Technique in Data Mining for Text Documents", International Journal of Computer Science and Information Technologies, vol. 3, no. 1, pp. 2943-2947, 2012.

[14]. A. Lamba and D. Kumar, "Survey on KNN and Its Variants", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 5, pp.430-435, May, 2016.