DOI 10.29042/2017-2007-2012

Helix Vol. 7(5): 2007-2012

Semantic Similarity based Web Document Clustering Using Hybrid Swarm Intelligence and FuzzyC-Means

[*1]J.Avanija, [2]G. Sunitha, [3]K. Reddy Madhavi
Sree Vidyanikethan Engineering College, Tirupati, Andhra Pradesh, India
Email: [1]avans75@yahoo.co.in, gurramsunitha@gmail.com, kreddymadhavi@gmail.com

Received: 2nd June 2017, Accepted: 15th June 2017, Published: 1st September 2017

Abstract
Information retrieval technology has been central to the success of the web. The volume of information stored and accessed on web is increasing continuously. This enlargement leads to the difficulties such as seeking and managing the existing information. The use of keyword based method in information retrieval processing is the reason behind this limitation and can be overcome by performing semantic search instead of using keywords. Optimization problem occurs even through the use of semantic technologies. Hybrid clustering algorithm combining FuzzyC-Means and PSO has been proposed to generate optimum number of clusters and to get better accuracy in the recovery of documents. Experimental results reveal that the proposed method performs better than hybrid approaches combining PSO and KMeans.

Keywords: Fuzzy C-Means, Clustering, Semantic Similarity, Ontology, K-Means,Particle Swarm Optimization

1 Introduction
Clustering algorithms are applied on a variety of fields. Appropriate document clustering can be provided through clustering methods **[1].** The use of Ontologies improves the keyword search mechanism. Grouping of documents performed based on similarity score using PSO based clustering to improve the document relevancy [2].Documents are categorized using ontology concept weights which improves the accuracy of documents. Similarity measure based clustering of documents along with ontology is proposed [3]. Fuzzy clustering based on semantic analysis using ontology is performed **[4].**FuzzyC-Means algorithm is the popular clustering method since it is efficient and the implementation is also simple. Even then FCM has drawbacks such as getting trapped to local optima and sensitive to initialization.The K-Means is a hard clustering process in which data is divided into distinct clusters, where each data element belongs to exactly one cluster. FuzzyC-Means also known as soft clustering is well suited for real data set clustering since there may be sharp boundaries between clusters. The membership grades of every document represents the degree by which a document belongs to a particular cluster [5]. To overcome the limitation faced by K-Means Algorithm, PSO is combined with FCM and to improve the speed, fuzziness parameter is included in PSO and combined with FCM. Performance analysis shows that FPSOFCM is better than PSOK-Means (PSOK).

Ontology based fuzzy clustering scheme along with semantic analysis is used in document recovery process [6]. Hierarchical clustering based on ontology and Fuzzy mechanism is used in document clustering**[1]**.PSO based clustering mechanism based on similarity score improves document relevancy[3][7] .According to "Potok et al" [8] clustering based on hybrid approaches like K-Means and PSO improves the relevancy of documents. Fuzzy ontology-based clustering of documents (the "FODC" methodology) is presented and compared with K-Means algorithm [9] [10]. The results shows that "FODC" based method performs better than "K-Means" clustering approach **[6**]. Analysis reveals that PSO performs better than ACO and for problems that needs crisp results and with predefined source and destination ACO is more applicable **[11].**

The remaining part of the paper is specified as follows: Division 2 specifies the similarity calculation methodology. Division 3 briefs about the clustering approach. Division 4 shows the results and discussion. Division 5 offers concluding remarks and future scope of research..

2 Methodology
2.1 System Architecture
Figure 2.1 specifies the system architecture of information retrieval using ontology based using FPSOFCM document clustering method. The web pages collected from internet are annotated to generate the semantic meaning for the given content. The annotated documents are then sent to the FPSOFCM based clustering method for optimal cluster generation. KIM plugin was used to perform semantic annotation of web pages. Clustering of annotated web pages is performed through FPSOFCM based document-clustering. A semantic-meaning disclosure file created through semantic annotation generation process for each annotated document. The annotated documents are given as input to FPSOFCM based clustering logic where

semantic similarity based clustering was done. Relevant document clusters are generated using PSO. The dataset can be selected through the user interface and passed to the semantic similarity based clustering logic which returns the relevant result to the user.

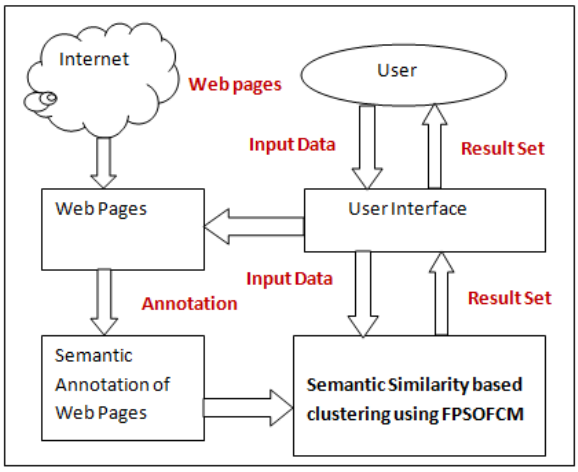

Figure 2.1 System Architecture

2.2 Document Representation

The documents to be clustered is specified as vector set $X=\{x_1, x_2, \ldots, x_n\}$, vector x_i represents single object termed as feature vector. The content of text document is represented through vector space model m represented by $d_j=(w_{1j}, w_{2j},\ldots,w_{nj})$ where w_k indicates weight of "k^{th}" term in document j. In "term-weight" calculation the frequency of occurence of the term within a document and in entire set is treated as part of key vocabulary extraction process of documents. TFxIDF occurs and is calculated as mentioned in Equations (2.1) and (2.2).

$$TF = \frac{freq_{c,d}}{\max_{l,d} freq_{l,d}} \quad (2.1)$$

$$IDF = \frac{\log |D|}{n_c} \quad (2.2)$$

$$w_c = TF * IDF \quad (2.3)$$

Concept weight is specified by the parameter w_c, $freq_{c,d}$ indicates the term frequency in document based on concept c, $\max_{l,d} freq_{l,d}$ specifies the maximum frequency of most frequently repeated concept in d. D represents total number of documents, and number of documents annotated with concept c is represented by n_c.

2.3 Semantic Similarity Measurement

"Wu and Palmer" similarity measure is used to find semantic similarity matrix**[12][13].**The depth of two concepts in wordnet and least common superconcept (LCS) is measured by similarity metric, and these figures are combined into a similarity score as specified in Equation (2.4).Based on the concept importance the weight assignment to concepts and relations was carried out. The depth from root to term is the depth of w_c , and LCS is the least common super concept of w_c and w_s. Term-reweighting is performed with the computed similarity score as given in Equation (2.5).

$$sim(w_c, w_s) = \frac{2*depth(LCS)}{depth(w_c)+depth(w_s)} \quad (2.4)$$

$$w_c' = w_c + \sum_{sim(w_c,w_s)\geq t}^{p} \frac{1}{n} sim(w_c, w_s) * w_s \quad (2.5)$$

"Term-reweight" is specified by w_c', t specifies user defined minimum threshold value,the number of terms is specified by p, the weight for concept term c is represented as w_c , and n specifies the number of hyponyms of each term.

3 Clustering Methodology

Before performing clustering the documents are annotated by "KIM" plugin. Then the documents are clustered using FPSOFCM and concept weight is calculated . Term weight is recalculated through the steps specified in SEMHYBODC algorithm as specified in Figure 3.1.Accuracy of clustering is computed using measures such as cluster purity and compared with various other hybrid approaches..

Main Procedure:

Algorithm: SEMHYBODC()

Input: "Web Documents"

Output: "Retrieval of Relevant documents"

1: Annotate web documents using "KIM" plugin.

2: Do Concept Extraction over annotated documents.

3: Find "term-weight" through dot product Of "Term-frequency" (TF) and "Inverse Document Frequency" (IDF) .

4: Calculate Semantic Similarity as in Equation (2.4).

5: Recalculate Concept Weight as in Equation (2.5).

6: Call "SEMFPSOFCMCLUS()" to cluster the documents

7: Calculate document relevancy by measuring "precision", "recall" and " F-Measure".

Figure 3.1 Steps for Ontology based Semantic Document Clustering using FPSOFCM

3.1 Hybrid FPSOFCM based Semantic document-clustering

FuzzyC-Means algorithm is the popular fuzzy clustering techniques because it is efficient as well as easy to implement[14]. However, the drawback of FuzzyC-Means is that it is sensitive to initialization and gets trapped in local optima easily**[8].** "Particle swarm optimization" (PSO) is an optimization technique which could be used to solve various optimization problems [9**].**Hybrid fuzzy clustering approach based on "Fuzzy C-Means" and "FuzzyPSO"(FPSO is proposed which uses the advantages of both algorithms. The proposed "SEMFPSOFCM" algorithm is shown in Figure 4.2.Experimental results over real time datasets like Reuters specifies that the proposed approach is

better and gives better results than FCM.In "PSO" the flow of algorithm begins with population of particles and their positions represent the solutions for the identified problem.. The velocity is initialized randomly and updated along with the position in each iteration to get optimal position. Fitness function in Equation (3.5) is used in determining the fitness value for each particle position. The velocity updation is performed using 2 best positions, "personal best position" (Pbest) and "global best position" (Pgbest). The velocity and position of particle is updated as mentioned in Equations (3.3),(3.4).

$$Pbest_i(t+1) = \begin{cases} Pbest_i(t), & f(X_i(t+1) \le f(X_i(t)) \\ X_i(t+1), & f(X_i(t+1) > f(X_i(t)) \end{cases} \quad (3.1)$$

where $X_i(t+1)$ represents particles current position , "P$best_i$" specifies the best personal position of particle and " P$best_i$(t +1)"specifies the new best position. After performing the calculation of particles personal best position the global best position of the particle is calculated using Equation (3.2).

$$Pgbest(t) = argmin_{i=0}^{n}\{f(Pbest_i(t)\} \quad (3.2)$$

where i represents the particle index from 0 till n where n represents the total number of particles . The "velocity" of particle is computed as specified in Equation (3.3).

$$V_i(t+1) = w * V_i(t) + c_1 * rand_1 * Pbest - X_i(t) + c_2 * rand_2 * Pgbest - X_i(t) \quad (3.3)$$

Where the new velocity of particle is represented by "$V_i(t+1)$" , current velocity specified by "$V_i(t)$" inertia weight represented by w, $c_1$and $c_{2\,t}$ termed as acceleration coefficients; the problem-space dimension denoted by d; the range for random values $rand_1$, $rand_2$ (0, 1). To avoid particle convergence at local optima inertia factor w is used .The position of the particle is updated using the position update Equation (3.4).

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (3.4)$$

$$f = \sum_{\substack{i=1 \\ No.of\ documents}}^{n} \sum_{\substack{j=1 \\ Features}}^{p} d(c_{li}, g_{ji}) \quad (3.5)$$

Particle swarm optimization (PSO) along with fuzzy set theory is termed as fuzzy particle swarm optimization (FPSO), proposed by [10].FPSO redefines particles velocity and position through the fuzzy relations and then applied for clustering problem**.** In this method, X represents the position of particle, fuzzy relations of data objects for cluster centers "z={z1,z2,…,zn}" can be calculated as specified in Equation (3.6).

$$X = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1c} \\ \vdots & \ddots & \vdots \\ \mu_{n1} & \cdots & \mu_{nc} \end{bmatrix} \quad (3.6)$$

in which μ_{ij} represents the membership function of "jth" object with "jth" cluster. Updation of particle velocity is performed using Equations (3.3),(3.4). In FPSO algorithm similar to other evolutionary algorithms, fitness function is computed to evaluate generalized solution as expressed in Equation (3.7).

$$f(x) = \frac{h}{k_m} \quad (3.7)$$

where h is referred as a constant and objective function of "FCM" is represented by k_m. For better clustering the value of k_m should be smaller the individual fitness function represented as f(x) should be higher. The condition for termination is reaching maximum number of "iterations" or there is no improvement in "Pgbest" in number of "iterations".

Steps for SEMFPSOFCM based Clustering

Input: "Annotated documents"

Output:"Clustered document set"

1:Initialize parameters for FPSO and FCM including population size P,w,m,c1 and c2.

2: Create swarm with "P" particles where X,pbest,gbest,V are nxc matrices.

3: Initialize "X,,pbest,V" for individual particle and "gbest" for swarm.

4: FPSO algorithm

- 4.1 Inherit the cluster center from Equation (3.6)
- 4.2 Calculate fitness function through Equation(3.7)
- 4.3 Calculate Pbest of particle through Equation(3.1)
- 4.4 Calculate Pgbest of swarm through Equation (3.2)
- 4.5 Update velocity matrix V_i(t+1) through Equation(3.3)
- 4.6 Update position matrix X_i(t+1) through Equation(3.4)
- 4.7 Go to "4" if termination condition not met.

5: FPSOFCM algorithm

- 5.1 Inherit the cluster center from Equation (3.6)
- 5.2 Compute Euclidean distance using Equation (3.5)
- 5.3 Update the membership Function using Equation (3.7)
- 5.4 Calculate Pbest of each particle using Equation (3.1)
- 5.5 Calculate Pgbest of each particle using Equation (3.2).

6: If terminating condition not met go to "5".

7: If FPSOFCM terminating condition not met go to "4".

Figure 3.2 Steps for SEMFPSOFCM based Clustering

FCM algorithm is faster than FPSO since it needs only minimal function evaluation but gets trapped to "local optima". To overcome this drawback "FCM" is combined with "FPSO" to form the hybrid clustering approach "FPSOFCM".

4 Results and Discussion

Efficient document-clustering is needed to help users to find relevant information within larger dataset. Partitional clustering algorithms like

FuzzyC-Means and K-Means are well-suited for large dataset. PSO is better in finding quality solutions but the convergence is slow for complex solutions. The drawback of FuzzyC-Means is that it may converge to local optima. To overcome this limitation, fuzziness parameter is included in PSO and then combined with FuzzyC-Means (FPSOFCM) to generate clustering solution. Experimental results reveal that the proposed semantic clustering process based on FPSOFCM performs better than the approaches like K-Means, FuzzyC-Means and PSOK. For better optimizing the performance of "FPSOFCM" fine tuning of parameters with best values is carried out. 20 random particles are generated, and the fitness value is calculated based on cluster centroids. The inertia weight w is selected as 0.9, and the value of c_1 and c_2 are selected as 0.2. In FuzzyC-Means, the terminating condition is met when the algorithm cannot improve the generated solution. The terminating condition for FPSO is that there is no improvement in "gbest" in further iterations or maximum iterations are reached. For FPSOFCM algorithm, the terminating condition is that the value of "Pgbest" cannot be improved in consecutive iterations. The initial centroid vector is selected randomly for every simulation. FPSOFCM based semantic document-clustering generates the optimal number of clusters when compared to K-Means, FuzzyC-Means, KPSO and PSOK algorithms.

The F-Measure values are the average of 100 runs and after 100 iterations the same cluster is obtained. The performance metrics such as F-Measure, cluster purity and CPU execution time are considered for performance evaluation of the system. Experimental results show that FPSOFCM performs better than other algorithms such as PSOK and KPSO. The average value of F-Measure mentioned in Table 4.1 shows that there is an improvement in the recovery of documents by SEMFPSOFCM method when compared to the other methods. Table 4.3 shows the comparison of clustering methods like K-Means, FCM, SEMPSO, SEMPSOK and SEMFPSOFCM based on computational time. Table 4.2 shows that the
cluster purity of SEMFPSOFCM based method performs well when compared to K-Means, FCM, SEMPSO and SEMPSOK.

Table 4.1 Performance Comparison based on F-Measure

No. of Documents	K-Means		FCM		SEMPSO		SEMPSOK		SEMFPSOFCM	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
100	0.61	0.63	0.63	0.65	0.65	0.61	0.64	0.68	0.66	0.69
200	0.62	0.64	0.65	0.64	0.67	0.65	0.68	0.72	0.7	0.73
300	0.59	0.62	0.62	0.66	0.7	0.68	0.71	0.74	0.74	0.78
400	0.7	0.72	0.73	0.75	0.72	0.7	0.74	0.75	0.75	0.77
500	0.65	0.67	0.7	0.72	0.74	0.72	0.74	0.77	0.76	0.79
600	0.72	0.75	0.74	0.75	0.76	0.74	0.77	0.8	0.8	0.83
700	0.74	0.77	0.74	0.76	0.78	0.76	0.8	0.83	0.84	0.88
800	0.69	0.71	0.67	0.69	0.8	0.78	0.82	0.85	0.85	0.89
900	0.66	0.7	0.6	0.64	0.82	0.8	0.84	0.88	0.88	0.92
1000	0.73	0.76	0.62	0.66	0.84	0.82	0.86	0.9	0.89	0.93

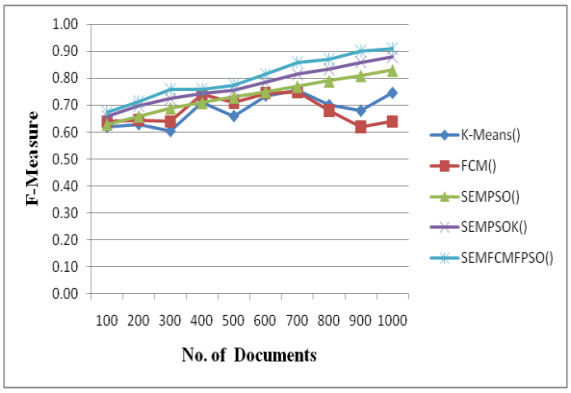

Figure. 4.1 Performance Analysis based on F-Measure

Table 4.2 Comparison based on Cluster Purity

No. of Clusters	K-Means()	FCM()	SEMPSO()	SEMPSOK()	SEMFPSOFCM()
3	0.65	0.73	0.76	0.78	0.82
5	0.67	0.78	0.81	0.83	0.86
15	0.69	0.83	0.85	0.88	0.9
20	0.7	0.85	0.88	0.91	0.94
Average	0.678	0.797	0.825	0.85	0.88

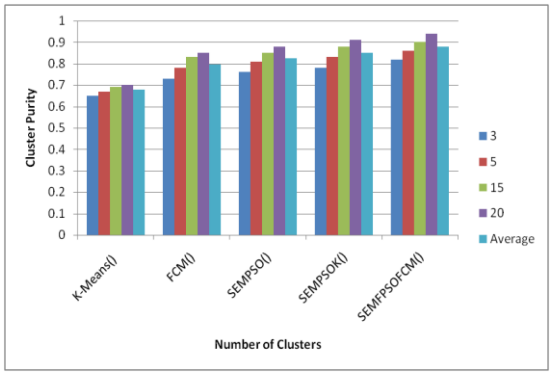

Figure.4.2 Performance Comparison based on Cluster Purity

Table 4.3 Comparison based on Computational Time in Milliseconds

No.of Documents	PSOK	KPSO	K-Means	FCM	FPSOFCM
100	1576	890	162	97	692
200	1824	953	196	121	724
300	2122	1005	240	147	835
400	2403	1102	322	172	893
500	2720	1165	450	223	931
600	3012	1196	510	261	972
700	3421	1210	552	311	1004
800	3862	1243	603	369	1045
900	4061	1276	645	409	1072
1000	4412	1302	672	443	15096

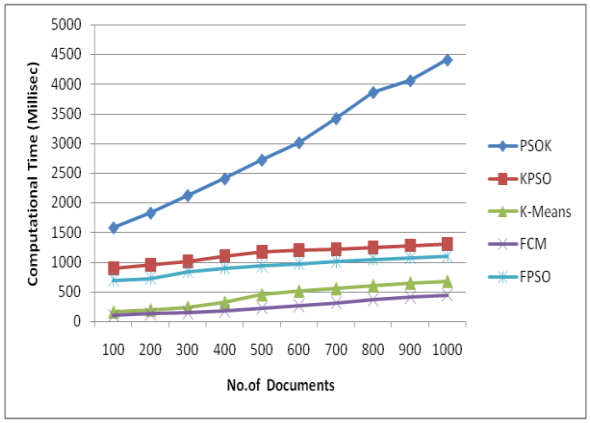

Figure. 4.3 Comparison based on Computational Time

The cluster purity values of SEMPSO, SEMPSOK and SEMFPSOFCM are very closer, but there is vast difference in computational time as in Figure 4.3 K-Means and FCM based clustering processes are faster, but the clustering results are less accurate when compared to SEMPSO, SEMPSOK and SEMFPSOFCM based clustering methods as shown in Table 4.3. Performance analysis based on cluster purity specified in Figure. 4.2 shows that there is better accuracy in SEMFPSOFCM based clustering when compared to methods such as K-Means and FCM. Experimental analysis shows that the proposed semantic document-clustering using FPSOFCM method is better when compared to state-of-the-art approaches.

5 Conclusion

Information retrieval relies on keywords for indexing and retrieving documents. Mostly keyword based retrieval returns inappropriate results since same concept can be described using different keywords in documents as well as queries. The background work related to semantic information retrieval indicates that semantic clustering has drawn attention in recent years since it provides accuracy, but efficient clustering techniques are still needed to improve accuracy of document recovery. Further, optimization problem also occurs during clustering process. To overcome this limitations evolutionary approaches like PSO and fuzzy are used along with semantic based clustering process to cluster documents.

Semantic document-clustering using swarm intelligence gives better accuracy, but the computational time is more and is not suitable for large dataset. To handle large dataset and to reduce the computational time, a hybrid approach combining clustering algorithms with evolutionary approaches such as PSO and fuzzy is proposed for clustering web documents. The hybrid approach of clustering combining PSO and K-Means provided optimal solution, but computational time is still higher. Considering this limitation FuzzyC-Means is combined along with PSO to cluster the documents. Document-clustering based on "FPSO+FCM" method shows improvement over algorithms like FuzzyC-Means K-Means and hybrid approaches like KPSO and PSOK. In the "FPSO+FCM" the ability of globalized searching of the PSO algorithm and quick convergence of FCM algorithm are combined. The result from FPSO is used as the initial seed for FCM algorithm, which is applied for refining and generating the final result. Future scope will be to successfully approach real world problems in diverse domains.

References

[1] ShengliSong,ZengxinGuo,Ping Chen. Fuzzy Document Clustering using Weighted Conceptual Model, Information Technology Journal Vol:10, 2011.

[2] Siddhartha Panda , Narayana Prasad Padhy. Comparison of Particle Swarm Optimization and Genetic Algorithm for TCSC based Controller Design, International Journal of Computer Science and Engineering, Vol.:1, 2008.

[3] Thangamani M,.Thangaraj P. Ontology Based Fuzzy Document Clustering Scheme, International journal of Modern Applied Science, Vol: 7, 2010.

[4]Takazumi Matsumoto , Edward Hung,"Fuzzy Clustering and Relevance Ranking of Web Search Results with Differentiating Cluster Label Generation" FUZZ-IEEE 2010, IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 2010.

[5] Runkler ,Katz , "Fuzzy Clustering by PSO", Proceedings of IEEE International Conference on Fuzzy Systems, 2006.

[6]Amy Trappey, Charles , Trappey,Fu-Chiang Hsu, and David W. Hsiao. A Fuzzy Ontological Knowledge Document Clustering Methodology. IEEE Transactions on Systems,Man and Cybernetics, Vol:39, 2009.

[7] Sandeep U. Mane, Pankaj G. Gaikwad. Hybrid Particle Swarm ptimization (HPSO) for Data Clustering, International Journal of Computer Applications),Vol: 97 ,2014.
[8] Hesam Izakian , Ajith Abraham, Fuzzy C-means and fuzzy swarm for fuzzy clustering problem , Expert Systems with Applications, Vol: 38, 2009.
[9] Kennedy J and R.C. Eberhart, "Particle swarm optimization", Proceedings of the IEEE international conference on neural networks , 2005.
[10] Mahamed Omran GH, Andries Engelbrecht, Ayed Salman. Dynamic Clustering using Particle Swarm Optimization with Application in Unsupervised Image Classification, Transactions on Engineering ,Computing and Technology, Vol:.9,2005
[11] Selvi, V , Umarani, R . Comparative Analysis of Ant Colony Optimization Techniques, International Journal of Computer Applications, Vol.5, 2010.
[12] Gang Liu, "A wordnet based semantic similarity measure enhanced by internet based knowledge", International Conference on Software Engineering and Knowledge Engineering, 2010.
[13] Punitha, SC, Mugunthadevi, K, Punithavalli. Impact of Ontology based Approach on Document Clustering. International Journal of Computer Applications, Vol: 22, 2011.
[14] Mehdizadeh, S, Sadi-Nezhad , Tavakkoli-Moghaddam. Optimization of Fuzzy Clustering Criteria by a HybridPSO and Fuzzy C-Means Clustering Algorithm, Iranian Journal of FuzzySystems, Vol: 5, 2008.
[15] Xiaohui Cui, Thomas E. Potok,Cui.X and Potok. Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm, Journal of Computer Sciences, Vol:4 , 2005.
[16] Yang Cheng, "Ontology-Based Fuzzy Semantic Clustering", Third International Conference on Convergence and Hybrid Information Technology, 2008.