Ontology-based semantic smoothing model for biomedical document clustering

S. Logeswari* and K. Premalatha

Department of Computer Science and Engineering,
Bannari Amman Institute of Technology,
Sathyamangalam, Erode – 638401, Tamil Nadu, India
Email: slogesh76@gmail.com
Email: kpl_barath@yahoo.co.in
*Corresponding author

Abstract: One of the major issues of data mining is the clustering of unstructured text documents. Traditional clustering algorithms are failing to prove the accuracy of the clustering process because of the characteristics of text documents such as high dimension, complex semantics, sparsity, etc. Recent researches focus on the clustering of text documents based on the semantic smoothing technique, which resolves the conflicts by general words and the sparsity of class-specific core words. In this work ontology-based semantic smoothing model is proposed which uses the domain ontology for concept extraction. It is a mixture of simple language model and a topic signature translation model. The results obtained from the proposed method shows a significant improvement in the clustering process than the existing methods in terms of cluster quality.

Keywords: smoothing; document clustering; ontology; Jaccard index; silhouette index; FM index.

Reference to this paper should be made as follows: Logeswari, S. and Premalatha, K. (2015) 'Ontology-based semantic smoothing model for biomedical document clustering', *Int. J. Telemedicine and Clinical Practices*, Vol. 1, No. 1, pp.94–110.

Biographical notes: S. Logeswari received her BE (Computer Science and Engineering) from Bharathiyar University, Coimbatore in 1997 and a ME (Computer Science and Engineering) from Anna University, Chennai in 2007. She is currently working as an Associate Professor at the Bannari Amman Institute of Technology, Erode, Tamil Nadu, India. Her areas of interest are data mining, compiler design, soft computing and optimisation techniques.

K. Premalatha is currently working as a Professor in the Department of Computer Science and Engineering at the Bannari Amman Institute of Technology, Erode, Tamil Nadu, India. She completed her PhD in Computer Science and Engineering (CSE) at the Anna University, Chennai, India. She did her Master of Engineering and Bachelor of Engineering in CSE at the Bharathiar University, Coimbatore, Tamil Nadu, India. Her research interests include data mining, image processing and information retrieval.

1 Introduction

Document clustering is one of the elementary functions in text mining. Clustering is the process of classifying a collection of text documents into diverse category groups so that documents in the each category group describe the same topic. Document clustering plays a vital role in real world applications of information retrieval domain, for example, grouping the web search results and categorising digital documents. Clustering of unstructured text data faces a number of new challenges. Huge volume, high dimensionality, sparsity of core words and complex semantics are the most important ones. Therefore there is a need for new clustering techniques that are scalable and able to handle more complex semantics involved in it (Hamzah et al., 2007; Jayabharathy et al., 2011).

Conventional clustering techniques for document clustering are mainly based on the term frequency within the documents. The keywords which are having highest frequency values are used as features for representing the documents and they are treated independently which can be easily applied to non-ontological clustering. In order to overcome the various limitations due to the sparsity of core words, a novel and an efficient solution which is scalable clustering with onslaught the improper feature using Ontology-based computing is proposed. This system comprises of a dynamic weight assignment technique based on the semantic relations between the terms as well as phrases within a document (Zhang and Wang, 2010).

Text document clustering based on the semantic smoothing approach is mostly seen as an objective method, which delivers one clearly defined result, which needs to be optimal in some way. Smoothing is used to capture the importance of the terms within the document by the adjustment of the maximum likelihood estimator (MLE) of a language model. With the rapid development and widely use of the internet, clustering of the documents should be based on the theme and it will be more significant than the traditional term-based method.

Generally, documents are having more general words which are not relevant to the search query and with less frequency of class dependent core words. Hence, clustering becomes more difficult. Recent works on document clustering (Zhou et al., 2007; Tu et al., 2010) show that the semantic smoothing approach based on the TF-IDF schema is an effective solution to the limitations of clustering due to the sparsity of the core words in the documents. The main objective behind the semantic smoothing technique is to discount the general words and give more emphasise for core words. The analysis of experimental results in Verma and Bhattacharyya (2009) show that the semantic smoothing models are more suitable for agglomerative clustering and not effective enough for partitional clustering.

In this paper, Medical Subject Heading (MeSH) ontology-based semantic smoothing model for biomedical document clustering is proposed. MeSH ontology is used as the domain reference for concept extraction. The experiments conducted based on this proposed method show that our method performs better than the existing methods in terms of cluster quality. Section 2 provides the related works in document clustering based on semantic smoothing approach. Section 3 presents the conventional clustering algorithm. General overview of the ontology-based semantic smoothing approach is discussed in Section 4. Section 5 presents the detailed experimental results and the performance comparison of term, concept and semantic smoothing approaches for document clustering.

2 Related works

In recent years, language modelling approach is playing a vital role in the information retrieval era. The clustering quality is getting improved by the language modelling approach due to its strong mathematical foundation and empirical effectiveness (Hamzah et al., 2007). It involves with two components, one is the construction of a language model for each document and the ranking of documents is done based on the likelihood of the query term according to the estimated language model.

Smoothing is the main problem in the language modelling approach. The inaccuracy in data sparse is handled by the smoothing technique by adjusting the MLE. The limitations of language model smoothing and its effectiveness in the information retrieval are studied in this work.

A context-sensitive semantic smoothing method is proposed for agglomerative clustering (Zhou and Hu, 2006) which can identify multiword phrases in a document automatically. Then, it maps the multiword phrases into individual terms of document using statistical methods. Multiword phrase is referred as a topic signature which defines a pair of two concepts which are semantically and syntactically equivalent to each other. Kullback-Leibler divergence (KL-divergence) metric is used as the similarity measure in this approach. The incorporation of semantic smoothing and KL-divergence similarity measure considerably improves the quality of clusters generated using agglomerative hierarchical clustering is exposed by the experimental results.

An improved semantic smoothing model is proposed which is suitable for both agglomerative and partitional clustering (Liu et al., 2007). It is framed based on the term frequency-inverse document frequency (TF-IDF) schema. It comprises of two context-sensitive semantic smoothing models named document model-based and cluster model-based techniques. In the proposed method, inverse document frequency (IDF) factor is used to enhance the ability of discounting general words in a document. The results drawn from the experiments confirm that the improved semantic smoothing model is more efficient than the existing methods.

3 Conventional clustering algorithms

The conventional partitional and hierarchical clustering algorithms are used to implement the term-based, concept-based and semantic smoothing-based clustering techniques. The functionalities of the partitional and hierarchical clustering techniques are discussed as follows:

3.1 Partitional clustering methods

Partitioning methods rearrange instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. These algorithms minimise the criterion function of the clustering process by iteratively relocating data points between the clusters until a locally optimal partition is obtained. The most commonly used methods are k-means and k-medoids. Convergence is local and the global optimal solution cannot be guaranteed in these popularly used partitional algorithms.

K-means is an iterative method that divides the given data set into K-disjoint groups. Each object can be thought of as being represented by some feature vector in an *n* dimensional space, *n* being the number of all features used to describe the objects to cluster. The algorithm then randomly chooses *k* points in that vector space, these point serve as the initial centres of the clusters. Afterwards all objects are each assigned to the centre they are closest to. Usually the distance measure is chosen by the user and determined by the learning task.

3.1.1 Algorithmic steps of K-means

1 Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2 Assign each object to the group that has the closest centroid.

3 When all objects have been assigned, recalculate the positions of the K centroids.

4 Repeat steps 2 and 3 until the centroids no longer move.

This produces a separation of the objects into groups from which the metric to be minimised can be calculated.

3.2 Hierarchical clustering methods

In data mining, a hierarchy of clusters will be produced by the method called hierarchical clustering. There are two approaches for clustering the datasets namely agglomerative and divisive methods. The bottom-up construction technique is used in agglomerative clustering for grouping smaller clusters into a larger one. The top-down approach which is splitting the larger cluster into smaller ones is called as divisive technique.

The linkage criterion is used to determines the distance between sets of observations as a function of the pairwise distances between observations in hierarchical clustering techniques. Single link, complete link and centroid-based techniques are the commonly used techniques in hierarchical agglomerative clustering.

Single link clustering defines the distance between two clusters as the minimum distance between their members. The complete link method uses the least similar pair between each of two clusters to determine the intercluster similarity; it is called complete link because all entities in a cluster are linked to one another within some minimum similarity. Small, tightly bound clusters are characteristic of this method. In the centroid method, each cluster as it is formed is represented by the coordinates of a group centroid, and at each stage in the clustering the pair of clusters with the most similar mean centroid is merged

4 Ontology-based semantic smoothing approach

This system is designed to perform semantic smoothing process based on the dynamic concept weight assignment which is supported by the ontology. This approach transforms the feature-represented documents into a concept represented one with the assistance of domain specific ontology. Therefore, the target documents will be clustered by extracting the keywords/phrases that are representing the concepts presented in the domain

ontology. In order to find the semantic relationship between the keywords to concepts and concept-to-concept, the domain ontology is used as the background knowledge.

The hybrid technique is the mixture of both the concept-based approach and the semantic smoothing technique. The MeSH ontology is used as the domain reference for finding the semantic relations such as identity, synonyms, hypernyms and meronyms between the descriptor terms contained in the documents (Zhang et al, 2008). The phrases within the documents that are having the contribution in the identification of the dominating concept are considered using n-gram techniques. The key idea of using a concept-based model with the semantic smoothing approach is to discount general words and assign reasonable counts to unseen core words. It is mainly based on the principle of TF-IDF factor. So, that the core words which are representing the concept are discriminated from the general words using this approach.

4.1 Concept identification through phrase structures

Concept-based model involves the analysis of complex semantic relations such as identity, synonyms, hypernyms and meronyms that are related to the concept queries. The terms that are representing the semantic relations within the ontology can be a single term or a multiword phrase. Such multiword phrases related to the concept are taken as topic signatures. Multiword phrase translation estimates the translation probability of each topic signature (i.e.) determining the probability of translating the given phrase to concept in the vocabulary.

4.1.1 Simple concept translation model through keywords

The documents which are pre-processed by applying the tokenisation and stop word removal are then compared with the ontology in order to find the relationship between the descriptive terms in the documents. The core words can be found by mapping the content of the pre-processed document onto the MeSH ontology. The importance of the concept in the document as well as in the corpus is determined by the dynamic weight allocation for the concept keywords and the semantic relationships. The weight allocation for the concept identification is as follows:

Table 1 Semantic relations and its weights

Semantic relation	*Initial weight*
Identity	1
Synonyms	1
Hypernyms	0.9
Meronyms	0.99

Since the hypernyms are representing more general form of the concept, all the upper level keywords that are contributing in the concept identification are assigned with the lesser values than the concept keyword. The meronyms are contributing more in the concept identification than the hypernyms. Therefore the meronyms are assigned with the closure values of concept which are not equal to 1.

Based on this mapping, the probability of the document may belong to the particular concept is identified. The first component of the semantic smoothing model is a simple

language model p_b(Concept | d). This model can be obtained by using the MLE document model p_{ml}(Concept | d) together with a background smoothing model p(Concept | Corpus) with the controlling coefficient α. The method of maximum likelihood identifies the set of values of the model parameters that maximises the likelihood function. The effect of TF-IDF scheme in convention document clustering methods is roughly equivalent to the background model smoothing. The likelihood of each concept of the document can be done using the following equations.

$$p_b\left(\text{Concept} \mid d\right) = (1-\alpha)p_{ml}\left(\text{Concept} \mid d\right) + \alpha p\left(\text{Concept} \mid \text{Corpus}\right) \tag{1}$$

$$p_{ml}\left(\text{Concept} \mid d\right) = \frac{\sum_{r \in R} \text{Freq}_r * \text{Weight}_r}{\text{Total no.of words in the document}} \tag{2}$$

$$p\left(\text{Concept} \mid \text{Corpus}\right) = \frac{\sum_{r \in R} \text{Freq}_r * \text{Weight}_r}{\text{Total no.of words in theCorpus}} \tag{3}$$

where

p_{ml}(Concept | d)	maximum likelihood probability of the document belonging to a particular concept
p(Concept | Corpus)	probability of the whole corpus belonging to that particular concept
Freq_r	denotes the frequency of the word occurring in the document
Weight_r	weight assigned to the concept relation.

The weight of each core word in the document is assigned based on the following conditions:

1 if the word is same as the identity of the query word (input) or one of its synonyms then $\text{Weight}_r = 1$

2 if the word is the parent element (since MeSH ontology is stored in a tree structure) to the input word then decrement the weight by 0.1 up one level

3 if the word is the child element of the input word then decrement the weight by 0.01 down one level.

4.1.2 Topic signature translation model

The second component of the semantic smoothing model is the topic signature (multiword phrase) translation model. Here, the probability p(Concept | t_k) of translating t_k to Concept is estimated in the training process using the following equations.

$$p_t\left(\text{Concept} \mid d\right) = \sum_k p\left(\text{Concept} \mid t_k\right) * p_{ml}\left(t_k \mid d\right) \tag{4}$$

$$p_{ml}\left(t_k \mid d\right) = \frac{\sum \text{Frequency of pharses}}{\text{Total no.of words in the document}} \tag{5}$$

$$p(\text{Concept}|t_k) = \frac{\sum_{r \in R} \text{Freq}_r * \text{Weight}_r}{\text{Total no. of words in the document}} \quad (6)$$

where

$p_{ml}(t_k \mid d)$ denotes the maximum likelihood of the phrase t_k, presenting in the document

$p(\text{Concept} \mid t_k)$ denotes the probability of translating the phrase t_k into a specific Concept after comparing it with the ontology.

The probability of translating the phrase t_k into a specific Concept is done in a sentence by sentence order. The variable Freq_r here denotes the number of core words found in the sentence being processed.

4.2 Building the ontology-based semantic smoothing model

The semantic smoothing model with ontology is a mixture model with two components, (i.e.) the simple concept translation model and the multiword phrase translation model. The influence of two components is controlled by the translation coefficient (λ) in the mixture model. The model is organised as follows

$$p_{bt}(\text{Concept} \mid d) = (1-\lambda)p_b(\text{Concept} \mid d) + \lambda p_t(\text{Concept} \mid d) \quad (7)$$

The first component is the simple concept translation model, which can be obtained using the MLE document model together with a background smoothing model with the controlling coefficient α. The second component of the document model is the multiword phrase translation model.

5 Experimental results and analysis

5.1 Datasets

The experiments are carried out for the 500 documents that are collected from PubMed based on five categories such as neoplasms, viral diseases, cardiovascular diseases, eye infection and respiratory diseases with 100 documents each.

5.2 Cluster validity measures

The quality of the clusters produced by the proposed approach is analysed with the measures such as Silhouette index, Fowlkes-Mallows (FM) index and Jaccard index.

5.2.1 Silhouette validity index

It is a measure used for verifying the accuracy in assignment of data points into the appropriate clusters. This technique computes the silhouette width for each data point, average silhouette width for each cluster and overall average silhouette width for the total

dataset (Ansari et al., 2011). To compute the silhouettes width of i^{th} data point, following formula is used:

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{8}$$

where a_i is average dissimilarity of i^{th} data point to all other points in the same cluster; b_i is minimum of average dissimilarity of i^{th} data point to all data points in other cluster. A value of S_i close to 1 indicates that the data point is assigned to a very appropriate cluster. If S_i is close to zero, it means that that data pint could be assign to another closest cluster as well because it is equidistant from both the clusters. If S_i is close to –1, it means that data is misclassified and lies somewhere in between the clusters. The overall average silhouette width for the entire data set is the average S_i for all data points in the whole dataset. The largest overall average silhouette indicates the best clustering. Therefore, the number of cluster with the maximum overall average silhouette width is taken as the optimal number of the clusters.

5.2.2 FM index

The FM index computes the similarity between the clusters returned by the clustering algorithm and the benchmark classifications. The value of the FM index is between 0 and 1, and a high value means better accuracy. It can be computed using the following formula:

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} \tag{9}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

5.2.3 Jaccard index

The Jaccard index is used to quantify the similarity between two datasets. The Jaccard index takes on a value between 0 and 1. An index of 1 means that the two dataset are identical, and an index of 0 indicates that the datasets have no common elements. The Jaccard index for the datasets A and B is defined by the following formula

$$J(A \cup B) = \frac{|A \cap B|}{|A \cup B|} \tag{10}$$

For the experimental analysis both hierarchical and partitional clustering algorithms with the distance measure Euclidean distance and Pearson correlation are used. The performance measures Silhouette index, FM index and Jaccard index are evaluated in clustering algorithms K-means and simple, complete and centroid linkage methods for hierarchical clustering which use Eucledian distance measure and Pearson correlation. Table 2 shows the weight assigned for a query term based on term, concept and smoothing from randomly selected three documents from the five document corpus. The comparison of cluster analysis for both partitional K-means and hierarchical algorithms with Euclidean distance measure is recorded in Table 3. Table 4 gives the consolidated

analysis of term-based, concept-based and semantic smoothing-based methods for both the partitional K-means and hierarchical algorithms using Pearson correlation measure.

Table 2 Term, concept and smoothing-based weights of sample documents

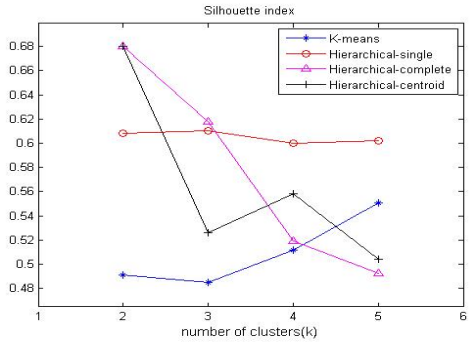
Domain	*Query term*	*Sample documents*	*Term-based weight TF-IDF*	*Ontology-based (dynamic weight allocation based on semantic relation)*	
				Concept weight	*Semantic smoothing model-based weight*
Neoplasm	Cancer	C_{r1}	0.016804	0.097452	0.058752
		C_{r2}	0	0.35	0.279318
		C_{r3}	0.097711	0.281111	0.167727
Respiratory diseases	Asthma	A_{r1}	0.041458	0.095597	0.048482
		A_{r2}	0	0.224648	0.113007
		A_{r3}	0	0.294568	0.149272
Cardiac diseases	Cardiovascular	H_{r1}	0	0.100686	0.051266
		H_{r2}	0.033968	0.212255	0.10705
		H_{r3}	0.008683	0.190132	0.095989
Eye infection	Conjunctivitis	E_{r1}	0	0.129412	0.065577
		E_{r2}	0.006242	0.237299	0.11931
		E_{r3}	0	0.178627	0.090165
Viral diseases	Dengue	D_{r1}	0.006989	0.2168	0.111368
		D_{r2}	0	0.243293	0.13064
		D_{r3}	0.02102	0.135263	0.070283

5.3 Result analysis

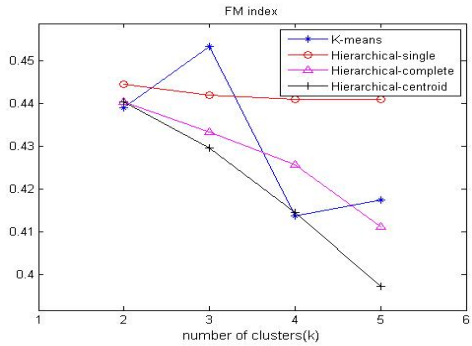
5.3.1 Term-based clustering

Figure 1 shows the performance of term-based clustering using Euclidean distance measure for Figure 1(a) Silhouette index, Figure 1(b) FM index and Figure 1(c) Jaccard index. It shows that for K-means clustering, the Silhouette index, FM index and Jaccard index are high for cluster size 5 compared to other clustering methods. This is because the five different document corpus are used for clustering. Regarding Silhouette index, hierarchical algorithms (single) produce better quality clusters than K-means. Whereas for FM index and Jaccard index, K-means outperforms hierarchical methods.

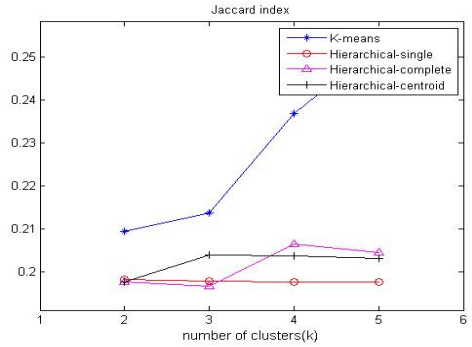
Figure 1 Term-based clustering with Euclidean distance measure (see online version for colours)



(a)



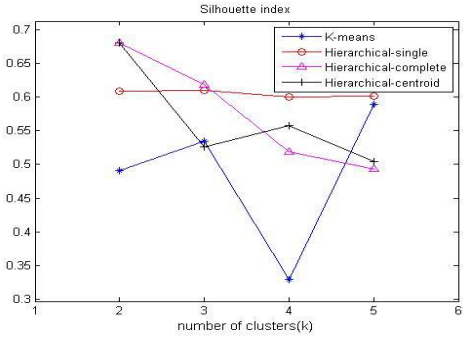
(b)



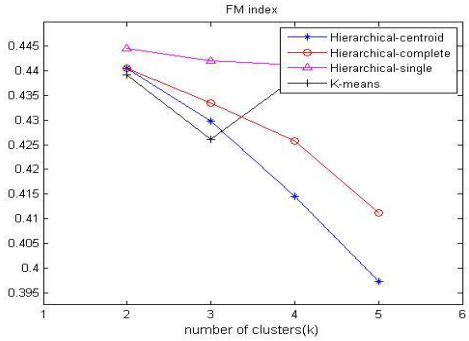
(c)

Figure 2 shows the performance of term-based clustering using Pearson correlation measure for Figure 2(a) Silhouette index, Figure 2(b) FM index and Figure 2(c) Jaccard index.

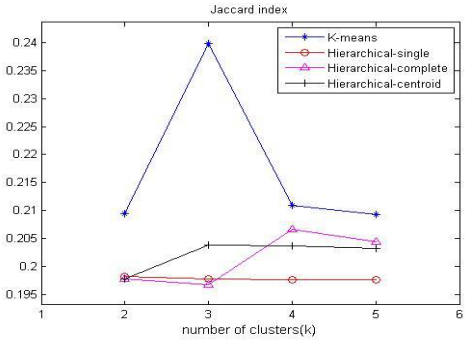
Figure 2 Term-based clustering with Pearson correlation (see online version for colours)



(a)



(b)

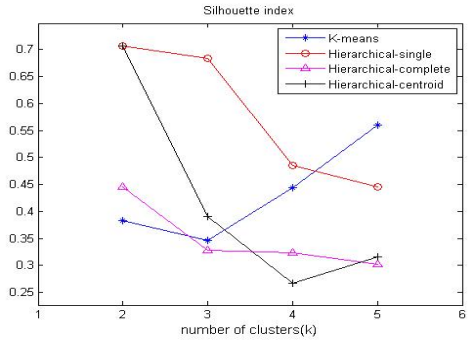


(c)

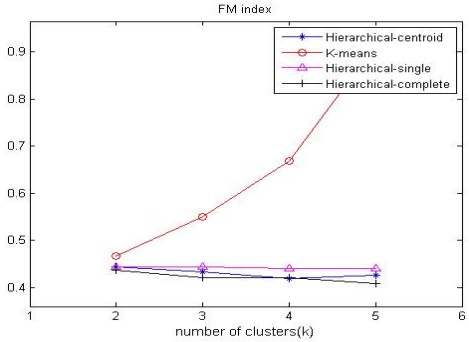
5.3.2 Concept-based clustering using ontology

Figure 3 shows the performance of concept-based clustering using Euclidean distance measure for Figure 3(a) Silhouette index Figure 3(b) FM index and Figure 3(c) Jaccard index. In this K-means clustering outperforms hierarchical clustering for all three indices.

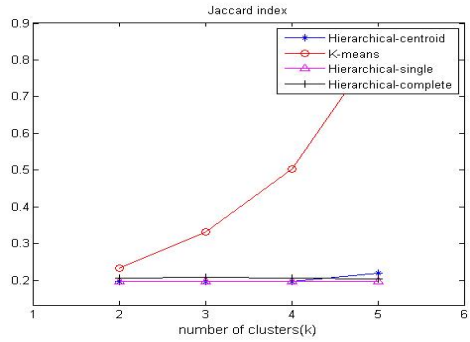
Figure 3 Concept-based clustering with Euclidean distance (see online version for colours)



(a)



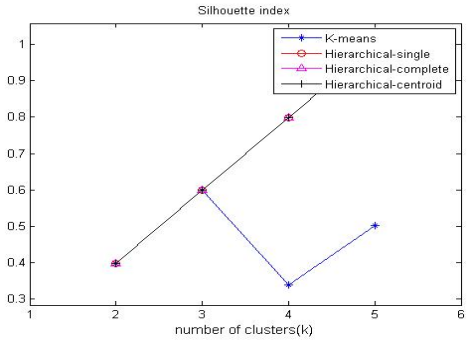
(b)



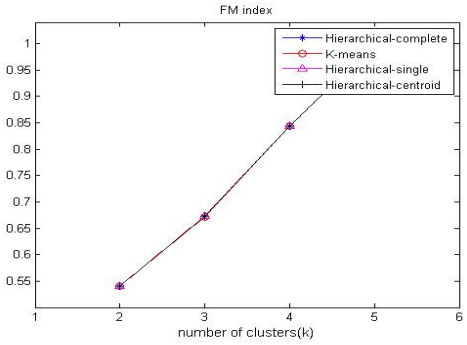
(c)

Figure 4 shows the performance of concept-based clustering using Pearson correlation as the similarity measures for Figure 4(a) Silhouette index, Figure 4(b) FM index and Figure 4(c) Jaccard index. The analysis on the clustering results shows that both hierarchical and partitional K-means are producing clusters with same quality for FM index and Jaccard index. Hierarchical clustering methods outperform K-means for Silhouette index.

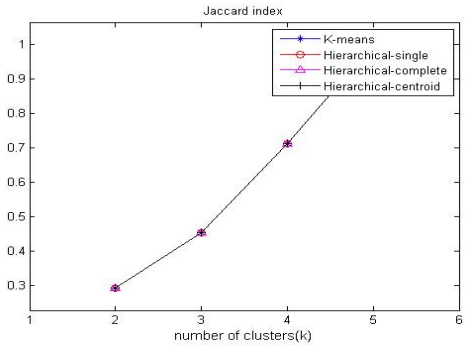
Figure 4 Concept-based clustering using Pearson correlation (see online version for colours)



(a)



(b)

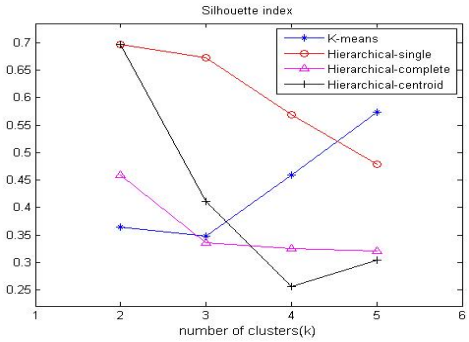


(c)

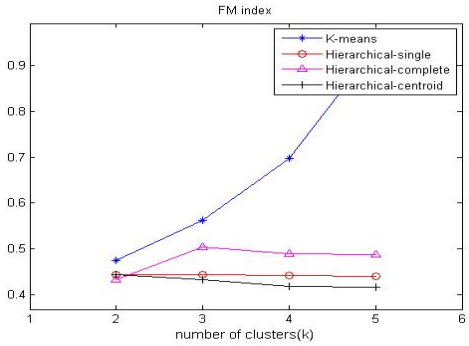
5.3.3 Semantic smoothing-based clustering using ontology

Figure 5 shows the performance of semantic smoothing-based clustering using Euclidean distance as the similarity measure for Figure 5(a) Silhouette index, Figure 5(b) FM index and Figure 5(c) Jaccard index. For all the three indices K-means outperforms hierarchical clustering algorithms.

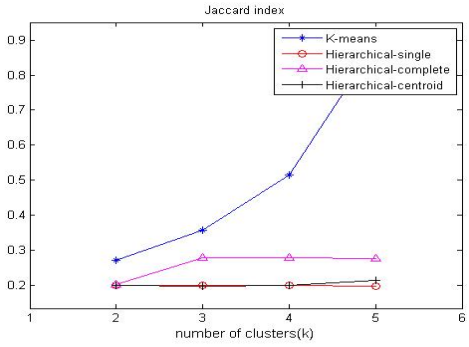
Figure 5 Semantic smoothing using Euclidean distance (see online version for colours)



(a)



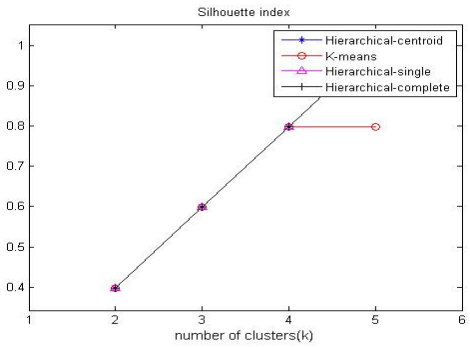
(b)



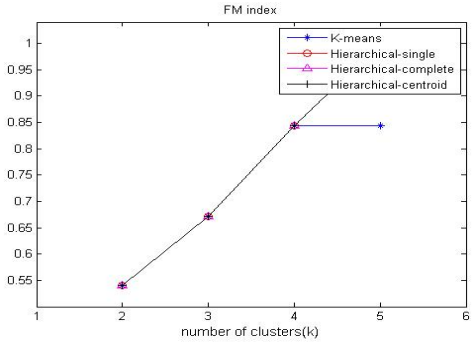
(c)

Figure 6 shows the performance of semantic smoothing-based clustering using Pearson correlation as the similarity measure for Figure 6(a) Silhouette index, Figure 6(b) FM index and Figure 6(c) Jaccard index. The analysis on experimental results based on the semantic smoothing approach shows that the hierarchical algorithms outperforms K-means algorithm for all the Silhouette, Jaccard and FM indices.

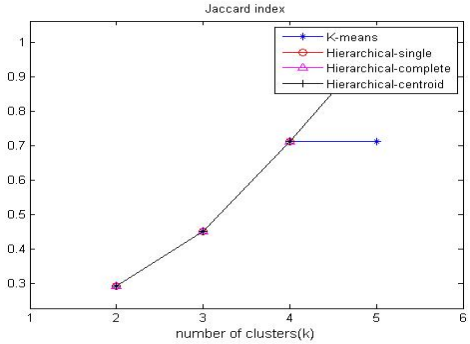
Figure 6 Semantic smoothing using Pearson correlation (see online version for colours)



(a)



(b)



(c)

Table 3 Result analysis for five clusters using Euclidean measure

Model	*Technique*	*Silhouette index*	*Jaccard index*	*FM index*
Term-based model TF-IDF	K-means	0.53396	0.25264	0.44519
	Hierarchical single	0.61018	0.19824	0.44453
	Hierarchical complete	0.68018	0.20657	0.44053
	Hierarchical centroid	0.68018	0.20383	0.44053
Ontology-based simple concept model	K-means	0.56011	0.84094	0.91363
	Hierarchical single	*0.70642*	0.19809	0.44365
	Hierarchical complete	0.44544	0.20784	0.43653
	Hierarchical centroid	*0.70642*	0.22046	0.44365
Ontology-based semantic smoothing	K-means	0.57426	*0.88944*	*0.94149*
	Hierarchical single	0.69675	0.1981	0.44366
	Hierarchical complete	0.45973	0.27727	0.50338
	Hierarchical centroid	0.69675	0.21376	0.44366

Table 4 Result analysis for five clusters using Pearson correlation measure

Model	*Technique*	*Silhouette index*	*Jaccard index*	*FM index*
Term-based model TF-IDF	K-means	0.35902	0.2467	0.35902
	Hierarchical single	0.32264	0.19824	0.44453
	Hierarchical complete	0.45276	0.45021	0.6344
	Hierarchical centroid	0.45276	0.45021	0.6344
Ontology-based simple concept model	K-means	0.7996	0.71141	0.84467
	Hierarchical single	*1*	1	1
	Hierarchical complete	1	1	1
	Hierarchical centroid	*1*	1	1
Ontology-based semantic smoothing	K-means	0.79899	0.71223	0.84394
	Hierarchical single	1	1	1
	Hierarchical complete	1	1	1
	Hierarchical centroid	1	1	1

5 Conclusions

An ontology-based semantic smoothing model is proposed in this work. Clustering is based on the domain knowledge with dynamic weight assignment. The results are analysed using partitional and hierarchical clustering algorithms. Silhouette index, Jaccard index and FM index are used to measure the performance of the clustering processes and the quality of resultant clusters. Hierarchical algorithms outperform K-means for the Pearson correlation. In Euclidean distance measure, K-means outperforms hierarchical methods for Jaccard and FM indices. This work shows the Eucledian measure gives better performance for K-means and Pearson correlation

provides better result for hierarchical clustering methods. The concept-based clustering outperforms the smoothing model for Silhouette index. The proposed smoothing model outperforms term-based and concept-based methods in Jaccard and FM indices.

References

Ansari, Z., Vinaya Babu, A., Azeem, M.F. and Ahmed, W. (2011) 'Quantitative evaluation of performance and validity indices for clustering the web navigational sessions', *World of Computer Science and Information Technology Journal*, Vol. 1, No. 5, pp.217–226.

Hamzah, A., Susanto, A., Soesianto, F. and Istyanto, J.E. (2007) 'Concept-based text document clustering', *Proceedings of the International Conference on Electrical Engineering and Informatics*, pp.210–213.

Jayabharathy, J., Kanmani, S. and Parveen, A.A. (2011) 'Document clustering and topic discovery based on semantic similarity in scientific literature', *IEEE International Conference on Communication Software and Networks*, pp.425–429.

Liu, Y., Cai, J., Yin, J. and Huang, Z. (2007) 'Document clustering based on semantic smoothing approach', *Advances in intelligent Web Mastering*, Springer Berlin Heidelberg, pp.217–222.

Tar, H.H. and Nyaunt, T.T.S. (2011) 'Enhancing traditional text documents clustering based on ontology', *International Journal of Computer Applications*, Vol. 33, No. 10, pp.38–42.

Tu, X., He, T., Chen, L., Luo, J. and Zhang, M. (2010) 'Wikipedia-based semantic smoothing for the language modeling approach to information retrieval', *Proceedings of the 32nd European Conference on Advances in Information Retrieval*, pp.370–381.

Verma, K.S. and Bhattacharyya, P. (2009) 'Context-sensitive semantic smoothing using semantically relatable sequences', *Proceedings of International Joint Conference on Artificial Intelligence*, pp.1580–1585.

Zhang, L. and Wang, Z. (2010) 'Ontology-based clustering algorithm with feature weights', *Journal of Computational Information Systems*, Vol. 6, No. 9, pp.2959–2966.

Zhang, X., Jing, L., Hu, X., Ng, M. and Xia, J. (2008) 'Medical document clustering using ontology-based term similarity measures', *International Journal of Data Warehousing and Mining*, Vol. 4, No. 1, pp.62–73.

Zhou, X. and Hu, X. (2006) 'Context-sensitive semantic smoothing for model-based document clustering', *Proceedings of ICDM*, pp.1193–1198.

Zhou, X., Zhang, X. and Hu, X. (2007) 'Semantic smoothing of document models for agglomerative clustering', *Proceedings of International Joint Conference on Artificial Intelligence*, pp.2922–2927.