Web Documents Clustering Using PSO Algorithm and Ontology

Wai Wai Lwin
University of Computer Studies, Yangon
wai2020.smile@gmail.com

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

Abstract

The largest shared information source, the World Wide Web has been increasing a tremendous proliferation in the amount of information rapidly. As a result of its huge sharing and highly dynamic data, there is a need for grouping the documents into clusters for faster information retrieval. Clustering web documents is collection of documents into groups such that the documents within each group are similar to each other. Document clustering is one of the main challenging tasks in web data mining and it is still requires an efficient clustering techniques. The typical way of representing a web document is a huge box of terms. The representation of these terms is often unsatisfactory as it does not exploit the semantics. This paper proposes Particle swarm Optimization (PSO) and Ontology for clustering of text documents. A domain Ontology is developed thus it enriched with semantic and synonyms for the representation of terms. Particle Swarm Optimization (PSO) algorithm is used to cluster high voluminous of data efficiently. This paper presents comparative results of using PSO algorithm only and PSO algorithm using Ontology for clustering web documents. The analysis result of test data is efficient enough in representation of terms by using Ontology and the performance of PSO clustering algorithm is high in intra cluster and inters cluster similarity.

Keywords: *document clustering, representation of terms, PSO, Ontology, inter cluster, intra cluster*.

1. Introduction

The World Wide Web continues to grow at a fantastic rate and the amount of information on the web is overwhelming. Due to its wide distribution, openness and highly dynamic data, the resources on the web are greatly scattered and they have no unified management and structure. Near about 90 % web data is unstructured and needed to be structure as it greatly reduces the efficiency in using web information. Web text feature extraction and clustering are the main challenging tasks in web data mining, which requires an efficient clustering technique [2].

Data mining is the process of extracting the implicit, previously unknown and potentially useful information from data. That information can be used to decrease search time and cuts costs. Digitized text documents are increasing exponentially. As such, clustering becomes imperative for ever increasing digitized data [13]. Clustering has been investigated to use in a number of different areas of text mining, document organization, data analysis and information retrieval.

Clustering is an useful technique that automatically organizes a collection with a substantial number of data objects into a much smaller number of coherent groups [1, 4]. 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 [2].

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) [3]. 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].

On the other hand, researchers had found some weaknesses in PSO clustering like it cannot filter web documents, convergence problem, searching for global optima is still not sufficient in high dimensional data and similarity measure on diverse data and extends, etc [2, 13]. The typical way of representing a text document is a huge box of terms. The representation of these terms is often unsatisfactory as it does not exploit the semantics [13].

In this system, Particle Swarm Optimization (PSO) and Ontology is applied to cluster the World's Gold prices and currency exchange rates among News Websites and Financial News documents. The system is aimed 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

Due to complex linguistics properties of the text
Documents, document clustering can be a huge challenging for web mining. Most of the clustering techniques are based on traditional collection of terms approach to represent the documents. Using traditional techniques for document representation based on word and/or phrase frequencies that cannot be handled text ambiguity, synonymy and semantic similarities.

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 [2, 13].

In [14] authors suggested a model to subtract web data by clustering with BRAPSO. The purpose is study of multi dimensional clustering techniques to achieve higher accuracy. This model shows that convergence problem of clustering is improving.

Some researchers make effort to change PSO and its variant for clustering high-dimensional data in [15]. Changing PSO variants lead to outstanding performance and better cluster formation but searching in global optima is still not sufficient.

In [17] authors revealed a case study of clustering high dimensional web Log data with RVPSO and PSO RTP models. The models showed that the velocity of PSO and global optimum were improved.

D. P. Rini, S. Mariyam, Shamsuddin, S. S. Yuhaniz in [16] surveyed PSO and they have found that changing of PSO variants improved clustering and optimized in performance but the convergence issue occurred.

Some researchers have presented a new approach using particle swarm optimization technique to improve term extraction precision in [18]. 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, GlossaryExtraction and TermExtractor).

In[19], 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 [20], 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 [23, 24]. 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 [21] showed that only the term+concept representation improved clustering performance.

In [22] 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.

A semantic text document clustering approach based on the WordNet lexical categories and Self Organizing Map (SOM) neural network is proposed in [23]. In this work, documents vectors are generated by using the lexical category mapping of WordNet after preprocessing the input documents.

In the [24], 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 [25] 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 [24,

25] 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 proposed system, a technique of clustering Web pages by using Ontology and PSO algorithm is introduced. Proper representation of document based on Ontology is extremely important so that they can be reducing 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 system 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 Web Document Clustering 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) below describes the architecture of the system.

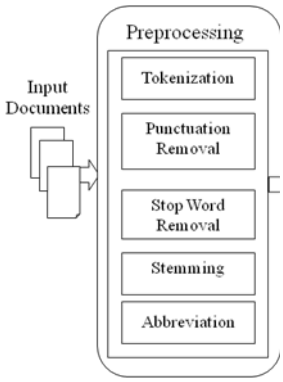

Figure 1: Approaches Showing Document Clustering

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:

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

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

i. To share common understanding of the structure of information among people or software agents

ii. To enable reuse of domain knowledge
iii. To make domain assumptions explicit
iv. To separate domain knowledge from the operational knowledge
v. To analyze domain knowledge

Often an ontology of the domain is not a goal in itself. Developing an ontology is akin to defining a set of data and their structure for other programs to use. Problem-solving methods, domain-independent applications, and software agents use ontologies and knowledge bases built from ontologies as data [9].

In this system, classes of the Ontology might be gold, currency, business_news, etc. The classes and sub_classes will consists of is_a and has_a relationships. The system will be developed a light_weigh Ontology to achieve high accuracy and similarity measures of the words.

3.3 Representation of Documents based on Ontology

For creating the vector space model, a document *d* is represented as n-dimensional document vector [wt_0,wt_1, . . .,wt_n], where t_0, t_1, . . .,t_n is a set of distinct terms present in given document and wt_i expresses the weight of term t_i in document *d*[7]. The weight of a term reflects the importance of term within a particular document.

A domain Ontology for Gold Price and Currency Exchange has been used to represent business documents. Ontology is a lexical database that provides the sense information. A term may have more than one sense. In this concept, ID is unique for each synset for each sense. All the synonyms in this synset share the same ID. For the synset concept, corresponding ID of the word is taken and vectors of IDs are prepared. Once the document vectors are completed in this way, weights are assigned to each word across the corpus using TF*IDF method [6], which is the combination of the term frequency (TF), and the inverse document frequency (IDF). TF*IDF is mathematically written as

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

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**.

After creating representation of documents vectors, the weight of terms are computed in the proposed system. This process reflects the importance of term within a particular document.

3.4 Similarity Measure

The similarity between two documents needs to be computed in a clustering analysis process. There are several similarity measures are available to compute the similarity between two documents like Euclidean distance, Manhattan distance, Cosine Similarity etc. Among these measurements, the proposed system is used cosine similarity measure [8] to compute the similarity between two documents in the experiments.

$$\cos(d1, d2) = \frac{(d1 \cdot d2)}{\|d1\| \|d2\|} \quad (2)$$

where · and || indicates dot product and length of vector respectively.

3.5 Particle Swarm Optimization

PSO algorithm and its concept of "Particle Swarm Optimization"(PSO) were introduced by James Kennedy and Russel Ebhart in 1995 [10]. 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 [11].

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 [12].The i^{th} particle is represented by a position denoted as $x_i = (x_{i1}, x_{i2}, . . . , x_{iD})$. 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.

$$v_{id}(t+1)=\omega.v_{id}(t) +c_1.rand().(p_{id}-x_{id}) +c_2.rand().(p_{gd}-x_{gd}) \quad (3)$$

$$x_{id}(t+1)=v_{id}(t+1)+x_{id}(t) \quad (4)$$

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:

$$x_i=(o_{i1},\ldots,o_{ij},\ldots,o_{iNc}) \quad (5)$$

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.

$$f = \frac{\sum_{i=1}^{N_c}\left\{\frac{\sum_{j=1}^{P_i} d(o_i, m_{ij})}{P_i}\right\}}{N_c} \quad (6)$$

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.

The PSO Clustering algorithm can be summarized as:

(1) Initially, each particle randomly selects k different document vectors from the document collection as the initial cluster centroid vectors.

(2) For t= 1 to t_{max} do
a)For each particle i do:
b)For each document vector m_p do
(i) Calculate the distance d (m_p,o_{ij}), to all cluster centroids C_{ij}
(ii) Assign each document vector to the closest centroid vector.
(iii) Calculate the fitness value based on equation (6).
c) Update the global best and local best positions
d) Update the cluster centroids using equations (3) and (4)

Where t_{max} is the maximum number of iterations.

After the similarity between two documents has been computed in a clustering analysis process, the PSO algorithm is applied to perform local and global optimal results of the proposed system. According to the analysis test results, the system can perform better intra cluster similarity and achieved an appropriate inter cluster similarity.

4. Experimental Results

In this paper, the dataset has collected from BBC, Channel_News_Asia, Oanda, Instaforex, Kitco, Goldprice.org and 24hgold.com. The terms in test pages consists of data from different domains such as literature, media, business etc. The system is tested with totally 60 pages of the above corresponding websites. These pages are tested with PSO algorithm only and Ontology+PSO algorithm.

According to the test results, 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 as1.47.

In this ongoing study, the proposed system has found that Intra-cluster similarities could be maximized, i.e. the distance between data vectors within a cluster could minimum; Inter-cluster similarity could be minimized, i.e. the distance between the centroids of the clusters could maximum.

The results of experimental testing of Ontology + PSO system performed on 60 different dataset are 82% of Intra-cluster similarity and 7% of Inter-cluster similarity. On the other hand, testing of only PSO system performed on same dataset got 65% of Intra-cluster similarity and 9% of Inter-cluster similarity.

5. Conclusion and Future Work

In this paper, the representation of documents based on domain Ontology and Particle Swarm Optimization (PSO) algorithm which is efficiently applied to web data is discussed. According to the analytical studies, it is found that the performance of the system is high in intra cluster similarity and an appropriate inter cluster similarity is found with the support of both Ontology and PSO. The research work is ongoing and the future work is to upgrade Ontology and to 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] 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

[3] 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.

[4] 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.

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

[7] D. Weiss, *"Descriptive Clustering as a Method for Exploring Text Collections,"* Ph.D Thesis.

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

[9] http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html

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

[11] 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

[12] 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

[13] 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

[14] J.Ghorpade and V.A.Metre, *"PSO based Multidimensional Data Clustering", International Journal of Computer Applications*, 87(16), pp.41-48, 2014

[15] A. A. A. Esmin, R.A. Coelho, S.Matwin, *"A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data", Springer*, pp.1-23, 2013

[16] D. P. Rini, S. M. Shamsuddin, S. S. Yuhaniz, *" Particle Swarm Optimization: Technique, System and Challenges"*, IJOCA, 2011, pp.19-27.

[17] S. Karol and V. Mangat, *"Survey On Particle Swarm Optimization Based Web Mining", IJIOME*, pp. 273-276, 2012

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

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

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

[21] 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

[22] 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

[23] 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

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

[25] 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