Text Mining Approach for Prediction of Tumor Using Ontology Based Particle Swarm Optimization with Clustering Techniques

[1]P. Jyotsna, [2]P. Govindarajulu

[1]Research Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh
[2]Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India

Abstract
Text mining with Particle Swarm Optimization (PSO) Clustering Techniques to build a tumor prediction scheme. The proposed prediction scheme is based on Historical medical Reports associated with Tumor data. This research approach provides Effective Clustering by using Semantic Similarity that is calculated in Historical medical Reports Annotation Process. The Clustering Techniques group the reports into unsupervised cluster based on the features of the medical Reports. The Document Clustering is done through PSO. A PSO with ontology model of Clustering Knowledge Representation based on Historical medical report documents is presented and Compared to the traditional Support vector machine (SVM) approach. The SVM Methods to carry out the Integration of Medical ontology and the Text mining techniques is accomplished of mining the potential patterns and categorize clinical medical reports. Proposed ontology based frame work provides improved performance and better clustering compared to the traditional SVM Clustering.

Keywords:
Tumor, Prediction, Text mining, Ontology, Clustering, Support vector machine

1.Introduction

Medical reports clustering can be applied to improve the retrieval process. Fast and high quality document clustering algorithms play an important role in effective navigating, summarizing and organizing information [3,4]. The clustering methods may improve the results of text data mining application. The goal of document clustering is to partition documents into clusters according to their similarities. Major steps of document clustering include tokenization, stemming, elimination of stop words, index term selection, the term document matrix setup and clustering. The PSO algorithm is a population based analyzed statistically optimization technique that can be used to find an optimal or near optimal solution to a numerical and qualitative problem [6]. PSO can be successfully applied to a wide range of applications and a survey of the method can be found in Kennedy et al [10].
The clustering problem is an optimization problem that locates the optimal centroids of the clusters. One of the most popular in clustering is K-means algorithm [15]. The drawback of K-means algorithm is that it may end in local optima. Therefore, it is necessary to use some other global searching algorithm for generating the clusters. The PSO can be used to generate the clusters and it can avoid being trapped in the local optimal solution. In the PSO document clustering algorithm, the multi-dimensional document vector space is modeled as problem space [3]. In Cui and Potok[4], the document clustering is evaluated using the mean distance of document to cluster centroid (MDDCC). Text Clustering is suitable Technique used to partition a huge set of Text based medical reports into a predetermined number of Clusters
The clustering problem can be viewed as an optimization problem. In the PSO algorithm, the particle's location represents one solution for the problem. A different problem solution can be generated based on the particle movements to a new location. In each movement of the particle, the velocity and direction will be altered. When while most of the clustering procedures perform local searching, the PSO algorithm performs a global searching for solutions. The objective of the PSO clustering algorithm is to discover the proper centroids of clusters for minimizing the intra-cluster distance as well as maximizing the distance between clusters.
Particle swarm optimization (PSO) is a population-based a random probability distribution of algorithm modeled on social behaviors observed in flocking birds . A particle flies through the search space with a velocity that is dynamically adjusted according to its own and its companion's historical behaviors. Each particle's position represents a solution to the problem. Different topology structure can be utilized in PSO, which will have different strategy to share search information for every particle. PSO algorithms have recently been shown to produce good results in a wide variety of real-world data
The PSO algorithm performs a globalized searching for solutions whereas most other partitioned clustering procedures perform a localized searching [4]. The optimized clustering model that uses annotation is proposed to inherit the advantages of semantics and the relationship between the terms. The method is used to enhance the effectiveness of document clustering algorithms. PSO algorithm can be applied on the annotated documents to find the optimal centroids of the

Manuscript received May 5, 2018
Manuscript revised May 20, 2018

clusters. The clustering performance is evaluated using particle swarm optimization.

2. Literature Review

Hotho et al[7] have used the approach of applying background knowledge during preprocessing in order to improve clustering results for selection between results.

Zhang et al [20] have discussed various semantic similarity measures of terms including path based similarity measure, information content based similarity measure and feature based similarity measure.

Song et al [18] have discussed how to cluster databases semantically. The semantic measure uses ontology and then a hybrid PSO algorithm is provided for databases clustering.

Dash et al [5] have used the hybrid clustering algorithm to find the centroids of a user specified number of clusters, where each cluster groups similar patterns.

Killani et al [11] present a comparative analysis of K-means and PSO clustering performances for text datasets. The result in the work shows the PSO approaches find better solution compared to K-means due to its ability to evaluate many cluster centroids simultaneously in any given time unlike K-means.Many methodologies, languages and tools are used to support ontological Text Mining process.

Niknam et al. [14] presented "Application of a New Hybrid optimization Algorithm on Cluster Analysis Data clustering" proposed an efficient hybrid evolutionary optimization algorithm based on a combination of Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO), so as to be called PSO-ACO, for optimally clustering N object into K clusters. In this algorithm, the decision making process of each particle for Select best particle Yes Generate initial particles Print the result Calculate Particle velocities Update particle positions Update the best particle Evaluate particle; The performance of the new PSOACO algorithm was compared with those of ACO, PSO and K-means clustering. The simulation results revealed that the proposed evolutionary optimization algorithm is suitable for handing data clustering.

Cui et al. [3] Presented" Document Clustering Analysis Based on Hybrid PSO+Kmeans Algorithm" it describes hybrid PSO based algorithm for document clustering. In this algorithm, they applied the PSO, K-means and a hybrid PSO clustering algorithm on four different text document datasets. The results have shown that the hybrid PSO algorithm can generate more compact clustering results than the K-means algorithm.

Sridevi and Nagaveni [19] presented "Semantically Enhanced Document Clustering Based on PSO Algorithm" presented a clustering algorithm that employs semantic similarity measure. They have proposed a model by combining ontology and optimization technique to improve the clustering In this model the ontology similarity is used to identify the importance of the concepts in the document and the particle swarm optimization is used to cluster the document.

D. Asir Antony Gnana Singh, E. Jebamalar Leavline, K. Valliyappan and M. Srinivasan [1] presented "Enhancing the Performance of Classifier Using Particle Swarm Optimization (PSO) - based Dimensionality Reduction".it proposes PSO and F-Score based feature selection algorithm for selecting the significant features that contribute to improve the classification accuracy. This proposed method employs the particle swarm optimization and the F-score feature selection metrics.

V. Krishnaiah et al[12] Presented "Diagnosis of Lung Cancer Prediction system using Data mining Classification Techniques" This is a prototype lung cancer disease prediction system using Data mining Classification techniques. This model predicts patients with lung cancer disease appears to be naviebayes followed by IF-THEN rule.

Pratiksha Y. Pawar and S. H. Gawande, Member, IACSIT [17] Presented "A Comparative Study on Different Types of Approaches to Text Categorization" this is focused on The documents can be classified by three ways unsupervised, supervised and semi supervised methods. Text categorization refers to the process of assign a category or some categories among predefined ones to each document, automatically. They present a comparative study on different types of approaches to text categorization. it can be concluded that SVM classifier has been recognized as one of the most effective text classification method in the comparisons of supervised machine learning algorithms.

Liliya Demidova, Evgeny Nikulchev, Yulia Sokolova [13] Presented "The SVM Classifier Based on the Modified Particle Swarm Optimization" it describes The idea of particles' «regeneration» is put on the basis of the modified particle swarm optimization algorithm. At the realization of this idea, some particles change their kernel function type to the one which corresponds to the particle with the best value of the classification accuracy. The offered particle swarm optimization algorithm allows reducing the time expenditures for development of the SVM classifier. The results of experimental studies confirm the efficiency of this algorithm.

3. Framework For Ontology-Based Reports Clustering

In this work we apply text mining techniques on a corpus of clinical reports to extract the potential designs from the text collection. In addition, we developed a tumor

knowledge base for supporting medical diagnosis. Here we describe the framework for discovery of the relationships between Tumor and probable factors from clinical medical reports. The proposed text mining approach consists of two levels. In the first level we applied an ontology-based keyword extracting process. After extracting keywords, they are used as index terms to encode the medical reports. Secondly we applied the optimized clustering approach is used to improve the relevance. Consequently, we analyzed the Historical medical reports and identified a set of issues to obtain potential terms associated with tumor diseases. The Experiment is conducted using classifiers namely Support Vector Machine (SVM) with ontology and PSO with K-means Clustering approaches. We applied a Support Vector Machines (SVM) classifier method [17] to supporting acquisition of relatedness among texts and biomedical literatures.

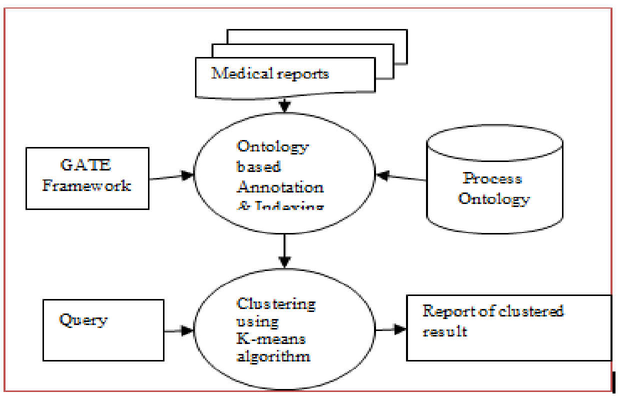

Fig. 1 Ontology –based model for Medical Reports Clustering

In contrast to the existing methods, this research approach provides effective clustering by using the semantic similarity that is calculated in REPORTS annotation process. The proposed approach is composed of two main parts. Firstly, the REPORTS annotation is done using ontology and secondly, the optimized clustering approach is used to improve the relevance. The novelty of this approach resides in the annotation and in applying PSO to retrieve optimized result. In the traditional retrieval system, keyword method cannot meet the need of user's requirements. The clustering techniques group the documents into unsupervised cluster based on the features of the documents. The document annotation model is utilized for the clustering of the documents.

The main features of the model proposed by this research, in comparison with the models such as the values of the variables are the truth values true and false, statistical and probabilistic are:

1. Combining of the conceptual features of documents and to develop an effective annotation
2. Using semantic similarity score to improve the concept relevance
3. Using optimization algorithm to improve the clustering performance.

3.1 PSO CLUSTERING

In the PSO system, particles move through the search space which allows the particles to explore more information space and better the chance of finding global optima. One particle's position in the swarm represents one possible solution for clustering the document collection. Therefore, a swarm represents a number of candidate clustering solutions for the document collection [4]. Each particle maintains a matrix

$$P1g = (cck_1, cc\ k_2, cck_{kh}, .., cckm),$$

Where cckm is the cluster centroid of kth cluster. The cckh is represented as position vector xkh. xkh is the hth element position of the Kthparticle. The particle in the swarm has its own position and velocity. The movement of the particle is directed by the velocity. The term Vkh represents the hth element of the velocity vector of the kth particle. For every generation, the particle's new location is computed by adding the particle's current velocity V-vector to its location X-vector. In Equation (1), the random values rand 1 and rand 2 are used for a wide search. The øfactor is the inertia random value. The values of c1 and c2 control the weight in deciding the particle's next movement.

$$V_{kh} = \acute{Ø} \text{ x } v_{kh} + c_1 \text{ x } rand_1 \text{ x } (Ppbest - x_{gh}) + c_2 \text{ x } rand_1 (Pgbest - x_{호}) \quad (1)$$

$$X_{gh} = x_{gh} + v_{gh} \quad (2)$$

At every generation, the particle's new location is computed by adding the particle's current velocity, vgh, to its location, xgh using equation (1) and equation (2) and thus the velocity of the particle is updated [9]. Each particle stores the personal and best position in the search space [6]. The personal best position of gth particle is the best position visited by the particle so far and is denoted as Ppbestg(t). Let 'f 'be the objective function and the personal best of particle at time step t is updated through Equation (3).

$$\text{If } (Ppbest_g(t+1) = Ppbest_g(t))$$

$$f\ (x_g(t+1) \geq f\ (Ppbest_g(t))$$

else

$$f\ (x_g(t+1) < f\ (Ppbest_g(t)) \quad (3)$$

The best solution is represented as Pgbest. The best particle is determined from the entire swarm by selecting the personal best position. The best is updated based on the best known position of the swarm using Equation (4).

Pgbest (t) € ($Ppbest_0$, $Ppbest_{1.....}Ppbest_k$)=min (f ($Ppbest_0$ (t),…$Ppbest_k$(t)) (4)

The PSO clustering algorithm needs a fitness function to evaluate each particle performance. Each particle maintains a matrix with the cluster centroid vectors. The inertia weight Ǿchanges the momentum of particles to avoid the stagnation of particles at the local optima. The initial iteration can be a global searching stage. After several iterations, the particle's velocity will reduce and the searches will gradually reduce while the particle will approach the optimal solution.

3.2 CLUSTERING USING PSO ALGORITHM

The PSO algorithm is given as follows:
Step 1: Randomly initialize the particle velocity and particle position. The K cluster centroids are randomly selected for each particle.
Step 2: Evaluate the fitness for each particle for the initial values
Step 3: Preserve the document cluster structure optimally.
The cluster quality can be measured using within cluster, between-cluster and mixture scatter matrices and it is given in Equation (5), Equation (6) and Equation (7) respectively.
The cluster quality is high if the cluster is tightly grouped, but well separated from the other clusters. Equation (5) defines the within cluster and the Equation (6) defines the similarity measure between clusters. The function preserves the cluster structure and maximizes the Equation (7).

$$C_W = \sum_{i=1}^{k} \sum_{j=Ni} (d_j - cc_i)\,(d_j - cc_i)^{\mathrm{T}} \quad (5)$$

$$C_b = \sum_{i=1}^{k} \sum_{j \in N_i} (c_j - pc)\,(c_j - pc)^{\mathrm{T}} \quad (6)$$

$$C_m = \sum_{j=1}^{n} (d_j - pc)\,(d_j - pc)^{\mathrm{T}} \quad (7)$$

Where dj is the document that belongs to the cluster cci. cci is the ithcluster centroid, pc is the process cluster centroid, K is the number of clusters and Ni is the number of documents in the cluster cci. Cwis the within cluster, Cb is between clustering and Cm is the mixture scatter matrices which is sum of Cw and Cb. The fitness value of the particle is evaluated using Cb. The goal of clustering is to attain high intra-cluster similarity and low inter-cluster similarity.
Step 4: The particle's best position found thus far, using Equation (3)
Step 5: The best position in the neighborhood of that particle using Equation (4)
Step 6: Apply velocity update for each dimension of the particle using Equation (1)
Step 7: The position and generate new particle's location using Equation (2)
Step 8: Repeat steps 2–8 until a stopping criteria, such as a sufficiently good solution is discovered or a highest n number of generations is completed. The individual particle that scores the best fitness value in the population is taken as optimal solution.
Step 9: The relevance of documents is evaluated

4. Experimental Result

4.1. Data set Description

The Medical data set is used for text clustering. The hospital historical Tumor data which is the subset of oncology departments which is used as a data set and it has a collection of Reports from medical oncology. Each category consists of 1000 Reports assigned to it. Each department is stored in a subdirectory, with each particle stored as a separate file. Some of the sign and symptoms are closely related with each other while some are highly unrelated. For the purpose of experimentation, clustering is done using up to causes of most tumor, they do know some of the risk factors that increase the likelihood of a tumor in different organs. Using the approach of the four datasets are extracted from the benign or malignant tumor datasets. The datasets T1 and T2 contain categories of different symptoms. The datasets T3, T4 consist of categories of similar symptoms collected from the collection.
Table 1 and Table 2 show the different symptoms organized into hospital historical medical reports. Table 1 describes the dataset used for evaluation and tumor is the one of the leading cause of benign and malignant in both women and men. Manifestation of tumor in the body of the patient reveals through early symptoms in most of the cases. The documents are annotated using ontology and it is stored using the GATE tool [2].

Table 1: Datasets used for evaluation

Dataset	No. of Document	No. of Classes	No. of unique terms
Historical Medical Reports	15828	20	15330

4.2. Cluster Evaluation

20 random particles are generated. The fitness value is calculated using the distance between the cluster centroid and the documents that are clustered. The fitness values are recorded for ten simulations. In the PSO module, the inertia weight ø initially set as 0.95 is chosen and the acceleration coefficient constants c1 and c2 are set to 2. If the fitness value is fixed, then it indicates the optimal solution. For each simulation, the initial centroid vectors of each particle are randomly selected. A set of 20 queries was prepared manually for comparative performance measurement. After 100 iterations the same cluster solution is obtained. The F-measure values are the average of 20 runs. The parameters and their values are shown in Table 2.

Experimental results show that the PSO clustering algorithm using ontology performs better than the PSO clustering algorithms without using ontology. To cluster the large document data sets, PSO requires more iteration to find the optimal solution.

Table 2: Particle Swarm Optimization parameters and their values

Parameter	Value
No. of clusters	10
No. of particles	20
Maximum no. of iterations	100
C1	2
C2	2
Ø	0.95
Rand1 and rand2	Uniformly distributed random variable between 0 and 1

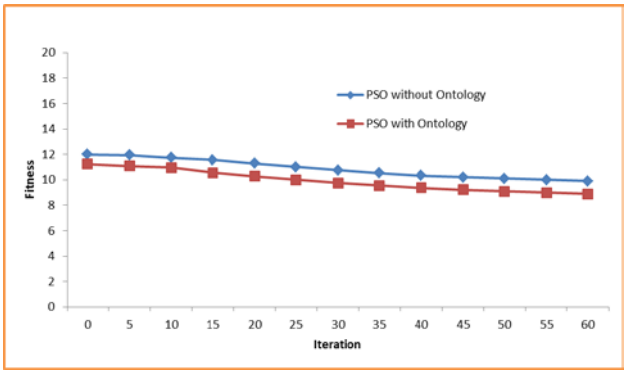

Fig. 2 Fitness curve comparisons

As shown in Figure 2, the PSO approach generates the clustering result with the lowest fitness value for all four datasets using the Euclidian similarity measurement and the ontology-based similarity measurement. In order to make fair comparisons, we applied the experiments over 20 runs. This number is selected based on the previous studies in text clustering domain which sufficient to evaluate the proposed method.

The convergence of PSO depends on the particle's best known position and the swarm's best known global position. If the global position remains constant throughout the optimization process then the algorithm converges to the optimal. For a large data set, the PSO requires more iteration before optimal clustering. The PSO clustering algorithm can generate more compact clustering results. The PSO approach in ontology-based SVM has improvements compared to the results of the SVM-based PSO. However, when the similarity measurement is changed to the ontology-based measurement, the PSO algorithm has a better performance.

Table 3: Performance comparison of K-mean and PSO

		Medical records value	
		K-means	**PSO**
D1	Euclidian	7.11	5.38
	Ontology	9.413	7.146
D2	Euclidian	5.252	3.518
	Ontology	8.468	6.112
D3	Euclidian	3.132	1.221
	Ontology	4.661	3.11
D4	Euclidian	7.111	5.234
	Ontology	11.772	8.334

The convergence behavior of K-means and PSO algorithm are given in Table 3. Since 100 iterations is not enough for the PSO algorithm to converge to an optimal solution, the result values in the Table 3 indicate that the PSO approach have improvements compared to the results of the K-means approach when using the Euclidian similarity measurement. However, when the similarity measurement is changed to the ontology similarity measurement, the PSO algorithm has a better performance than the K-means algorithm.

4.3. Discussion on the Clustering Quality

For clustering, two measures of cluster "goodness" or quality are used. One type of measure allows to compare the different sets of clusters without reference to external knowledge and is called an internal quality measure. The other type of measures permits an evaluation of how well the clustering is working by comparing the groups produced by the clustering techniques to known classes. This type of measure is called an external quality measure. The commonly measured clustering quality F-measure and entropy are used to evaluate the quality of the clusters produced by the clustering algorithm.

Entropy is a commonly used external validation measures for K-means clustering [8]. As external criteria, entropy uses external information — class labels in this case. Indeed, entropy measures the purity of the clusters with respect to the given class labels. Thus, if every cluster consists of objects with only a single class label, the entropy is 0. However, as the class labels of objects in a cluster become more varied, the entropy value increases. To compute the entropy of a set of clusters, we first

calculate the class distribution of the objects in each cluster, i.e., for each cluster j we compute pij , the probability that a member of cluster j belongs to class i. Given this class distribution, the entropy of cluster j is calculated using the standard entropy

$$E_j = -\sum_{j=1}^{K} p_{ij} \log(p_{ij}) \quad (8)$$

where the sum is taken over all classes and the log is log base 2. The total entropy, for a set of clusters is computed as the weighted sum of the entropies of each cluster, as shown in equation (9)

$$E = -\sum_{j=1}^{k} \frac{n_j}{n} E_j \quad (9)$$

Where nlj is the size of cluster j, K is the number of clusters, and n is the total number of all data points. The entropy considers the overall distribution of all the categories in a given cluster. In general, the smaller the entropy value, the better the quality of the cluster is. At each step of the iteration process, the entropy value is given in Figure 3.

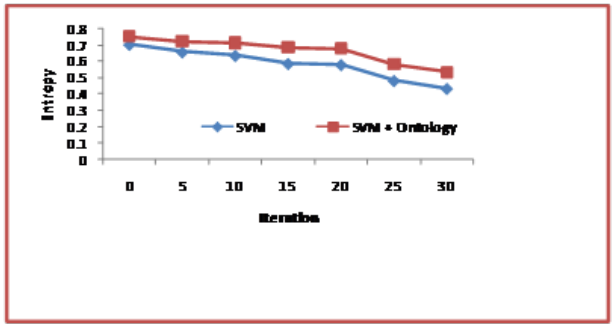

Fig. 3 Comparison of clustering quality using entropy

The cluster quality for the four data sets from two type of medical records is given in Figure .4. For different data sets in Lung tumor sign and symptoms the F-measure performance of ontology-based SVM increases greatly when compared to the SVM method. It shows consistent performance in F- measure in ontology based SVM and is still the best among these methods. The aim of the algorithm is to maximize the F-measure. The higher the overall F-measure, the better the clustering accuracy achieved. The semantic similarity measure outperforms the keyword-based similarity measure model.

$$p = \frac{TP}{TP+FP} \quad (10)$$

$$R = \frac{TP}{TP+FN} \quad (11)$$

$$F = \frac{2.\ P.R}{P+R} \quad (12)$$

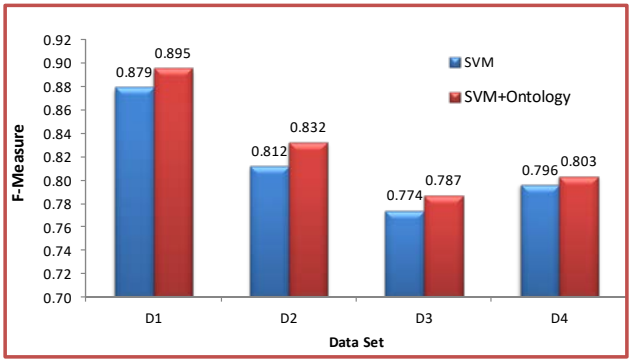

Fig. 4 Comparison of clustering quality

As shown in Table 4 to evaluate the performance of the proposed method in terms of cluster quality, the experiment is conducted using classifiers namely support vector machine(SVM) and ontology-based SVM method produces better results. It is obvious that the SVM model with ontology similarity plays an important role in accurately judging the relation between documents.

Table 4: Performance evaluation of document representation model

Dataset	SVM		SVM+ Ontology	
	F-Measure	Entropy	F-Measure	Entropy
D1	0.879	0.069	0.895	0.063
D2	0.812	0.051	0.832	0.049
D3	0.774	0.471	0.787	0.46
D4	0.796	0.373	0.803	0.343

The mean relevance for the top 20 retrieved documents using ontology is 3.1 and the same using the keyword based search is 2.2. SVM + Ontology obtain the best clustering performance in terms of the results of Precision and Recall. Evaluation results on the chosen process domain and the annotation ontology have shown improvements of retrieval performance compared to simple keyword matching approaches. The performance of ontology-based SVM shows a better improvement than traditional SVM approach. The clustering results, produced by the semantic similarity approach, have a higher quality than those produced by a keyword based similarity approach. The evaluations results indicate that the semantic annotation, indexing and retrieval approach improve the performances of retrieval not just in terms of Recall, but also in terms of Precision.

Based on the clustering results [15], the clustering quality is compared with the standard K-means and PSO in terms of the ontology distance measure. The traditional SVM is compared with Number-based Term Similarity (NBTS) method and Ontology-based NBTS. The standard K-means and PSO on the traditional SVM are compared for their relative effectiveness on the clustering quality. **REFS**

is relative effectiveness in F-measure and **REEN** is relative effectiveness in entropy measure. Table5 shows the relative effectiveness in F-measure and Entropy.

Table 5: Relative Improvement of F-score and Entropy

Methods	Measure	Percentages			
		D1	D2	D3	D4
K-means	RE F-score	5.80	8.43	1.71	5.87
	RE Entropy	6.43	10.79	5.01	14.23
PSO	RE F-score	5.55	8.10	1.95	6.14
	RE Entropy	8.34	11.30	6.21	16.17

***RE –Relative effectiveness**

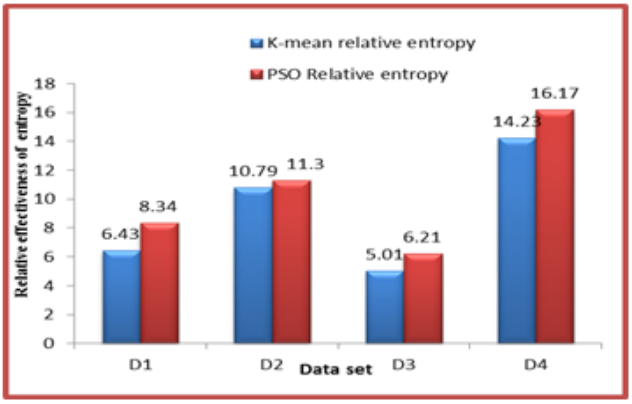

Fig. 5 Comparison of Relative Effectiveness of entropy

Several observations from the comparison between these methods are worth discussing:

1. Our ontology-based SVM clustering results are more accurate than the traditional SVM clustering.
2. The semantic similarity performs consistently and significantly better than the keyword-based term frequency methods and achieves the best performance in all experiments.
3. The PSO clustering approach combined with semantic similarity performs significantly the best among the traditional clustering methods in all the experiments of this research.

5. Conclusion

The proposed model improves the text clustering model by combining the annotation weights. The aim of the current research work is to provide qualitative improvement over SVM-based search by using ontology.
The document clustering is done through Particle Swarm Optimization. A PSO-based ontology model of clustering knowledge documents is presented and compared to the traditional vector space model. The proposed ontology-based framework provides improved performance and better clustering compared to the traditional SVM model. It also overcomes the problems existing in the SVM model commonly used for clustering. The clustering result based on semantic similarity has higher fitness values than those based on the traditional SVM model. It is worth pointing out that the above observations are made by combining semantic similarity and PSO clustering in terms of F-measure.

References

[1] D. Asir Antony Gnana Singh, E. Jebamalar Leavline, K. Valliyappan and M. Srinivasan presented “Enhancing the Performance of Classifier Using Particle Swarm Optimization (PSO) - based Dimensionality Reduction”. International Journal of Energy, Information and Communications Vol.6, Issue 5 (2015), pp.19-26

[2] Cunningham, H. “GATE, a General architecture for text engineering”. Computers and the Humanities, Vol. 36, pp. 223-254, 2002.

[3] Cui, X. and Potok, T.E. “Document clustering analysis based on hybrid PSO + K-means algorithm”, Journal of Computer Sciences (Special Issue), pp. 27-33, 2005.

[4] X. Cui, T. E. Potok, P. Palathingal, “Document clustering using particle swarm optimization” In Proceedings of the IEEE swarm intelligence symposium, SIS 2005 Piscataway: IEEE Press, pp. 185191.

[5] Dash, R., Mishra, D., Rath, A.K. and Acharya, M. “A hybridized K-means clustering approach for high dimensional dataset”, International Journal of Engineering, Science and Technology, Vol. 2, No.2, pp. 59-66, 2010.

[6] Eberhart, R.C. and Kennedy, J. “A new optimizer using particle swarm theory”, Proceeding of Sixth International Symposium on Micromachine and Human Science, pp. 39-43, 1995.

[7] Hotho, A., Maedche, A. and Staab, S. “Ontology-based text document clustering”, Kunstliche Intelligenz, Vol. 16, No. 4, pp. 48-54, 2002.

[8] HuiXiong, Junjie Wu, and Jian Chen. K-means clustering versus validation measures: a data distribution perspective. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 779–784, 2006.

[9] Kennedy, J. and Eberhart, R.C. “Particle swarm optimization”, Proceeding of IEEE International Joint Conference on Neural Networks, Vol. 4, pp. 1942-1948, 1995.

[10] Kennedy, J., Eberhart, R.C. and Shi, Y. “Swarm intelligence”, Morgan Kaufmann Publishers, San Francisco, 2001.

[11] Killani, R., Rao, S. K., Satapathy, S. C., Pradhan, G. and Chandran, K. R. “Effective document clustering with particle swarm optimization”, Proceedings of First International Conference on Swarm, Evolutionary, and Memetic Computing, Lecture Notes in Computer Science, Vol. 6466, pp. 623-62, 2010.

[12] V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques” IJCSIT Vol.4 (1), 2013, 39- 45.

[13] Liliya Demidova, Evgeny Nikulchev, Yulia Sokolova Presented “The SVM Classifier Based on the Modified

Particle Swarm Optimization" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016

[14] Niknam,T; Nayeripour, M; Firouzi, BB(2008). Application of a New Hybrid optimization Algorithm on Cluster Analysis Data clustering. World Academy of Science, Engineering and Technology.

[15] Niknam, T., Amiri, B., Olamaei, J. and Arefi, A. "An efficient hybrid evolutionary optimization algorithm based on PSO and SA for clustering", Journal of Zhejiang University Science, Vol. 10, No. 4, pp. 512-519, 2009.

[16] NadanaRavishankar.T, Shriram.R Ontology based clustering algorithm for information retrieval https://www.researchgate.net/publication/271419783 On 7 March 2015.

[17] Pratiksha Y. Pawar and S. H. Gawande, Member, IACSIT "A Comparative Study on Different Types of Approaches to Text Categorization" International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012

[18] Song, L., Ma, J., Yan, P., Lian, L. and Zhang, D. "Clustering deep web databases semantically". Proceedings of the 4th Asia Information Retrieval Conference on Information Retrieval Technology, Lecture Notes in Computer Science, Springer-Verlag, pp. 365-376, 2008.

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

[20] Zhang, X., Jing, L., Hu, X., Ng, M. and Zhou, X. "A comparative study of ontology-based term similarity measures on PubMed document clustering", Proceedings of the 12th International Conference on Database Systems for Advanced Applications, pp. 115-126, 2007

P.JYOTSNA received Master of Computer Applications degree from Sri Venkateswara University, Tirupati, AP and. Pursuing Ph.D in the department of Computer Science, Sri Venkateswara University, Tirupati. Her research area are Databases and Data Mining, Her research focus is on Text Mining Techniques for Detection of Tumors Using Ontology Based Particle Swarm Optimization with Clustering Approaches.

P. GOVINDARAJULU, Professor, Department of Computer Science, Sri Venkateswara University, Tirupathi, AP, India. He received his M. Tech., from IIT Madras (Chennai), Ph. D from IIT Bombay (Mumbai), His area of research are Databases, Data Mining, Image processing, Intelligent Systems and Software Engineering.