Ontology-based Semantic Text Documents Clustering Using Particle Swarm Optimization

Wai Wai Lwin, Nang Saing Moon Kham
University of Computer Studies, Yangon
wai2020.smile@gmail.com, moonkham.ucsy@gmail.com

Abstract

The World Wide Web, the largest shared information source is growing exponentially and the amount of business news on the web is overwhelming and need to be handled properly. As such, grouping the web document into cluster for speedy information retrieved becomes imperative. Clustering technique organizes a large quantity of unordered text documents into a small number of meaningful and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanisms. The quality of clustering result depends greatly on the representation of documents and the clustering algorithm. In traditional document representation methods, the frequency count of the document terms is used for the feature vector representing the documents. But traditional document representation methods cannot identify related terms semantically. Documents written in human language contain contexts and the words used to describe these contexts are generally semantically related. Motivated by this fact, domain ontology is developed to promote the enrichment of semantic representation of terms. Then, Particle Swarm Optimization (PSO) clustering algorithm is used to cluster the web documents efficiently. The paper constitutes the comparative results of using PSO algorithm only and PSO algorithm with Ontology for clustering web documents. According to the analytical results, the representation of terms by using Ontology is significantly efficient and the implementation of PSO algorithm achieves better performance in intra cluster and inters cluster similarity.

Keywords: *document clustering, representations, PSO, Ontology, inter cluster, intra cluster.*

1. Introduction

Data mining is the process of extracting the implicit, previously unknown and potentially useful information from data [13]. That information can be used to decrease search time and cuts costs. With the ever increasing volume of information, fast and high-quality document clustering algorithms play an important role to efficiently navigate, summarize and organize the information. Web text feature extraction and clustering are the main challenging tasks in web data mining, which requires an efficient clustering technique.

Clustering in general is an important and useful technique that automatically organizes a collection with substantial number of data objects into a much smaller number of coherent groups [1]. Clustering algorithms define a similarity metric that determines the distance from a document to a point that represents a cluster. The notion of a "cluster" varies between algorithms and is one of the many decisions to take when choosing the appropriate algorithm for a particular problem [3].

Clustering has been investigated to use in a number of different areas of text mining, document organization, data analysis and information retrieval. Document clustering can be defined as grouping text documents within a cluster have similar concepts. The representation of documents should be as compact as possible in order to allow efficient processing of large documents collections, yet it should contain all of the relevant information [17]. Vector Space Model is a widely used method for document representation in information retrieval. In this model, each document is represented by a feature vector. The unique terms occurring in the whole document collection are identified as the attributes (or features) of feature vector. There are some common used term weighting methods such as binary method, tf (term frequency) method, and tf-idf (term frequency-inverse document frequency) method etc. in the vector space model. The “bag of words” feature representation is not able to reflect the semantic content of a document because of the synonym problem and polysemy problem [20].

PSO is a bio-inspired swarm intelligence algorithm introduced by Kennedy and Eberhart in 1995 as a population-based stochastic search and optimization process [5]. It is originated from the computer simulation of the individuals (particles or living organisms) in a bird flock or fish school,

which basically show a natural behavior when they search for some target (e.g., food) [7]. To converge to the global optima of some multidimensional and possibly nonlinear function or system is the main goal of PSO.

PSO comply with the same way of other evolutionary algorithms (EAs), such as Ant Colony Optimization algorithm (ACO) and genetic algorithm (GA). PSO algorithm is used to cluster multi-dimensional data clustering and high voluminous of data efficiently. Many researchers found that it has better performance than conventional algorithms such as better cluster formation; accuracy is fast and high, etc [2].

In this paper, we present a semantic text document clustering approach that using a domain ontology and PSO algorithm to cluster the Business related news of News Websites and Financial Websites. We generate a domain ontology for representing the term with enriched semantic and synonyms and then clustering of the documents using PSO clustering technique to achieve the semantic concepts of terms, to filter web documents precisely and a better performance for inter and intra cluster similarity.

The rest of the paper is presented as follows. In section 2, it describes the related work of the system. Section 3 describes details of the architecture of the system. Section 4 represents the discussion about the analysis results of the system. Finally, section 5 is the conclusion and future work of the proposed system.

2. Related Work

Most of the clustering techniques are based on traditional collection of terms approach to represent the documents. By using the traditional techniques based on word and/or phrase frequencies for document representation cannot be handled text ambiguity, synonymy and semantic similarities.

In [21], the authors proposed a semantic similarity based model to capture the semantic of the text. The proposed model in conjunction with lexical ontology solves the synonyms and hypernyms problems. The proposed model use semantic weights added to the term frequency weight to calculate the semantic similarity between terms.

Conceptual features in text representation were introduced in [10, 22]. They proposed three methods to include concept features in VSM, namely, (i) adding concept features to the term space (i.e., term+concept); (ii) replacing the related terms with concept features and (iii) reducing the VSM to only concept features. Experimental results in [2] showed that only the term+concept representation improved clustering performance.

A semantic text document clustering approach based on the WordNet lexical categories and Self Organizing Map (SOM) neural network is proposed in [18]. In this work, documents vectors are generated by using the lexical category mapping of WordNet after preprocessing the input documents.

In [12], the traditional vector space model represents several features present in document have been used for computing the document similarity cannot account for the semantics of the document. It proposed to define a semantic-based similarity measure by utilizing text annotation through external thesauruses like WordNet (a lexical database).

Gad and Mohamed S. Kamel [20] proposed a semantic similarity based model to capture the semantic of the text. The proposed model in conjunction with lexical ontology solves the synonyms and hypernyms problems. The proposed model utilizes WordNet as ontology to reflect the relationships by the semantic weights added to the term frequency weight to represent the semantic similarity between terms.

Many researchers found a modern algorithm in the previous years, the PSO algorithm which is aimed to converge to the global optima of some multidimensional and possibly nonlinear function or system. Researchers found that it has better performance than conventional algorithms such as better cluster formation, fast and high accuracy.

But, PSO clustering has some weaknesses such as filtering in web documents, centroids convergence problem, global optima searching is still not sufficient in high dimensional data and similarity measure on diverse data and extends, etc [16].

Some researchers have presented a new approach using particle swarm optimization technique to improve term extraction precision in [10]. Experimental results showed the use of particle swarm optimization technique can improve the precision of the extracted terms compared with four other known algorithms (TFIDF, Weirdness, Glossary Extraction and Term Extractor).

In [8], the authors combined the mutual information matrix and the traditional vector space model to design a new data model (considering the correlation between terms).

In [9], the authors presented a hybrid Particle Swarm Optimization, Subtractive + (PSO) clustering algorithm that performed fast clustering. The results illustrated that the Subtractive + (PSO) clustering algorithm can generate the most compact clustering results as compared to other algorithms.

In [19] authors showed that combination of ontology and optimization to improve the clustering performance. They proposed a ontology similarity measure to identify the importance of the concepts in the document. Ontology similarity measures are defined using WordNet synsets and the particle swarm optimization is used to cluster the document. Even though WordNet is a popular lexical dictionary, it doesn't have concept at all. The proposed system will be developed a specific domain Ontology instead of using WordNet.

In the [22], a hybrid algorithm of PSO represents two functions, the PSO and the K-means algorithms. The hybrid algorithm first executes PSO clustering algorithm to find points close to the optimal solution by global search and simultaneously avoid high computation time. In this case PSO clustering is terminated when the maximum number of iterations is exceeded. This showed that the result of the PSO algorithm is then used as initial centroid vectors of the K-means algorithm. The K-means algorithm is then executed until maximum number of iterations is reached.

The K-means algorithm tends to converge faster which leads less function evaluation than the PSO, but less accurate clustering [11] and PSO can conduct a globalized searching for the optimal clustering, but requires more computation time than the K-means algorithm. The hybrid PSO algorithm in [22, 11] achieves the advantage of both the algorithms which shows better globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm**.**

In this paper we proposed ontology-based semantic text document representation to reduce the complexity of terms which compact the representation of documents vector to decrease processing time. Then the system uses PSO clustering algorithm to execute to find points close to the optimal solution by global search and simultaneously avoid high computation time. This paper achieves compact representation of documents vector which decrease the iteration time of PSO and the performance of inter and intra cluster similarity of PSO is high.

3. Architecture of the System

The earliest state of the proposed system is started with preprocessing. And then, the system performs generating document vector for document dataset to apply clustering algorithms on web document dataset. So proper representation of document based on Ontology is extremely important that they can be represented easily and reduces complexity. The Figure: 1 describes the architecture of the system.

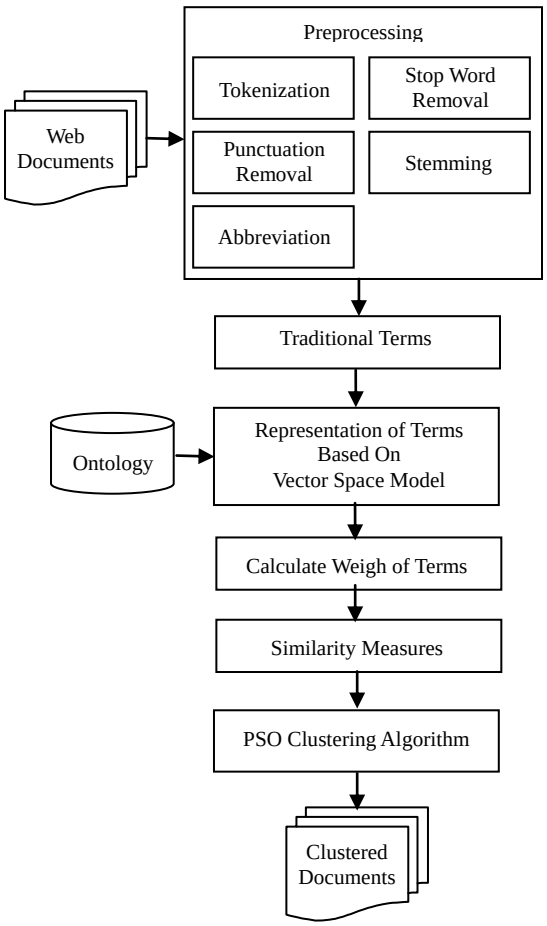

Figure 1: Architecture of Web Document Clustering using Domain Ontology and PSO clustering Algorithm

3.1 Preprocessing of Documents

Preprocessing of documents leads to gain the good quality of clustering result. Business related test pages are used in the work and detail processes of preprocessing stages are as follow.

Tokenization: break an item into atomic words e.g., break "goldprice" into "gold" and "price", break "US_Dollar" into "US" and "Dollar".

Punctuation Words: remove the punctuation words e.g., comma, semicolon, hyphen, question mark,

dash, slash, bracket, quotation mark, parenthesis, etc.
Stop word removal Stemming: remove a, and, the, is, at, which, on, about etc.
Stemming: stemming is defined as the technique of finding root/ stem of a word.
For example: Plays, playing and played are holding a single root word *'play'.* Some of stemming methods are:

i. *Remove Ending:* elimination of suffixes like "es" from "goes" to "go".
ii. *Transform words:* the root words are derived by adding a transform suffixes like "ies" instead of "y" which affect effectiveness of word matching. For example- "try" derived to "tries".

Abbreviations: expand abbreviations and acronyms to their full words. In this proposed system, the currency symbols consists of and they are expanded as the following samples e.g., "$" "Dollar", "Euro" to "Europe", etc.
After this steps standardization of words have done such as irregular words are standardized to a single form, e.g., "Dollar" derived to "dollar", "colour" derived to "color".

3.2 Ontology

Ontology defines a common vocabulary for a particular domain to share information. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. Domain ontology can define as the concepts relevant to a particular topic or area of interest, for example, information technology or computer languages, or particular branches of science.

An ontology is a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions)).

Ontology together with a set of individual instances of classes constitutes a knowledge base. In reality, there is a fine line where the ontology ends and the knowledge base begins. Classes are the focus of most ontologies. Classes describe concepts in the domain.

Developing an Ontology includes:

- defining classes in the ontology,
- arranging the classes in a taxonomic (subclass–superclass) hierarchy,
- defining slots and describing allowed values for these slots,
- filling in the values for slots for instances.

The reasons why the user wants to develop an Ontology are:

- To share common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

An Ontology might be considered the most complete and powerful model for information representation. It is not only a classification of concepts but also of individuals. Ontology is an attempt to define an exhaustive and formal hierarchy within a given domain of knowledge that contain for the relevant terms, the relationship between them and the rules within the domain. The model consists of a set of types with their properties, with relationships of different types between them, a set of rules and instances.

In this paper, a light weight business news Ontology is developed and it is included three parent classes named economy, business of sport and global education. These classes inherit subclasses, for example, economy class is sub divided into metal, currency and petrol.

In this domain ontology, economy class consists of metal, currency and petrol sub classes. These sub classes are disjointed each other. Platinum is_a_kind_of metal and ecomony class has_a_kind_of business_sports. In the domain ontology, the sub class region consists of individuals such as Hong Kong, India, America and Myanmar, etc. Instances belong to broader and to narrower terms. If instances can only be assigned to one subtype, then they belong to a disjoint partition. For example, an individual of class region class cannot be both Myanmar and America.

Figure: 2 represents part of the business news Ontology which includes super class-subclass hierarchy, is-a and has-a relationship between classes.

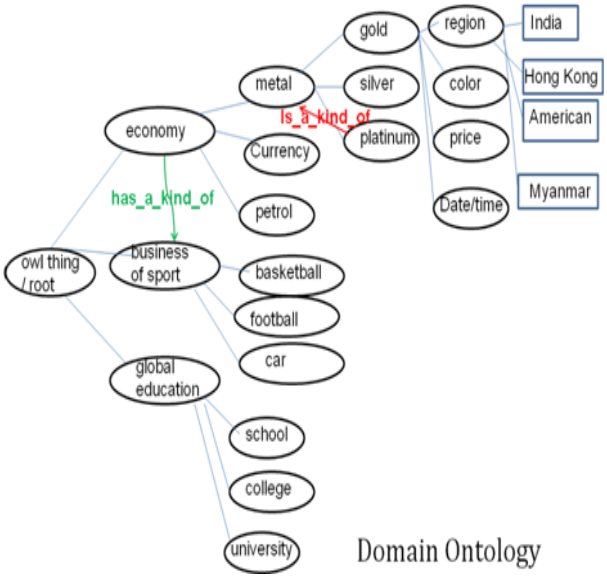

Figure 2: Chart of Domain Ontology for proposed system

3.3 Representation of Documents based on Ontology

A document is commonly represent as a vector of terms in a vector space model (VSM) [4]. The basis of the vector space corresponds to distinct terms in a document collection. Each vector represents one document. The components of the document vector are the weights of the corresponding terms that represent their relative importance in the document and the whole document collection.

In this paper, we create a domain Ontology as a lexical database in which terms are organized in so-called synsets consisting of synonyms and thus representing a specific meaning of a given term. In our ontology, ID is unique for each synset for each sense of term. All the synonyms in this synset share the same Id. Sense tagged data have been used to extract the exact sense of a word. For the synset concept, corresponding ID of the word is taken and vectors IDs are prepared. Once the document vectors are completed in this way, weights are assigned to each words across the corpus using TF*IDF method, which combination of the term frequency (TF), and the inverse document frequency (IDF).

TF*IDF is mathematically written as

$$W_{ij} = tf_{i,j} * \log (N / df_i) \qquad (1)$$

Where W_{ij} is the weight of the term **i** in document **j**.

$tf_{i,j}$ is the number of occurrences of term **i** in document **j**.

N is the total number of documents in the corpus, df_i is the number of documents containing the term **i**.

However, text documents exhibit the rich linguistic and conceptual structures. Based on these considerations we need to improve the conceptual background knowledge available to text mining. Therefore, we use the algorithms which take advantages of concept background knowledge.

This paper based on mining the term information with the aid of the conceptual background knowledge given by domain ontology. It combines the background knowledge into the traditional term-based vector space model and modify the term vector according with the following methods [8].

A set of documents X in traditional term-based VSM is defined by $X=\{x_1,x_2,x_3,\ldots\ldots\ldots, x_n\}$. One document is represented by $x_j=(x_{j1i}, x_{j2}, \ldots\ldots x_{jm})$. Terms in the vocabulary extracted from the corresponding document collection are represented by $\{t_1, t_2,t_3,\ldots\ldots,t_m\}$.

Where x_{ji} is the weight of term t_i in document x_j and usually determined by the number of times t_i appears in x_j(known as the *term frequency*).

For each term t_{i1}, we check whether it is semantical correlated to the other term t_{i2} with Ontology. $\delta_{i_1 i_2}$ uses to indicate the semantic information between two terms. If t_{i_2} appears in the synsets of t_{i_1} (*only synonym and hypernym synsets are considered*), $\delta_{i_1 i_2}$ *will be treated in a same level for different* t_{i1} and t_{i2}, otherwise, $\delta_{i_1 i_2}$ will be set zero. With $\delta_{i_1 i_2}$, *the weight* x_{ji} of term t_{i1} in each document X_j will be changed by equation 2.

$$\tilde{x} = x_{ji_1} + \sum_{\substack{i_2=1 \\ i_2 \neq i_1}}^{m} \delta_{i_1 i_2} x_{ji_2} \qquad (2)$$

This updates the original *term-base VSM* by considering the semantic relationship between each pair of terms, and the new text representation is called by *ontology-based VSM*[8]. Table 1 and Table 2 show a simple example for the modification

of text representation with Ontology. Table 1 gives two documents in the traditional *term-base VSM*. According to Ontology, terms *gold*, *silver* and platinum are semantic related to each other, so we uses equation 2 [8] updating the term weights for each document as shown in Table 2 (where δ is assigned to 0.8)

Table 1: A simple example for traditional term-based VSM

	gold	silver	platinum	school
d1	5	0	3	2
d2	0	4	1	0

Table 2: The representation for Table 1 data in ontology based VSM

	gold	silver	platinum	school
d1	7.4	6.4	7	2
d2	4	4.8	4.2	0

3.4 Similarity Measure

The mutual information between two terms t_1 and t_2 can be calculated on the basis of the ontology based VSM[8]. This paper adopted *cosine similarity* to measure the term mutual information between their corresponding vectors:

$$\cos\left(\angle\left(t_1, t_2\right)\right) = \frac{t_1 . t_2}{\|t_1\| . \|t_2\|} \quad (3)$$

$$= \frac{\sum_{j=1}^{n} \tilde{x}_{j1} \tilde{x}_{j2}}{\sqrt{\sum_{j=1}^{n} \tilde{x}_{j1}^2} \sqrt{\sum_{j=1}^{n} \tilde{x}_{j2}^2}} \quad (4)$$

Where $\tilde{x}_{j1}$ and $\tilde{x}_{j2}$ represents the term weights of t_1 and t_2 in the document $\tilde{X}_j$ in the *ontology-based VSM*.

In addition, the proposed system also uses cosine similarity measure [23] to compute the similarity between two documents. The similarity between two documents needs to be computed in a clustering analysis process.

3.5 Particle Swarm Optimization

PSO algorithm and its concept of "Particle Swarm Optimization"(PSO) were introduced by James Kennedy and Russel Eberhart in 1995 [5]. PSO is a population-based stochastic search algorithm which is modeled after the social behavior of a bird flock. A swarm refers to a number of potential solutions to the optimization problem, where each potential solution is referred to as a particle. The aim of the PSO is to find the particle position that results in the best evaluation of a given fitness (objective) function [14].

Each individual in the particle swarm is composed of three D-dimensional vectors, where D is the dimensionality of the search space. These are the current position x_i, the previous best position p_i, and the velocity v_i [15]. The i^{th} particle is represented by a position denoted as $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$. In a PSO system, each particle flows through the multidimensional search space, adjusting its position in search space according to its own experience and that of neighboring particles. To evolve towards an optimal solution a particle uses a combination of the best position realized by itself and the best position realized by its neighbours. The standard PSO method updates the velocity and position of each particle according to the equations given below.

$$v_{id}(t+1)=\omega.v_{id}(t)+c_1.rand().(p_{id}-x_{id})+c_2.rand().(p_{gd}-x_{gd}) \quad (5)$$

$$x_{id}(t+1)=v_{id}(t+1)+x_{id}(t) \quad (6)$$

where c_1 and c_2 are two positive acceleration constants, rand() is a uniform random number in (0,1), p_{id} and p_{gd} are the best positions found so far by the i^{th} particle and all the particles respectively, t is the iteration count and ω is an inertia weight which is usually, linearly decreasing during the iterations. The inertia weight ω plays a role of balancing the local and global search.

In the context of clustering, a single particle represents the N_c cluster centroid vectors. That is, each particle x_i is constructed as follows:

$$x_i=(o_{i1},\ldots,o_{ij},\ldots,o_{iNc}) \quad (7)$$

Where o_{ij} refers to the j^{th} cluster centroid vector of the i^{th} particle in cluster C_{ij}.

Therefore, a swarm represents a number of candidate clusters for the current data vectors. The fitness of particles is measured using the equation given below.

$$f = \frac{\sum_{i=1}^{N_c}\left\{\frac{\sum_{j=1}^{P_i} d(o_i, m_{ij})}{P_i}\right\}}{N_c} \quad (8)$$

where m_{ij} denotes the j^{th} document vector, which belongs to cluster i; o_i is the centroid vector of the i^{th} cluster; $d(o_i, m_{ij})$ is the distance between document mij and the cluster centroid o_i; P_i stands

for the number of documents, which belongs to cluster C_i and N_c stands for the number of clusters.

Figure: 3 represents the step by step procedures of the PSO clustering algorithm.

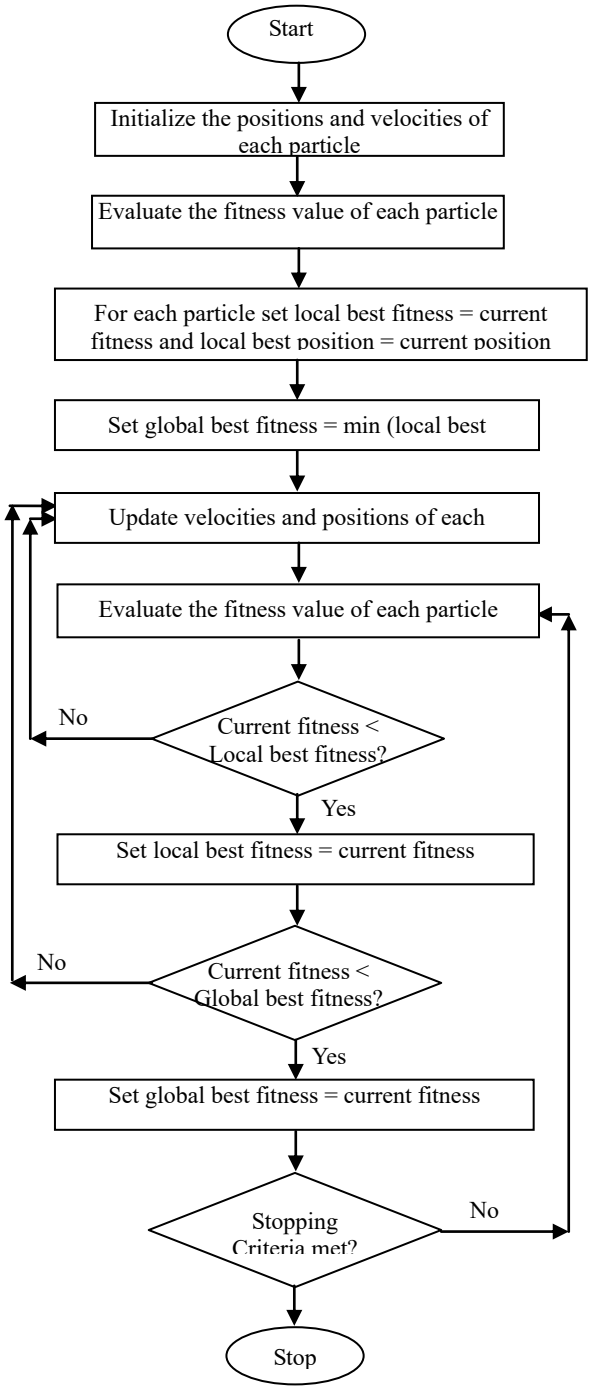

Figure 3: Flow diagram of the PSO clustering algorithm

4. Experimental Results

Dataset are collected from BBC, Channel_News_ Asia, Goldprice.org, Instaforex, Kitco, Oanda, and 24hgold.com. The system is used 100 test pages of these websites containing different data from different domains such as literature, media, business etc. The system is tested with PSO algorithm only and Ontology + PSO algorithm.

The results show that a domain Ontology can fully support to decrease ambiguity to terms. And also the reasonable representation of documents leads PSO clustering algorithm to conduct an appropriate globalized searching for the optimal clustering. In the PSO clustering algorithm, the system is chosen 10 particles, the inertia weight ***w*** is initially set as 0.89 and the acceleration coefficient constants ***c1*** and ***c2*** are set as 1.47.

In this ongoing study, the system will be obtained that Intra-cluster similarities (i.e. the distance between data vectors within a cluster) could be maximized and Inter-cluster similarity (i.e. the distance between the centroids of the clusters) could be minimized.

The test results are shown in the table 3.

Table 3: Comparison results of using Ontology+ PSO algorithm and PSO algorithm

Test Data Set	Test Method	Algorithm	Intra Cluster	Inter Cluster
100	Based on Term	Ontology + PSO	83%	6.83%
		PSO	63%	9.2%

5. Evaluation of cluster

There are numerous evaluation measures to validate the cluster quality. To evaluate the clustering results of proposed system, precision, recall, and F-measure has been used.

For cluster *j* and cluster *i*:

$$\text{Recall } (i,j) = n_{ij}/n_i \quad (9)$$

$$\text{Precision } (i,j) = n_{ij}/n_j \quad (10)$$

where n_{ij} is the number of members of class *i* in cluster *j*, n_j is the number of members of cluster *j* and n_i is the number of members of class *i*. F-measure is computed using precision and recall as below:

$$F(i,j) = \frac{2 * recall\,(i,j) * precision\,(i,j)}{(precision\,(i,j) + recall\,(i,j)} \quad (11)$$

Table 4 shows the F-values for varying number of

clusters with only PSO algorithm and PSO algorithm with Ontology.

Table 4: Comparison result of F-values

No. of Clusters(K)	F- measure	
	PSO	PSO + Ontology
K=2	0.59	0.75
K=3	0.46	0.725
K=4	0.49	0.69
K=5	0.465	0.66

6. Conclusion and Future Work

In this paper, the experimental results show higher results using applying domain Ontology and Particle Swarm Optimization (PSO) algorithm rather than using PSO algorithm only. By using Ontology-based Semantic representation and PSO clustering algorithm, the performance of intra cluster similarity is higher than the previous works of proposed system. Moreover, it is found that the inter cluster similarity achieves the appropriate results also. The proposed system gives the better result of F-measure values as compare to PSO only algorithm on Business News Group dataset (100 documents) so that the better is the clustering solution.

The proposed system is ongoing work and the future work of the system will upgrade the Ontology and make effort to change PSO's variant for clustering web document data.

References

[1] A. Huang,*"Similarity Measures for Text Document Clustering" NZCSRSC 2008*, April 2008, Christchurch, New Zealand.

[2] A. Hotho, S. Bloehdorn *"Text classification by boosting weak learners based on terms and concepts".In Proceedings of the IEEE international conference on data mining*, Brighton, UK, pp 72-79, 2004

[3] J. Ghorpade-Aher, R. Bagdiya, *"A Review on Clustering Web Data using PSO"*, International Journal of Computer Applications (0975 – 8887) Volume 108 – No. 6, December 2014

[4]G. Salton, and M. J. McGill, *Introduction to Modern Information Retrieval*, *McGraw-Hill* Inc., 1983

[5] J. Kennedy and R.C. Eberhart, *"Particle Swarm Optimization," Proc. IEEE, International Conference on Neural Networks. Piscataway*. Vol. 4, pp 1942-1948,1995

[6] J. Sedding and D. Kazakov, *"WordNet-based Text Document Clustering," ROMAND*, page104, 2004

[7] Kiranyaz, Serkan, et al. *"Fractional particle swarm optimization in multidimensional search space." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions*, 40.2 (2010): 298-319.

[8]Liping Jing, Lixin Zhou, Michael K.Ng, and Joshua Zhexue Huang *"Ontology-based Distance Measure for Text Clustering"*

[9] M. El-Tarabily, *"A PSO-Based Subtractive Data Clustering Algorithm ", IJORCS* , pp.1-9, 2013

[10] M. Syafrullah and N. Salim, *"Improving Term Extraction Using ParticleSwarm Optimization Techniques"*, *JOC* , pp. 116-120, 2010

[11] M. Omran, A. Salman, A.P. Engelbrecht, *"Image Classification using Particle Swarm Optimization," Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning*, Singapore, 2002

[12] M. Rafi and M. Shahid Shaikh, "An improved semantic similarity measure for document clustering based on topic maps," *Computer Science Department, NU-FAST, Karachi Campus, Pakistan*

[13] R. Bhagel, and R. Dhir, *"A Frequent Concept Based Document Clustering Algorithm", IJCA*, vol 4, no.5, 2010

[5] R. C. Eberhart and J. Kennedy, *"A new optimizer using particle swarm theory," in Proc. 6th Int. Symp. Micro Machine and Human Science*, 1995, pp. 39-45.

[14] S.C. Satapathy, N. VSSV P B. Rao, JVR. Murthy, R. P.V.G.D. Prasad, *"A Comparative Analysis of Unsupervised K-means, PSO and Self-Organizing PSO for Image Clustering," International Conference on Computational Intelligence and Multimedia Applications*,2007

[15] S. Sarkar, A.Roy, B.S.Purkayastha, *"Application of Particle Swarm Optimization in*

data clustering : A survey," International Journal Of Computer Applications (0975- 8887) Volume 65-No.25, 2013 International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014

[16] S.Sarkar, A. Roy and B. S. Purkayastha,*"A Comparative Analysis of Particle Swarm Optimization and K-means Algorithm For Text Clustering Using Nepali Wordnet", International Journal on Natural Language Computing (IJNLC)*, Vol. 3, No.3, June 2014

[17] Sunita Sarkar et al, *"Clustering of Documents using Particle Swarm Optimization and Semantics Information", International Journal of Computer Science and Information Technologies (IJCSIT)*, Vol. 5(3),2014, 4175-4180

[18] T.F Gharib, M. M Fouad, A. Mashat, I.Bidawi, *" Self Organizing Map based Document Clustering Using WordNet Ontologies". IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 1, No. 2, 2012

[19] U. K. Sridevi and . N. Nagaveni. *"Semantically Enhanced Document Clustering Based on PSO Algorithm", European Journal of Scientific Research*, ISSN 1450-216X Vol.57 No.3, pp.485-493, 2011

[20] Walaa K. Gad and Mohamed S. Kamel (2009) Enhancing Text Clustering Performance Using Semantic Similarity, ICEIS, LNBIP 24, pp. 325-335, 200

[21]W. K. Gad and M. S. Kamel, *"Enhancing Text Clustering Performance Using Semantic Similarity"*, *ICEIS, LNBIP 24*, pp. 325–335, 2009

[22]X. Cui, T.E. Potok, P. Palathingal, *"Document Clustering using Particle Swarm Optimization," IEEE*,2005

[23] X. Rui, *"Survey of Clustering Algorithms", IEEE transactions on Neural Networks*, 16(3), pp.634-678, 2005