Semantic Based Model for Text Document Clustering with Idioms

B. Drakshayani *draksha_m@yahoo.co.in*
Lecturer in CME
Govt. Polytechnic
Nalgonda, 508001, India

E.V.Prasad *profevprasad@yahoo.com*
Rector, JNTUK
Kakinada, India

Abstract

Text document clustering has become an increasingly important issue in recent years because of the availability of tremendous amount of unstructured data in various forms such as the web, social networks, and other information networks. Clustering is a very powerful data mining technique to organize the large amount of information on the web for easy access. Traditionally, document clustering methods do not consider the semantic structure of the document. This paper addresses the task of developing an effective and efficient clustering methodology to take care of semantic structure of the text documents. A method has been developed that performs the following sequence of operations : tagging the documents for parsing, replacement of idioms with their original meaning, semantic weights calculation for document words and apply semantic grammar. The similarity measure is obtained between the documents and then the documents are clustered using Hierarchical clustering algorithm. The method adopted in this work is evaluated on different data sets with standard performance measures and the effectiveness of the method to develop in meaningful clusters has been proved.

Keywords: Document Clustering, Idiom, POS Tagging, Semantic Weight, Semantic Grammar, Hierarchical Clustering Algorithm, Chameleon, Natural Language Processing.

1. INTRODUCTION

The World Wide Web [1] services are huge, widely distributed and acts as a global information service centre for news, advertisements, consumer information, financial management, education, government, e-commerce information and many other information services. Many organizations and societies place most of their public accessible information on the web. With the rapid growth of the World Wide Web (www), it becomes a critical issue to design and organize the vast amounts of on-line documents on the web according to their topic. Even for the search engines it is very important to group similar documents in order to improve their performance when a query is submitted to the system. Clustering is useful for taxonomy design and similarity search of documents on such a domain [2]. There has been extensive research activity on the construction and use of the semantic web and set of methods and technologies to make machines to understand the meaning - or "semantics" - of information on the World Wide Web. NLP techniques [3] help to deal with syntax, semantics, discourse context and pragmatics to structure the data. NLP-based techniques have recently attracted more attention in applications such as content-based search engines and systems for automatic reviewing, biomedical text mining, text summarization and spam detection.

Most of the documents clustering methods are based on Vector space model. Document clustering methods are accepting input data either numerical or categorical form. Documents [4]

are represented at the lexical and semantic levels by the words they contain. This creates an independent representation called bag-of-words. To create this representation documents are segmented into tokens based on white space, paragraph separators and punctuation marks. Then all words are extracted and stemmed, stop words are removed and the number of occurrences of each word is counted.

The standard document representation technique is the vector space model (VSM) [5].Vector Space Model represents each document as a feature vector of the terms in the document. The document vector d_j is $d_j = (w_{j1}, w_{j2}, \ldots, w_{jn})$,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. This representation of text excludes any grammatical analysis and any concept of distance between words. The existing vector space model is 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. The VSM representation creates problems during retrieval due to polysemy- one word can have different meanings, synonymy's are not unified and semantic connections between words are neglected, which not only encompasses the synonymy and polysemy relations but extends to the more general sense of two words being semantically related [4]. These document clustering methods are not suitable for the semantic web searching process.

The Semantic Model (SM) method [6] only concentrates on the compositional semantics of the grammar. Compositional semantics signifies a system of constructing logical forms for sentences or parts of sentences in such a way that the meanings of the components of the sentences (phrase) are used to construct the meanings of the whole sentence. But it is insufficient to get the original semantic meaning of the documents. The semantic model does not concentrate on the grammar sentences that contain idioms. The document sentences contain idiom phrases does not have compositional semantics, the words collectively do not give the original meaning. Compositional semantics are useful for common sentences or phrases. For example a phrase like "kick the bucket" (meaning is die) does not have compositional semantics as the meaning of the whole is unrelated to the meanings of the component words. The authors have considered compositional semantics, disambiguity and idioms. For example idiom phrases like:

- Kick the bucket: to die
- Dog Days of summer: The hottest days of the summer season
- Raining Cats and Dogs: A very loud and noisy rain storm
- The Ball Is In Your Court: It is your decision this time

It does not have compositional semantics as the meaning of the whole is unrelated to the meanings of the component words. Compositional semantics does not consider idioms. The authors have considered compositional semantics and idioms. The rest of the paper is organized as follows: section 2 presents the related work, section 3 presents the proposed Idiom based Semantic Based model, performance studies are explained in section 4 and the final conclusions are in section 5.

2. RELATED WORK

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 sentences. 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 [3]. Syntax analysis or parsing is a very important task in NLP or text mining and the partial syntactical information can help to solve many other NLP tasks such as information retrieval, information extraction, text summarization etc. Syntax analysis 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 and focuses on the relation between words and their literal meaning. In linguistics, semantics is the study of relationship between different linguistic units: Homonymy, Synonymy, Polysemy, Hypernymy, Hyponymy, Meronymy, and Holonymy [7]. The greatest source of difficulty in natural language is identifying its semantics. Corpus based computational linguistics [3] computes statistics over large text collections in order to discover useful patterns. These patterns are used to inform algorithms for various sub problems within NLP, such as Parts-Of-Speech (POS) tagging, word sense disambiguation. The proposed model mainly concentrates on the documents that consisting of idioms.

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 [8]. 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 document processing like stop words elimination, stemming, POS tagging. Other higher order features include Semantic grammar, semantic relation between words and similarity measure. Once the features are extracted the text is represented as structured data, and traditional data mining techniques like clustering can be used. The proposed model is developed to concentrate on idioms processing and semantic knowledge. This model will be helpful to enhance the performance of the search engines.

3. PROPOSED MODEL

The entire model is represented in terms of document representation, similarity measure, clustering algorithm followed by clustering measures. The model of the proposed system is shown in Fig.1. The proposed model consists of five components: Idiom processing, POS Tagging, Document pre-processing, Semantic weights calculation, Document representation model using Semantic grammar, Document similarity and Hierarchical clustering algorithm has given below.

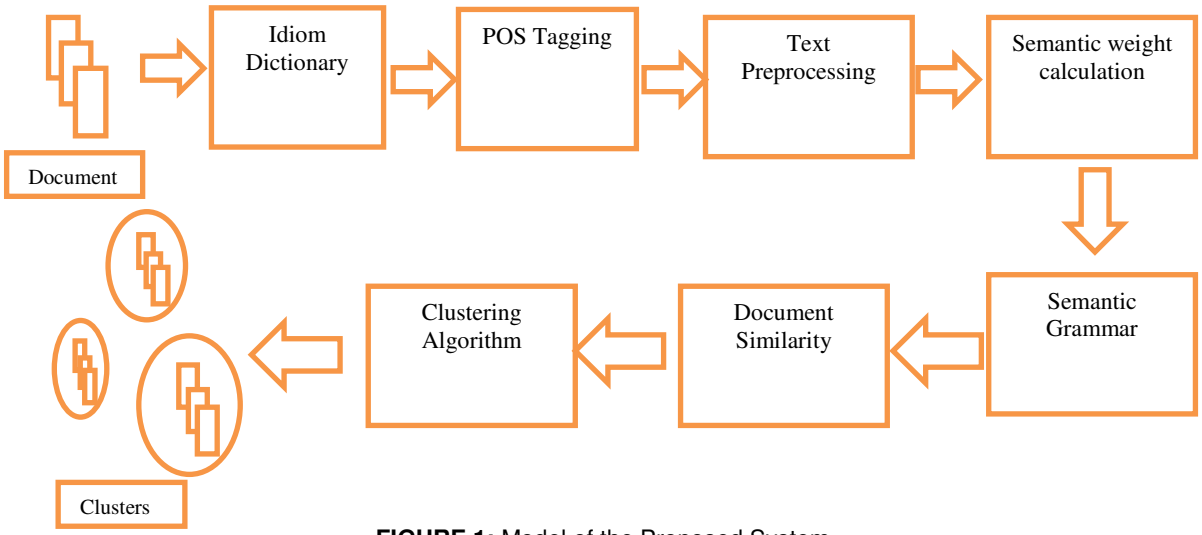

FIGURE 1: Model of the Proposed System.

3.1. Idiom Processing

An idiom is a common word or phrase with a culturally understood meaning that differs from what its composite words' denotations would suggest. Idioms add more complexity to identify the meaning of the written text. So the authors have considered compositional semantics with idioms. As an example, consider the sentence: “Raining Cats and Dogs”, the meaning of this sentence has nothing to do with the words “raining”, “cats” and “dogs” appearing on it. The meaning is a

very loud and noisy rain storm. The idiom phrase checking is very much useful for semantic web design and search engines, to create meaningful clusters. The use of idiom is ubiquitous in natural language text and it is a serious bottleneck in automatic text understanding. In technical documents idioms are not been used but when we consider documents related to literatures, novels, news articles, magazines idioms are relevant [10]. Due to the high frequency use of idioms, a system capable of interpreting idiomatic expressions in unrestricted text would become an invaluable component of any semantics oriented NLP application. Initially the documents processed against the idiom dictionary. An idiom dictionary contains commonly used idiom phrases and their meanings. The idioms phrases in the documents are compared with the dictionary phrases and matched idiom phrases are replaced by equivalent original meaning in the corresponding phrase, otherwise the documents are returned as usual. As this methodology checks only the verb of the document sentence the time consumed for processing is very small. Though simple in form, the idiom demands a complex interpretation of the relationship, revealing subtle correspondences between the two documents compared and to improve the search engine results. In this paper we have considered idioms to improve the semantic relations between the documents through wordNet. Usage of idioms is a special case of semantic web mining. Datasets which are used to evaluate in this model are verified with the idiom dictionary. We have included idiom phrases into the documents and applied the methodology. It impacts in evaluation measures are shown in the later sections. Finally, idiom processing improves the overall quality of the document clusters; this has been verified through the evaluation measures purity and entropy.

3.2. POS Tagging

Documents can be parsed, by using any standard parsers, for generating the syntactic structure also called parts of speech (POS) tagging [9] 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.

3.3. Document Preprocessing

Standard information retrieval methods for indexing involve a small amount of language specific processing [11,12]. The text is processed further for eliminating of stop words and to perform stemming. In computing, stop words are the words which are filtered out prior to or after processing of natural language data. Some example of stop words include “the”, “is”, “who”, “it”, “on” etc. Standard stop word list is available but sometimes it is necessary to retain some of the stop words for retention of their usual meaning of the sentences. Hence the authors have created their own stop word list.

Stemming is the process of removing suffixes and prefixes of a word to get the root word and standard stemming algorithm like Porter Stemmer can be used [13]. 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 “information”, “informing”, “informer”, “informed”, would be stemmed to their common root “inform”. Many times the stemmers perform stemming by losing the meaning of a word. To retain the original meaning of a word, it is essential to have stemming rules separately framed for verb phrases.

3.4. Semantic Weight

Semantic weight [7] of term $t_{i1}$is defined as in terms of semantic term frequency is:

$$stf(j,t_{i1}) = tf(j,t_{i1}) + \sum \mathrm{tf}(j,t_{i2})\mathrm{SIM}(t_{i1},t_{i2})$$

Where $tf(j,t_{i1})$ is the frequency of term t_{i1}in document j, $SIM(t_{i1},t_{i2})$ is the semantic relatedness between terms t_{i1} and t_{i2} using the extended gloss overlaps measure and n is the number of terms in document j.

Given two words a and b, the semantic similarity between them can be calculated as:

$$SIM(a,b) = \sum_{s=1}^{m} \frac{SIM(a_s,b)}{m}$$

Where SIM (a_s,b) is the similarity between word a_s and phrase b can calculated as follows:

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

The extended overlap measure is used to calculate the semantic similarity between two words a_i and b_i. Both a_i and b_i are represented by their WordNet synsets as inputs. The output is a numeric value that quantifies their degree of semantic relatedness. The relatedness degree of two words is defined by the number of overlaps in their glosses. The semantic relatedness between a_i and b_i is computed by comparing the glosses of synsets that are related to a_i and b_i through explicit relationships of WordNet. The semantic similarity score between a_i and b_i is defined as :

$$sim(a_i, b_i) = \sum_{\forall (R)} score(R(a_i), R(b_i))$$

Where R is a set of defined relations between a_i and b_i in word Net. The score function accepts two glosses, finds overlap between them, and returns their relatedness score. For example {hypernym, hyponym} is a set of relations between a_i and b_i the relatedness between a_i and b_i is calculated as follows:

Sim (a_i , b_i) = score (hypernym (a_i) , hypernym (b_i)) + score (hyponym (a_i) , hyponym (b_i))

WordNet is a lexical database or lexical reference system organized into taxonomic hierarchies and grouped into synonyms sets (synsets) [11,12]. 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, antonym, troponymy etc. 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 hypornym of canine

- Holonym: y is a holonym of x if x is a part of y
  E.g. building is a holonym of window

- Meronym: y is a meronym of x if y is a part of x
  e.g. window is a meronym of building

3.5. Semantic Grammar

A grammar developed with the intention of handling semantics is called semantic grammar [8]. One of the main motivations behind the use of semantic grammar is dealing with idioms. The design of semantic grammar follows the above process. Semantic grammar from the field of Natural Language Processing (NLP) is a triplet form i.e. verb (subject, object) form. A document sentence is taken in the form of verb (subject, object). For example if we take the sentence “she bought laptop” is translated into bought (she, laptop). Weightage is given to the verb of a sentence and also consider the subject and object of the sentences. Semantics can be the meaning of individual words, in a sentence or how individual words combine to give meaning to a sentence. Compositional semantics only consider the meaning of individual words, whereas this model use idiom dictionary to get the original meaning of the sentence.

3.6. Document Similarity

Cosine is among most commonly used similarity measure. Cosine measure gives the cosine of the angle between the document vector and query. The cosine similarity is used in this paper to calculate the cosine of the angle between the two document vectors d_{j1} and d_{j2} .

$$\cos(\mathbf{d}_{j1}, \mathbf{d}_{j2}) = \frac{\overrightarrow{\mathbf{d}_{j1}} . \overrightarrow{\mathbf{d}_{j2}}}{\|\mathbf{d}_{j1}\| . \|\mathbf{d}_{j2}\|}$$

3.7. Hierarchical Clustering

Hierarchical clustering algorithms can usually find satisfiable clustering results [14]. A hierarchical clustering is able to obtain different clustering results for different similarity requirements. However, most of those hierarchical algorithms are very computationally intensive and require much memory space. In recent years, with the requirement for handling large scale data sets in data mining and other fields, many new hierarchical clustering techniques have appeared and greatly improve the clustering performance. Typical examples include Chameleon. Chameleon algorithm from the hierarchical method is used in this work.

3.7.1. Chameleon Algorithm

Chameleon algorithm is an agglomerative hierarchical clustering algorithm based on the k-nearest neighbor graph, in which an edge is eliminated if both vertices are not within the k-closest points related to each other [15,16]. At the first step, Chameleon divides the connectivity graph into a set of sub clusters with the minimal edge cut. Each sub graph should contain enough nodes in order for effective similarity computation. By combining both the relative interconnectivity and relative closeness make the Chameleon flexible enough to explore the characteristic of the potential clusters. Chameleon merges these small subsets , and thus comes up with the ultimate clustering solutions.

Algorithm Chameleon:

1. Construct a k-nearest neighbor graph

2. Partition the k-nearest neighbor graph into many small sub clusters using partitioning algorithm.

3. Merge those sub clusters into final clustering results on the basis of chameleon interconnectivity principle.

Chameleon uses a dynamic modeling framework to determine the similarity between pairs of clusters by looking at their relative interconnectivity (RI) and relative closeness (RC). Chameleon selects pairs to merge for which both RI and RC are high. That is, it selects clusters that are well interconnected as well as close together.

Relative Interconnectivity:

Clustering algorithms typically measure the *absolute* interconnectivity between clusters C_i and C_j in terms of *edge cut*—the sum of the weight of the edges that straddle the two clusters, which we denote $EC(C_i, C_j)$. *Relative* interconnectivity between clusters is their absolute interconnectivity normalized with respect to their internal interconnectivities. To get the cluster's *internal* interconnectivity, we sum the edges crossing a min-cut bisection that splits the cluster into two roughly equal parts. Recent advances in graph partitioning have made it possible to efficiently find such quantities. Thus, the relative interconnectivity between a pair of clusters C_i and C_j is

$$RI(C_i, C_j) = \frac{|EC(C_i, C_j)|}{|EC(C_i)| + |EC(C_j)|_{/2}}$$

By focusing on relative interconnectivity, Chameleon can overcome the limitations of existing algorithms that use static interconnectivity models. Relative interconnectivity can account for differences in cluster shapes as well as differences in the degree of interconnectivity for different clusters.

Relative closeness:
Relative closeness involves concepts that are analogous to those developed for relative interconnectivity. The *absolute closeness* of clusters is the average weight (as opposed to the sum of weights for interconnectivity) of the edges that connect vertices in C_i to those in C_j. Since these connections come from the *k*-nearest-neighbor graph, their average strength provides a good measure of the affinity between the data items along the interface layer of the two clusters. At the same time, this measure is tolerant of outliers and noise. To get a cluster's *internal closeness*, we take the average of the edge weights across a min-cut bisection that splits the cluster into two roughly equal parts. The *relative closeness* between a pair of clusters is the absolute closeness normalized with respect to the internal closeness of the two clusters:

$$RC(C_i, C_j) = \frac{\bar{S}EC(C_i, C_j)}{\frac{|C_i|}{|C_i| + |C_j|}\bar{S}EC(C_i) + \frac{|C_j|}{|C_i| + |C_j|}\bar{S}EC(C_j)}$$

where $SEC(C_i)$ and` $SEC(C_j)$ are the average weights of the edges that belong in the min-cut bisector of clusters $C_i$ and $C_j$, and` $SEC(C_i, C_j)$ is the average weight of the edges that connect vertices in C_i and C_j. Terms $|C_i|$ and $|C_j|$ are the number of data points in each cluster. This equation also normalizes the absolute closeness of the two clusters by the weighted average of the internal closeness of C_i and C_j. This discourages the merging of small sparse clusters into large dense clusters. In general, the relative closeness between two clusters is less than one because the edges that connect vertices in different clusters have a smaller weight. By focusing on the relative closeness, Chameleon can overcome the limitations of existing algorithms that look only at the absolute closeness. By looking at the relative closeness, Chameleon correctly merges clusters so that the resulting cluster has a uniform degree of closeness between its items. Results have been carried out by varying the document representation of the proposed model with the vector space model and semantic model.

4. EXPERIMENTAL RESULTS AND EVALUATION

4.1. System Details

The effectiveness of the proposed model has been proved by conducting set of experiments using wordNet, the lexical database, nltk tool kit and compared with vector space model and Semantic model. The experiments were performed on Python, Windows-XP, Pentium 4, 3.0GHz CPU with 2 GB RAM.

4.2. Data Sets

To ensure, the experimental results are independent of one special test collection. We used three collections to test our proposed method. They are Reuters-Transcribed test, Reuters-21578 and mini newsgroups. They are available from the UCI KDD archive [17]. Chameleon also has been executed considering the three different document representation models.

4.3. Evaluation Measures

To prove the superiority of the semantic based structure, we have considered the information retrieval measures for evaluation. Our model is suited for semantic web search process. So we have taken Precision, Recall, Purity and Entropy [7]. The precision and recall of a cluster c ϵ C for a given class $l \in L$ are defined as:

$$\text{Precision} \quad P = \frac{|c \cap l|}{|c|} \qquad \text{Recall} \quad R = \frac{|c \cap l|}{|l|}$$

The purity measure overall value is computed by taking the weighted average of maximal precision value.

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

The entropy measure is how homogeneous a cluster is. The higher the homogeneity of a cluster, then lower entropy is and vice versa. Entropy of a cluster c is

$$E(c) = -\sum_{l \in L} P(c, l).\log(c, l)$$

All the measures used to be maximized to satisfy the users of the semantic web site.

The Table I, II and III list the values for the evaluation measures purity (Pu) and entropy (En) on datasets Reuter's transcribed, Mini_news_group and Reuters-21578 document datasets of our experiments. The Vector space method clustering has found to exhibit poorly and the proposed method has obtained the best performance indice. The proposed method has proven its superiority by obtaining better values both of purity and entropy. The results of purity and entropy measures are shown in Fig.2 to Fig.7 on the datasets. The comparison study of purity and entropy values obtained with the values obtained by Vector Space Model, Semantic Model and the Proposed Model on Reuters Transcribed dataset (Fig.2 and Fig. 3),on Mini-news-group dataset (Fig.4 and Fig.5) and on Reuters-21578 dataset (Fig.6 and Fig.7) establishes the cluster quality of clusters obtained with the proposed model and undoubtedly out performs on the vector space model and semantic model.

Table I: Purity and Entropy Values For Reuters-transcribed Data Set.

No. of Documents	Proposed Model		Semantic Model		Vector Space Model	
	Purity	Entropy	Purity	Entropy	Purity	Entropy
20	93.89%	2.64%	89.45%	3.64%	79.88%	12.39%
40	93.21%	7.65%	88.01%	10.84%	78.65%	35.97%
60	92.43%	19.75%	85.72%	33.09%	76.35%	59.97%
80	91.25%	25.43%	83.45%	39.98%	74.48%	69.85%
100	90.98%	38.95%	81.87%	47.45%	71.47%	78.92%
120	87.38%	47.87%	79.12%	53.06%	67.98%	86.78%
140	83.54%	56.88%	76.85%	59.95%	63.62%	89.98%
160	79.87%	66.65%	69.89%	64.17%	59.98%	91.65%
180	75.68%	77.89%	65.25%	69.14%	55.78%	96.85%
200	73.98%	85.97%	60.89%	73.16%	49.63%	98.65%

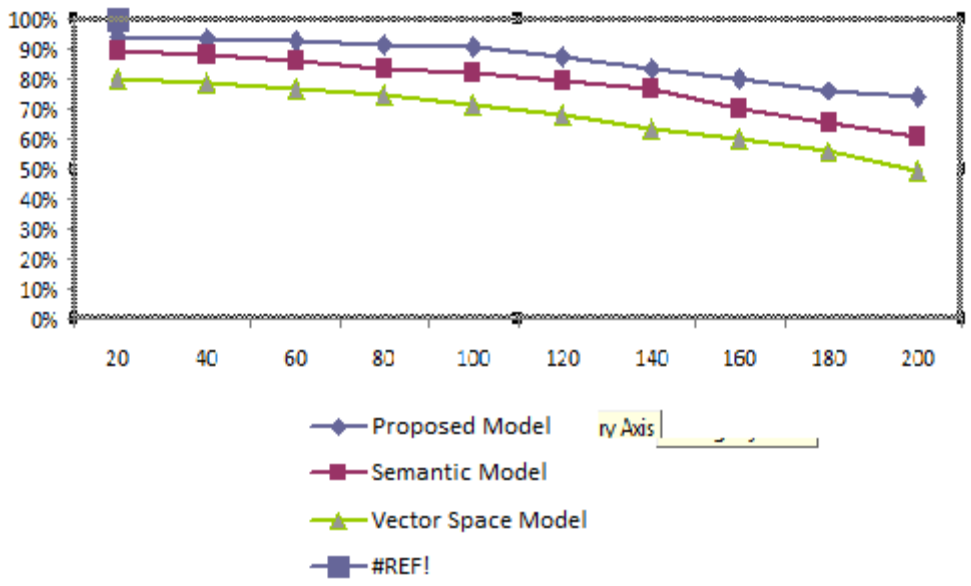

FIGURE 2: Purity Measure – Reuters-Transcribed-Dataset.

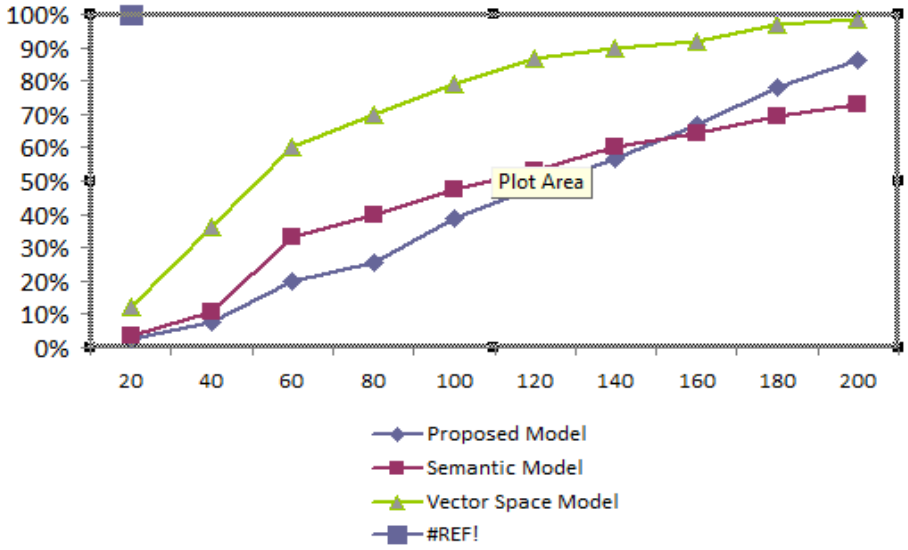

FIGURE 3: Entropy Measure – Reuters-Transcribed-Dataset.

Table II: Purity and Entropy Values For – Mini_news_group Dataset.

No. of Documents	Proposed Model		Semantic Model		Vector Space Model	
	Purity	Entropy	Purity	Entropy	Purity	Entropy
100	93.74%	0.57%	91.85%	0.87%	78.28%	0.75%
200	93.26%	2.84%	90.78%	3.65%	77.34%	2.43%
300	92.77%	3.96%	88.99%	9.89%	74.21%	3.67%
400	92.25%	5.89%	85.84%	15.44%	72.09%	6.56%

500	91.87%	6.87%	79.96%	22.76%	70.34%	7.45%
600	90.98%	7.98%	75.98%	29.97%	68.23%	10.62%
700	89.98%	9.85%	71.85%	35.59%	65.71%	16.87%
800	88.21%	11.98%	69.84%	40.78%	63.89%	24.34%
900	86.85%	15.96%	68.78%	49.45%	60.01%	33.45%
1000	85.96%	21.56%	64.81%	57.56%	55.29%	46.87%
1250	83.74%	35.69%	61.98%	64.24%	51.19%	54.98%
1500	81.98%	43.79%	59.78%	73.57%	47.19%	66.45%
1750	79.87%	59.87%	54.90%	79.54%	42.02%	75.89%
2000	77.95%	72.85%	52.67%	89.56%	36.89%	84.67%

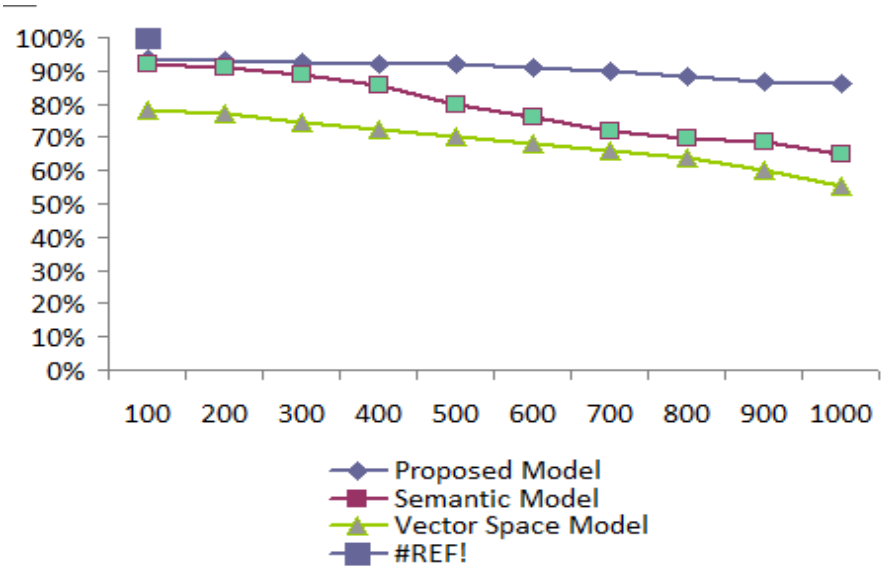

FIGURE 4: Purity Measure – Mini_news_group Dataset.

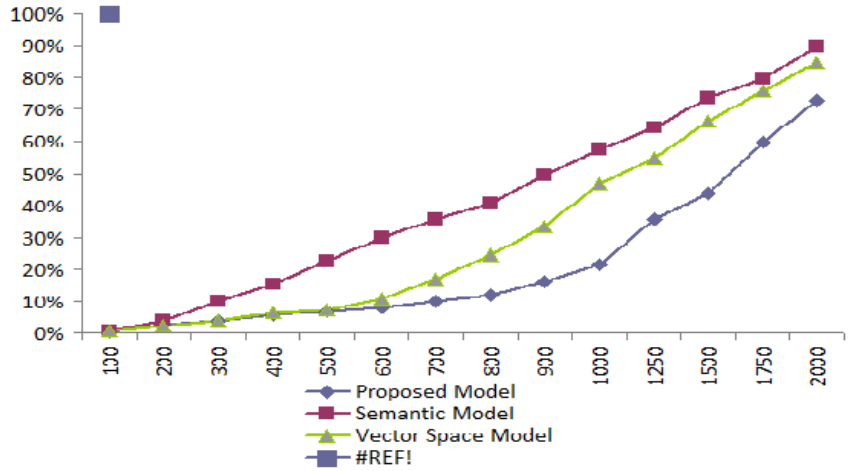

FIGURE 5: Entropy Measure – Mini_news_group Dataset.

Table III: Purity and Entropy Values For Reuters-21578 Data Set.

No. of Documents	Proposed Model		Semantic Model		Vector Space Model	
	Purity	Entropy	Purity	Entropy	Purity	Entropy
1000	91.45%	14.67%	88.57%	20.67%	76.87%	48.09%
2000	90.67%	17.78%	86.34%	21.87%	75.34%	49.24%
3000	89.56%	21.89%	85.56%	22.99%	74.78%	52.25%
4000	88.45%	24.07%	83.98%	27.34%	72.98%	55.89%
5000	86.12%	29.67%	80.49%	32.89%	60.56%	59.56%
6000	84.32%	32.66%	77.56%	37.33%	65.12%	62.02%
7000	81.45%	37.82%	74.71%	41.55%	61.88%	65.11%
8000	78.34%	43.91%	69.35%	46.78%	55.09%	69.27%
9000	74.56%	49.56%	66.91%	57.28%	50.43%	74.92%
10000	71.32%	56.45%	63.19%	65.99%	46.87%	77.99%
12500	68.57%	62.99%	54.87%	72.32%	40.34%	81.46%
15000	63.45%	68.32%	49.61%	78.02%	36.78%	88.88%
17500	60.76%	76.33%	44.78%	83.44%	29.76%	90.98%
20000	58.34%	82.91%	40.89%	88.07%	23.32%	93.78%
20500	52.87%	89.34%	34.24%	92.89%	19.67%	95.89%

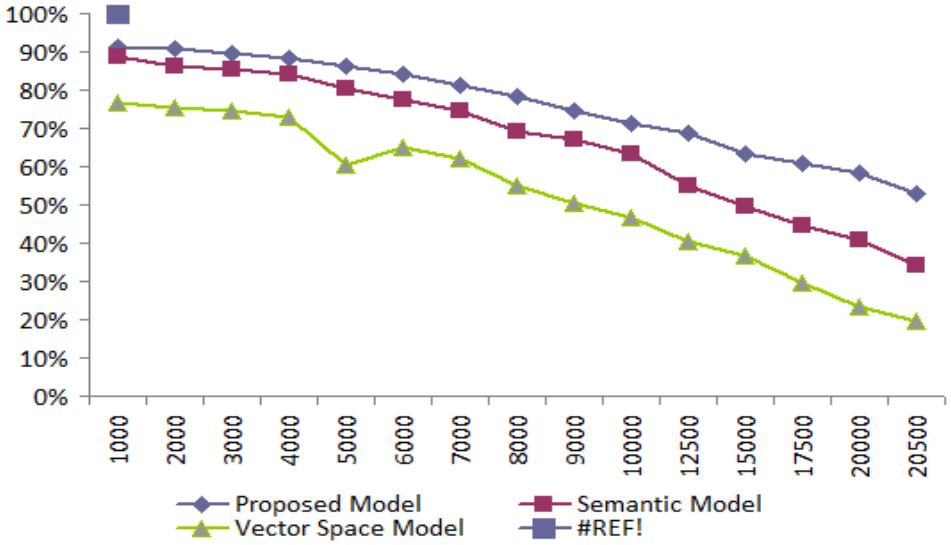

FIGURE 6: Purity Measure – Reuters-21578 Document Set.

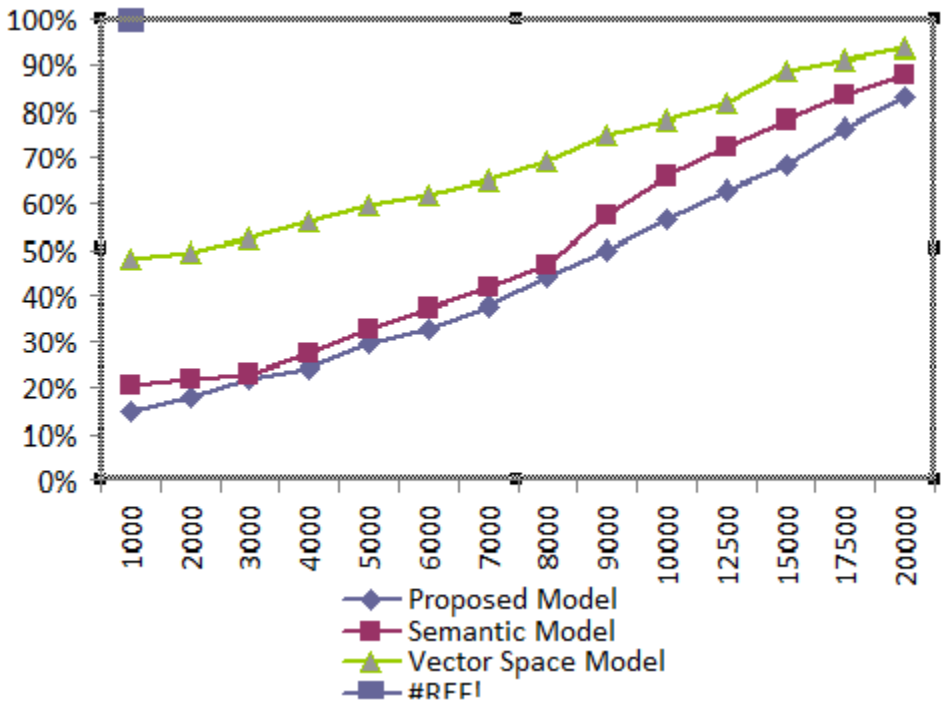

FIGURE 7: Entropy Measure- Reuters-21578 Document Set.

5. CONCLUSION AND FUTURE WORK

In this paper the proposed Idiom Semantic Based Mining Model, the documents are clustered based on their meaning using the techniques of idiom processing, semantic weights using Chameleon clustering algorithm. The enhanced quality in creating meaningful clusters has been demonstrated and established on three different datasets, with idiom based documents, with the use of of performance indices,entropy and purity.The results obtained with the vector space model and semantic model are compared and presented in graphs to show the improved performance of the proposed method. The further work need to be concentrated on data documents consisting of metaphors and ellipses. Adopting a multilevel or hybrid clustering may like to improve cluster quality and justification of time complexity need to be made.

6. REFERENCES

[1] A K Jain, "Data clustering : 50 Years Beyond K-Means," in International Conference in Pattern recognition, Pattern Recognition Letters, 31, Issue 8, pp. 651-656, June 2010.

[2] D.Wunsch II, and R.Xu, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, vol. 16, No. 3, pp. 46-51, May 2005, DOI:10.1109/TNN.2005.845141.

[3] U.S.Tiwari, T.Siddiqui, Natural Language Processing and Information Retrieval., Oxford University Press.

[4] L.Huang, D.Milne, E.Frank and L.H.Witten, " Learning a Concept-Based Document Similarity Measure", Journal of the American Society for Information Science and Technology, 63(8):1593-1608, July 2012,DOI: 10.1002/asi.22689.

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

[6] Supreethi.K.P and E.V.Prasad, "A Novel Document Representation Model for Clustering," International Journal of Computer Science &Communication, vol. 1, no. 2, pp. 243-245, December 2010.

[7] W.K.God, M.S.Kamel, “PH-SSBM: Phrase Semantic Similarity Based Model for Document Clustering”, IEEE Second International Symposium on Knowledge Acquisition and Modeling, 978-0-7695-3888-4/09, April 2009, DOI: 10.1109/kam.2009.191.

[8] Supreethi.K.P and E.V.Prasad, “ Web Document Clustering using Case Grammar Structure”, International Conference on Computational Intelligence & Multimedia Applications, vol.2, pp. 98-102, Dec 2007, DOI: 10.1109/ICCIMA.2007.245.

[9] David Holmes, “Idioms and Expressions “, a method for learning and remembering idioms and expressions.

[10] POS Tagging-The Stanford Parser, nlp.stanford.edu/software/lex-parser.shtml.

[11] S. Staab, and G.Stumme A.Hotho, "Wordnetimprovetext document clustering," in proceedings of the Semantic web workshop *SIGIR*, 2003, pp. 541-544.

[12] Z.Elberrichi and M.Simonet Abdelmalek Amine, “Evaluation of Text Clustering Methods using WordNet", International Arab Journal of Information Technology, vol. 7, no. 4, Oct 2010.

[13] M.F.Porter, “An algorithm for suffix stripping”, Program: electronic library and information systems, Vol. 14 Iss: 3, pp.130 – 137, 1980, DOI:10.1108/eb046814.

[14] F.Murtagh "A Survey of Recent Advances in Hierarchical Clustering Algorithms ", in the Computer Journal, vol. 26, no. 4, Jan1983, pp. 354- 359.

[15] G.Karypis, Eui-Hong Han, Vipin Kumar, “Chameleon: Hierarchical Clustering using Dynamic Modeling “, IEEE International Journal of Computer, Aug1999, vol.32, Issue 8, pp.68-75, DOI: 10.1109/2.781637.

[16] M.A.Abbas and A.A.Shoukry, “Clustering Using Shared Reference Points Algorithm Based on a Sound Data Model”, International Journal of Data Engineering(IJDE), Volume 3, Issue 2, 2012.

[17] UCIKDD ARCHIVE, kdd.ics.uci.edu