International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-4, February 2019

Distributed Web Usage Mining Based Ecommender System in Big Data Analytics using Hybrid Firefly Algorithm

AC. Priya Ranjani, M. Sridhar

Abstract: ***One of the fast upcoming data mining disciplines that deal with large, unstructured complex data is Big data analysis. Web usage mining is a primary area of research that has been focusing on the valuable information derived from web server logs. Not having any explicit ratings of the users, the large data volume and its sparse nature have been posing challenges to the techniques of collaborative filtering with respect to performance and scalability. Techniques like clustering are dependent on the discovery of offline patterns from the user transactions and are used to improve scalability in terms of collaborative filtering but at reduced cost and recommendation accuracy. To improve the situation, this work has been taken up on the basis of nature inspired, meta heuristic algorithms Firefly and Teaching Learning Based Optimization (FA-TLBO). This FA-TLBO was hybridized using the K-Means algorithm (FA-TLBO with K-Means) in order to obtain optimal cluster centres. There were numerical experiments which indicated the fact that novel FA-TLBO with K-means was more efficient compared to TLBO algorithm.***

Index Terms: Big Data Analysis, Web Usage Mining, Recommender System, Clustering, K-Means Algorithm, Firefly Algorithm (FA) and Teaching Learning Based Optimization (TLBO).

I.INTRODUCTION

Big data has become a very important abstraction applied to the data and which will not comply with any structure available in the traditional databases. For example, any data which is machine-derived will multiply quickly and also will consist of a rich and diverse content which has to be discovered. One more example is the data of social media which is human-generated. This is textual in nature but has some valuable insights overloaded with various implications. The analytics of Big data will reflect the facts that are unstructured, vast and fast changing and cannot be effectively managed by the traditional methods. The information deduced from large amounts of such data has now become extremely important to all organizations. The performance of business organizations and increase in their market share is critical and strongly depends on the insights drawn from data analysis. Big data analysis helps to predict the future trends of an organization. The tools that are available for handling different characteristics of Big data [1] such as variety, velocity and volume have been recently improving.

Manuscript published on 28 February 2019.
*Correspondence Author(s)
AC. Priya Ranjani, Research Scholar, Dept. Of Computer Science, Acharya Nagarjuna University ,Guntur, Andhra Pradesh, India.
Dr.M.Sridhar, Associate Professor, Dept. Of Computer Applications, RVR & JC College of Engineering, Guntur, Andhra Pradesh

Hadoop is an open-source framework used to process large data across commodity clusters of computers using high level data processing languages. The modules also comprise of languages that are easy to use along with some graphical interfaces or tools for administration in order to manage petabytes of data found in different computers. Hadoop Distributed File System and Map Reduce were widely used for data processing. The former is a large-scale framework for distributed storage of data and the later was a processing tool which supports huge chunks of data by making use of a programming model which is simple. The Apache Hadoop project incorporates HDFS, Map Reduce, Pig, YARN, Hive, HBase and several other modules for harnessing the power of processing on clustered computing and the failures are duly managed at the node level itself [2]. Around 30 million new web pages were posted daily, and the World Wide Web (WWW) is used as a primary source of knowledge. For holding onto the users in a profitable manner, in case of these surroundings which are rising hastily, there is a need for a website to be built in a manner in which user personalization can be well-supported. For achieving this, the focus needs to be on keeping track of the user activities. Even though there are several tools used to analyse data, the data which is provided will have to give sufficient information for the website and not to the designer. One way to overcome this issue is adapting data mining techniques for applications on web [3]. Web mining – denotes the utilization of data mining techniques for discovering patterns from WWW. It focuses on integration of all gathered information using the traditional techniques and methodologies and usage of these techniques for automatic discovery and extraction from the services and the web documents. Web mining can be classified based on usage of web data, content of web pages and hyper link structures. For the purpose of this work, Web Usage Mining has been chosen. This aims at capturing the users that interact with the web and is a process which is non-trivial for discovering the valid and novel information from the web by using techniques of data mining. The data that is stored in the usage logs are used to solve problems in navigation, for improving web search, suggesting the websites, recommending various queries and for enhancement of search engine performance. The sources that are used for web mining includes web data repositories such as : web server logs, proxy server logs and the browser logs. The knowledge obtained from web usage mining includes : the actual number of hits, the platform, cookies, the visitor IP address, the Path analysis, the duration and time, the visitor referral website, the visitor referring website and so on.

There are many such actions that are carried out once the web log files are analysed. The most common among those actions are: the shortening of paths for pages of high visits, the redesigning of pages for optimization of the search engine, the redesigning of pages in order to help in user navigation and the elimination or combining of the pages of low visit. The main phases found in the mining of web usage are: pre-processing of data, discovery of patterns and analysis of discovered patterns. There are many applications of web mining which are found at a faster rate mainly owing to the interest in the business in the websites in ecommerce and the applications of web-marketing. Also, the growth in interest in the field of web semantics and web semantic mining has brought about newer perspectives in applications that are related to web usage mining [4] [5]. There is a widespread acceptance to the recommender systems by masses for more than a decade. They alleviate any complexities among products of tasks of service selection and will be used for overcoming all issues in information overload [6]. These recommender systems will collate information on the user preference and go through large volumes of information scattered all over the web and pick the information that suits user's preferences. These systems are useful for various domains like providing music, movies and books, e-commerce, information filtering and personalization of the Web. The personalisation of the Web environment by means of providing a list of the related items based on user interest is a very important aspect. Normally the demographic, collaborative-based and content-based filtering are the techniques employed for generation of recommendations [7]. The applications of Web-personalization have imposed some challenges that are conflicting in nature. This type of anonymous web personalization on the system's online components needs to be run in real time and isolate user's needs within the current session. Furthermore, the volume of the per site generation of navigated data is generally quite high [8]. At the same time, all the information available about each individual customer should be considered, if not the customization will be rendered inefficacious. Efficient customization demands a careful compromise between accuracy and scalability based on the specifications of the application. The primary purpose behind clustering in a web usage mining was the aggregation of sessions that are similar. Algorithms like Self-Organized Maps (SOM), EM-Fuzzy C Means (EM-FCM), graph partitioning, K-Means with the Genetic Algorithm (GA) and the Ant-based techniques are widely used in clustering of the sessions. Clustering is primarily of two types which are usage and page clustering. The former will discover the user with similar patterns of browsing wherein the page cluster will collate the pages that are content related together. This is normally performed by means of a transformation of the session into various vectors of the n elements in which n may refer to the number of pages or the page views. Once this is done, the distance measure is employed to identify the similarity that may exist between various sessions. Clustering is employed to personalizing websites since it can identify users that have similar behaviour [9]. For the purpose of this work, there is a recommender system that is proposed which uses the hybrid FA-TLBO together with K-means clustering algorithm. Section 2 has discussed the related works in the literature. The various methods employed are explained in Section 3. The experimental results are discussed in Section 4 and the work is duly concluded in Section 5.

II.RELATED WORKS

A frequent page set mining along with an Association Rule Mining (ARM) algorithm has suffered from issues of data repository and execution time. This is due to mining of frequent page sets based on their minimum support threshold. So, in order to analyse the level of usage of the web, useful and quality-oriented mining has to be performed using the Weighted ARM (WARM) on the weblogs that are employed by Malarvizhi and Sathiyabhama [10]. The WARM also reduces the runtime and the storage since all frequent pages on weighted support with association rules are based on their weighted confidence. The proposed T + weight tree highlights the dwelling time of user visited pages. These pages will be assigned along with weights that are based on the dwelling time that show the significance and the interest of the user. The T + weight tree algorithm will identify the page sets that are frequently based on the weights within a single database scan. Both clustering and data mining are critical elements for different applications in this field. A primary area in which clustering is used in web intelligence frequently, is greatly researched today. The data which is collected from across the web have been very complex, large, without structure and dynamic. Thus, the traditional techniques of clustering are not sufficient to be used. Tuba and et al., [11] had made a proposal of a swarm intelligence algorithm which was combined with the K-Means clustering and bare bones fireworks algorithm for the clustering of data of web intelligence. The method proposed had been compared to the other approaches and as per the results obtained, it was found that the characteristics of clustering and execution time were very promising. Katarya and Verma [12] had made a presentation of a novel recommender system which was web-based for obtaining the sequential information on the web page navigation of the user. The work had received some top N clusters at the time a Fuzzy C-Mean (FCM) clustering was applied. The authors also found usage for that specific user and further evaluated its weight for every web page. It has further attempted at solving the recommender system problems since it had offered the system to be able to forecast the subsequent web page visit. For this work, the authors had proposed another new system that generated recommendations to users keeping in mind all sequential information in the patterns of usage. Fuzzy clustering was employed in order to give the recommender system an approach which was sequential. It further calculated the weights for every category that was opted and forecasted recommendations for that of the target user. Tripathi, Sharma and Bala [13] had further presented another efficient method of clustering method, the Map Reduce based Enhanced Grey Wolf Optimizer (MR-EGWO) and this was used for the purpose of clustering datasets of a large scale.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

This method further introduced some novel variants of the Grey Wolf Optimizer (GWO) and the Enhanced GWO (EGWO) in which the grey wolf's strategy of hunting along with a binomial crossover that included the levy flight steps had been inducted for enhancing the capability of the search for prey. Also, this variant proposed was used for the optimization of the process of clustering. The consistency and the behaviour of convergence of that of the EGWO were also validated using the boxplots and the convergence graph. Furthermore, this EGWO proposed was parallelized on the model of Map Reduce within the framework of Hadoop and this was called the MR-EGWO for handling datasets of a large scale. Lin et al., [14] had designed another recommendation system which was personalized that took into consideration three different objectives conflicting in nature which are the accuracy, the diversity and its novelty. For letting this system to be able to provide a better comprehensive method, it had presented the Multi-Objective Personalized Recommendation algorithm that made use of the Extreme Point Guided Evolutionary Computation (called MOEA-EPG). This MOEA-EPG proposed was guided using three different extreme points and the crossover operator was designed for satisfying the user demand. The results of the experiment validated the MOEA-EPG and its effectiveness on being compared to the algorithms of recommendation that were state-of-the-art in connection with recommendation novelty, diversity and accuracy. In [15], the authors had presented another Hybrid approach which combined the Bees Swarm Optimization with the TABU Search (HBSO-TS) and this proved to outperform other bio-inspired approaches. The limitation of HBSO-TS was that the intensification was improved using the TABU Search (TS), and diversification continues to be unchanged on being compared to the Bees Swarm Optimization (BSO) (BSO)-ARM, and this is the very first approach that was proposed using the BSO for the ARM. So, to ensure a proper balance between diversification and intensification , the work has proposed two other new heuristics to determine the bees and their search space. There was an evaluation that was conducted on the instances of data which were well-known in order to show the heuristics to improve the actual performance of the HBSO-TS. Furthermore, it had also shown the heuristics and its usefulness in the mining association rules from that of a diversified data like a Weblog mining. Heraguemi, Kamal and Drias [16] had presented another cooperative multi-swarm Bat Algorithm used for the ARM (BAT-ARM). This was based on an algorithm which was Bat-inspired which was further adapted to the problem of rule discovery (BAT-ARM). The latter has a problem of the absence of communication among the bats in a population that can bring down the search space exploration. But this also had a powerful process of rule generation that resulted in a powerful local search. So, in order to maintain a proper trade-off between both diversification and intensification. For this, the authors had proposed some cooperative strategies among swarms which had proved their efficiency in the case of a multi-swarm optimization algorithm (the Ring or the Master-slave). The experimental results proved that this proposal has out performed all other similar approaches in connection to time and rule. Weblog mining was primarily used to get the navigation patterns of the users from the weblogs for web personalization. Wei et al., [17] had defined another new concept known as the "interest pheromone" with a set of models for user navigation paths. The authors further proposed another simple algorithm that relies on improved Ant Colony Optimization (ACO) for mining the dynamic interest of users. For the purpose of the algorithm, there are three different browsing time will access the frequency and the time taken for measuring and "interest pheromone", that will reflect the real interest of users. Lastly, it will conduct the experiments of simulation in contrast to the navigation accuracy and their patterns that are analysed by this approach and their current approaches. The outcomes of the experiment had illustrated that this paradigm proposed will be able to capture the browsing preferences of the users. Agarwal and Nanavati [18] had considered the optimality of Pareto which was an excellent trade-off between these conflicting parameters of disproportionate performance, their comprehensibility, confidence and interesting of these mined rules. Both, the Genetic Algorithm (GA) and the Particle Swarm Optimisation (PSO), are methods of population-based stochastic search which is found in a strong base of the rules of association mining. The authors further proposed another ARM scheme that made used of multi-objective hybridisation in the GA-PSO algorithm. The main advantage of this system was the multiple objective GA hybridization along with the multi objective-PSO thus resulting in the extraction of some interpretable and accurate mined rules.

III. METHODOLOGY

Web personalization, based on web usage mining identifies the needs of individual users. The process will contain two different components. The first was an off-line component that mines the access patterns from the training datasets to learn about various users-accesses. Next was when the access model has been identified, it made use of an online component for interpreting the behaviour of navigations for all active users. An online component was needed for the anonymous identification of the needs of users in real-time. The association rules were used to find correlation among the web pages that appear in the browsing sessions. The Apriori algorithm [19] was a very popular algorithm expressing the co-occurrence of the web pages. The rules further helped in visiting the sites of the web designers in order to restructure the websites. For the purpose of this section, there was a K-Means clustering, the hybrid FA along with the TLBO and the TLBO methods had been discussed.

A. K-Means Algorithm

K-Means clustering method makes use of the k clusters for characterizing data [20]. They will be determined by means of minimizing the actual sum of the squared errors as per equation (1):

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} || X_n - M_k ||^2 \quad (1)$$

where X denotes a data matrix and M denotes the cluster's centroid, k the count of clusters, n the data points, and ||.|| the Euclidean distance. The algorithm is as follows :

1. Initially Select k centres (data points) randomly from X and save in matrix C

2. Look out for the nearest neighbour: every data point in X will be assigned to C, the nearest centre resulting in K clusters.
3. Update centroid s- for every cluster, the centre or median was recomputed.
4. Repeat steps 2, 3 until the centroid matrix C changes

B. Firefly Algorithm with K-Means Algorithm

FA is duly based on the behaviour of communication of the tropical butterflies and also their idealized flashing patterns. It makes use of three different idealized rules for building a mathematical model to this [21]

All the fireflies will be unisex and one firefly gets attracted to that of the other irrespective of their sex;

- The attractiveness of the fireflies corresponds to their brightness and so for any two of them the one that is less brighter moves towards the one that is more brighter. They decrease with an increase in the distance;

- The objective function and landscape determines the brightness of a firefly. (So for a problem of maximization, the brightness may be proportional to the objective function value).

In case of a standard FA, two important points are considered. One will be the light intensity formulation and the other will be the attractiveness change [22]. It may be assumed that the firefly's brightness may be determined using the encoded landscape function. Next, it will have to define the light intensity variation and then formulate the attractiveness and its change. As known already, in nature the intensity of light will decrease as the distance between fireflies increases and the media absorbs the light to ensure the simulation of the intensity of light I will vary with that of the distance r and the parameter of light absorption γ both monotonically and exponentially. That will be (2):

$$I = I_0 e^{-\gamma r^2} \quad (2)$$

Where I_0 is the original intensity of light at the source (which means the distance r = 0) and the γ denotes the coefficient of light absorption. It is evident that in the case of simulation this will denote the firefly and its attractiveness which is proportional to the intensity of light I. Thus, it will be able to define the light attractive coefficient of the firefly β in a manner which is similar to the coefficient of light intensity I. This is as per (3)

$$\beta = \beta_0 e^{-\gamma r^2} \quad (3)$$

Where β_0 denotes its original light attractiveness which is r = 0. Now its Cartesian distance will be employed for the purpose of calculating the actual distance between that of the fireflies i and j at xi and xj as in (4)

$$r_{ij} = || x_i - x_j || = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{k,j})^2} \quad (4)$$

Wherein d denotes the dimensions and the actual amount of the movement of a firefly i was another attractive (brighter) firefly which is as in (5)

$$x_{i=} x_{i+} \beta_0 e^{-\gamma r^2} (x_j - x_i) + \alpha \varepsilon_i \quad (5)$$

In which the three terms denote the present location of the firefly i, the attraction and the randomization of the random variables and their vector εi respectively. The third term, α denotes the scaling parameter controlling the size of the step and this has to be linked with the problems and their interest.

Input :: Dataset D

Output :: Dataset D1

1. i=0, o*= NULL;
2. P(0) = Initialize_ fireflies();
3. while (i < I) do
4. ComputeFitness(p(i),f(o));
5. P(i+1) = Move_Fireflies(p(i));
6. o* = Find_Bestfirefly(p(i),f(o));
7. i++
8. end while
9. Run Kmeans(o*);
10. Return x

Once the iterations are pre-specified, the firefly that has the highest level of fitness or the smallest squared error will be named as the best among the solutions of initial clustering and this will be passed to the K-Means algorithm and will then reassign data points to the centroid that is the closest until the time of coverage. FA combined with the K-means algorithm [23] will begin by an initialization of the firefly population (as per line 2) and either the brightness or the fitness values of the fireflies which are evaluated (as per line 4). The firefly movements are based on their brightness so that those which are darker will get attracted by the ones that are brighter and will move towards them (as per line 5). After this, there is a re-evaluation of the brightness and the steps are iterated. Once the predetermined iterations i are complete the best one o* having the fitness value which is the smallest (o*) will be passed to the K-means algorithm for a local and optimal clustering (as per line 9).

C. Teaching Learning Based Optimization (TLBO) with K-Means Algorithm

As in the case of all other algorithms which are inspired by nature, the TLBO is a method making use of a population of such solutions for moving towards a global one where the population is taken to be a group of learners. For the optimization algorithms, there are various design variables. In the case of the TLBO, there are various design variables which are analogous to the subjects that are offered to the learners and this results in the analogous to 'fitness', as in the case of other techniques.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The teacher is taken to be the best solution [24]. The TLBO has been split into two phases. The first one is the 'Teacher Phase' and the next the 'Learner Phase'. The former denotes gaining knowledge from a teacher and the latter denotes learning by interacting with other learners in the group. **The Teacher phase**: this has been included to be the first segment among the TLBO in which the learners get knowledge from their teachers. Here the teacher tries to increase the class room's mean value from any of the value $mean_1$ to the echelon I_A. This is, however not very promising and the teacher may be able to move the mean belonging to the classroom $mean_1$ to another value $mean_2$ that is healthier compared to $mean_1$ based on competence [25]. If $mean_j$ is considered to be the mean and the I_i the teacher at iteration I, the teacher I_i attempts at improving the currently existing $mean_j$ towards it so that a new man is I_i which is designated as the new mean, and the difference between current mean and the new mean has been given below (6):

$$diverged\_mean_i = r_i(mean_{new} - T_F * mean_j) \quad (6)$$

Wherein the T_F was the teaching factor which fixes the actual mean value that has to be changed and the r_i will denote the random number within a range [0, 1], which has been used for supporting a teaching factor. T_F value may be either 1 or 2 and this is a step which is interrogative and determined in a random manner that is an equivalent probability as per (7):

$$T_F\ round\lfloor 1 + rand(0,1)\{2-1\}\rfloor \quad (7)$$

The teaching factor has been arbitrarily produced in the TLBO inside the scope of 1–2, wherein 1 will compare to the no increase in the level of learning and 2 will relate to an exchange of knowledge and all intermediate values denote the knowledge exchange. This shifting knowledge level may be dependent on the competence of the learner.

On the basis of the diverged_mean, this existing solution will be updated based on the expression as in (8):

$$\alpha_{new,i} = \alpha_{old,i} + diverged\_mean_i \quad (8)$$

The Learner phase: this has been included in the second segment in which the learners will improve knowledge by means of communication among themselves. The learner further adapts to the new things if in case the other learner has better knowledge. This type of learning trend in the phase has been articulated as per equation (9 and 10). For any iteration i, two distinct learners α_i and α_j are considered where $i \neq j$

$$\alpha_{new,i} = \alpha_{old,i} + \gamma_i(\alpha_i - \alpha_j) \quad \text{if} \quad f(\alpha_i) < f(\alpha_j) \quad (9)$$

$$\alpha_{new,i} = \alpha_{old,i} + \gamma_i(\alpha_j - \alpha_i) \quad \text{if} \quad f(\alpha_j) < f(\alpha_i) \quad (10)$$

For the purpose of this work, the TLBO has been used with K-means algorithm for clustering of data into the user-specified numbers. It further shows that this can be used to identify the centroids and a hybrid algorithm was implemented to attain better clustering results [26].

In the concept of clustering, every particle will show an N_k cluster centroid vector [27]. This means every particle Y_i will be taken to be as in (11)

$$Y_i = (C_{i1}, C_{i2}, C_{i3}, \ldots\ldots, C_{ij}, \ldots\ldots C_{iNk}) \quad (11)$$

Where in C_{ij} will indicate the j^{th} cluster centroid for the i^{th} particle in the cluster. In order to calculate the particle's fitness, the error of quantization was used and measured based on the formula (12)

$$J_e = \sum_{vDC} c_{i,j} d(d_v c_j) \quad (12)$$

Wherein 'd_v' denotes the data vector, d the Euclidian distance of the data vector, the centroid and $|S_{ij}|$ denotes the actual number of vectors that are part of a cluster S_{ij}, which is the cluster's frequency.

Application of the TLBO method leads to clustering of all data points as described below:

1. Load every particle in a way that it contains N_k in a randomly selected centroid cluster.
2. Let the t = 1 to t_m do
   (a) For every particle i do
   (b) For every data vector d_v

i. Compute Euclidean distance d (d_v, C_{ij}) to centroids in S_{ij}

ii. Accredit the dv to cluster S_{ij} so that as in (13)

$$d(d_v, C_{ij}) = \min_{Vc=1} \ldots\ldots Nk\{d(d_v, C_{ik})\} \quad (13)$$

iii. Compute fitness with the help of equation (12).

(c) Update the mean values

(d) Updating of the cluster centroids with (8), (9), and (10).

Wherein the t_m denotes the maximum number of such iterations.

D. Proposed Hybrid FA-TLBO Algorithm

For any of the algorithms of metaheuristic optimization that are based on the population where the process of search includes two different stages which are the exploration and the exploitation. There are operators included for exploring every path in the search domain with some randomized movements. The stage is called exploration in which the search space has been explored for getting a better solution. The stage of exploitation will be followed by the stage of exploration and this identifies and locally searches in areas which are promising. There is a need to have a right balance between both to get an approximation. The balance between both will be a task which is for the development of an algorithm which is meta heuristic owing to its stochastic nature [28]. For the FA-TLBO algorithm, the FA and the TLBO are merged to compensate the deficiencies of each other in which the FA is employed to explore its search space and the TLBO is for accelerating the process of exploitation. In FA-TLBO, FA along with the TLBO will compete for iteration based on the self-adaptive Selection Rate (SR). This SR will change dynamically based on a newly obtained solution which is better compared to the worst solution of a population in a T cycle. In the initial stages, the FA will be able to obtain better opportunities to explore all unknown regions. Where there is a region that consists of a solution which is globally optimal is identified in the later stage.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Here the TLBO obtains some more opportunity to exploit solutions of high performance [29]. In the case of new and unknown problems, exploring unknown spaces in the early stages will be considered for an FA-TLBO algorithm, that has FA as an ideal choice. There may be a high SR ($\geq$ 0.95) which is assigned to the FA algorithm to explore unknown regions. In case the global region is not identified it may by quite expensive to get a high SR. But if the FA is able to obtain an SR which is higher to explore the unknown areas, and the FA gets a low rate of selection (where the TLBO has a high probability), the population diversity is quickly lost owing to its quick convergence. So, there is a need to maintain the population diversity to a particular level and in the next half of this stage, the FA will be given more than about 0.3 probabilities for it to run. The primary issue with the FA with the K-means was that it can be easier for a single data point to change among the two neighbouring indices. One more problem with that of the FA with the K-means algorithm which is at the time of the process of moving the clustering indices that turn into decimals that have to be rounded up or down to the integers which are the closest. So, the hybrid FA-TLBO algorithm along with the K-means clustering which is proposed. This method will deliver its final clustering to K-Means owing to a higher speed of execution. At the same time, a hybrid FA-TLBO algorithm which has solved the issue of falling of both the FA and the TLBO within the local optimization which eliminates the dependence of the methods which are partition space like the K-means to its initial starting centres.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

The experiments were realised by making use of web links and the data from simulated e-commerce websites based on Hadoop framework and it was evident that the proposed system achieved better results. In this section, the TLBO and FA-TLBO methods are used. Experiments are carried out using window size=2, 3 and 4. The precision and coverage as shown in tables I to IV and Figs. 1 to 4.

Table I. Precision for TLBO

Recommendation Threshold @ Support = 0.06	Window size=2	Window size=3	Window size=4
0.1	0.42	0.43	0.45
0.2	0.45	0.47	0.49
0.3	0.54	0.55	0.58
0.4	0.58	0.61	0.62
0.5	0.6	0.65	0.66
0.6	0.63	0.65	0.68
0.7	0.66	0.68	0.74
0.8	0.66	0.7	0.75
0.9	0.7	0.72	0.75
1	0.7	0.73	0.78

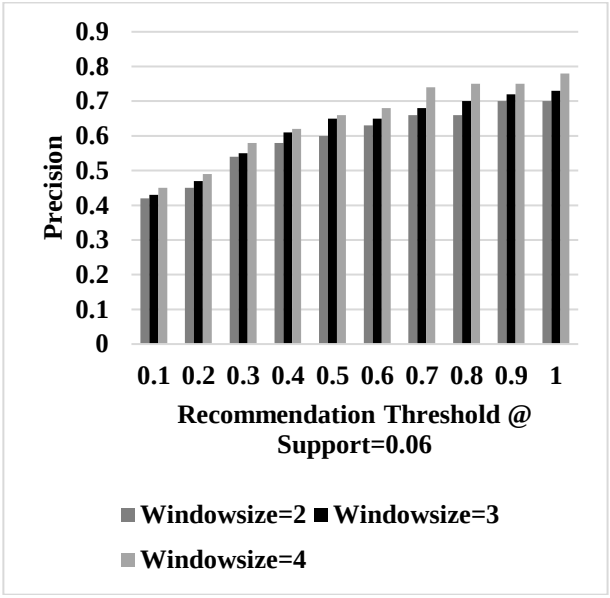

Figure 1. Precision for TLBO

From Fig. 1 it is evident that the precision (TLBO) for window size 4 performs better by 6.89% and by 4.54% at recommendation threshold 0.1 than window size 2 and 3 respectively. The precision (TLBO) for window size 4 performs better by 9.52% and by 1.52% at recommendation threshold 0.5 than window size 2 and 3 respectively. The precision (TLBO) for window size 4 performs better by 10.81% and by 6.62% at recommendation threshold 1 than window size 2 and 3 respectively.

Table II. Precision for FA-TLBO

Recommendation Threshold @ Support = 0.06	Window size=2	Window size=3	Window size=4
0.1	0.47	0.48	0.5
0.2	0.51	0.52	0.53
0.3	0.59	0.61	0.63
0.4	0.64	0.66	0.68
0.5	0.67	0.7	0.72
0.6	0.68	0.7	0.76
0.7	0.73	0.76	0.81
0.8	0.74	0.76	0.82
0.9	0.76	0.79	0.84
1	0.78	0.79	0.86

From the Fig. 2 shows that the precision (FA-TLBO) for window size 4 performs better by 6.18% and by 4.08% at recommendation threshold 0.1 than window size 2 and 3 respectively. The precision (FA-TLBO) for window size 4 performs better by 7.19% and by 2.81% at recommendation threshold 0.5 than window size 2 and 3 respectively. The precision (FA-TLBO) for window size 4 performs better by 9.75% and by 8.48% at recommendation threshold 1 than window size 2 and 3 respectively.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

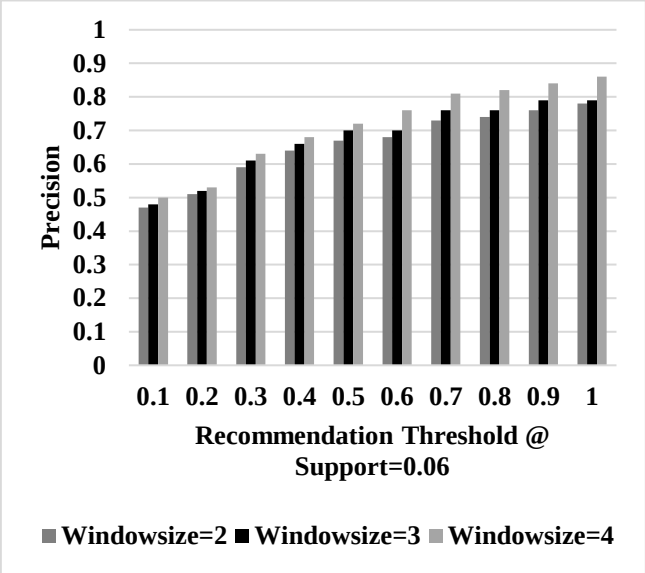

Figure 2. Precision for FA-TLBO

From Fig. 3, it is known that the coverage (TLBO) for window size 2 performs better by 15.95% and by 41.09% at recommendation threshold 0.1 than window size 3 and 4 respectively. The coverage (TLBO) for window size 2 performs better by 27.84% and by 53.52% at recommendation threshold 0.5 than window size 3 and 4 respectively. The coverage (TLBO) for window size 2 performs better by 9.52% and by 31.57% at recommendation threshold 1 than window size 3 and 4 respectively.

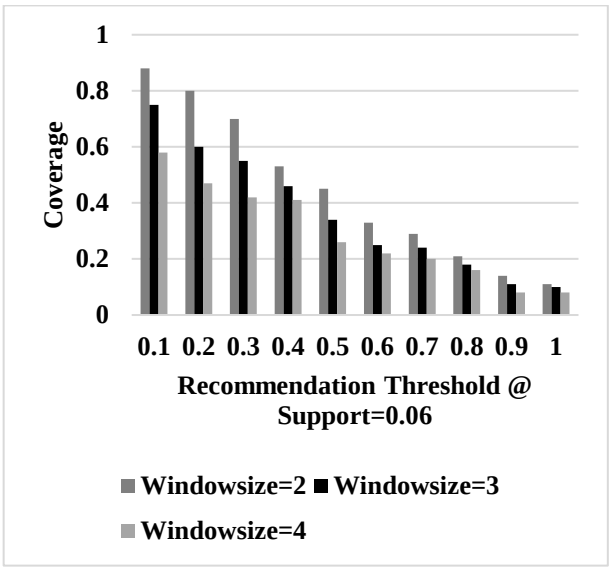

Figure 3. Coverage for TLBO

Table III. Coverage for TLBO

Recommendation Threshold @ Support = 0.06	Window size=2	Window size=3	Window size=4
0.1	0.88	0.75	0.58
0.2	0.8	0.6	0.47
0.3	0.7	0.55	0.42
0.4	0.53	0.46	0.41
0.5	0.45	0.34	0.26
0.6	0.33	0.25	0.22
0.7	0.29	0.24	0.2
0.8	0.21	0.18	0.16
0.9	0.14	0.11	0.08
1	0.11	0.1	0.08

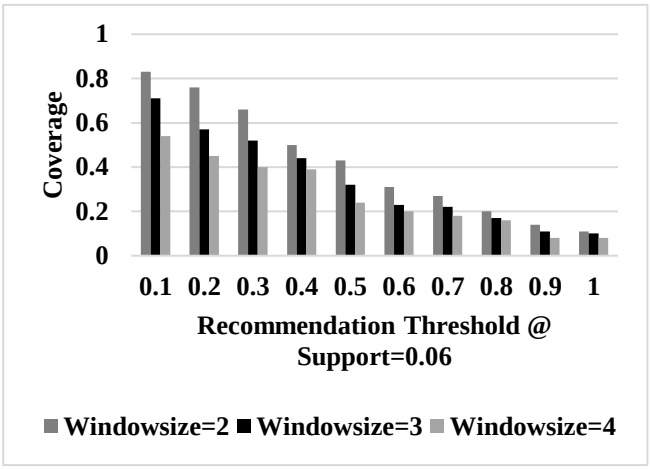

Figure 4 Coverage for FA-TLBO

From the Fig. 4 shows that the coverage (FA-TLBO) for window size 2 performs better by 15.58% and by 42.33% at recommendation threshold 0.1 than window size 3 and 4 respectively. The coverage (FA-TLBO) for window size 2 performs better by 29.33% and by 56.71% at recommendation threshold 0.5 than window size 3 and 3 respectively. The coverage (FA-TLBO) for window size 2 performs better by 9.52% and by 31.57% at recommendation threshold 1 than window size 3 and 4 respectively

Table IV. Coverage for FA-TLBO

Recommendation Threshold @ Support = 0.06	Window size=2	Window size=3	Window size=4
0.1	0.83	0.71	0.54
0.2	0.76	0.57	0.45
0.3	0.66	0.52	0.4
0.4	0.5	0.44	0.39
0.5	0.43	0.32	0.24
0.6	0.31	0.23	0.2
0.7	0.27	0.22	0.18
0.8	0.2	0.17	0.16
0.9	0.14	0.11	0.08
1	0.11	0.1	0.08

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

V. CONCLUSION

The primary challenge in web mining was to design an effective system of web personalization that improves collaborative filtering and its scalability. Also, the system's effectiveness has to be measured based on coverage and its accuracy or precision. For the purpose of this work, a novel hybrid FA with TLBO algorithm integrated with the K-Means clustering is taken. The results proved that the actual precision (FA-TLBO) for a window size 4 has performed better by about 6.18% and by about 4.08% at a threshold of recommendation 0.1 than the window size 2 or 3 respectively. The coverage (FA-TLBO) for the window size of 2 will perform better by about 9.52% and further by 31.57% at a recommendation threshold of 1 than a window size of 3 and 4. The advantages of FA-TLBO include choosing an elite solution from population, avoids trapping in local minima and eliminates duplicate values using hybrid solutions.

REFERENCES

1. J. Zakir, T.Seymour, and K.Berg, "Big Data Analytics," Issues in Information Systems, 2015, .16(2), pp. 81-90.
2. V.Dagade, M.Lagali, S.Avadhani, and P. Kalekar, "Big Data Weather Analytics Using Hadoop," in IJETCSE, 14(2), pp.847-851.
3. R.Ali, "Cluster Optimization for Improved Web Usage Mining", in IJRITCC, 2015,3(11), pp.6394-6399.
4. M. Sajwan, K.Acharya, and S.Bhargava, "Swarm intelligence based optimization for web usage mining in recommender system," 2014, IJCATR, 3(2), pp.119-124.
5. J.Vellingiri, S.Kaliraj, S.Satheeshkumar, and T.Parthiban . "A novel approach for user navigation pattern discovery and analysis for web usage mining,". JCS, 2015, 11(2), pp.372-382.
6. Abbas, L.Zhang, and S.U.Khan, "A survey on context-aware recommender systems based on computational intelligence techniques," In Computing, Springer, 97(7), pp.667-690.
7. M.Jafari, F.S.Sabzchi, and A.J.Irani, "Applying web usage mining techniques to design effective web recommendation systems: A case study", Advances in Computer Science: an International Journal, 3(2), 2014, pp.78-90.
8. 8 .C.Shahabi and F.Banaei-Kashani, "Efficient and anonymous web-usage mining for web personalization", INFORMS Journal on Computing, 2003, pp.123-147.
9. 9..A.G.Abdalla, T.M.Ahmed and M.E.Seliaman, "Web Usage Mining and the Challenge of Big Data: A Review of Emerging Tools and Techniques", In Handbook of Research on Trends and Future Directions in Big Data and Web Intelligence, . IGI Global, 2015, (pp. 418-447.
10. 10. S.P.Malarvizhi, and B.Sathiyabhama, "Frequent page sets from web log by enhanced weighted association rule mining", Cluster Computing, 2016, 19(1), pp.269-277.
11. 11.E.Tuba, R.Jovanovic, R.C.Hrosik, A. Alihodzic and M.Tuba, "Web Intelligence Data Clustering by Bare Bone Fireworks Algorithm Combined with K-Means". In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, (2018, June)
12. (article 7). ACM
13. 12.R.Katarya and O.P.Verma, "An effective web page recommender system with fuzzy c-mean clustering. Multimedia Tools and Applications", ,2017, 76(20), pp.21481-21496.
14. A.K.Tripathi, K.Sharma and M.Bala, "A Novel Clustering Method Using Enhanced Grey Wolf Optimizer and MapReduce. ",Big Data Research, Elsevier, 2018
15. Q.Lin, X.Wang, B.Hu, L.Ma, F. Chen, J.Li, and C.A.Coello Coello, "Multiobjective Personalized Recommendation Algorithm Using Extreme Point Guided Evolutionary Computation", Hindawi Complexity, 2018.
16. 15.Y.Djenouri, Z.Habbas, D.Djenouri, and M.Comuzzi, "Diversification heuristics in bees swarm optimization for association rules mining," In Pacific-Asia Conference on Knowledge Discovery and Data Mining ", Springer, 2017, pp. 68-78.
17. 16.K.E.Heraguemi, N.Kamel, and H.Drias, "Multi-swarm bat algorithm for association rule mining using multiple cooperative strategies," Applied Intelligence, 2016, 45(4), pp.1021-1033.
18. 17.X.Wei, Y.Wang, Z.Li, Z., Zou, T., & Yang, G. "Mining users interest navigation patterns using improved ant colony optimization. Intelligent Automation & Soft Computing", 2015, 21(3), pp.445-454.
19. A.Agarwal, and N.Nanavati, "Association rule mining using hybrid GA-PSO for multi-objective optimisation," In Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on pp. 1-7.
20. 19.J.Umarani, R.Sivaprakash, and G.Thangaraju, "Web Usage Mining Analysis for Big Data Applications in Government Sectors of India", International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), 2016, 23 (5), pp.201-211.
21. 20.S.Burla, "High Dimensional Data Clustering Using Hybridized Teaching-Learning-Based Optimization", Journal of Computer and Mathematical Sciences, 2013, 4(3), pp.135-201.
22. S.X.Yang, and X.He "Firefly algorithm: recent advances and applications",2013, arXiv preprint arXiv: pp.1308.3898.
23. L.Zhang, L.Liu, S.X.Yang, andY. Dai, "A novel hybrid firefly algorithm for global optimization," 2016, PloS one, 11(9), e0163230.
24. L.Zhou, and L.Li,(2018). "Improvement of the Firefly-based K-means Clustering Algorithm," International Conference on Data Science,2018, pp.157-162.
25. R.V.Rao, V.J.Savsani, and D.P.Vakharia,, "Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems," 2011, Computer-Aided Design, 43(3), pp.303-315.
26. 25.R.R.Kurada, K.K.Pavan, and A.A.Rao, "Automatic teaching–learning-based optimization: A novel clustering method for gene functional enrichments", In Computational Intelligence Techniques for Comparative Genomics, Springer, Singapore, pp. 17-35.
27. P.K.Mummareddy, and S.C.Satapaty, "An hybrid approach for data clustering using K-means and teaching learning based optimization," In Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI , Springer, Cham**,** 2015, Vol. 2, pp. 165-171.
28. R.Singh, H.Chaudhary, and A.K.Singh, "A new hybrid teaching–learning particle swarm optimization algorithm for synthesis of linkages to generate path, In Sadhana, 2017, 42(11), pp.1851-1870.
29. R.Singh, H.Chaudhary, and A.K.Singh, "A new hybrid teachinglearningparticle swarm optimization algorithm for synthesis of linkages to generate path, In Sadhana, 2017, 42(11), pp.1851-1870.
30. S.Tuo, L.Yong, Y.Li, Y.Lin and Q.Lu, "HSTLBO: A hybrid algorithm based on Harmony Search and Teaching-Learning-Based Optimization for complex high-dimensional optimization problems", PloS one, 2017, 12(4), e0175114.

AUTHORS PROFILE

Mrs. A.C.Priya Ranjani, has completed M.Sc in Computer Science from PB.Siddhartha College of Arts & Science , Vijayawada. Later attained M.Tech degree from VR.Siddhartha College of Engineering, Vijayawada. At present working as Assistant Professor in CSE department in Vijaya Institute of Technology for Women, Vijayawaada. She is doing her research at Acharya Nagarjuna University, Guntur. Her areas of interest include Data Mining & Big Data Analytics.

Dr. Sridhar Mandapati, obtained his master's degree in Computers Applications from S.V University, Tirupathi. He received his Ph.D. in Computer Science & Engineering at Acharya Nagarjuna University, Guntur. He is currently working as Associate Professor in the Department of Computer Applications, R.V.R. & J.C College of Engineering, Guntur. He has 20 years of teaching experience. He has more than 40 International and National publications to his credit. His research interests include Data Mining, Cloud Computing and Information Security.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)