# Self Organizing Map -based Document Clustering Using WordNet Ontologies

**Tarek F. Gharib[1,2], Mohammed M. Fouad[3], Abdulfattah Mashat[1], Ibrahim Bidawi[1]**

**[1] Faculty of Computing and Information Technology, King Abdulaziz University Jeddah, Saudi Arabia**
**[2]Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt**

**[3] Faculty of Informatics and Computer Science, The British University in Egypt (BUE) Cairo, Egypt**

**Abstract**

With the rapid development of web content, retrieving relevant information is difficult task. The efficient clustering algorithms are needed to improve the results of the retrieval. Document clustering is a process of recognizing the similarity or dissimilarity among the given objects and forms subgroups sharing common characteristics**.** In this paper, we propose a semantic text document clustering approach that using WordNet lexical and Self Organizing Maps. The proposed approach uses the WordNet to identify the importance of the concepts in the document. The SOM is used to cluster the document. We use this approach to enhance the effectiveness of document clustering algorithms. The approach takes the advantages of the semantics available in knowledge base and the relationship between the words in the input documents. Some experiments are performed to compare efficiency of the proposed approach with the recently reported approaches. Experiments show advantage of the proposed approach over the others.

***Keywords:*** *Text Document Clustering; WordNet Lexical Categories; Self Organizing Map (SOM)*

## 1. Introduction

With the recent growth and diversity of electronic data on the World Wide Web (www), it becomes more difficult for Internet users to find the useful information from these huge amounts of data. Search engines and recommender systems help people to reduce the information overload by finding relevant information on their search topic. Clustering of documents is one of the techniques used in search engines and in recommender systems for efficiently finding documents that have similar topics [1], for improving the performance of information retrieval systems [2], for assisting users on a web site [3] and for personalization of search engine results [4]. Formally, document clustering is an optimization problem where the input of the problem is a set of documents and a (dis)similarity measure between these documents. Thus, similarity plays an important role in document clustering.

Text document clustering provides an effective navigation mechanism to organize this large amount of data by grouping their documents into a small number of meaningful classes. Text document clustering can be defined as the process of grouping of text documents into semantically related groups[5]. Most of the current methods for text clustering are based on the similarity between the text sources. The similarity measures work on the syntactically relationships between these sources and neglect the semantic information in them. By using the vector-space model in which each document is represented as a vector or 'bag of words', i.e., by the words (terms) it contains and their weights regardless of their order [6].

Vector space model is a popular model for document representation in document clustering including the above methods. Documents are represented by vectors of weights, where each weight in a vector denotes importance of a term in the document. In the standard VSM, however, semantic relations between terms are not taken into account. Two terms with a close semantic relation and two other terms with no semantic relation are both treated in the same way. This unconcern about semantics could reduce quality of the clustering result.

Many well-known methods of text clustering have two problems: first, they don't consider semantically related words/terms (e.g., synonyms or hyper/hyponyms) in the document. For instance, they treat {Vehicle, Car, and Automobile} as different terms even though all these words have very similar meaning. This problem may lead to a very low relevance score for relevant documents because the documents do not always contain the same forms of words/terms.

Second, on vector representations of documents based on the bag-of-words model, text clustering methods tend to use all the words/terms in the documents after removing the stop-words. This leads to thousands of dimensions in

the vector representation of documents; this is called the "Curse of Dimensionality". However, it is well known that only a very small number of words/terms in documents have distinguishable power on clustering documents and become the key elements of text summaries. Those words/terms are normally the concepts in the domain related to the documents [7].

There are some approaches that employ WordNet based semantic similarity to enhance the performance of document clustering [8, 9]. They modified the VSM model by readjusting term weights in the document vectors based on its relationships with other terms co-occurring in the document.

In this paper, we propose a semantic text document clustering approach that using WordNet lexical and Self Organizing Maps. The proposed approach uses the WordNet to identify the importance of the concepts in the document. The SOM is used to cluster the document. We use this approach to enhance the effectiveness of document clustering algorithms. The clustering performances are evaluated versus K-means and bisecting k-means algorithms. The approach takes the advantages of the semantics available in knowledge base and the relationship between the words in the input documents. Some experiments are performed to compare efficiency of the proposed approach with the recently reported approaches. Experiments show advantage of the proposed approach over the others.
The rest of this paper is organized as following; recent related work is discussed and presented in section 2. In section 3, we show the proposed semantic text clustering approach. In section 4 a set of experiments is presented to compare the performance of the proposed approach with current text clustering methods. Finally, conclusion and future work are given in section 5.

## 2. Related Work

In the recent years, text document clustering has been introduced as an efficient method for navigating and browsing large document collections and organizing the results returned by search engines in response to user queries [10]. Many clustering techniques are proposed like bisecting k-means [11], FTC and HFTC [12] and many others. From the performed experiments in [11] bisecting k-means overcomes all these algorithms in the performance although FTC and HTFC allows to reduce the dimensionality if the data when working with large datasets.

WordNet is used by Green [13-14] to construct lexical chains from the occurrences of terms in a document: WordNet senses that are related receive high higher weights than senses that appear in isolation from others in the same document. The senses with the best weights are selected and the corresponding weighted term frequencies constitute a base vector representation of a document.

Dave and Lawrence [15] use WordNet synsets as features for document representation and subsequent clustering. But the word sense disambiguation has not been performed showing that WordNet synsets decreases clustering performance in all the experiments. Hotho et al. use WordNet in an unsupervised scenario taking into account the WordNet ontology and lexicon. They used some strategy for word sense disambiguation which achieved improvements for the clustering results [16].

In [9] the authors explore the benefits of partial disambiguation of words by their PoS and the inclusion of WordNet concepts; they show how taking into account synonyms and hypernyms, disambiguated only by PoS tags, is not successful in improving clustering effectiveness because the noise produced by all the incorrect senses extracted from WordNet. Adding all synonyms and all hypernyms into the document vectors seems to increase the noise.

Reforgiato[17] presented a new unsupervised method for document clustering by using WordNet lexical and conceptual relations .In this work, Reforgiato uses WordNet lexical categories and WordNet ontology in order to create a well structured document vector space whose low dimensionality allows common clustering algorithms to perform well. For the clustering step he has chosen the bisecting k-means and the Multipole tree algorithms for their accuracy and speed.
Friedman et al. [18] introduced FDCM algorithm for clustering documents that are represented by vectors of variable size. The algorithm utilizes fuzzy logic to construct the cluster center and introduces a fuzzy based similarity measure which provided reasonably good results in the area of web document monitoring.

Hung and Wermter [19] proposed three novel text vector representation approaches for neural network based document clustering. The first is the extended significance vector model (ESVM), the second is thehypernym significance vector model (HSVM) and the last is the hybrid vector space model (HyM). ESVM extracts the relationship between words and their preferred classified labels. HSVM exploits a semantic relationship from the WordNet ontology. HyM is a combination of a TFxIDF vector and a hypernym significance vector, which combines the advantages and reduces the disadvantages from both unsupervised and supervised vector representation approaches. According to their experiments,

the self-organizing map (SOM) model based on the HyM text vector representation approach is able to improve classification accuracy and to reduce the average quantization error.

Sridevi and Nagaveni [20] proposed a model by combining ontology and optimization technique to improve the clustering. The proposed model uses the ontology similarity in identifying the importance of the concepts in the document. The particle swarm optimization is used to the cluster the document.

## 3. Semantic Text Document Clustering

In this section we describe in details the components of the proposed semantic text clustering approach. There are two main processes: Document Preprocessing that generated output document vectors from input text documents using WordNet[1] lexical information is introduced in the first step. The second step is Document Clustering that applies SOM neural network on the generated document vectors to obtain output clusters as illustrated in fig. 1.

### 3.1 Document Preprocessing

The first step in the proposed approach is document preprocessing which aims to represent the corpus (input documents collection) into vector space model. Data preprocessing is a very important and essential phase in an effective document clustering. The first part of feature extraction is preprocessing the lexicon and involves removal of stop words and stemming [6]. The stop words removal accounts to 20% to 30% of total words counts while the process of stemming reduces the number of terms in the document. Both the process helps in improving the effectiveness and efficiency of text processing as they reduce the indexing file size.

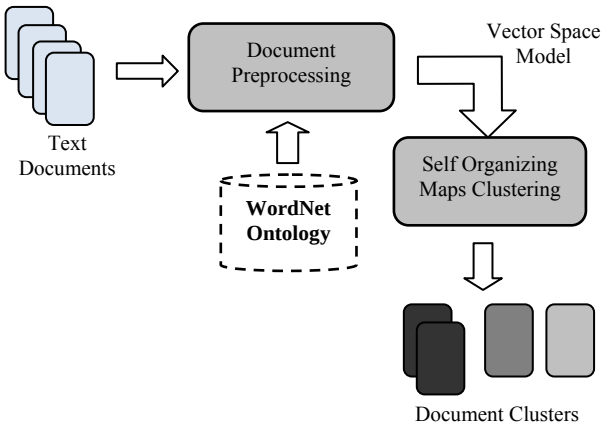


Fig. 1 Diagrammatic representation of the proposed approach

#### 3.1.1 Stopwords Removal

This is the first step in preprocessing which will generate a list of terms that describes the document satisfactorily. The document is parsed through to find out the list of all the words. The next process in this step is to reduce the size of the list created by the parsing process, generally using methods of stop words removal.

#### 3.1.2 Stemming

Stemming is process of linguistic normalization in which the variant forms of a word is reduced to a common form. For example: the word, connect has various forms such as connect, connection, connective, connected, etc., Stemming process reduces all these forms of words to a normalized word connect. Porter's English stemmer algorithm is used to stem the words for each of the document in our stemming process. This step aims to reduce the extracted frequent word list to optimize the next step for WordNet mapping. In our implementation we use minimum support value set to 10%, which means that the words found in less than 10% of the input documents is removed from the extracted word list.

#### 3.1.3 WordNet Lexical Category Mapping

As proposed in [17], we use WordNet lexical categories to map all the stemmed words in all documents into their lexical categories. We use WordNet 2.1 that has 41 lexical categories for nouns and verbs as shown in tables 1 and 2. For example, the word "dog" and "cat" both belong to the same category "noun.animal". Some words also has multiple categories like word "Washington" has 3 lexical categories (noun.location, noun.group, noun.person) because it can be the name of the American president, the city place, or a group in the concept of capital.

[1] WordNet project: http://wordnet.princeton.edu/

Some word disambiguation techniques are used to remove the resulting noise added by multiple categories mapping which are: disambiguation by context and concept map which are discussed in details in [13].

Table 1: WordNet nouns lexical categories

| | | | |
|---|---|---|---|
| Act | Animal | Artifact | Attribute |
| Body | Cognition | Communication | Event |
| Feeling | Food | Group | Location |
| Motive | Object | Person | Phenomenon |
| Plant | Possession | Process | Quantity |
| Relation | Shape | State | Substance |
| Time | Tops | | |

Table 2: WordNet verbs lexical categories

| | | | |
|---|---|---|---|
| Body | Change | Cognition | Communication |
| Competition | Creation | Contact | Perception |
| Emotion | Motion | Weather | Consumption |
| Social | Stative | Possession | |

The output vectors that are generated based on the number of words that found in each lexical category. The generated document vector D for each document d in the input text document is defined as in Eq. (1).

$$D^T = [X_1, X_2, ....., X_{41}] \quad (1)$$

We calculate $X_i$ as the number of words in document d that belongs to the $i^{th}$ lexical category in the WordNet lexical categories for the output vector.

## 3.2 Document Clustering

Clustering is one technology to find intrinsic structures in data sets. Text clustering method usually uses the document vector space model to split the document into vectors in high dimensional space, and then make clustering of these vectors. Text clustering can generally be divided into partitioned clustering algorithms and hierarchical clustering algorithms.

After generating the documents' vectors for all the input documents using feature extraction process, we start the clustering process as shown in fig. 1.

The problem of document clustering is defined as follows. Given a set of n documents called DS, DS is clustered into a user-defined number of k document clusters D1, D2,…Dk, (i.e. {D1, D2,…Dk} = DS) so that the documents in a document cluster are similar to one another while documents from different clusters are dissimilar.
In this stage we apply three different clustering algorithms which are k-means (partitioning clustering), bisecting k-means (hierarchical clustering) and SOM neural network. These algorithms are most commonly used in the document clustering step in the recent researches.

### 3.2.1 K-means and Bisecting k-means

We have implemented the k-means and bisecting k-means algorithms as introduced in [11]. We will state some details on bisecting k-means algorithm that begins with all data as one cluster then perform the following steps:

Step1: Choose the largest cluster to split.
Step2: Use k-means to split this cluster into two sub-clusters. (Bisecting step)
Step3: Repeat step2 for some iterations (in our case 10 times) and choose the split with the highest clustering overall similarity.
Step4: Go to step1 again until the desired k clusters are obtained.

### 3.2.2 Self Organizing Maps (SOM)

Self-organizing maps (SOM) learn to classify input vectors according to how they are grouped in the input space. They differ from competitive layers in that neighboring neurons in the self-organizing map learn to recognize neighboring sections of the input space. Thus, self-organizing maps learn both the distribution (as do competitive layers) and topology of the input vectors they are trained on.

In this paper we focus on using SOM to perform the document clustering. The two reasons for using SOM rather than other clustering methods are that it is topologically preserving and clustering is performed non-linearly on the given input data sets. The topologically preserving property allows the SOM applied to document clustering, to group similar documents together in a cluster and organize similar clusters close together unlike most other clustering methods.

In our proposed approach, we use the implementation of self organizing maps in MATLAB (*Neural Network Toolbox*). We construct a 1-D SOM neural network that takes the generated document vector as input. The size of the network (number of hidden neurons) is based on the desired number of clusters. The network then is trained on the input document vector for about 250 epochs. The output from the network is the weights that define the centers of each cluster. Then we assign each document into its appropriate cluster to be evaluated after that.

Here we list some of the MATLAB-Neural Network Toolbox functions that used in this implementation:

- **newsom**: Create 1-D SOM neural network.
- **train**: Apply SOM training algorithm on input document vectors.
- **sim**: Assign each document vector to its cluster center.

### 3.2.3 Silhouette Coefficient

For clustering, two measures of cluster "goodness" or quality are used. One type of measure allows us to compare different sets of clusters without reference to external knowledge and is called an internal quality measure. The other type of measures lets us evaluate how well the clustering is working by comparing the groups produced by the clustering techniques to known classes which called an external quality measure [7].

In our application of document clustering, we don't have the knowledge of document classes in order to use external quality measures. We will investigate silhouette coefficient (SC Measure) as one of the main internal quality measures.

To measure the similarity between two documents d1 and d2 we use the cosine of the angle between the two document vectors. This measure tries to approach the semantic closeness of documents through the size of the angle between vectors associated to them as in Eq. (2).

$$dist(d_1, d_2) = \frac{d_1 \bullet d_2}{|d_1| \cdot |d_2|} \quad (2)$$

Where $(\bullet)$ denotes vector dot product and (| |) is the dimension of the vector. A cosine measure of 0 means the two documents are unrelated whereas value closed to 1 means that the documents are closely related [18].

Let $\mathbf{D_M} = \{\mathbf{D_1}, .., \mathbf{D_k}\}$ describe a clustering result, i.e. it is an exhaustive partitioning of the set of documents DS. The distance of a document $d \in DS$ to a cluster $\mathbf{D_i} \in \mathbf{D_M}$ is given as in Eq. (3).

$$dist(d, D_i) = \frac{\sum_{p \in D_i} dist(d, p)}{|D_i|} \quad (3)$$

Let further consider $a(d, D_M) = dist(d, D_i)$ being the distance of document d to its cluster Di where $(\mathbf{d} \in \mathbf{D_i})$. $b(d, D_M) = \min_{d \notin D_i} dist(d, D_i) \; \forall D_i \in D_M$ is the distance of document d to the nearest neighbor cluster. The silhouette S (d, DM) of a document d is then defined as in Eq. (4).

$$S(d, D_M) = \frac{b(d, D_M) - a(d, D_M)}{\max(b(d, D_M), a(d, D_M))} \quad (4)$$

The silhouette coefficient (SC Measure) is defined as shown in Eq. (5).

$$SC(D_M) = \frac{\sum_{p \in DS} S(p, D_M)}{|DS|} \quad (5)$$

The silhouette coefficient is a measure for the clustering quality that is rather independent from the number of clusters. Experiences, such as documented in [18], show that values between 0.7 and 1.0 indicate clustering results with excellent separation between clusters, viz. data points are very close to the center of their cluster and remote from the next nearest cluster. For the range from 0.5 to 0.7 one finds that data points are clearly assigned to cluster centers. Values from 0.25 to 0.5 indicate that cluster centers can be found, though there is considerable "noise". Below a value of 0.25 it becomes practically impossible to find significant cluster centers and to definitely assign the majority of data points.

## 4. Experimental Results

The experiments were conducted on three text document datasets EMail1200, SCOTS and Reuters with the three algorithms. There are two main parameters to evaluate the performance of the proposed approach, which are clustering quality and running time.

Document preprocessing step is implemented in Java using NetBeans 5.5.1 and Java API for WordNet Searching (JAWS Library) to access WordNet 2.1 lexical. All the clustering algorithms (k-means, bisecting k-means and SOM neural network) are implemented in MATLAB (Version 7.6.0.324). All experiments were done on Processor P4 (3GHz) machine with 1GB main memory, running the Windows XP Professional® operating system and all times are reported in seconds.

### 4.1 Text Document Datasets

We evaluate the proposed semantic text document clustering approach on three text document datasets: EMail1200, SCOTS and Reuters text corpuses. These datasets vary in the numbers of documents in each dataset, the total number of words, and the average numbers of words in single document. EMail1200 corpus contains test email documents for spam email detection with about 1,245 documents with about 550 words per document. SCOTS corpus (Scottish Corpus Of Text and Speech) contains over 1100 written and spoken texts, with about 4 million words of running text. 80% of this total is made up of written texts and 20% is made up of spoken texts. SCOTS dataset contains about 3,425 words per document. Reuters corpus contains about 21,578 documents that appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by

personnel from Reuters Ltd. and Carnegie Group, Inc. in 1987. All the three datasets are used in the text mining testing studies and they are available online for download in [22, 23, 24] respectively.

### 4.2 Clustering Quality

Fig. 2, 3, and 4 show the silhouette coefficient values for the three datasets respectively. In all experiments SOM neural network outperforms k-means and bisecting k-means algorithms in the overall clustering quality using silhouette measure.

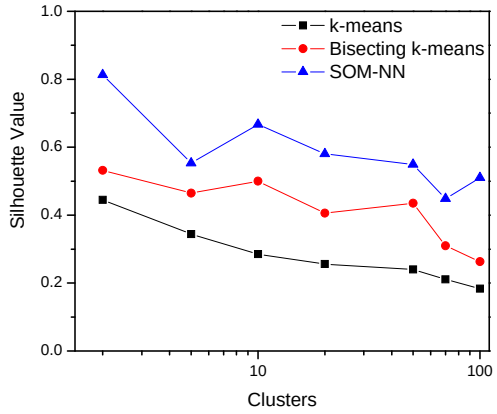


Fig. 2 Silhouette values comparing all clustering algorithms – EMail1200 Dataset

From these figures we notice the good clustering quality results obtained by SOM-NN with comparison to other algorithms. For example, at number of clusters (k = 2), we found that SC value for SOM for EMail1200 dataset is about 0.813 which considered an excellent clustering result with well separated clusters. If we check the other algorithms results, we found that bisecting k-means overcomes basic k-means algorithm with SC value equal to 0.532 which means that the data points are clearly assigned to cluster centers.

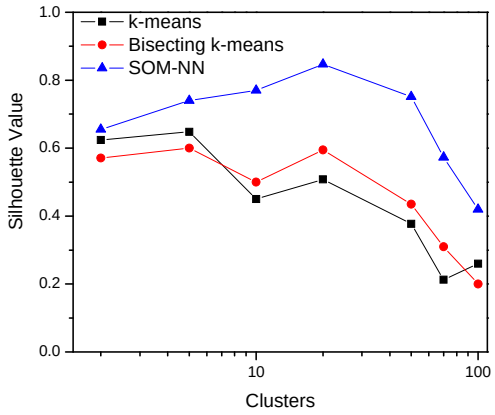


Fig. 3 Silhouette values comparing all clustering algorithms – SCOTS Dataset

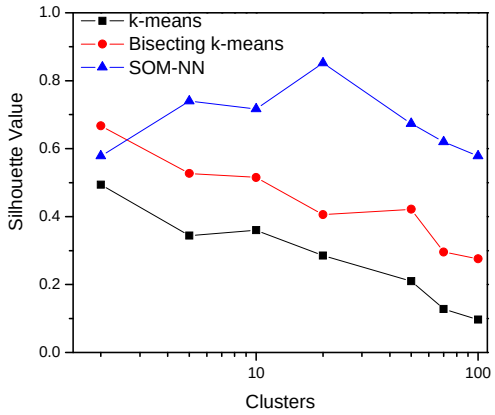


Fig. 4 Silhouette values comparing all clustering algorithms – Reuters Dataset

For SCOTS dataset, as in fig. 3, we found that k-means and bisecting k-means algorithms nearly generates the same clusters. However SOM outperforms other algorithms at k=20. SOM-NN achieves silhouette value equal to 0.847 where other algorithms obtain about 0.595 and 0.508 respectively. The last experiment results in fig. 4 show the great performance optimization between SOM-NN and other algorithms in Reuters datasets.

We have performed two more experiments to show the effect of using WordNet lexical categories with SOM neural network on the final clustering quality results. We measure SC value for SOM on both SCOTS and Reuters datasets in two cases: first using traditional bag-of-words technique, second using WordNet lexical categories. Fig. 5 and 6 show the silhouette coefficient values for the two datasets.

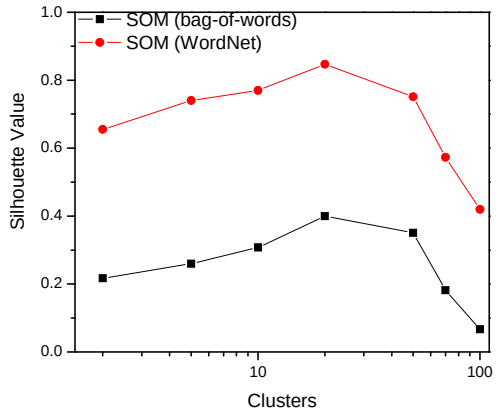


Fig. 5 WordNet improves SOM clustering results using SCOTS dataset

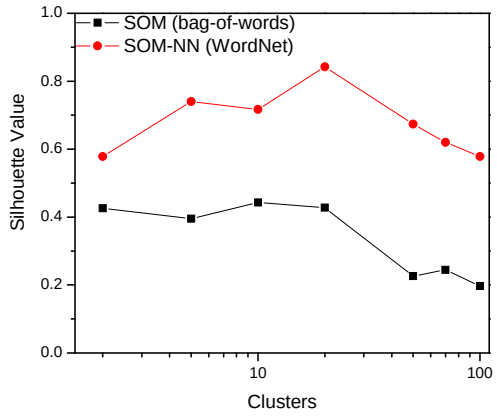


Fig. 6 WordNet improves SOM clustering results using Reuters dataset

For SCOTS dataset the clustering results is very good, because the proposed approach overcomes the traditional approach with about 3 times. The clustering results for Reuters dataset is also positive. The proposed approach achieves about twice clustering quality than traditional technique. This experiment shows that using WordNet lexical categories in the feature extraction process improves the overall clustering quality of the input dataset document than traditional approaches that uses bag-of-words technique.

### 4.3 Running Time

Reuters dataset, as mentioned early in this section, contains about 21,578 documents. This is considered a real challenge task that faces any clustering approach because of "Scalability". Some clustering techniques that are helpful for small data sets can be overwhelmed by large data sets to the point that they are no longer helpful. For that reason we test the scalability of our proposed approach with the different algorithms using Reuters dataset. This experiment shows that the SOM neural network performs a great running time optimization with comparison to other two algorithms. Also, according to the huge size of Reuters dataset, the proposed approach shows very good scalability against document size.

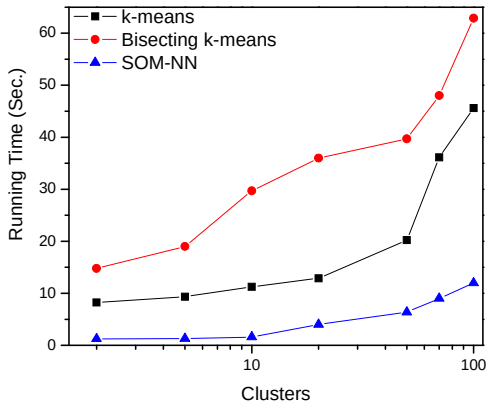


Fig. 7 Scalability of all clustering algorithms on Reuters dataset

Fig. 7 depicts the running time of the different clustering algorithms using Reuters dataset with respect to different values of desired clusters. The overall process of document clustering using WordNet lexical categories is done in a very low time in comparison with other two approaches. SOM neural network achieves speed-up ratio 10 times faster than bisecting k-means algorithm and about 5 times faster than basic k-means algorithm for Reuters dataset.

## 5. Conclusion

In this paper we proposed a semantic text document clustering approach based on the WordNet lexical categories and SOM neural network. The proposed approach generates documents vectors using the lexical category mapping of WordNet after preprocessing the input documents. We apply three different clustering algorithms, SOM neural network, k-means, and bisecting k-means to the generated documents vectors. The output clusters in each case are evaluated using silhouette coefficient measure to test the performance of the proposed approach. The results show that SOM neural network achieves higher clustering quality than other two clustering algorithms k-means, and bisecting k-means. Also, the results show that by using WordNet lexical categories in the feature extraction process for text documents improves the overall clustering quality. Finally, the proposed approach shows good scalability against the huge number of documents as in Reuters dataset along with different values of desired clusters.

## References

[1] R. Saraçoğlu, K. Tütüncü, and N. Allahverdi, (2007) "A fuzzy clustering approach for finding similar documents using a novel similarity measure," Expert Systems with Applications, vol. 33, no. 3, pp. 600–605.

[2] H. X. W. Wu and S. Shekhar, (2003)Eds., Clustering and Information Retrieval. Kluwer.

[3] K. Bade and A. Nurnberger,( 2006) "Personalized hierarchical clustering," in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC, USA: IEEE Computer Society, pp. 181–187.

[4] Z. Jiang, A. Joshi, R. Krishnapuram, and L. Yi, (2000) "Retriever: Improving web search engine results using clustering," University of Maryland Baltimore County, Technical Report.

[5] J. Sedding and D. Kazakov, (2004) "WordNet-based Text Document Clustering", COLING 3rd Workshop on Robust Methods in Analysis of Natural Language Data, pp. 104–113, Geneva, Switzerland.

[6] M. Lan, C.L. Tan, H.B. Low and S.Y. Sung, (2005) "A Comprehensive Comparative Study on Term Weighting Schemes", Proceedings of the 14th International World Wide Web (WWW2005) Conference, Japan, pp.1032–1033.

[7] B.B. Wang, R.I. McKay, H.A. Abbass and M. Barlow, (2002) "Learning text classifier using the domain concept hierarchy", In Proceedings of International Conference on Communications, Circuits and Systems, China, pp. 1230–1234.

[8] W.K. Gad and M.S. Kamel, (2009) "Enhancing text clustering performance using semantic similarity", Lecture Notes in Business Information Processing, 24 LNBIP, pp. 325-335.

[9] L. Jing, M.K. Ng and J.Z. Huang, (2010) "Knowledge-based vector space model for text clustering", Knowledge and Information Systems, 25 (1), pp. 35-55.

[10] O. Zamir, O. Etzioni, O. Madani, and R.M. Karp, (1997)"Fast and intuitive clustering of web documents", In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, KDD97, pp. 287–290.

[11] M. Steinbach, G. Karypis and V. Kumar, (2000)"A Comparison of Document Clustering Techniques", Department of Computer Science and Engineering, University of Minnesota, Technical Report , #00-034.

[12] F. Beil, M. Ester and X. Xu, (2002) "Frequent term-based text clustering", Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD02), Edmonton, Alberta, Canada, pp. 436–442.

[13] S.J. Green, (1999) "Building hypertext links by computing semantic similarity", IEEE Transactions on Knowledge and Data Engineering, Vol.11, pp.713–730.

[14] S.J. Green, (1997) "Building hypertext links in newspaper articles using semantic similarity", The 3rd Workshop on Applications of Natural Language to Information Systems, NLDB 97, pp. 178–190.

[15] D.M.P.K. Dave and S. Lawrence, (2003) "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews", Proceedings of the 12th International World Wide Web Conference, Budapest, Hungary, pp. 519–528.

[16] A. Hotho, S. Staab and G. Stumme, (2003) "Wordnet improves text document clustering", ACM SIGIR 2003 Workshop on Semantic Web, pp. 541–544.

[17] D. Reforgiato, (2007) "A new unsupervised method for document clustering by using WordNet lexical and conceptual relations", Journal of Information Retrieval, Vol. 10, pp.563–579.

[18] M. Friedman. A. Kandel, M. Schneider, M. Last, B. Shapka, Y. Eloviciand O. Zaafrany, (2004) "A Fuzzy-Based Algorithm for Web Document Clustering. Fuzzy Information", Processing NAFIPS '04, IEEE Annual Meeting of the North American, Vol. 2, pp. 524–527.

[19] C. Hung and S. Wermter, (2004) "Neural Network-based Document Clustering Using WordNet Ontologies", International Journal of Hybrid Intelligent Systems, Vol. 1, pp. 127-142.

[20] U.K. Sridevi and N. Nagaveni (2011) "Semantically Enhanced Document Clustering Based on PSO Algorithm", European Journal of Scientific Research, Vol.57, No.3, pp. 485-493.

[21] L. Kaufman and P.J. Rousseeuw, (1999) "Finding Groups in Data: an Introduction to Cluster Analysis", Published by John Wiley & Sons, USA.

[22] EMail1200 dataset: http://boole.cs.iastate.edu/book/acad/bag/data/lingspam

[23] SCOTS dataset: http://www.scottishcorpus.ac.uk/

[24] Reuters dataset: http://www.daviddlewis.com/resources/testcollections/reuters21578/