# Text Document Clustering based on Semantics

B.Drakshayani
Department Of Cse, Ssce
Srikakulam (A.P), India

E V Prasad
Director, Iste, Jntuk,
Kakinada (A.P), India

## ABSTRACT
Text document clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing large sets of documents into a small number of meaningful clusters. Clustering is a very powerful data mining technique for topic discovery from text documents. The partitional clustering algorithms, such as the family of K-means, are reported performing well on document clustering. They treat the clustering problem as an optimization process of grouping documents 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. In order to deal with the problem, we integrate core ontologies as background knowledge into the process of clustering text documents. This model combines phrases analysis as well as words analysis with the use of WordNet as background Knowledge and NLP to explore better ways of document representation for clustering. The Semantic based analysis assigns semantic weights to both document words and phrases. The new weights reflect the semantic relatedness between the documents terms and capture the semantic information in the documents to improve the web document clustering. The method adopted has been evaluated on different data sets with standard performance measures to develop meaningful clusters has been proved.

## General Terms
The following submitted material is related to Data mining, Text mining and Clustering concepts. This is very much useful to represent documents in a way that is used for clustering. This is applicable for search engines.

## Keywords
Document Clustering, K-means, Semantic Weights, Semantic Similarity, POS tagging, Ontologies, WordNet, NLP, Similarity measure.

## 1. INTRODUCTION
Text document clustering plays an important role in intuitive navigation and browsing mechanisms by organizing large sets of documents into a small number of meaningful clusters. Data mining includes many techniques that are used to structure the data. Clustering is one of these techniques deals with data sets to group them into a set of clusters. Text databases store large collections of documents from various sources [1]. Text mining concentrates on text databases are rapidly growing due to the increasing amount of information available in electronic form. Text mining attempts to discover new, previously unknown information by applying techniques from Natural Language Processing (NLP) and Data mining. NLP is both a modern computational technology and a method of investigating and evaluating about human language itself. Text mining shares many concepts with traditional data mining methods. Clustering is a very powerful Data Mining technique for topic discovery from text document .A cluster is a collection of data object that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A great number of clustering methods viz. Partitioning, hierarchical are employed for the improvement of text document clustering [1].Among which k-means comes from the family of partitioning method is versatile used for clustering .However due to unsatisfactory conditions arises in k-means such as Bag of words representation relationship between important terms phrase analysis leads to the new method of clustering. There exist two major problems in text clustering technology: one is that pure statistical methods usually fail to recognize the semantic information in the texts; the other is that in the clustering analysis stage, it's very difficult to accurately and effectively evaluate the semantic similarity of different texts by merely considering statistics such as frequency of words/phrases in the texts. The proposed model, adds a new semantic weight to document terms (words and phrases) to extract the semantic relatedness of terms. In particular, we analyze our novel clustering technique in depth in order to find explanations of when background knowledge may help. In this paper we are introducing Ontologies as background knowledge using WordNet [2] to enhance the functioning of the web in many ways. WordNet [3] is a lexical database for English used to maintain relationships among the phrases [4]. The relationships like Holonymy, Hyponymy, Homonymy, and Meronymy are considered [2]. This model uses phrases, which is useful in information rather than words. The phrase based analysis adds Semantic weights to both documents and phrases. This model improves the performance of the web document clustering over other techniques interns of document preprocessing, phrase analysis, semantic weights with semantic similarity measure is considered.

## 2. MATHEMATICAL APPROACH
Most of the documents clustering methods are based on vector space model and Latent indexing model.

### *2.1 Vector Space Model*
Vector space model was proposed in the late 60s by Gerald Salton et. al. [5] to represent texts by vectors .The vector space model is the most used approach to represent textual documents. We represent a text by a numerical vector obtained by counting the most relevant lexical elements present in the text. The document vector

$$d_j = (w_{j1}, w_{j2}, \ldots\ldots w_{jn})$$

where $w_{ji}$ is the frequency weight which is the number of occurrences of word i in document j, n is the number of terms in document j. The similarity between two documents is measured by the cosine similarity measure [1]. Vector space model is an algebraic model for representing text documents as vectors of identifiers such as index terms. It is the basic model where have certain limitations like requiring lot of processing time, documents with similar content but different vocabularies may result in a poor inner product, Improper wording. However, it is widely accepted that words as features are superior to characters and phrases.

### *2.2. Latent Semantic Model*

Latent Semantic Indexing (LSI) model uses singular value decomposition; a mathematical approach to construct a term document matrix represents documents and words [6]. A similarity measure is used between documents to find the most similar documents. LSI seeks to uncover the most representative features rather than the most discriminating documents with different semantics, which is the ultimate goal of clustering. These existing methods are well suited for the search engines and websites based on keywords. Keyword based search engines such as Google, Yahoo, Msn, Ask and Bing are the main tools to use the web. These search engines take the users query displays the web links that are relevant to the query. A document is said to be relevant, if the words in the users query match with the document words. Relevance is subjective in nature, only the user can tell the true relevance. Precision and Recall measures of information retrieval system are used for assessing the quality of text document retrieval. Most of the current information retrieval models are based on keyword representation. This representation creates problems during retrieval due to polysemy, homonymy and synonymy. Another problem associated with key word based retrieval is that it ignores semantic and contextual information in the retrieval process.

The existing text document clustering methods have concentrated on the syntax of the sentence in a document, rather than semantics. The syntax analysis is used to find the syntactic structure of the sentence and it is the process of analyzing a text made of sequence of tokens(words) to determine its grammatical structure with respect to a given document sentence. Syntax analysis or parsing refers to the way that human beings rather than computers analyze a sentence or phrase in terms of grammatical constituents, identifying the parts of speech, syntactic relations. Semantics is the study of meaning, focuses on the relation between words and their literal meaning. In linguistics, semantics is the study of relations between different linguistic units: Homonymy, Synonymy, Polysemy, Hypernymy, Hyponymy, Meronymy, and Holonymy [2]. The greatest source of difficulties in natural language is identifying its semantics.

## 3. PROPOSED MODEL

The proposed model mainly concentrates on the semantic relation between documents. The proposed model(PM) consists of document preprocessing methods like Parts-Of-Speech(POS) tagging, stop word elimination and stemming, phrase analysis, and semantic weights. Words are classified into categories called Parts-of-speech [7]. These are sometimes called word classes or lexical categories. These lexical categories are usually defined by their syntactic and morphological behaviors. The most common lexical categories are nouns and verbs. Other lexical categories include adjectives, adverbs, and conjunctions. Word classes are further categorized as open and closed word classes. Open word classes constantly acquire new members while closed do not. Nouns, verbs (except auxiliary verbs), adjectives, adverbs and interjections are open word classes. Prepositions, auxiliary verbs, delimiters, conjunction and particles are closed word classes. Documents are parsed to fetch parts of speech (POS) tagging for the sentences in the document. POS tagging is the process of assigning a parts of speech such as a noun, verb, pronoun, preposition, adverb and adjective to each word in a sentence. The input to a tagging algorithm is the sequence of words of a natural language sentence. The output is a single best POS for each word. Stanford parser is used for generating the syntactic structure i.e., POS tagging text. For example, the English word ‘book’ can be a noun as in “I am reading a good book “or a verb as in “The police booked the snatcher”. The collection of tags used by a particular tagger is called a tag set. Most POS tag sets make use of the same basic categories i.e. noun, verb, adjective and preposition. POS tagging is an early stage of text processing in many applications including information retrieval, information extraction, speech synthesis and machine translation. In information retrieval, POS tagging can be used for indexing, parsing and for disambiguating word senses

This text is processed further, stop words are removed and stemming performed. In computing, stop words are the words which are filtered out prior to or after processing of natural language data. Search Engines generally ignore stop words. Some example of stop words include “the”, “is”, “who”, “it”, “on” etc. Most search engines do not consider extremely these common words in order to save disk space or to speed up search results. Standard stop words are used at the backend to save the disk space. Sometimes it is necessary to retain their meaning of the sentences, so the authors have to create their own stop word list.

Stemming is the process of removing suffixes and prefixes of a word to get the root word. Standard stemming algorithms like Porter stemmer is used. Unfortunately, the words that appear in documents often have many morphological variants. This is not only means that different variants of a term can be conflated to a single representative form, it also reduces the dictionary size i.e. the number of distinct terms needed for representing a set of documents , that results in a saving of storage space and processing time. For example the words System, Systematic, Systematically, Systematics, Systematise, Systematised, Systematism, Systematist, and Systematists are stemmed to the word System. Many times the stemmers perform stemming by losing the meaning of a word. To retain the original meaning of a word the authors have framed stemming rules exclusively for the verb phrases.

After preprocessing, the text would be presented by a set of words:

$D=\{w_1,w_2\ldots.w_n\}$.

Text collection, in general lacks the imposed structure of a traditional database. The data mining techniques are essentially designed to operate on structured databases. When the data is structured it is easy to define the set of items and hence, it becomes easy to employ the traditional mining techniques [1]. Identifying individual items or terms is not so obvious in a textual database. Thus, unstructured data particularly free running text, places a new demand on data mining methodology. Specific text mining techniques have to be developed to process the unstructured textual data to aid in knowledge discovery. For an unstructured document, features are extracted to convert it to a structured form. Some of the important features are stop words, stemming, POS tagging and other. Once the features are extracted the text is represented as structured data, and traditional data mining techniques like clustering can be used. Information Retrieval is querying against a set of documents to find a subset of relevant documents. In recent years various similarity measures have been proposed, but each has its own limitations and advantages. Most of these similarity measures do not consider semantic aspect of terms in the sentence. In this paper we have tried to overcome from these limitations. In general the weight of a word or term in a document can be calculated using the traditional frequency weights called tf-idf (term frequency-inverse document frequency) measure. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. A corpus is a repository for a collection of natural language material such as text, paragraphs, and sentences from one or many languages. Two types of corpuses have been used in query transaction: parallel and comparable. The importance increases propositionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a documents relevance given a user query. But the main drawback with tf-idf measure is that, it does not consider the semantic relations like Homonymy, Synonymy, Polysemy, Hypernymy, Hyponymy, Meronymy, and Holonymy between words. For instance, in a document terms like “beef”, “fork” and “meat” are found to be similar, where beef and fork are sub concepts of meat. So the word meat has given more weight in the document [8]. To consider semantic relations between words semantic weight is calculated for each word in a document. It uses the extended gloss overlaps measure to calculate the semantic relationships between pairs of terms using WordNet as background knowledge. WordNet [4] is a lexical database or lexical reference system developed at Princeton University. WordNet is organized into taxonomic hierarchies and grouped into synonyms sets (synsets). Each synset has a gloss that defines the concept that it represents. The synsets are connected to each other by lexical and semantic relations. Lexical relations occur between word forms (i.e. senses) and semantic relations between word meanings. These relations include synonymy, Hypernymy/hyponymy/Meronymy/holonymy,antonymy,tropo nymy etc. A word may appear in more than one synset and in more than one parts-of-speech. The meaning of a word is called sense. WordNet [3] lists all senses of a word, each sense belonging to a different synset. However, semantic similarity between entities changes overtime and across domains. For example, apple is frequently associated with computers on the web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries. A user, who searches for apple on the web, might be interested in this sense of apple and not apple as a fruit. New words are constantly being created as well as new senses are assigned to existing words. WordNet sense-entries consist of a set of synonyms and a gloss. A gloss consists of a dictionary style definition and examples demonstrating the use of a synset in a sentence. The WordNet relations only connect word senses that are used in the same part of speech [8]. These relations for instance

Hypernym: y is a hypernym of x if every x is a (kind of) y

E.g.: canine is a hypernym of dog

Hyponym: y is a hypernym of x if every y is a (kind of) x

E.g.: dog is a hypernym of canine

Holonym: y is a holonym of x if x is a part of y

E.g. building is a holonym of window

Meronymy: y is a meronymy of x if y is a part of x

E.g. window is a meronymy of building [2].

Based on phrase analysis we represent document terms as $T=\{t_1,t_2,\ldots.t_n\}$ where n is the number of terms and $t_i$ is defined as follows.

$T_i = p_i$ where $p_i=w_{i1}, w_{i2}\ldots.w_{im}$

We define the semantic weight of phrase $p_i$ as follows [9] [10].

$$ptf(j,p_{i1}) = tf(j,p_{i1}) + \sum_{\substack{i2=1 \\ i2\neq i1}}^{n} tf(j,p_{i2}).Sim(p_{i1},p_{i2})$$

Where $tf(j,p_{i1})$ is the frequency weight of phrase $p_{i1}$ in document j, $sim(p_{i1},p_{i2})$ is the semantic relation between phrase $p_{i1}$ and $p_{i2}$ using adapted lesk [11] measure and n is the number of phrases in document j. A phrase based similarity measure based on matching phrases at the document and semantic weights with phrases are considered rather than individual terms (words).

$Sim(a_s,b)= \max(sim(a_s,b_1)\ldots.sim(a_s,b_n)$

We use the Adapted lesk measure to calculate the semantic similarity [12]between two phrases a and b. both a and b are represented by their wordNet synsets as inputs.

The semantic relation [9] between phrases $a_i$ and $b_j$ is computed as

$$Sim(a_i,b_i)= \sum_{YR} score(R(a_i),R(b_j))$$

Where R is a set of defined relations between $a_i$ and $b_i$. In wordNet synsets like Meronymy, Holonymy, Polysemy, Synonymy, Hyponymy, Hypernym. For example terms like "beef" and "fork" are found to be similar because both are sub concepts of meet in wordNet, finger is a meronymy of hand or hand is a holonym for finger. These relations are shown in the given formula.

$Sim(a_i,b_i)=score(Holonymy(a_i),(Holonymy(b_i))+score(Meronymy(a_i),Meronymy(b_i))$

The document similarity is defined as follows.

$d_j = (ptf(j,p_1), ptf(j,p_2),\ldots.ptf(j,p_n))$

Where ptf (j, $p_1$) is the semantic weight of term $t_i$ in

Document *j*, and *n* is the number of terms in $d_j$.

We adopt the cosine similarity measure to calculate the

Cosine of the angle between the two document vectors $d_{j1}$

and $d_{j2}$ is:

$$\cos(d_{j1}, d_{j2}) = \frac{d_{j1}^{'} . d_{j2}^{'}}{||d_{j1}|| . ||d_{j2}||}$$

K-means: Partitional clustering algorithms assign a set of objects into k clusters. In principle the optimal partition is based on some specific criterion function [13]. One of the important factors in partial clustering is the criteria function. The sum of squared error function is one of the most widely used criteria. Suppose we have a set of objects $x_j \in c_w$, j=1,2,…..N and we want to organize them into k subsets c={$c_1$,….$c_k$}. The squared error criterion is defined as

$$J(T,M) = \sum_{i=1}^{K}\sum_{j=1}^{N}\gamma_{ij} \left\|x_j - m_i\right\|^2$$

Where T= A partition matrix

$\gamma_{ij}$= 1 if $x_j$ ϵ cluster i with $\sum_{i=1}^{K}\gamma_{ij}$=1¥j;

= 0 otherwise

M cluster prototype or centroid matrix; [ $m_1$,……$m_k$]

$m_i$ = sample mean for the $i^{th}$ cluster

The k-means algorithm is the best known squared error based clustering algorithm.

1) Initialize a k-partition randomly or based on some prior knowledge. Calculate the cluster prototype matrix M=[ $m_1$,……$m_k$]
2) Assign each object in the data set to the nearest cluster $c_w$ , i.e. $x_j \in c_w$, if <
for j=1,….N, i≠w ,and i=1,….K
3) Recalculate the cluster prototype matrix based on the current partition
4) Repeat steps 2)-3) until there is no change for each cluster.
The K-means algorithm is very simple and can be easily implemented in solving many practical problems. It can work very well for compact and hyper spherical clusters. The time complexity of K-means is O(NKd)

The entire model should be represented in a diagram is shown in Fig.1:

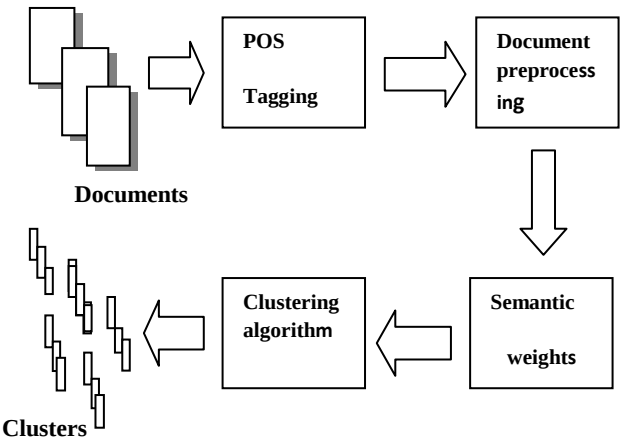


**Fig.1: Semantic Web document Clustering Model**

## 4. RESULTS & DISCUSSION

In order to test the effectiveness of phrase matching, in determining an accurate measure of similarity between documents, we conducted a set of experiments using our proposed data model and similarity measure.

Each evaluation result described in the following denotes an average from 20 test runs performed on the given corpus for a given combination of parameter values with randomly chosen initial values for Bi - Section- K-Means. Without background knowledge, averaged purity values for PRC-min15-max100 ranged from 46.1% to 57%. For clustering using background knowledge using WordNet , we have also performed pruning and we have investigated how inverse purity, F-measure and entropy would be affected for the best baseline that is in terms of purity and typically good strategy based on background knowledge. We used Reuters Transcribed and 20 News groups datasets to assess the quality of clustering, because of its compatibility with the wordNet. We evaluated the performance of the proposed model using three clustering quality measures F-Measure, Purity and Entropy. Table I summarizes the characteristics of all the test data sets used for our experiments.

**TABLE 1: DATA SET PARAMETERS**

| Source | Dataset | No. of doc | No. of Classes |
|---|---|---|---|
| Reuters | RT-1 | 145 | 5 |
| Reuters | RT-2 | 285 | 4 |
| 20NewsGroups | NG-1 | 265 | 5 |
| 20NewsGroups | NG-2 | 341 | 3 |
| 20NewsGroups | NG-3 | 132 | 4 |

The F-measure combines precision and Recall measures. The precision and recall of a cluster c € C for a given class x € X are defined as :

$$P(c,x) = \frac{|c \cap x|}{|c|}$$

$$R(c,x) = \frac{|c \cap x|}{|x|}$$

respectively. Where |c∩x| is the number of documents belonging to cluster c and class x, |c| is the size of the cluster c, |x| is the size of class x.

The F-Measure of a class x is defined as:

$$F(c,x) = \frac{2PR}{P+R}$$

The second measure is the purity. The purity is computed by taking the weighted average of maximal precision values:

$$Purity(C,X) = \sum_{c \in C} \frac{|c|}{|D|} \max_{l \in L} P(c,x)$$

The third measure Entropy measures how homogeneous a cluster is. Entropy of a cluster ic is :

$E(c) = \sum_{x \in X} P(c,x) \cdot \log(c,x)$

**TABLE I : COMPARISON OF ENTROPY VALUES**

| Source | Dataset | Entropy | | |
|---|---|---|---|---|
| | | VSM | LSI | PM |
| Reuters | RT-1 | 0.41 | 0.39 | 0.29 |
| Reuters | RT-2 | 0.38 | 0.32 | 0.26 |
| 20 NGs | NG-1 | 0.52 | 0.5 | 0.42 |
| 20 NGs | NG-2 | 0.42 | 0.36 | 0.31 |
| 20 NGs | NG-3 | 0.36 | 0.33 | 0.27 |

**TABLE II : COMPARISON OF F-MEASURE VALUES**

| Source | Dataset | F-Measure | | |
|---|---|---|---|---|
| | | VSM | LSI | PM |
| Reuters | RT-1 | 0.68 | 0.65 | 0.79 |
| Reuters | RT-2 | 0.56 | 0.56 | 0.71 |
| 20 NGs | NG-1 | 0.54 | 0.63 | 0.7 |
| 20 NGs | NG-2 | 0.75 | 0.71 | 0.81 |
| 20 NGs | NG-3 | 0.55 | 0.54 | 0.71 |

**TABLEIII: COMPARISON OF PURITY VALUES**

| Source | Dataset | Purity | | |
|---|---|---|---|---|
| | | VSM | LSI | PM |
| Reuters | RT-1 | 0.61 | 0.71 | 0.82 |
| Reuters | RT-2 | 0.53 | 0.61 | 0.78 |
| 20 NGs | NG-1 | 0.58 | 0.64 | 0.76 |
| 20 NGs | NG-2 | 0.66 | 0.78 | 0.86 |
| 20 NGs | NG-3 | 0.51 | 0.59 | 0.73 |

The Table II, III, IV shows the results of three measures for VSM, LSI and Proposed Model(PM) with 5 different datasets. The K-Means clustering is chosen for testing the effectiveness of the model. We compared the results of our proposed model , to the vector space model and Latent semantic indexing model. Table II, III, IV shows the performance improvement in the clustering quality obtained by the proposed model. This improvement is gained by combining POS tagging, preprocessing, semantic weights, semantic similarity measure and similarity measure. The proposed model outer performs the VSM and the LSI in terms of F-Measure, Entropy, and Purity. The Fig.2, Fig.3 and Fig.4 show the performance improvement of clustering quality in terms of purity, F-measure and entropy using the proposed model with the VSM and LSI. Fig.2 shows the performance improvement of purity of the proposed model. Fig.3 shows the performance improvement of F-measure of the proposed model. Fig.4 shows the performance improvement of entropy.

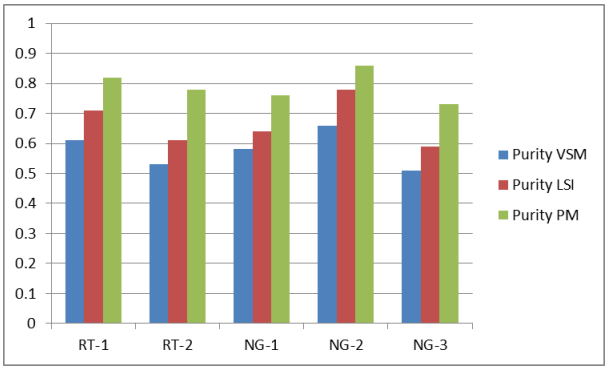


**Fig.2: Comparison of Purity**

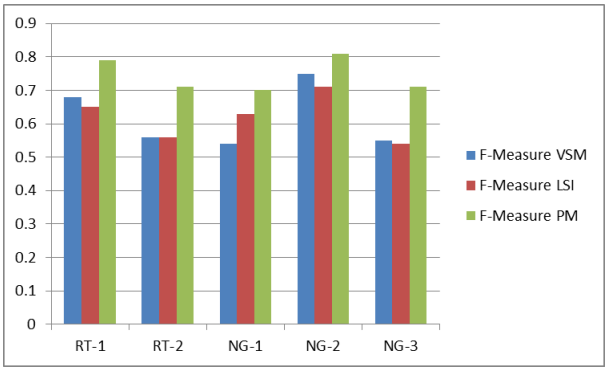


**Fig.3: Comparison of Purity**

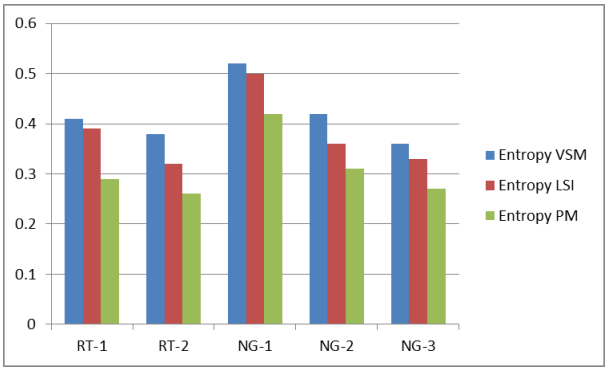


**Fig.4: Comparison of Entropy**

## 5. CONCLUSIONS

In this paper, we proposed a new model for text document representation. The proposed model follows parsing, preprocessing and assignment of semantic weights to Document phrases to reflect the semantic similarity between phrases and k-means clustering algorithm. We evaluated the proposed model using 5 different datasets in terms of F-Measure, Entropy, and Purity for K-Means clustering algorithm. The results demonstrate a performance improvement compared to the traditional vector space model and latent semantic indexing model. More NLP techniques may be included to enhance the performance of the text document clustering.

## 6. REFERENCES

[1] Rui Xu, Donald Wunsch II, « Survey Of Clustering Algorithms » in IEEE Transactions on Neural Networks, Vol. 16, No.3, May 2005.

[2] James Z. Wang, William Taylor, « Concept Forest : A New Ontology-assisted Text Document Similarity Measurement Method » in 2007 IEEE /WIC/ACM International Conference on Web Intelligence.

[3] Abdelmalek Amine, Zakaria Elberrichi and Michel Simonet, « Evaluation Of Text Clustering Methods Using WordNet », International arab Journal of Information Technology, Vol.7, No. 4, October 2010.

[4] A.Hotho, S. Staab, and G.Stumme, « Wordnet improve text document clustering », in SIGIR 2003 Semantic Web Workshop, 2003, pp. 541-544.

[5] A.Wong, C S Yang G Salton, "A vector space model for Automaticindexing ," Communication ACM, vol. 18, no. 11, pp. 112-117, 1975.

[6] S.Dumais ,S T Landauer Deerwester, "Indexing by Latent Semantic analysis," Journal of the Society for Information Science, pp. 391-407, 1990.

[7] Thorstan Brants, "statistical POS tagger," in NLP conference, 2000.

[8] Jung Ae Kwak and Hwan- Seung Yong, « Ontology Matching Based On Hypernym, Hyponym, Holonym, and Meronym Sets in WordNet » in International Journal of Web & Semantic Technology(IJWesT), April 2010.

[9] K.Hammouda and M.Kamel, « Efficient Phrase based document indexing for web document clustering », IEEE Transactions on Knowledge and Data Engineering, vol.16, no.10, pp.1279-1296, October 2004.

[10] Walaa K.Gad, Mohamed s. Kamel, «PH-SSBM : Phrase Semantic Similarity Based Model for Document Clustering » in 2009 second International Symposium on Knowledge Acquisition and Modeling.

[11] S.Benerjee and T. Pederson, « Adapted Lesk algorithm for word sense disambiguation using wordnet », in Computational Linguistics and Intelligent Text Processing, Feb.2002.

[12] W.Gad and M.Kamel, « New Semantic Similarity based model for text clustering using extended gloss overlaps, » in International Conference on Machine Learning and Data Mining, July 2009, pp.663-677.

[13] Tapas Kanungo, Nathan S.Netanyahu, Angela Y.Wu, « An efficient k-means clustering algorithm : Analysis and implémentation »