A Review on Clustering Web Data using PSO

Jayshree Ghorpade-Aher
Dept. of Computer Engineering
MITCOE
Pune, India

Roshan Bagdiya
Dept. of Computer Engineering
MITCOE
Pune, India

ABSTRACT
There is a tremendous proliferation in the amount of information available on the largest shared information source, the World Wide Web. 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. Data mining tasks require fast and accurate partitioning of huge unstructured data which may come with a variety of dimensions and attribute. In our paper we are focusing on the different clustering techniques, helpful for web data clustering. For such novel approach we perform a literature survey and depicted an evolutionary bio-inspired Swarm Intelligence algorithm called Particle Swarm Optimization (PSO) for optimized clustering result. In order to preprocess input data for improving the accuracy and optimize keyword searching, stop word removal and stemming methods are used. PSO algorithm will greatly improve the efficiency of web texts processing, and such evolutionary clustering techniques are used for web text data clustering.

General Terms
Data Mining, Soft Computing, Artificial Intelligence.

Keywords
Particle Swarm Optimization, Clustering, Evolutionary Algorithm, Web data

1. INTRODUCTION
The Internet continues to grow at a phenomenal rate and the amount of information on the web is overwhelming. This web data is growing exponentially and need to be handled properly. Thus, text mining and clustering the huge volume of data is the main challenge for web data. For handling such huge amount of data we are dealing with preprocessing technologies and perform clustering on this preprocessed data using PSO algorithm. Clustering is defined as grouping of similar type objects (particle) in the same group (called a cluster). Clustering is one of the main task of data mining, and a common technique for statistical data analysis, but before this data need to be preprocessed. Preprocessing is the term related with data mining domain, which serves to remove stop word and stemming the word suffixes for topic detection. To perform clustering we are focusing of soft computing domain based evolutionary clustering technology known as Particle Swarm Optimization (PSO). PSO is a bio-inspired swarm intelligence algorithm introduced by Kennedy and Eberhart in 1995 as a population-based stochastic search and optimization process [2]. 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) [32]. The goal is, therefore, to converge to the global optima of some multidimensional and possibly nonlinear function or system. Henceforth, PSO follows the same path of other evolutionary algorithms (EAs), such as genetic algorithm (GA) and Ant Colony Optimization algorithm (ACO). Figure 1 shows the review of various clustering algorithm [1].

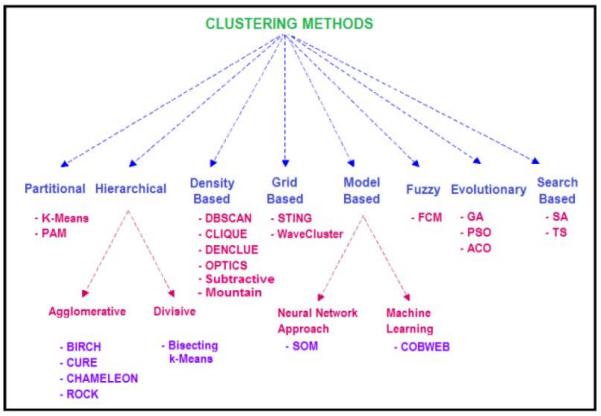

Fig 1: Clustering Methods

The various clustering approaches are [28].

- **Partitional Clustering** - Construct different partitions and then evaluate them based on given criterion.
- **Hierarchical Clustering** - depicts hierarchical decomposition of the set of data.
- **Density-based Clustering** - depend on connectivity and density functions.
- **Grid-based Clustering** – it is based on a multiple-level granularity structure.
- **Model-based Clustering** - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other.
- **Fuzzy Clustering** -based on fuzzy logic.
- **Evolutionary Clustering** - In artificial intelligence, an evolutionary algorithm (EA) is a subset of evolutionary computation, which posses' population-based metaheuristic optimization algorithms like ACO, GA and PSO [1]. Here the candidate solutions of individual among the population plays important role in optimization problem, and the fitness function determines the quality of the solutions. Generally Evolutionary algorithm gives better result as compared to other clustering techniques. A good clustering method always keen to produce high quality clusters having high intra-class similarity (cohesive within clusters) and low inter-class similarity (distinctive between clusters). PSO gives better result than GA, a comparison between GA, PSO and ACO has been shown in further discussion.

2. LITERATURE SURVEY

The following Table 1 shows the literature survey of various papers related to web data clustering techniques.

Table 1. Literature Survey

Sr. No	Publication /Author	Paper Title	Algorithm/ Tech.	Purpose	Datasets	PROS	CONS
1	IJCA, 2014 Jayshree Ghorpade, Vishakha Metre [1].	PSO based Multidimensional Data Clustering: A Survey	Subtractive Clustering with BRAPSO	Study of multi dimensional clustering techniques to achieve higher accuracy.	Vowel Iris CMC Cancer Glass	1.Improved convergence performance.	n/a
2	Springer Science + Business Media, 2013 Ahmed et al.[3].	A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data	PSO and its variant	Clustering high-dimensional data	n/a	1.Outstanding performance 2.Better cluster formation	Searching for global optima is still not sufficient
3	IJCSNS, 2007 Song Liangtu, Zhang Xiaoming [4].	Web Text Feature Extraction with Particle Swarm Optimization	VSM PSORTP Web text feature extraction algorithm	To search the multidimensional complex space efficiently and To improve the efficiency of web texts Processing.	HTML pages	1. Penalty function lead the particles to best region 2.Improve the efficiency of web texts processing	Web document filtering not possible
4	IEEE, 2013 Xing Huang, Qing Wu [6].	Micro-blog Commercial Word Extraction Based On Improved TF-IDF algorithm	TF-IDF KNN classificatio n algorithm	Fast and high accuracy extraction terms	Micro-blog in JSON object format.	1.Fast and high accuracy extraction terms, fast data processing	Only Json format support.
5	IEEE 2005 Stefan Janson, Martin Middendorf [7].	A Hierarchical Particle Swarm Optimizer and Its Adaptive Variant	H-PSO	H-pso on set of test functions (Sphere, Rosenbrock, Rastrigin, Ackley etc)	n/a	1.Uses dynamic hierarchy.	Branching degree affects solution
6	IEEE 2012 Mita K. Dalal et al. [8]	Automatic text classification of sports blog data	TF-IDF	Automatic Classification Of Sports Blog Data	sports blogs data	1.Improved accuracy	n/a
7	IJOCA 2011 Dian Palupi Rini, Siti Mariyam, Shamsuddin, Siti Sophiyati Yuhaniz [9].	Particle Swarm Optimization: Technique, System and Challenges	PSO and its variant	Study of PSO algorithm and literature survey	n/a	1.Improved clustering 2.Optimize performance	Convergenc e problem
8	IEEE 2013 Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle [10].	Clustering Heterogeneous Web Usage Data Using Hierarchical Particle Swarm Optimization	HPSO	hierarchical agglomerative manner clustering web usage data	Data from web server of Dept. of Computer Science, Univ. of Auckland	1.Similarity measure on web usage data.	Similarity measure on more diverse data and extend is not that efficient
9	IJIOME 2012 Stuti Karol, Veenu Mangat	Survey On Particle Swarm Optimization Based Web Mining	RVPSO PSO RTP Web usage Mining	High-dimensional Web Log File clustering	Web Log	1.Improve velocity of PSO Improve global optimum	n/a

	[11].						
10	JOC 2010 Mohammad Syafrullah,Naomie Salim [12].	Improving Term Extraction Using Particle Swarm Optimization Techniques	PSO Term Extraction Precision	to improve term extraction precision	Quran Reuters-21578 Gold Standard	1.Provide appropriate weight and best score for each term	Premature convergence problem May occur
11	IEEE 2009 Hongbo LI, Yunming Ye [13].	Improved Blog Clustering Through Automated Weighing of Text blocks	K-means type Langrangian method	Improved clustering	356 blog files from Windows live spaces	1.Better than k-means algorithm. 2.Work with real blog data	Link and time feature not experimented
12	IJCSE 2011 Mrs.G. Sudhamathy, Dr. C. Jothi [14].	Web Log Clustering Approaches – A Survey	PSO Fuzzy	Web Log Clustering	Web Log and Usage	1.PSO gives better result	n/a
13	IJORCS 2013 Mariam El-Tarabily [15].	A PSO-Based Subtractive Data Clustering Algorithm	Hybrid Subtractive Clustering PSO	To achieve fast and efficient clustering	Iris, Wine, Yeast	1.Dimension reduction approach is not required 2.Better result than basic PSO	Overlapping cluster membership problem.
14	IEEE 2007 Ziqiang Wang, Qingzhou Zhang, Dexian Zhang [16].	A PSO-Based Web Document Classification Algorithm	PSO	to investigate the capability of the PSO algorithm to web document classification	Reuters corpus TREC-AP	1.Better performance than other conventional algorithms	Convergence problem
15	IEEE 2005 Xiaohui Cui, Thomas E. Potok [17].	Document Clustering using PSO	K-means PSO HPSO ADDC	Document Clustering	TREC Dataset	1.PSO generate more compact clustering than K-means	Inefficient clustering huge vol. document.
16	IEEE 2008 Shouning Qu , Sujuan Wang, Yan Zou [18].	Improvement of Text Feature Selection Method based on TFIDF	TFIDF	Text classification	Intelligent Q&A Database from Univ. of Jinan	1. Available and doable method. 1.Removes traditional TF-IDF disadvantages	n/a
17	IEEE 2012 Tang Rui, Simon Fong, Xin-She Yang, Suash Deb [23].	Nature-inspired Clustering Algorithms for Web Intelligence Data	K-means C-PSO C-Firefly C-Cuckoo C-Bat C-WSA	To investigate nature-inspired optimization algo for performing clustering over Web Intelligence data	Page block IU(Internet usage) IP(Ipod ebay) SB(spam base)	1.Nature-inspired Clustering Algorithms 2.outperform existing algorithm	Not give best result in all cases only enhance individual parameter

3. PREPROCESSING

Preprocessing is defined as a technique processing raw data into proper format i.e. understandable format. Web data are often unstructured, incomplete, and inconsistent. Such issues can be resolved using preprocessing. There are many special techniques for pre-processing text documents to make them suitable for mining. Most of these techniques are from the field of 'Information Retrieval'. Some of the commonly used techniques are:

3.1 Stop Words

Many of the most frequently used words in English are worthless, which are term as 'stop words' in data mining [8]. In a document word count 'Stop words' account 20-30 % of total count and have a large number of counting hits. So there is need of removing stop word for getting reduced indexing file size and improve the efficiency as stop words are not useful in searching.

Example of stop words: a, and, the, is, at, which, on, about etc.

3.2 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-

(a) *Remove Ending:* elimination of suffixes like 'es' from 'goes' to 'go'.

(b) *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'.

After preprocessing we are applying PSO algorithm on web data for clustering purpose. A comparison of PSO algorithm with other to evolutionary clustering technique GA and ACO is mention in Table 2 [5] [8] [25] [26] [27].

Table 2 Comparison between GA, PSO and ACO

Sr.no	**Genetic algorithm (GA)**	**Particle Swarm Optimization (PSO)**	**Ant Colony Optimization (ACO)**
1	GA proposed by John Holland in 1962.	PSO was proposed by Eberhart and Kennedy in 1995.	ACO was proposed by Dorigo et al. in 1990.
2	Begins with a population of random chromosomes.	Social sharing of information among individuals of a population.	Inspiration from the interactive behavior of social species.
3	Three Operators 1. Selection 2. crossover 3. mutation	No Operator 1. Pbest or local best 2. global best 3. Velocity & Displacement	Multi agent 1. edge-based construction 2. edge weights
4	High computational cost	Low computational cost	Low computational cost
5	High memory requirement	Low memory requirement	Higher Search space for initial iteration and then decrease as computation goes on.
6	Population consists of solutions or chromosomes	Population consists of solutions or particles	Population consists of ants
7	More complex mechanism	Simple mechanism	Simple mechanism
8	More time to determine the results	Less time to determine the results	Better than GA in many cases

4. PSO ALGORITHM

Particle Swarm Optimization (PSO) uses the concept of social interaction for problem solving. It has been applied successfully to a wide variety of search and optimization problems. In PSO, a swarm of number of individuals communicates either directly or indirectly with each other in search directions.PSO is a simple but powerful search technique. Flow diagram for PSO algorithm is given in Fig 2 [3] [9][11].

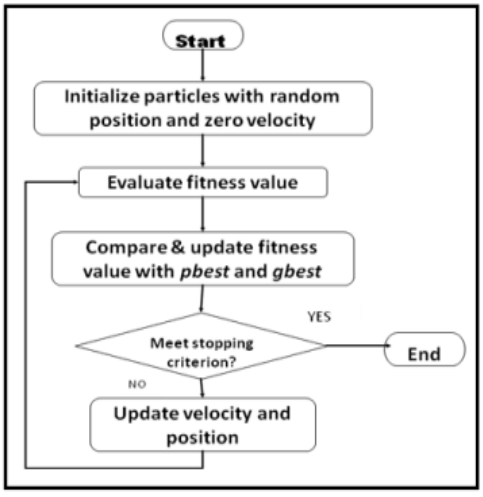

Fig 2: Flow of PSO algorithm

Initially all the preprocessed web data will act as particle in PSO algorithm and thus they are randomly initialized with random position and zero velocity. Each particle has its own fitness value and velocity to direct these particles. The fitness function is evaluated for each particle in the swarm and is compared to the fitness of the best previous position for that particle *pbest* and to the fitness of the global best particle among all particles in the swarm i.e. *gbest* [15]. To produce optimized clusters, a continuous updation in fitness function is seen based on number of iterations and termination criteria. Updation in fitness value updates velocity as well as position equation shown in equation 1 and equation 2 respectively. Meet Stopping criterion mention about the termination states if optimize clusters are formed or number of iteration are over system will move to end state. If stopping criteria is not meet, update velocity and position and loop back to fitness function evaluation for further optimization. Even though PSO seems to be an optimized technique, original PSO approach has limitation while optimizing the solution using global best there is chance to trap in local area.

Velocity is defined as rate of change of the position of a particle in a given search space, specifying its speed and direction of motion. This velocity need to be continuously optimized as per the number of iterations. Thus we are denoting equation for update velocity.

Velocity update equation [1] [9].

$$V_i^{t+1} = \underbrace{w.V_i^t}_{\text{Inertia}} + \underbrace{c_1.r_1(P_{best} - X_i^t)}_{\text{Cognitive Component}} + \underbrace{c_2.r_{12}(G_{best} - X_i^t)}_{\text{Social Component}} \quad (1)$$

Position Update equation [1][9].

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (2)$$

Where, parameters for PSO velocity and position updation equation are:

$V_i(t)$ - Velocity of the ith particle.
P_{best} - Personal best position of the ith particle.
G_{best} - Global best position of the particles.
$X_i(t)$ - Current position of the ith particle.
c_1 & c_2 - Acceleration Constants.
r_1 & r_{12} - Random function in the range [0, 1].
w - Inertia weight.

Each particle has its inertia or momentum which serves as a memory of the previous particle direction, preventing the particle from drastically changing direction.

Cognitive Component model defines the tendency of particles to return to its previously found best positions as each respective particle keeps an account of their earlier optimal location [26].

Social Component gives the best performance of a particle relative to its neighbors. It updates the particles velocity and position by considering the group as a whole keeping group to attain globally best result relative to their neighbor [26].

While considering velocity, Velocity Clamping is an important parameter in PSO, it clamps particles velocities on each dimension as it determines 'fineness' within which regions are searched-

(a) If velocity is too high, can move to past optimal solutions.

(b) If velocity is too low, can get stuck in local minima.

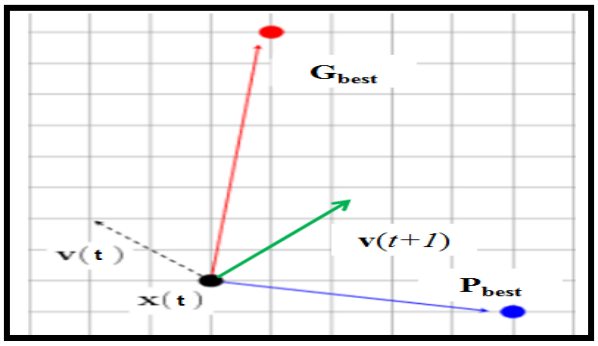

Fig 3: Motion Of Particle

In the above graph Fig 3 represents the direction of moving particle in 2D search space after the update of velocity and position. v(t+1)represent the updated velocity at t+1 iterations.

Kennedy identifies 4 types of PSO based on 'c1 & c2' acceleration constant values [29].

- Full Model (c1>0, and c2> 0)
- Cognition Only (c1 > 0 and c2 = 0)
- Social Only (c1 = 0 and c2 > 0)
- Selfless (c1 = 0, c2 > 0, and g ≠ i)

5. APPLICATION

Web data clustering has wide domain of applicability. Some of the applications where PSO clustering method can be efficiently utilized such as in ranking algorithms, search engine company, resemblance system, market analysis and in processing unstructured data.

6. CONCLUSION

In this paper we discussed about the evolutionary computational Particle Swarm Optimization algorithm which is efficiently applied to web data clustering and it also provides a comparison between three evolutionary algorithms GA, ACO and PSO. A literature survey related to web data clustering, aims to serve as a source of ideas for researchers working on exponentially growing internet data. Preprocessing techniques such as stop words removal, stemming is justified for processing raw data. On such preprocessed data a novel approach of PSO clustering has been discussed to obtain optimize result. The future development can be the implementation of variant of PSO to avoid convergence and local optima problem.

7. REFERENCES

[1] Jayshree Ghorpade and Vishakha Arun Metre. Article: PSO based Multidimensional Data Clustering: A Survey. International Journal of Computer Applications 87(16), 2014, pp.41-48.

[2] 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.

[3] Ahmed A. A. Esmin, Rodrigo A. Coelho, Stan Matwin, "A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data", Springer, 2013, pp.1-23.

[4] Song Liangtu, Zhang Xiaoming, "Web Text Feature Extraction with Particle Swarm Optimization", IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.6, June 2007, pp.132-136.

[5] Rania Hassan, Babak Cohanim, Olivier de Weck, "A comparison of PSO and GA", American Institute of Aeronautics and Astronautics,2004.

[6] Xing Huang, Qing Wu, "Micro-blog Commercial Word Extraction Based On Improved TF-IDF Algorithm", IEEE, 2013, pp.1-5.

[7] Stefan Janson and Martin Middendorf , "A Hierarchical Particle Swarm Optimizer and Its Adaptive Variant", Ieee Transactions On Systems, Man, And Cybernetics, Dec 2005, pp.1272-1282.

[8] Mita K. Dalal, Mukesh A. Zaveri, "Automatic text classification of sports blog data", IEEE, 2012, pp.219-222.

[9] Dian Palupi Rini, Siti Mariyam Shamsuddin, Siti Sophiyati Yuhaniz," Particle Swarm Optimization: Technique, System and Challenges", IJOCA, 2011, pp.19-27.

[10] Shafiq Alam,Gillian Dobbie, Yun Sing Koh, Patricia Riddle, "Clustering Heterogeneous Web Usage Data

Using Hierarchical Particle Swarm Optimization", IEEE, 2013, pp. 147-154.

[11] Stuti Karol and Veenu Mangat, "Survey On Particle Swarm Optimization Based Web Mining", IJIOME, 2012, pp. 273-276.

[12] Mohammad Syafrullah and Naomie Salim, "Improving Term Extraction Using ParticleSwarm Optimization Techniques", JOC , 2010, pp. 116-120.

[13] Hongbo LI Yunming Ye, "Improved Blog Clustering Through Automated Weighing of Text blocks", IEEE, 2009, pp. 1586-1591.

[14] Mrs. G. Sudhamathy, Dr. C. Jothi Venkateswaran, "Web Log Clustering Approaches – A Survey", IJCSE 2011, pp. 2896-2903.

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

[16] Ziqiang Wang, Qingzhou Zhang, Dexian Zhang, "A PSO-Based Web Document Classification Algorithm", IEEE, 2007, pp. 659-664.

[17] Xiaohui Cui, Thomas E. Potok, "Document Clustering using PSO", IEEE, 2005, pp. 185-191.

[18] Shouning Qu ,Sujuan Wang,Yan Zou, "Improvement of Text Feature Selection Method based on TFIDF", IEEE, 2008, pp.79-81.

[19] Huo Ling Yu1, Liu Bingwu, Yan Fang, "Similarity Computation of Web Pages of Focused Crawler" , International Forum on Information Technology and Applications, 2010, pp 499-505

[20] Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle, "Web Bots Detection Using Particle Swarm Optimization Based Clustering", IEEE, 2014, pp 2955-2962.

[21] Tien-Chi Huang, Shu-Chen Cheng, Yueh-Min Huang, "A blog article recommendation generating mechanism using an SBACPSO algorithm" , Expert Systems with Applications 36,2009, pp 10388–10396.

[22] Ching-Yi Cheo, Fun Ye, "Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis" , IEEE, 2004, pp 789-794.

[23] Tang Rui, Simon Fong, Xin-She Yang, Suash Deb, "Nature-inspired Clustering Algorithms for Web Intelligence Data", IEEE, 2012, pp.147-153.

[24] Wiak, Sławomir, Andrzej Krawczyk, and Ivo Dolezel, " Intelligent Computer Techniques In Applied Electromagnetics" vol 119, 2008, pp.1-291.

[25] F. Moussouni et al.: Comparison of Two Multi-Agent Algorithms: ACO and PSO for the Optimization of a Brushless DC Wheel Motor, Studies in Computational Intelligence (SCI) 119, 2008, pp.3–10.

[26] Marco A. Montes de Oca, Thomas Stützle, Mauro Birattari and Marco Dorigo, "Frankenstein's PSO: A Composite Particle Swarm Optimization Algorithm", IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 13, NO. 5, OCTOBER 2009,pp 1-30.

[27] Emad Elbeltagi, Tarek Hegazy, Donald Grierson, Comparison among five evolutionary-based optimization algorithms, Advanced Engineering Informatics, Volume 19, Issue 1, January 2005, pp 43-53.

[28] [JIA 06] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", published by Morgan Kauffman, 2nd Ed, 2006.

[29] [SER] Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj, "Multidimensional Particle Swarm Optimization For Machine Learning And Pattern Recognition", Springer Adaptation, Learning, And Optimization Volume 15.

[30] [MAU 06] Maurice Clerc, "Particle Swarm Optimization", © ISTE Ltd, 2006.

[31] [PSOL] http://www.particleswarm.info

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