Enhance Clustering Approach using PSO-A* for E-Commerce

Pankaj Pandey
SIRTE
Bhopal

Sneha Soni
SIRTE
Bhopal

ABSTRACT
In the current trends most of the people uses online searching and purchasing the items. So web data size is drastically increases and after some time not possible to put all data in memory. So there is a need to manage data access speed and also increase the speed of searching. In the existing system most of clustering algorithm uses only basic criteria for machine learning and data mining algorithms. Initially it is not able to select the good center point so it will do to increase the number of iteration. Second problem is to access of full database and third problem is needed to increase searching speed. The proposed approach uses PSO-A* algorithm for clustering and searching that is used for select the good center point initial for clustering. After that only clustered data is access not full database that make partial database access to manage memory management and speed. Finally third approach is uses for increase searching speed. Now the proposed approach is to perform well and improve the e-commerce website performance.

Keywords
Data Mining, Clustering, PSO, A*, PSO-A*

1. INTRODUCTION
Data Mining is the process of extracting hidden and interesting patterns or characteristics from very large datasets and using it in decision making and prediction of future behavior. This increases the need for efficient and effective analysis methods to make use of this information. In this one of these tasks is clustering.

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Here data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities. Here data is mined to find anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

PSO is a robust stochastic optimization technique based on the movement and intelligence of swarms. It uses a number of agents (particles) that constitute a swarm moving around in the search space looking for the best solution. Each particle is treated as a point in a N-dimensional space which adjusts its "flying" according to its own flying experience as well as the flying experience of other particles. This value is called personal best, pbest. Another best value that is tracked by the PSO is the best value obtained so far by any particle in the neighborhood of that particle. This value is called gbest.

In A* path planning algorithm the cost evaluation function uses an 'exact + heuristic' cost, which is given by the equation

$$f(n) = g(n) + h(n),$$

where, g(n) is the exact cost from start node s to current node n, and h(n) is the heuristic cost from the current node n to target node.

Best First search has f(n)=h(n)

Uniform Cost search has f(n)=g(n)

The Euclidean distance between the points p(x1,y1) and q(x2,y2) is calculated as Manhattan Distance. The Manhattan distance between the same points is calculated as

$$d(p,q) = |x1 - x2| + |y1 - y2|.$$

2. LITERATURE REVIEW
Rashmi P. Dagde (2017).The Bisecting K-mean algorithm has some drawback like it will not find the centroid for these the clustering not found proper manner and to remove this drawback used the PSO algorithm. The particle swarm optimization algorithm is removing the drawback of the clustering. It will find the optimal path and hybrid model increase the accuracy of the clustering.

Diego Vallejo-Huanga (2017). It has a aims to propose modifications in traditional clustering algorithms to incorporate size constraints in each cluster.

Bilal Sowan and Hazem Qattous (2017). It proposed approach is a flexible data mining approach that employs variety techniques. The flexibility means that a dependent variable of a numeric data type in a dataset is not only considered for a regression task.

Liwen Peng (2018). It propose a multilabel feature selection algorithm as a preprocessing stage before Multilabel Classification (MLC). It combines feature selection with an overlapping clustering algorithm.

Dr. S.K. Jayanthi (2018). The proposed numerical statistic approach called TF-IDF has been proosed to determine the relevance of word to a document corpus.

3. PROBLEM IDENTIFICATION
Clustering is one of the most important factors in the current research for measure the similarity to determine how close two patterns are to one another. It groups data vectors into a predefined number of clusters, based on Euclidean distance as similarity measure.

It is having the following problems –

1) Clustering algorithm initially does not select good center point for clustering the items. It will execute maximum number of iterations and does not produce good clustering. So there is need to choose good approach to select better center point for clustering.

2) Most of the algorithm access full database so there is need for partial database access by good clustering approach.

3) For searching the associated items it will take more time so there is need to select good approach to search the associated items, which consume less memory.

4. PROPOSED SOLUTION

In the proposed algorithm initial good centroids are selected by PSO based on average distance. After that apply modified A* approach for reduce the execution cost. So now clustered items is used to access partially not full database. At last choose A* algorithm for finding the associated items by using Open-List and Close-List in short span of time.

The following proposed solution of the problem -

1) First select PSO algorithm which is used to find the good center point for clustering the particle position that results in the best evaluation of a given fitness objective function.

2) This clustering approach is used to reduce the cost of execution time for searching the items. It is beneficial for partial database access of items.

3) Next process apply the A* algorithm for searching the associates items in less execution time and memory with accuracy. It maintains two lists Open List and Close List. At the initial all the particles are store in Open List. After removing the obstacles, the remaining particles (selected particles) are storing the Close List. The Close List is used to find many paths from the original source to original destination. So select the optimal path based on the total distance covered by all paths.

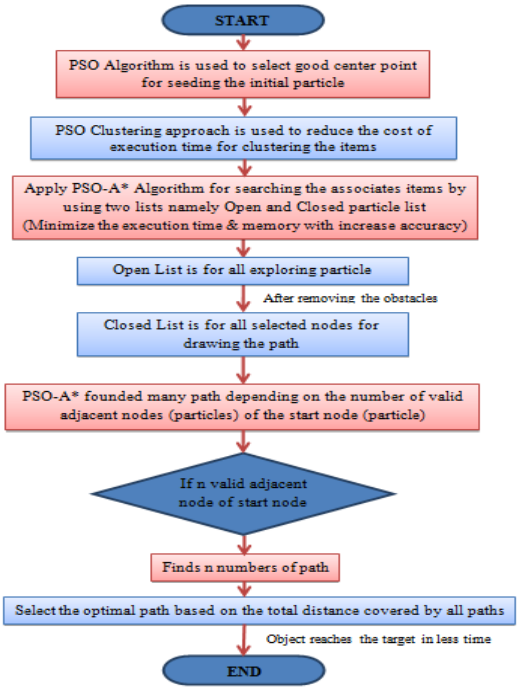

Fig. 4.1: Flow Chart of proposed PSO-A* Algorithm

Fig 4.1 is displaying the flow chart of proposed PSO-A* algorithm where it first taking the number of records after that it cluster the items for partial database access. Next step it maintain open list and close list for search the next targeted item from source item to target item in the form of node with minimum execution of time.

5. IMPLEMENTATION

The application is having three pages. Two main pages are PSO page and PSO-A* Page. The third page is PSO help page which describe the PSO algorithm in the form of text. The default home page of the proposed application is having some menus.

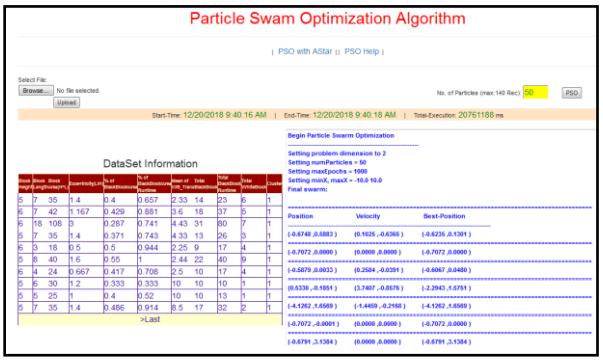

Fig. 5.1: Execution of PSO Page with 50 Records using Page-Blocks Dataset

The application is having PSO help page which describe theoretically with the formula. It is displaying the details of PSO algorithm step by step for users.

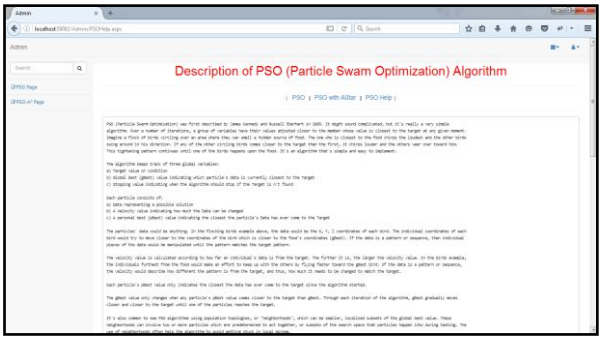

Fig. 5.2: PSO Help Page

The PSO-A* page uses number of records with the start node, end node and obstacle node. So it is used to search the shortest path from starting node to target node with minimum time of execution.

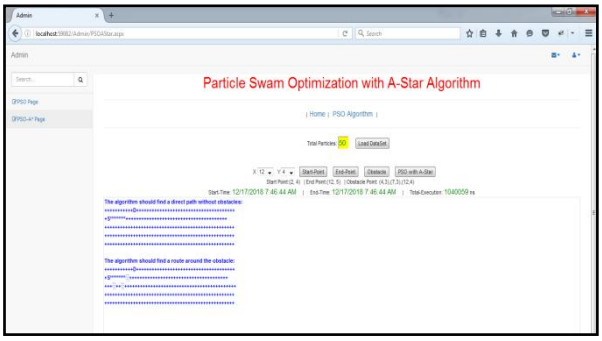

Fig. 5.3: Execution of PSO-A* Page with 50 Records using Page-Blocks Dataset

6. RESULT ANALYSIS

6.1 Execution Time of PSO with different Size using Page-Blocks Dataset

In the experiment execution it is having various record sizes just like 10, 20, 50, 100 and 150. Here the number of record size items is appearing for this experiment. First every particle or item search its current position and find the personal best position using updating of velocity. After that it collect the one best position for every item and by using this best position every item is compare with this position and finally also set the it's global best position in respect to all items.

The following Table 6.1 is displaying the Total Execution Time (in ms) for the PSO algorithm with different record sizes using Page-Blocks Dataset.

Table 6.1: Total Execution Time (in ms) for PSO Algorithm with Different Record Size using Page-Blocks Dataset

Record Size	PSO Algorithm (in ms)
10	8900509
20	10080577
50	19121094
100	40702329
150	61753533

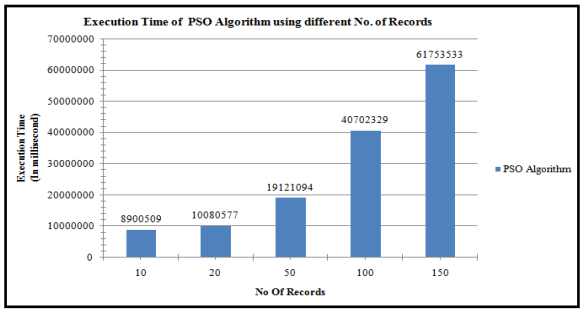

Fig. 6.1: Execution Time (in ms) for PSO algorithm with different Record Size using Page-Blocks Dataset

The Figure 6.1 is displaying the execution time of PSO algorithm using different record size using Page-Blocks Dataset. The execution time of with different record size 10, 20, 50, 100 and 150 are 8900509, 10080577, 19121094, 40702329, and 61753533.

6.2 Execution Time of PSO-A* with different Size using Page-Blocks Dataset

In the experiment execution it is having various record sizes just like 10, 20, 50, 100 and 150. Here the number of record size items is appearing for this experiment. First every particle or item search its current position and find the personal best position using updating of velocity. After that it collect the one best position for every item and by using this best position every item is compare with this position and finally also set the it's global best position in respect to all items.

For the searching the path from source item to target item there is need to select number of items with the start point, obstacle and end point. So it will find the best shortest path from source item to target item.

All the items excepted obstacles are collect in open list. After that each item pick up from open list and search the next node for movement. It will find the next node using the best short path but without using the obstacle.

The following Table 6.2 is displaying the Total Execution Time (in ms) for the PSO-A* algorithm with different record sizes using Page-Blocks Dataset.

Table 6.2: Total Execution Time (in ms) for PSO-A* Algorithm with Different Record Size using Page-Blocks Dataset

Record Size	PSO-A* Algorithm (in ms)
10	3590205
20	8180468
50	17143273
100	37024342
150	58193456

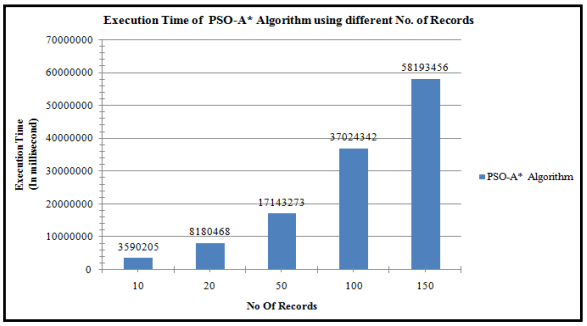

Fig. 6.2: Execution Time (in ms) for PSO-A* algorithm with different Record Size using Page-Blocks Dataset

The Figure 6.2 is displaying the execution time of PSO-A* algorithm using different record size using Page-Blocks Dataset. The execution time of with different record size 10, 20, 50, 100 and 150 are 3590205, 8180468, 17143273, 37024342, and 58193456.

6.3 Total Execution Time for PSO and PSO-A* Algorithm using different Records Sizes using Page-Blocks Dataset

The following Table-6.3 is displaying the comparison between PSO and the proposed PSO-A* Algorithm with number of records using Page-Blocks Dataset. Now, it is utilize different record sizes just like 10, 20, 50, 100, and 150 in the experiment of execution.

Table-6.3: Execution Time of PSO and PSO-A* Algorithm with Different Number of Records using Page-Blocks Dataset

Record Size	PSO Algorithm (in ms)	Execution Time (In ms) PSO-A* Algorithm	Percentage Improvement in Execution Time (in ms)
10	8900509	3590205	1.47 %
20	10080577	8180468	0.23 %
50	19121094	17143273	0.11 %
100	40702329	37024342	0.09 %
150	61753533	58193456	0.06 %

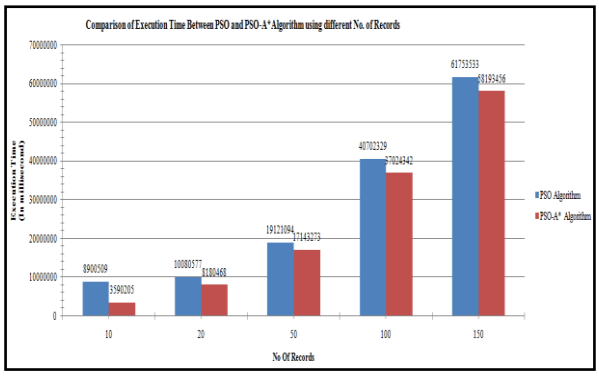

Fig. 6.3: Execution Time of PSO and PSO-A* Algorithm with Different Number of Records using Page-Blocks Dataset

The Figure 6.3 is displaying the total execution time (in ms) of PSO and the proposed PSO-A* by using Page-Blocks Dataset with using different record sizes (10, 20, 50, 100 and 150). If number of records are 10, 20, 50, 100 and 150 then PSO and PSO-A* is having Execution time (in ms) are 8900509 ms and 3590205 ms, 10080577 ms and 8180468 ms, 19121094 ms and 17143273 ms, 40702329 ms and 37024342 ms, 61753533 and 58193456 ms.

The performance of the proposed PSO-A* algorithm is performed well as compared with PSO over using various record sizes (eg. 10, 20, 50, 100 and 150) using Page-Blocks Dataset. In the proposed research approach is utilize the number of sequential traversal patterns for e-Commerce website for maintaining the quality of frequent pattern. It is always support to find the latest trends in the market. It is also safe the memory utilization using cluster or partitioned data access.

7. CONCLUSION

The proposed approach is maintaining the clustering and searching the path from one item to other item using partial database access. In these experiments variation is taken as record size. So it may vary record sizes then this approach perform better improvement in execution time is 1.47% when record size is 10. It is also displaying lest improvement in the execution time 0.06% when record size is 150. In the future work may include the investigation on different swarm intelligence algorithm such as, Ant Colony optimization (ACO), Artificial bee colony (ABC), Firefly algorithm (FA) etc. also the Genetic algorithm.

8. REFERENCES

[1] Hereford, J.M., Siebold, M., Nichols, S.: Using the Particle Swarm Optimization Algorithm for Robotic Search Applications. In: Proceedings of the 2007 IEEE Swarm Intelligence Symposium, SIS (2007).

[2] Li Lu, Dunwei Gond, "Robot Path Planning in Unknown Environments Using Particle Swarm Optimization", Fourth International Conference on Natural Computation, IEEE, pp. 422-426, 2008.

[3] Amin Zargar Nasrollahy, Hamid Haj Seyyed Javadi, "Using Particle Swarm Optimization for Robot Path Planning in Dynamic Environments with Moving Obstacles and Target", Third UK Sim European Symposium on Computer Modeling and Simulation, IEEE, pp. 60-65, 2009.

[4] M. Lotfi Shahreza and M.R. Delavar, "Anamaly detection using a self-organizing map and particle swarm optimization", Scientia Iranica, pp. 1460-1468, 2011.

[5] Roham Shakiba, Mohd. Reza Najafipour, "An improved PSO-based path planning algorithm for humanoid soccer playing robots", IEEE, 2013.

[6] Narendra Singh Pal, Sanjay Sharma, "Robot Path Planning using Swarm Intelligence: A Survey", International Journal of Computer Applications, Vol 83, No. 12, pp. 0975-8887, Dec. 2013.

[7] Manisha Sajwan, Kritika Acharya, Sanjay Bhargava, "Swarm Intelligence Based Optimization for Web Usage Mining in Recommender System", International Journal of Computer Applications Technology and Research, Vol. 3, Issue 2, pp. 119-124, 2014.

[8] Rashmi P. Dagde, Snehlata Dongre, "A Review on Clustering Analysis based on Optimization Algorithm for Datamining", International Journal of Computer Science and Network, Volume 6, Issue 1, pp. 36-41, 2017

[9] Diego Vallejo-Hanga, Paulina Morillo, and Cesar Ferri, "Semi Supervised Clustering Algorithms for Grouping Scientific Articles", ELSEVIER, pp. 325-334, 2017.

[10] Bilal Sowan and Hazem Qattous, "A Data Mining of Supervised learning Approach based on K-means Clustering", International Journal of Computer Science and Network Security, Vol. 17, No.1, pp. 18-24, Jan 2017.

[11] Pasi Franti, "Efficiency of Random Swap Clustering", Journal of Big Data, Springer, Vol. 5, Issue 13, pp. 2-29, 2018.

[12] Liwen Peng, Yongguo Liu, "Feature Selection and Overlapping Clustering-Based Multilabel Classification Model", Hindawi, pp. 1-13, 2018.

[13] Dr. S.K. Jayanthi, C. Kavi Priya, "Clustering Approach for Classification of Research Articles based on Keyword Search", IJARCET, Vol. 7, Issue 1, pp. 86-90, 2018.