Mining Web Data using PSO Algorithm

Ammulu K.
Research Scholar
Rayalaseema University, Kurnool

Venugopal T.
Associate Professor & Head
JNTUH College of Engineering, Sultanpur, Medak, Telangana

Abstract

Web is the fundamental source for the generation of information or data in tremendous amount. However, each and individual site classify their own data but fetching the classified data from the multiple website is not possible. Clustering the web data is the main challenge in the web data mining where an efficient approach is needed to cluster it. In the proposed system, multiple webpage are fetched by web crawling technique then the data are extracted, classified using the PSO algorithm. The fitness value gives good classification result and provides a novel searching technique. The experimental setup is carried out in java language and the accuracy of this approach is 80%.

Keywords: Crawler; PSO algorithm; mining; classification; webpage; hyperlink; website

I. INTRODUCTION

Nearly 90% web data is in unstructured formats available in the web which needs to be structured in order to utilize it efficiently [1]. Web Crawlers plays a vital role in the search engine technique. It is the fundamental approach for gathering the information from the Internet where the information growth is rapid. A web crawler is the process built by a software program which automatically traverses the websites by retrieving the content by following the link from page to page. The Focused web crawler is one type of the web crawling technique which is used to retrieve the document by fetching the hyperlink by following the home link. The main advantage of this approach is cost-effective in hardware resources, better search technique and reduces the amount of network traffic while downloading [6].

Web mining is similar to the data mining technique, in data mining the data are retrieved from the database whereas in web mining the data from the web pages or documents is discovered. The web mining is classified into three types they are web content mining, web structure mining, web usage mining. The process of extracting the data from the web into structured form, index the data it results to fast retrieval. Mainly it focuses on the structure of the inner documents which contains text, images, video, audio, and structured records such as tables and lists. Web structure mining is the process of extracting the hyperlink of the web document or pages. The objective of this process is to generate the complete structure of the websites. This is performed at both hyperlink level and document level. Web usage mining is used to extract the helpful data and navigation patterns from the web present in the server logs, agent logs, referrers log, client-side cookies, meta-data and user profiles [2].

Web content mining is the process of extracting, mining, integration of needed information or data from the web pages which is similar to the data mining and text mining. Web data are mainly semi-structured and unstructured whereas the data mining is structured and the text mining is unstructured. The approaches in the web content mining are unstructured data mining techniques, structured data mining techniques, semi-structured data mining techniques, multimedia data mining techniques. The web content mining tools are rapid miner, screen scaper, automation anywhere, web info extractor, mozenda, web content extractor.

However, there are several issues and challenges arises during the web content mining such as peculiar kind of data extraction, web information integration and schema matching, opinion extraction from online sources, knowledge synthesis, segmenting web pages and detecting noise. Two main issues are tried to sort out in this are as follows:

Extraction of Data/Information: Usually the content in the web pages are in structured format which means the information appears in the frontend is arranged using the tags. So the extraction of the data from the web page is crucial task. This needs machine learning algorithm to solve this issue.

Segmenting web pages and detecting noise: Each and every page contains numerous data including advertisements, image, navigation links, copyright notices. Extracting the main content from the web page is difficult task.

The process of mining, extracting and integration of useful data, information, and knowledge from the web content is known as the web content mining. The web content mining is generally carried out after the completion of the crawling of web pages [3]. Web content mining is referred as the text mining where the scanning and mining of the text, pictures and graphs of a web page. In addition to that customer reviews and forum postings to discover consumer sentiments. There are two types of web content mining, they are agent based approach and database approach. The Agent based approach is further divided into intelligent search agents, information filtering/categorizing agent, personalized web agents [4]. The process of searching the information based on the query from the user query and domain behaviours. The preprocessing step is carried out in each intelligent agent by utilizing number of approaches. The personalized web agents obtain knowledge from the user activities and then extract the files related to their user profile history. The database approach consists of database framework that is structured by attributes, domains and schemas.

All rights reserved by *www.ijirst.org*

The main objectives of this work are as follow:

- Designing a crawler for deep search which fetch huge websites.
- Extract the content from the websites related to the input keyword provided.
- Classify the data using the particle swarm optimization algorithm.

This paper contains related work at section II, proposed work at section III, Section IV shows the experimental results, Section V represent the Conclusion part.

II. RELATED WORK

In paper [5], Nisha Pawar et al., proposed an approach which is designed to search the web pages by utilizing the web crawler for ayurvedic medicinal domain system. The initial query is preprocessed and given as the input to the crawler. These related documents are retrieved from the frontier to classify the web document by utilizing the naive bayes classification algorithm. The features are extracted based on title text, meta-description, anchor text, URL tokens. Dataset is the Indian Ayurvedic medical plant. The Naive Bayes Classification is used to classify the web pages. According to their experimental result the classification accuracy is 90%.

In paper [7], Dipali Kharche and Anuradha Thakare proposed a hybrid algorithm by combining the ant colony and PSO algorithm. The initial centroids value is obtained from the ant colony system, and then the PSO algorithm is applied to search the optimal cluster from the fitness value obtained from the XB index, Sym index, DB index, Connected DB index, Connected Dunn index, and Mean square distance. The input dataset is the iris dataset and the performance measures are F-measure, purity, entropy, rand, jaccord.

In paper [8], Alexandre Szabo and Leandro Nunes de Castro proposed a innovative concept in particle swarm algorithms which is specially designed to solve the issues during classification. The PSO is modified into two phase as: PSClass(Particle Swarm Classifier) and cPSClass(Constructive Particle Swarm Classifier). The PSClass search for the available groups in the database in unsupervised approach by adjusting the prototypes position using the LVQ1 method, in supervised approach is used to minimize the misclassification error. The cPSClass follows the PSClass approach, in unsupervised approach the particles are found dynamically.

In paper[9], Sotiris Batsakis et al., proposed a novel crawler approach which is inspired from the Hidden Markov Model crawler. This crawler also has the same baseline implementation where the priority assignment changes for each crawler. Classic focused crawlers combining page content and link the anchor text and semantics or training sets of crawler. In paper [10], focused crawler is classified into five categories: priority base crawler, structured base crawler, Context base crawler, Leaning base crawler and other focused crawler. This approach provides the search spamming and ranking function. The precision and recall are improved in this process by comparing it with the existing approach. This took less time, money and effort for processing.

In paper[11], Girma S. Tewolde and Darrin M. Hanna proposed the PSO method for the single and multisurface data separation. The input data is breast cancer database collected from the UCI machine learning repository. In single separating surface system, the PSO is used to develop an optimized hyperplane which is used to divide the dataset into two classes. The initial fitness value is randomly selected to separate the dataset by assigning the attribute in the equation. Iteration continues by changing the fitness value. The linear programming package is used for multiple separating hyperplanes. Classification done by several stage with the paired parallel hyperplanes.

In paper[12], the particle swarm optimization technique is modified as evolutionary particle swarm optimization based clustering. The parameters in EPSO are particle id, particle current position, distance vector, associated data vectors, pBest position. Generally particles are initialized in the first generation and after each generation the swarm gets stronger by adapting the weaker ones. The strong particle generation is obtained by number of generations, number of iterations or minimum number of data vectors in a cluster.

III. PROPOSED SYSTEM

Classify the product

- Crawl the website by getting the input(Link)
- Traverse through the multiple website to collect the data(input keywords)
- Classify the products(based on products, brands, rate,)

Provide efficient output for searching technique

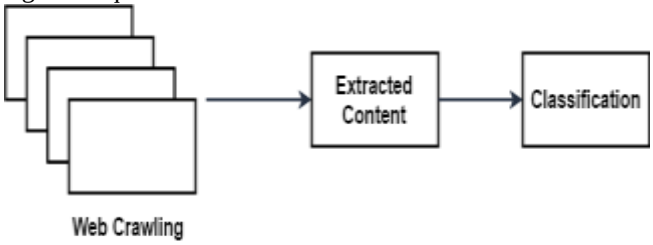


Fig. 1: Represent the entire system structure

All rights reserved by *www.ijirst.org*

A. Web Crawling

A standard crawler crawls through all the pages using the breadth first strategy. The Focused crawler crawls through the domain specific pages. The pages which are not related to the particular domain are not considered. The focused crawler tries to retrieve the web pages relevant to the input query. The relevancy factor is obtained by assigning the weightage to the keyword. The web pages which are not having the weightage will be removed from the queue. The input for the crawler is starting URL and topic description which includes the description as list of keywords.

B. Data Preprocessing

The content of the web pages contains number of useless information such as tags, advertisements, grammatical words and so on, but these maximize the difficulty in retrieving the main content.
Tags:
<a>, <script>, <noscript>, <style>, <meta>, <!—->, <param>, <button>, <select>, <optgroup>, <option>, <label>, <textarea>, <fieldset>, <legend, <input>, <image>, <map>, <area>, <form>, <iframe>, <embed>, <object>

Generally, the data preprocessing includes the data cleaning, data integration, data transformation, and data reduction. Data Cleaning is the initial step in the preprocessing of web content. The href specifies the links where the unvalued links are removed. The content with the extension such as jpeg, jpg, gif, png, tif, bmp, mp3, css, js, swf, ico, cgi are removed. The pages having the error code 400, 403, 404, 407, 500, 501, 502, 503, 504 are removed where it won't have any web content in it.

Table - 1
Description of Extension and Error Code

File Type/ Error Code	*Description*
.jpeg, .jpg, .gif, .png, .tif, .bmp	*Image file*
.css	*Cascading style sheet*
.swf	*Flash animation file*
.cgi	*Common gateway interface*
.mp3	*Audio file*
.js	*Java script file*
.ico	*Icon Image File Format*
400	*Bad Request*
403	*Forbidden*
404	*Not Found*
407	*Proxy Authentication required*
500	*Internal Server Error*
501	*Not Implemented*
502	*Bad Gateway*
503	*Server Unavailable*
504	*Gateway timeout*

Data Integration is the process of storing the data extracted from different server which includes the content of the web pages. Data transformation is the process of arranging the data in a unique format for further processing. Data reduction is the process of selecting the exact attributes from the information collected so far.

C. PSO Algorithm

PSO is an evolutionary algorithm inspired from the flocks of birds or schools of fish in coordinated motion. In PSO, individuals are called particles and the population is called a swarm. Each and every particle search for the best point and this is based on the particle movement and intelligence. Thus, each particle motion is to find the particle current location (lbest), particle best location (pbest), sum of best location (gbest). The current location of the particle is estimated by the fitness function which is obtained from the fitness value.
Steps:

1) Find the Objective(target to be achieved),
2) Let as Assume the Fitness value as 1 by Objective
3) Initialize Velocity and number of Iteration
   a) For each iteration calculate the local best from the population.
   b) Compare the local best with the previous local best to update the current lbest and velocity
4) Recalculate the Globalbest.
5) Compare with the fitness if reached stop the iteration
6) Else continue to the next step

PSO provides a valuable high level data points for the initial selection for further classification. Particles or potential solutions are represented having a position and rate of the change in d-dimensional space. In PSO, a number of solutions are encoded as a swarm of particles in search space. The initial values of a particle are randomly chosen. Each particle maintains a record of its best achieved since the beginning of the iteration. Also each particle has a defined neighborhood. Particles make decision based on the performance of its neighbor and itself.

All rights reserved by *www.ijirst.org*

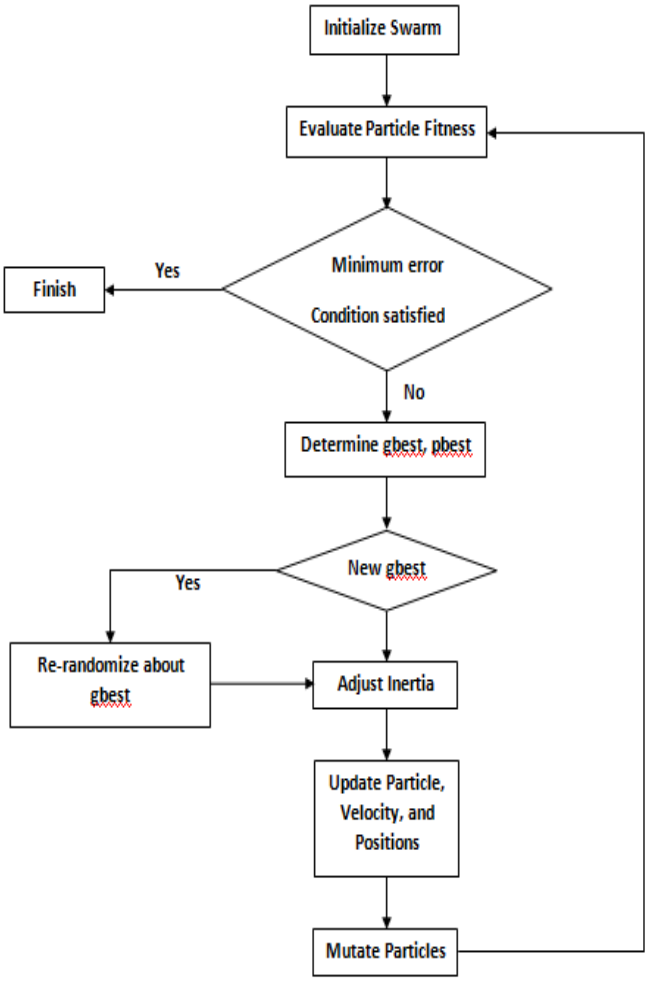


Fig. 2: The flow chart of the PSO algorithm

IV. EXPERIMENTAL RESULT

The online web pages are collected using the web crawling technique with the input seed URL and the searching keyword to fetch the content of the website. The extracted content is classified using the PSO technique. The extracted content includes attributes such as product id, product name, brand name, item description (product specification), category, quantity, price, payment method, rating, and hyperlink. The experimental process is carried out in java language, netbeans tool. The entire process is carried out three times to improve the performance.

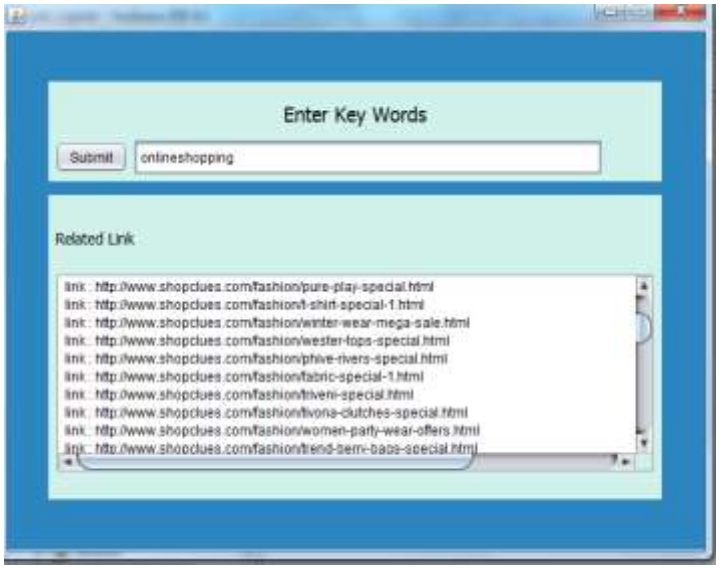


Fig. 3: Shows the extraction of website link from the keyword.

All rights reserved by *www.ijirst.org*

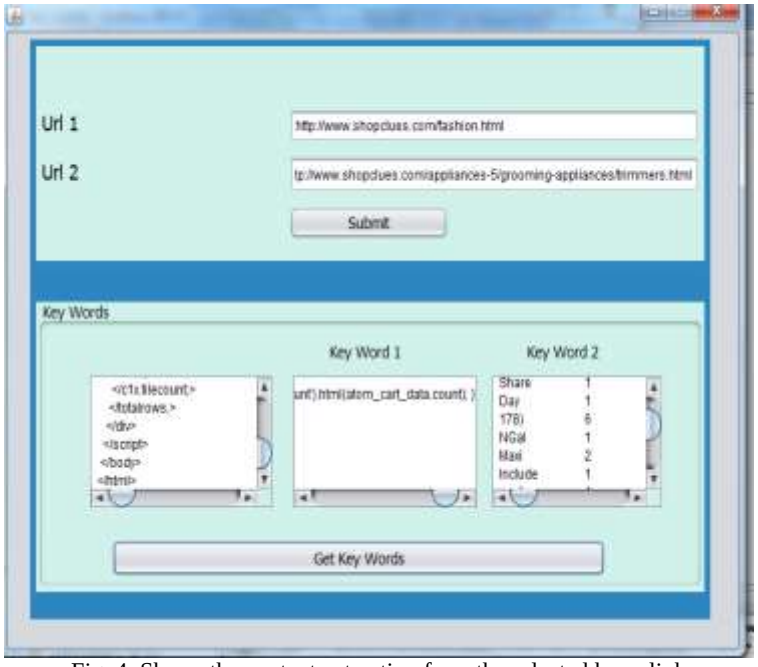
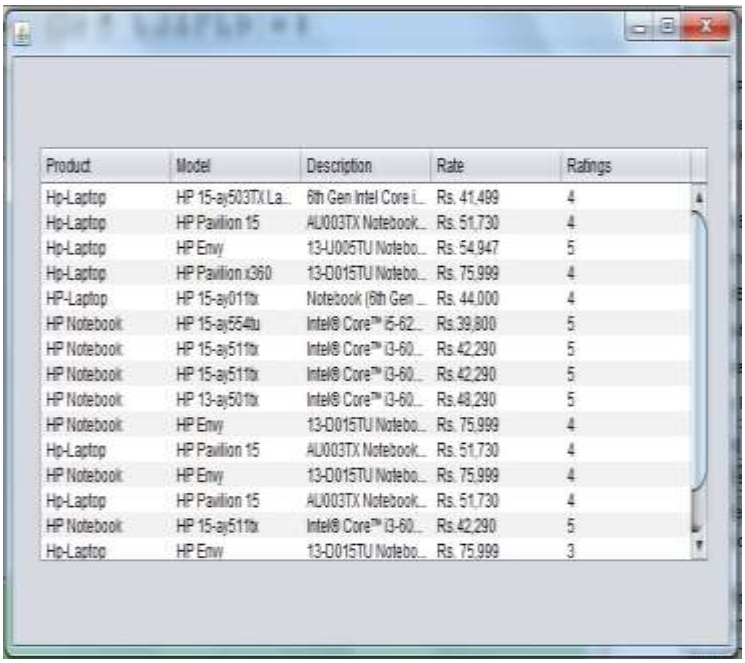


Fig. 4: Shows the content extraction from the selected hyperlink



Product	Model	Description	Rate	Ratings
Hp-Laptop	HP 15-ay503TX La...	6th Gen Intel Core i...	Rs. 41,499	4
Hp-Laptop	HP Pavilion 15	AU003TX Notebook...	Rs. 51,730	4
Hp-Laptop	HP Envy	13-U005TU Notebo...	Rs. 54,947	5
Hp-Laptop	HP Pavilion x360	13-D015TU Notebo...	Rs. 75,999	4
HP-Laptop	HP 15-ay011tx	Notebook (6th Gen ...	Rs. 44,000	4
HP Notebook	HP 15-ay554tu	Intel® Core™ i5-62...	Rs.39,800	5
HP Notebook	HP 15-ay511tx	Intel® Core™ i3-60...	Rs.42,290	5
HP Notebook	HP 15-ay511tx	Intel® Core™ i3-60...	Rs.42,290	5
HP Notebook	HP 13-ay501tx	Intel® Core™ i3-60...	Rs.48,290	5
HP Notebook	HP Envy	13-D015TU Notebo...	Rs. 75,999	4
Hp-Laptop	HP Pavilion 15	AU003TX Notebook...	Rs. 51,730	4
HP Notebook	HP Envy	13-D015TU Notebo...	Rs. 75,999	4
Hp-Laptop	HP Pavilion 15	AU003TX Notebook...	Rs. 51,730	4
HP Notebook	HP 15-ay511tx	Intel® Core™ i3-60...	Rs.42,290	5
Hp-Laptop	HP Envy	13-D015TU Notebo...	Rs. 75,999	3

Fig. 5: Represent the attributes such as product, model, description, rate, ratings from the extracted content.

A. Confusion Matrix

The confusion matrix is used to evaluate the performance of the classification algorithm. Each column in the matrix is indicates from the examples in the predicted class and each row in the matrix denotes the actual class. This will be easier to determine the misclassification due to the classification process to provide good accuracy result. The confusion matrix entries can be defined as follows:

- True positive (tp) is the number of positive instance grouped as positive.
- False positive (fp) is the number of negative instance grouped as positive.
- False negative (fn) is the number of positive instance grouped as negative.
- True negative (tn) is the number of negative instance grouped as negative.

The confusion matrix is used to estimate the value of the accuracy, precision, recall and F1 score.

Recall = $\frac{number\ of\ true\ positives}{total\ number\ of\ positives}$

Precision=$\frac{number\ of\ true\ positives}{number\ of\ true\ positive+number\ of\ false\ positive}$

All rights reserved by *www.ijirst.org*

Accuracy= $\frac{number\ of\ true\ positive + number\ of\ true\ negative}{total\ value}$

F1 Score = $2*\frac{precision*recall}{precision+recall}$

The test dataset contains 175 instances.

Table - 2
Represent of the Confusion Matrix

	True Positive	*True Negative*
Predicted Positive	*121*	*08*
Predicted Negative	*07*	*39*

Recall =0.9453
Precision =0.9380
Accuracy =0.9143
F1 score =0.9416

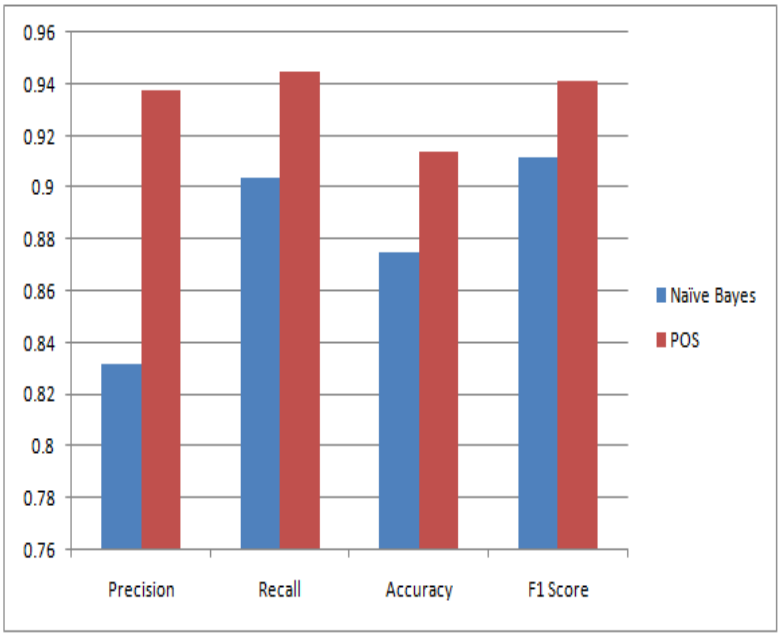


Fig. 6: The above graph represents the difference between the Naïve Bayes and POS algorithm for parameter such as Precision, Recall, Accuracy, F1 score.

V. CONCLUSIONS

Traditionally the online shopping for the users is done by searching each and every website and they need to finalize the product by comparing the same product in different website. This proposed work combines the multiple website by crawling the content and provide the classification by utilizing the particle swarm optimization technique. This provides better searching technique from the combination of the crawling and the classification technique. Finding the best fitness value results in good accuracy for classification. This provides better accuracy by comparing it with the existing naive bayes algorithm in classification.
In future, the stock market dataset can be processed in this framework by modifying the PSO algorithm.

REFERENCES

[1] Jayshree Ghorpade-Aher, Roshan Bagdiya, "A Review on Clustering Web Data using PSO", International Journal of Computer Applications (0975 – 8887), Volume 108 – No. 6, December 2014.(referred)
[2] Simranjeet Kaur, Kiranbir Kaur, "Web Mining and Data Mining: A Comparative Approach", International Journal of Novel Research in Computer Science and Software Engineering, Vol. 2, Issue 1, pp: (36-42), Month: January - April 2015. (referred)
[3] Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013.
[4] Faustina Johnson, Santosh Kumar Gupta, "Web Content Mining Techniques: A Survey", International Journal of Computer Applications (0975 – 888), Volume 47– No.11, June 2012.
[5] Nisha Pawar; K. Rajeswari; Aniruddha Joshi, "Implementation of an efficient web crawler to search medicinal plants and relevant diseases", IEEE Conference Publications, pp: 1-4, 2016.
[6] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik, "Study of Web Crawler and its Different Types", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 1, PP 01-05, Feb. 2014.
[7] Dipali Kharche, Anuradha Thakare, "ACPSO:Hybridization of Ant Colony and Particle Swarm Algorithm for Optimization in Data Clustering using Multiple Objective Functions" Proceedings of 2015 Global Conference on Communication Technologies(GCCT 2015), IEEE publisher, December 2015.
[8] Alexandre Szabo and Leandro Nunes de Castro, "A Constructive Data Classification Version of the Particle Swarm Optimization Algorithm", Mathematical Problems in Engineering, hindawi, Volume 2013 (2013).

All rights reserved by *www.ijirst.org*

[9] Sotiris Batsakis, Euripides G.M. Petrakis, Evangelos Milios, "Improving the Performance of Focused Web Crawlers", in Data & Knowledge Engineering 68(10):1001-1013 · October 2009.
[10] Anish Gupta, Priya Anand, "Focused Web Crawlers And Its Approaches", 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management, IEEE Xplore, 13 Junly 2015.
[11] Girma S. Tewolde, Darrin M. Hanna, "Particle Swarm Optimization for Classification of Breast Cancer Data using Single and Multisurface Methods of Data Separation", IEEE International Conference on Electro/Information Technology, 2007.
[12] Amreen Khan, Prof. Dr. N.G.Bawane, Prof. Sonali Bodkhe, "An Analysis of Particle Swarm Optimization with Data Clustering-Technique for Optimization in Data Mining", International Journal on Computer Science and Engineering, Vol. 02, No. 07, 2010.
[13] Sunita Sarkar, Arindam Roy, Bipul Shyam Purkayastha, "Application of Particle Swarm Optimization in Data Clustering: A Survey", International Journal of Computer Applications (0975 – 8887), Volume 65– No.25, March 2013.
[14] Martin Hlosta, Rostislav Striz, Jaroslav Zendulka, Tomas Hruska, "PSO-based Constrained Imbalanced Data Classification", International Scienti• c Conference INFORMATICS 2013, November 5-7, 2013.
[15] Navid Khozein Ghanad,Saheb Ahmadi, "Combination of PSO Algorithm and Naive Bayesian Classification for Parkinson Disease Diagnosis", Advances in Computer Science: an International Journal, Vol. 4, Issue 4, No.16 , July 2015.
[16] Priya I. Borkar and Leena H. Patil, "Web Information Retrieval Using Genetic Algorithm-Particle Swarm Optimization", International Journal of Future Computer and Communication, Vol. 2, No. 6, December 2013.
[17] Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle, "Web Bots Detection Using Particle Swarm Optimization Based Clustering", IEEE Congress on Evolutionary Computation (CEC), July 6-11, 2014.
[18] Sarita Mahapatra, Alok Kumar Jagadev and Bighnaraj Naik, "Performance Evaluation of PSO Based Classifier for Classification of Multidimensional Data with Variation of PSO Parameters in Knowledge Discovery Database", International Journal of Advanced Science and Technology, Vol. 34, September, 2011.
[19] Vishal Jain, Gagandeep Singh Narula and Mayank Singh, "Implementation Of Data Mining In Online Shopping System Using Tanagra Tool", International Journal of Computer Science and Engineering (IJCSE), ISSN 2278-9960, Vol. 2, Issue 1, Feb 2013, 47-58.
[20] Gomathi A, Jayapriya J, Nishanthi G, Pranav K S,Praveen Kumar G, "Ontology Based Semantic Information Retrieval Using Particle Swarm Optimization", International Journal on Applications in Information and Communication Engineering, Volume 1: Issue 4: April 2015.
[21] Debajyoti Mukhopadhyay, Arup Biswas, Sukanta Sinha, "A New Approach to Design Domain Specific Ontology Based Web Crawler", 10th International Conference on Information Technology, January 2008.

All rights reserved by *www.ijirst.org*