Journal of Web Engineering, Vol.12, No.3&4 (2013) 249-264

A HYBRID APPROACH USING PSO AND K-MEANS FOR SEMANTIC CLUSTERING OF WEB DOCUMENTS

J. AVANIJA

Velammal College of Engineering &Technology,Tamilnadu,India.

avans75@yahoo.co.in

Dr.K. RAMAR

Einstein College of Engineering,Tamilnadu,India.

kramar.einstein@gmail.com

Received May 16, 2012

Revised January 27, 2013

With the massive growth and large volume of the web it is very difficult to recover results based on the user preferences. The next generation web architecture, semantic web reduces the burden of the user by performing search based on semantics instead of keywords. Even in the context of semantic technologies optimization problem occurs but rarely considered. In this paper Document clustering is applied to recover relevant documents. We propose a ontology based clustering algorithm using semantic similarity measure and Particle Swarm Optimization(PSO), which is applied to the annotated documents for optimizing the result. The proposed method uses Jena API and GATE tool API and the documents can be recovered based on their annotation features and relations. A preliminary experiment comparing the proposed method with K-Means shows that the proposed method is feasible and performs better than K-Means.

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

Communicated by: D. Schwabe & S. Murugesan

1 Introduction

With the massive growth of web content retrieving relevant information becomes a difficult task. Efficient clustering algorithms are needed to improve the recovery of documents. Document clustering is the process of identifying similarity or dissimilarity between the objects and form groups based on the common characteristics shared between objects. The main objective of document clustering is to avoid the recovery of non relevant documents. David. A. Grossman and Ophir Frieder (2004) discussed the importance of document clustering to group the documents based on the contents reducing the search space for the given query. The keyword based on methodologies to cluster the documents is not convenient since it does not capture the semantic structure of documents. Moreover the keyword based methodologies for document clustering is not effective. To overcome the problems faced by the keyword based methodologies document clustering is performed by combining ontology with optimization technique like PSO and KMeans clustering process.

Salton et.al(1989) shows that most of the document clustering approaches use Vector Space Model(VSM) for document representation. But using VSM ignores the semantic relatedness among documents. For example having “Fruits” in one document and “Apple” in another document does not contribute to similarity measurement unless semantic relatedness is considered. Semantic relationship is not included in most of the clustering approaches. According to Hotho, Maedche, and Staab (2002) use of ontology provides a good background knowledge and improves document clustering. Recent works has shown that ontology is useful to improve the performance of text clustering in these situations.

Currently a challenge when querying information using semantics offered by ontology is how to extract information from ontology more efficiently [2]. Semantic annotation is about assigning to the entities in the text links to their semantic description [15]. Annotation provides additional information about web contents so that better decision on content can be made. Annotation ontology tells us what kind of property and value types should be used in describing a resource. The usage of domain ontologies are used for annotation. The manual annotation of document is of high cost and error prone task. However there is still some work to do achieve a complete automation of annotation. The classical model is incapable of supporting logical inference.

In this paper we propose an ontology based information retrieval model which uses hybrid approach combining PSO with K-Means to improve clustering of web documents. Ontology similarity is used to identify the importance of concepts in the document. Particle Swarm Optimisation is used to cluster the documents since it is effective for global search in finding solutions to nondeterministic problems. Moreover Particle Swarm Optimisation method enhances adaptability of meta searching. Performance of PSO and K-Means based clustering is evaluated using K-Means algorithm. The

proposed model uses semantics and relationship available in the knowledge base to improve the relevancy of documents.

The rest of the paper is organized as follows: Section 2 highlights the previous research in the related area. Section 3 describes the document representation and the methodology used for similarity calculation. Section 4 describes the clustering approach. Section 5 describes experimental results and discussion. Finally, we offer concluding remarks and describe future directions of our research work.

2 Related Work

Ranking of documents is combined along with clustering by ordering the web pages in the form of clusters based on the query given by the user[19]. The performance of the ordered result is measured based on relevancy. Since the classical clustering methods are not dealing with the semantics of the objects a new methodology was derived to incorporate knowledge in to clustering process[18]. Fuzzy clustering scheme is combined with semantic analysis mechanism along with relevant attributes in the ontology[23]. Hierarchical clustering along with fuzzy logic approach is used to cluster knowledge documents along using ontology[3].Use of clustering methods provide appropriate document retrieval[21].

Keyword based search mechanism is improved by the use of Ontologies. PSO based clustering mechanism is used to group documents based on their similarity score which improves the relevancy of documents [22] .According to Potok et al Hybrid PSO and K-Means based clustering improves document relevancy. Documents are categorized by measuring the ontology concept weights which can improve the accuracy and performance of text documents.Clustering of documents based on similarity measure combined with ontology to improve relevancy of documents[25].

Wu and Palmer similarity metric measures depth of two concepts in the WorldNet taxonomy. Accurate measurement of semantic similarity is still a challenging issue [5]. Semantic similarity between words is measured using page count and snippets retrieved from web search engine[5].Clustering data vectors using hybrid approach combining PSO and K-Means gives better convergence and reduces quantization errors[7].Ontology concepts can be used as a relevancy measure to re-rank the recovered web documents to reduce the ranking errors[1]. According to S.Kalyani et.al K-Means algorithm based on PSO improves accuracy in power system assessment[12].

Various information retrieval models are Boolean model, vector space model, probabilistic model and hyper link model . There have been works which employ semantic web technology for information and retrieval such as KIM [4]. Currently a challenge when querying information using semantics offered by ontology is how to extract information from ontology more efficiently [2]. Semantic annotation is about assigning to the entities in the text links to their semantic description [4]. Annotation provides additional information about web contents so that better decision on content can be made. Annotation ontology tells us what kind of property and value types should be

used in describing a resource. The usage of domain ontologies are used for annotation. The manual annotation of document is of high cost and error prone task. However there is still some work to do achieve a complete automation of annotation. The classical model is incapable of supporting logical inference.

3 **Overview of System Methodology**

3.1 System Architecture

Figure 1 shows the system architecture for ontology based information retrieval through PSO based clustering methodology. The crawler program collects the web pages on the internet with its semantic markup and corresponding ontology, described in an OWL document. The collected web pages are transported to web page database for future use.

Semantic annotation of web pages are performed using KIM plugin. Hybrid PSO and K-Means based document clustering is then applied for the annotated web pages. Semantic annotation generation process creates a semantic meaning disclosure file for each annotated document. Through the semantic meaning disclosure file, any ontology-aware machine agent can understand the target document. Moreover the KIM platform consists of a formal KIM ontology and a KIM knowledge base, a KIM Server (with an API for remote access or embedding), and front-ends that provide full access to the functionality of the KIM Server. The KIM ontology is a light-weight upper level ontology that defines the entity classes and relations of interest. The annotated documents are passed to hybrid PSO and K-Means based clustering logic where clustering was done based on semantic similarity score and PSO algorithm which in turn generates relevant document clusters.PSO is applied to get optimal clusters where the swarm represents number of candidate clustering solutions for the document collection. The user interface allows for the definition of query by the user which is passed to the clustering logic based on PSO and K-Means. The ordered relevant result set in the form of optimal clusters generated by this module and returned to the user.

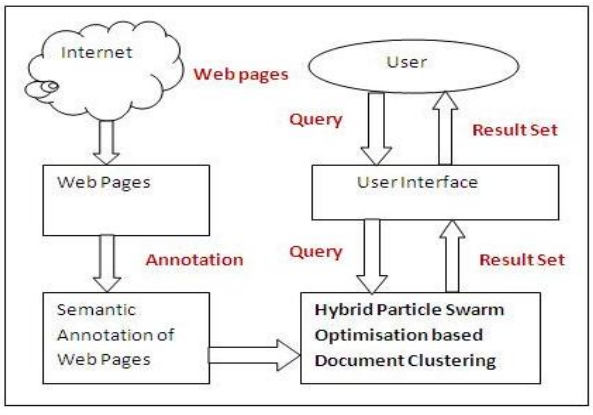

Figure 1 System Architecture

3.2 Document Representation

The most widely used document representation model in information retrieval is vector space model. Vector space model is used for effective representation of documents. Each document is identified as a n-dimensional feature vector and each term is associated with a weight. The dataset to be clustered is represented as vectors $X=\{x_1, x_2, ..., x_n\}$, *where the vector* $x_i$represents single object called as feature vector. Feature vector includes the feature of the object which can be represented using Vector Space Model(VSM). For weight calculation document j is represented as $d_j=(w_{1j}, w_{2j},...,w_{nj})$ where w_{kj} is the weight of k^{th} term in the document j. The term weight value represents the significance of a term in a document. To calculate the term weight, the occurrence frequency of the term within a document and in the entire set of documents must be considered.As part of key vocabulary extraction process of form documents.TFxIDF takes place. The terms t_k in the document is represented as document-term frequency matrix as shown in Table(1).

Table 1 – Document-Term Frequency Matrix

D_j/T_k	T_1	T_2	…	T_k
D1	tf_{11}	tf_{12}	…	tf_{1k}
D2	tf_{21}	tf_{22}	…	tf_{2k}
D3	tf_{31}	tf_{32}	…	tf_{3k}

3.3 Semantic Similarity Measurement

Semantic Similarity measurement is used to compute the similarity between the concepts but not the lexically similar terms. Semantic similarity is computed by mapping terms to ontology and examining their relationships in that ontology. The document collection on the semantic web is referred as D={d1,d2…dn}.Annotations of the documents represents the weights of the concept. The weights of the document is measured based on the importance of the concepts. Ontology based annotation is to improve the relevance of the documents.

The senses of document is represented using Wordnet (http://www.princeton.edu) .Wordnet defined relations between synsets and relation between word senses.The similarity between two documents can be measured based on the occurrences of instance in a document. In ontology indexing process the terms are mapped with the concepts based on the ontology similarity score. The importance of the terms are measured using Salton measure considering Term Frequency(TF) and Inverse Document Frequency(IDF) as mentioned in equation(3).TF is a measure of how often a term is found in a collection of documents. TF is combined with inverse document frequency (IDF) as a means of determining which documents are most relevant to a query. TF is also used to measure how often a word appears in a specific document. The TFIDF weight (term frequency–inverse document

frequency) is a numerical statistic which reflects how important a word is to a document in a collection or corpus.

$$TF = \frac{freq_{c,d}}{max_{l,d}\ freq_{l,d}} \quad (1)$$

$$IDF = \frac{\log |D|}{n_c} \quad (2)$$

$$w_c = TF * IDF \quad (3)$$

Concept weight is represented as w_c. $freq_{c,d}$ represents the frequency of terms in document based on concept c, $\max_{l,d} freq_{l,d}$ represents the maximum frequency of most repeated concept in d. D is the total number of documents and n_c represents the number of documents annotated with concept c. The semantic similarity matrix is computed using Wu and Palmer similarity measure (1994). The similarity metric measures the depth of the two concepts in the WordNet taxonomy, and the depth of the least common superconcept (LCS), and combines these figures into a similarity score as mentioned in equation(4).The weights are assigned to concepts and relations based on the importance of concept. The depth of $w_c$is the depth from the root to the term and LCS is the least common superconcept ofw$_c$and w_s. With the computed similarity score term reweighting is performed as given in equation(5).

$$sim(w_c, w_s) = \frac{2*depth(LCS)}{depth(w_c)+depth(w_s)} \quad (4)$$

$$w_c' = w_c + \sum_{sim(w_c,w_s)\geq t}^{p} \frac{1}{n} sim(w_c, w_s) * w_s \quad (5)$$

Term reweight is represented as w_c', User defined minimum threshold value is represented by t, p represents the number of terms and w_c represents the weight for concept term c and n represents the number of hyponyms for each term. The algorithm to calculate term reweight is represented in Figure 2.

4 Clustering Methodology

Based on the semantic similarity score obtained clustering was done on the annotated documents using hybrid approach based on PSO and K-Means. Initially the documents were annotated using KIM plugin. Based on the instance of the annotated document concept weight is calculated (w_c). Semantic similarity measure is used to recalculate the term weight as represented in SEMPSO algorithm shown in Figure 4. Then clustering of documents is done using 2 algorithms like normal PSO based clustering

and hybrid approach with PSO and K-Means as mentioned in Figure 2 and Figure 4 which returns the relevant documents. Experiments were conducted and the results obtained were evaluated considering K-Means and normal PSO.

4.1. PSO Based Clustering Algorithm

Particle swarm optimization first introduced by Kennedy and Eberhart [13,6], as an optimization technique based on the movement and intelligence of the swarm. It inspired by social behaviour and dynamics of movement of birds and fish. PSO uses a number of particles that constitute a swarm moving around in the search space to find the best solution. Each particle is treated as a point in the search space which adjusts its flying according to its own flying experience and other particles flying experience. A particle's location in the multi-dimensional problem space represents one solution for the problem. When a particle moves to a new location, a different problem solution is generated. This solution is evaluated by a fitness function that provides a quantitative value of the solution's utility. In PSO, the swarm is initialized to a random solution set. The particles then start moving through the solution space by maintaining a velocity value V while keeping track of its best previous position achieved so far. This value is known as its personal best position (Pbest). Global best (Pgbest) is another best value which is the best fitness achieved by any of the particles. The fitness of each particle or the whole swarm is evaluated by a fitness function. Sometimes the particle maintains another value called the local best, which is related to the best neighborhood fitness. The update equation of minimization for its best solution with dependence upon time step *t* is represented by equation(6).

$$Pbest_i\,(t+1) = \begin{cases} Pbest_i\,(t), & f(X_i(t+1) \le f(X_i(t)) \\ X_i\,(t+1), & f(X_i(t+1) > f(X_i(t)) \end{cases} \quad (6)$$

where$X_i(t+1)$ is the current position of the particle, $pbest_i$is the personal best position of the particle achieved so far and $pbest_i(t+1)$ is the new best position. After calculating the personal best position of the particle the next step is to calculate the global best position using equation(7).

$$Pgbest(t) = argmin_{i=0}^{n}\{f\,(Pbest_i\,(t)\} \quad (7)$$

Where *i* is the index of each particle ranging from 0 to the total number of particles *n* . The velocity of a particle influenced by the social component and the cognitive component is calculated using the equation(8).

$$V_i\,(t) = w * V_i\,(t) + c_1 * rand_1 * Pbest - X_i(t) + c_2 * rand_2 * Pgbest - X_i(t) \quad (8)$$

where $V_i(t)$ represents the current velocity, $V_i(t+1)$ is the new velocity of the particle, w is the inertia weight, $c_1$and $c_2$are constants and are known as acceleration coefficients; *d* denotes the dimension of the problem space; $rand_1$, $rand_2$are random values in the range of (0, 1). The inertia weight factor *w* provides the necessary diversity to the swarm by changing the momentum of particles to avoid the

stagnation of particles at the local optima. The position of the particle is updated using the position update equation(9).

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (9)$$

It is possible to view the clustering problem as an optimization problem that locates the optimal centroids of the clusters rather than finding an optimal partition. This view offers us a chance to apply PSO optimal algorithm on the clustering solution. A single particle 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. Each particle maintains a matrix $X_i = (C_1, C_2, \ldots, C_i, \ldots, C_k)$, where $C_i$represents the ith *cluster* centroid vector and *k* is the cluster number. At each iteration, the particle adjusts the centroid vector' position in the vector space according to its own experience and those of its neighbors. The average distance between a cluster centroid and a document is used as the fitness value to evaluate the solution represented by each particle. The fitness value is measured by the equation (10).

$$f = \sum_{\substack{i=1 \\ No.of\ Clusters}}^{k} \sum_{\substack{i=1 \\ No.of\ documents}}^{n} \sum_{\substack{j=1 \\ Features}}^{p} d(c_{l,1}, g_{j,1}) \quad (10)$$

where k is the cluster number, n represents the number of documents and m indicates the features in each cluster. d($c_{l,1}$,$g_{j,1}$) represents the distance between jth object and cluster centroid of lth cluster on ithfeature.PSO based clustering algorithm is represented in Figure 2.

Steps for PSO based Clustering

Algorithm:PSOCLUS()

Input: Annotated documents

Output:Clustered document set

Step 1: At the initial stage, each particle randomly chooses k numbers of document vectors from the document collection as the cluster centroid vectors.

Step 2:For each particle assign each document vector in the document set to the closest centroid vector.

Step 3:Cluster quality measured by sum squared error representing data points between 2 cluster
- Fitness value calculated using equation(10)

Step 4: Update personal best using equation (6)

Step 5: Update global best using equation (7)

Step 6: Apply velocity update for each dimension of particle using equation(8)

Step 7: Generate new particles location using equation(9)

Step 8: Repeat steps 2 – 8 until i)stopping criterion reached till good solution got (or) ii)maximum number of generations completed

Figure 2 PSOCLUS – PSO based Clustering Algorithm

4.2 PSO and K-Means Based Clustering(PSOK)

PSO algorithm is applied for clustering since the problem with normal clustering process is that it locates optimal centroids of clusters rather than finding optimal partition. PSO performs globalized searching while only localized search can be performed using K-Means. Moreover PSO clustering algorithm could generate more compact clustering clustering results than normal K_Means . However, the problem is that for larger dataset PSO will take more number of iterations to converge to optima. But K-Means algorithm tends to converge faster than PSO. Due to the problems addressed with PSO and K-Means it is very difficult to finalize whether to use PSO or K-Means for clustering of web documents. Based on this reason, we propose a hybrid approach combining PSO and K-Means for web document clustering.

In the hybrid PSO+K-means algorithm, the multidimensional document vector space is modeled as a problem space. Each term in the document dataset represents one dimension of the problem space. A single particle 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. Partitioning algorithms starts with an initial k partitions and then uses an iterative process to optimize the cluster quality.K-means algorithm is one of the most popular and widely used partition clustering method.PSO can conduct a globalized searching for the optimal clustering, but requires more iteration numbers and computation than the K-means algorithm does. The K-means algorithm tends to converge faster than the PSO algorithm, but usually can be trapped in a local optimal area. The PSO+K-means algorithm combines the ability of the globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm and can avoid the drawback of both algorithms.

Algorithm: SEMPSOODC()
Input: Web Pages
Output: Retrieval of Relevant documents
Main Procedure:
Step 1: Annotate Web Pages using KIM plugin.
Step 2: Perform Concept Extraction for the annotated documents.
Step 3: Compute term weight through dot product of Term Frequency (TF) and
Inverse Document Frequency (IDF) .
Step 4: Calculate Semantic Similarity using equation(4).
Step 5: Recalculate the Concept Weight using equation(5).
Step 6: Cluster the documents by calling either PSOCLUS() or DCPSO()
Case 1: Call PSOCLUS() Case 2: Call PSOK()
Step 7: Calculate relevancy of documents by measuring precision,recall and
F-Measure as mention in equations(12,13 &14).

Figure 3 – SEMPSOODC Semantic Similarity based Clustering Algorithm using PSO

The hybrid PSO and K-Means algorithm includes two cases, one is PSO another is K-Means. At the initial stage PSO algorithm represented in Figure 2 is executed for a short period to identify cluster centroids and then K-Means algorithm is executed to find optimal clusters. The K-means module will inherit the PSO module's result as the initial clustering centroids and will continue processing the optimal centroids to generate the final result. In PSO and K-Means algorithm optimal solution is discovered by global search through PSO and faster convergence is achieved by applying K-Means algorithm.The result from PSO is used as initial seed for K-Means algorithm which is used to refine the results.

Steps for PSO and K-Means based Clustering (PSOK)

Algorithm:PSOK()

Input: Annotated documents

Output:Clustered document set

Step 1. Inherit cluster centroid from PSO algorithm by calling PSOCLUS().

Step 2. Assign each document vector to closest cluster centroid.

Step 3. Recalculate cluster centroid using equation(11).

Step 4. Repeat steps 2 and 3 until convergence is achieved.

Figure 4 PSO and K-Means based Clustering Algorithm

$$x_{i,k}(t) = \begin{cases} 0, \ if \ r_k(t) \geq p_{ini}, \\ 1, \ if \ r_k(t) < p_{ini} \end{cases} \quad (11)$$

Where d_j denotes document vector belonging to cluster S_j

c_j – denotes centoid vector

n_j- denotes number of document vectors belonging to cluster S_j

5 Experiments and Discussion

5.1 Experimental Setup

After performing document preprocessing like stemming and stopword removal is done. The documents are annotated using KIM plugin (http://www.ontotext.com)and it is represented using GATE tool API which is an open source tool for information retrieval. 20 usenet newsgroup dataset is used from which of 20000 messages are taken. Mini newsgroups the subset in newsgroup dataset is used for clustering. A set of queries were prepared manually for performance measurement.

5.2 Evaluation Metrics

Speed of response and the size of the index are factors in user happiness. It seems reasonable to assume that relevance of results is the most important factor: blindingly fast, useless answers do not

make a user happy. However, user perceptions do not always coincide with system designers' notions of quality. To measure ad hoc information retrieval effectiveness in the standard way, we need a test collection consisting of three things:

1. A document collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standard a binary assessment of either relevant or non relevant for each query-document pair.

The standard approach to information retrieval system evaluation revolves around the notion of relevant and non relevant documents. With respect to a user information need, a document in the test collection is given a binary classification as either relevant or non relevant. This decision is referred to as the goldstandard or ground truth judgment of relevance. In an information retrieval scenario, the instances are documents and the task is to return a set of relevant documents given a search term and to assign each document to one of two categories, "relevant" and "not relevant". The "relevant" documents are simply those that belong to the "relevant" category. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

$$V = (c * N(2,1) + 1) * \frac{intra}{inter} \tag{12}$$

$$intra = \frac{1}{N_p} \sum_{k=1}^{K} \sum_{u \in C_k} ||u - m_k||^2 \tag{13}$$

$$inter = \min \left\{ || m_k - m_{kk} ||^2 \right\} for\ all\ \text{k=1,2…,K-1 and kk=k+1,…,K} \tag{14}$$

5.3 Experimental Results and Discussion

20 random particles are generated and the fitness value is calculated based on cluster centroid. In PSO clustering the inertia weight is selected as 0.9 and the value of c_1 and c_2 are selected as 0.2 based on the theoretical studies of convergence performed in [8,11]. For every simulation the initial

centroid vector is selected randomly. The F-measure values are the average of 100 runs and after 100 iterations same cluster is obtained.

Based on the experiment conducted it is noted that the K-means clustering algorithm can converge to a stable solution within 20 iterations and PSO needs to repeat for more than 100 iterations to generate stable solution. PSO algorithm is executed for 25 iterations and the result is taken as the initial seed for K-means algorithm to generate final result. The global best solution result from PSO algorithm is used as the initial cluster centroid for K-means algorithm. The total executing iterations for PSO and K-means is 50.The POS and K-means algorithm generates the highest clustering compact result in the experiments. In PSO+K-means algorithm clustering experiment although 25 iterations is not enough for PSO to discover optimal solution it has high possibility that one particles solution is located in the vicinity of the global solution.

Three performance metrics such as Purity of cluster, F-Measure and CPU execution time is taken in to account to evaluate the performance of the proposed system. Experimental results show that combing PSO based clustering algorithm with ontology performs better than normal clustering algorithms without using ontology.The output of Hybrid PSOK-Means based clustering algorithm is the optimal number of clusters and relevancy of documents are calculated by applying evaluation metrics such precision, recall and FMeasure as mentioned in equations[1,4,2]. K-means algorithm is used as base algorithm for testing.

Table 2 Performance Comparison of K-Means and SEMPSOODC based on F-Measure

No. of Documents	**K-Means**	**SEMPSOODC Case 1: PSOCLUS()**	**SEMPSOODC Case 2: PSOK()**
100	0.60	0.62	0.63
200	0.61	0.64	0.66
300	0.63	0.66	0..69
400	0.64	0.68	0.71
500	0.66	0.69	0.73
600	0.68	0.71	0.75
700	0.70	0.72	0.77
800	0.72	0.74	0.79
900	0.73	0.75	0.81
1000	0.75	0.79	0.83
Average	0.672	0.70	0.736

While considering the time taken or speed of clustering it was found that Hybrid approach based on PSO and K-means takes 485 milliseconds on average to converge while tested with the Reuters dataset. At the same time when the order of execution is changed as K-Means then PSO(KPSO) the time taken reduces to 128 ms but the accuracy is very less compared to PSOK-Means based

approach. The reason for not getting optimal clusters using KPSO is that applying PSO initially returns best cluster center but K-Means will not return best cluster center since it selects cluster center randomly. This leads to reduction in execution time while applying K-Means first and then PSO(KPSO),but accuracy is less compared with PSOK. After performing analysis based on time and accuracy to recover optimal clusters hybrid approach based on PSO and K-Means(PSOK) is considered.

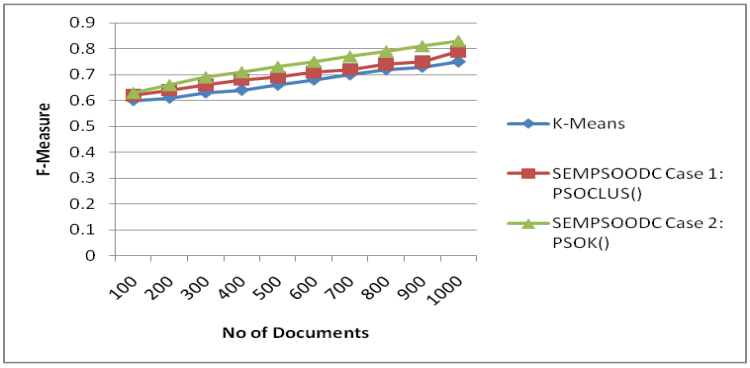

Figure 5 Performance Comparison of K-Means and SEMPSOODC based on F-Measure

Table 3 Difference Between SEMPSOODC and Keyword based K-Means Clustering

No	SEMPSOODC	Keyword based K-Means approach
1.	Extracts information based on meaning.	Extracts information based on key phrases.
2.	Ontology contains meaning and relations.	Keyword based methodology is less meaningful.
3.	Different views of documents are provided through concepts.	View of documents is provided based on keyword.
4.	SEMPSOODC F-Measure is higher than K-Means approach.	F-Measure of K-Means is lower than SEMPSOODC.
5.	SEMPSOODC outperforms K-Means approach.	K-Means approach is not as effective as SEMPSOODC.

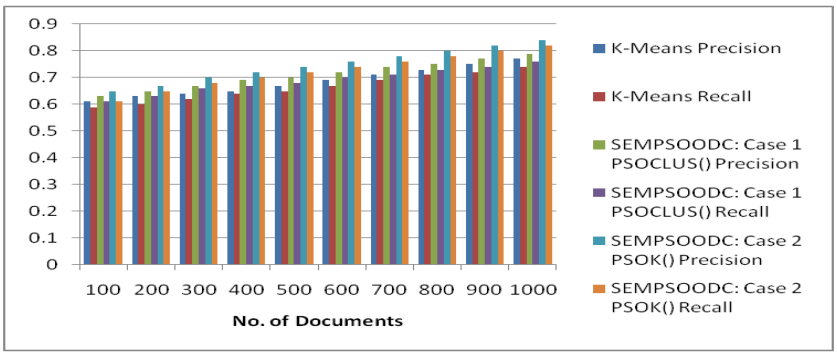

Figure 6 Performance Comparison of K-Means and SEMPSOODC based on Precision and Recall

Table 4 Comparison based on Purity of Cluster

No.of Clusters	**K-Means**	**SEMPSOODC Case 1: PSOCLUS()**	**SEMPSOODC Case 2: PSOK()**
3	0.65	0.73	0.76
5	0.67	0.78	0.81
15	0.69	0.83	0.85
20	0.7	0.85	0.88
Average	0.678	0.797	0.825

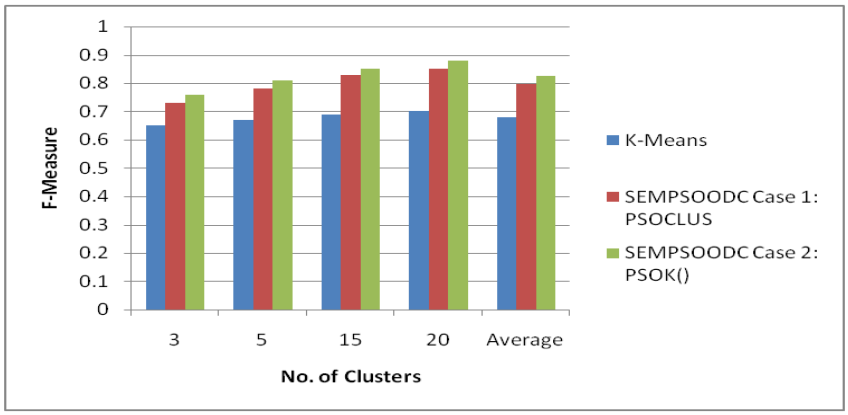

Figure 7 Comparison based on Number of Clusters

Performance analysis based on purity of cluster as in Figure 7 shows that the recovery of relevant documents is improved by 11.9% when SEMPSOODC is used instead of K-Means in clustering process. Table(3) shows the difference between SEMPSOODC with both the cases like normal PSO based clustering(PSOCLUS()), Hybrid clustering based on PSO and K-Means (PSOK()) and keyword based K-Means clustering process. Performance of PSOK is better since it returns optimal number of clusters even with minimal user interference. Precision and recall values Figure 6 shows that reweighting process based on ontology indexing along with dynamic clustering PSO shows better performance than normal K-Means process. Moreover traditional K-Means method cannot identify the semantic relationship between words. Accuracy of the proposed algorithm based on the number of clusters is also better than K-Means algorithm as mentioned in Table 4.While considering the time taken or speed of clustering, it was found that the ontology-based algorithm is fast and takes only 19.37 minutes on average while tested with the Reuters dataset. The K-means algorithm took 24.36 minutes, which is slow when compared with ontology based algorithm. All these results from the various experiments show that the clustering algorithm that uses semantics of the documents, that is, ontology-based clustering produces significant improvement in clustering results when compared with traditional existing algorithm.

6 Conclusions and Future Work

As the volume of information continues to increase, there is growing interest in helping people better find, filter and manage these resources. Ontology-based computing is emerging as a natural evolution of existing technologies to cope with the information onslaught. The proposed ontology based clustering system combined with semantic annotation is to improve the clustering process. The objective of our work is to improve the relevancy of documents over keyword based search. Hybrid clustering based on PSO and K-Means method shows performance improvement when compared to the baseline algorithm K-Means and normal PSO based clustering. Experimental results shows that the proposed methodology gives improved performance and better clustering than K-Means. Future work is to use Fuzzy logic in clustering and combining it with ranking of documents using ontology concepts to get better search results.

References

1. Ahmad Kayed, Eyas El-Qawasmeh & Zakariya Qawaqneh. (2010,December). Ranking Web Sites Using Domain Ontology Concepts. International Journal of Information & Management. Vol 47 pp.350-355.
2. Aleman-Meza, B., Halaschek, C., Arpinar, I., & Amith Sheth.(2003).A Context_Aware Semantic Association Ranking. Proc.First Int'l Workshop Semantic Web and Databses pp.33-50.
3. Amy, J.C., Trappey, Charles, V., Trappey,Fu-Chiang Hsu, & David Hsiao, W. (2009,June).A Fuzzy Ontological Knowledge Document Clustering Methodology.IEEE Transactions on Systems,Man and Cybernetics.Vol. 39, pp.806-814.
4. Atanas Kiryakov, Borislav Popov, Damyan Ognyanoff, Dimitar Manov, Angel Kirilov, & Miroslav Goranov. (2004,December).Semantic Annotation, Indexing, and Retrieval. Elsevier's Journal of Web Semantics. Vol 2 pp.49-79.
5. Danushka Bollegala, Yutaka Matsuo, & Mitsuru Ishizuka.(2011,July).A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. IEEE Transactions on Knowledge and Data Engineering,Vol 23 pp.977-990.
6. David, A., Grossman,& Ophir Frieder.(2004).Information Retrieval: Algorithms and Heuristics, Springer.
7. DW Van Der Merwe.(2003,December).Data Clustering using Particle Swarm Optimization. The 2003 Congress on Evolutionary Computation, 2003. CEC '03. Vol.1 ISSN: Print ISBN: 0-7803-7804-0 .
8. Grigoris Antoniou, & Frank Vanhamln.(2010). Semantic Web Primer,PHI Learning Pvt Ltd.
9. Hmway Hmway Tar, & Thi Thi Soe Nyunt.(2011).Ontology -Based Concept Weighting for Text Documents. International Conference on Information Communication and Management, IACSIT Press, Singapore IPCSIT. Vol.16 .
10. Ioan Cristian Trelea. (2003).The particle swarm optimization algorithm: convergence analysis and parameter selection. Inf. Process. Letter. 85(6) pp.317-325.
11. John Hebler, Mathew Fisher, &Ryan Blaces.(2009).Semantic Web Programming ,Wiley India.

12. Kalyani , S., & Swarup,K.S.(2011). Particle swarm optimization based K-means clustering approach for security assessment in power systems Expert Systems with Applications. Vol.38(9).pp.10839-10846.
13. Kennedy, J.,& Eberhart, R.C.(1995,Nov/Dec).Particle Swarm Optimization. Proceedings of the IEEE international conference on neural networks IV. Vol 4 pp. 1942–1948.
14. Kuncheva, L., & Bezdek, J. (1998).Nearest Prototype Classification:Clustering, Genetic Algorithms, or Random Search?. IEEE Transactionson Systems, Man, and Cybernetics-Part C: Applications and Reviews.Vol. 28(1), pp.160-164.
15. Maedche, A., Staab S., Stojanovic N., Studer R., & Sure Y.(2003).Semantic Portal: The SEAL Approach.Spinning the Semantic Web. pp.317-359.
16. Mahamed, G.H., Omran, Andries P., Engelbrecht,& Ayed Salman.(2005). Dynamic Clustering using Particle SwarmOptimization with Application in Unsupervised Image Classification, World Academy of Science, Engineering and Technology. Vol 9 ISSN:1307 6884,pp.199-204.
17. Maurice Clerc,&James Kennedy.(2002). The Particle Swarm - Explosion, Stability, and Convergence in a Multidimensional Complex Space. IEEE Trans. Evolutionary Computation. 6(1), pp.58-73.
18. Montserrat Batet, Aida Valls, & Karina Gibert.(2008,December).Improving classical clustering with ontologies.IASC,Japan.
19. Neelam, A.,& Sharma K.(2010,August).A Novel Approach for Organizing Web Search Results using Ranking and Clustering. International Journal of Computer Applications(0975 – 8887). 5(10).
20. Punitha , Mugunthadevi, & Punithavalli.(2011,May).Impact of Ontology based Approach on Document Clustering.International Journal of Computer Applications (0975 – 8887).22(2).
21. Stefania Gallova.(2007,November).Fuzzy Ontology and Information Access on the Web. IAENG International Journal of Computer Science,IJCS_34_2_11. 34(2).
22. Sridevi, U.K., Nagaveni. N.(2011).Semantically Enhanced Document Clustering Based on PSO Algorithm. European Journal of Scientific Research ISSN 1450-216X. Vol.57 pp. 485-493.
23. Thangamani, M., & Thangaraj, P.(2010,July).Ontology Based Fuzzy Document Clustering Scheme.International Journal of Modern Applied Science. Vol. 4.
24. Turi, R.H. (2001).Clustering-Based Colour Image Segmentation, PhD Thesis, Monash University, Australia.
25. Xiaohui Cui, Thomas E., Potok, Cui, X.,& Potok, T.K.(2005).Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm. Journal of Computer Sciences. pp. 27-33.
26. Yang Cheng.(2008). Ontology-Based Fuzzy Semantic Clustering. Third International Conference on Convergence and Hybrid Information Technology. pp.128-133.
27. Zongmin Ma.(2006).Soft Computing in Ontologies and Semantic Web ISBN-10 3-540-33472-6 Springer Berlin Heidelberg New York.

.