Hybrid Cluster based Collaborative Filtering using Firefly and Agglomerative Hierarchical clustering

Spoorthy G
Department of Computer Science & Engg.
NIT Warangal
Warangal, India
Email: gspoorthy6 [AT] gmail.com

Sriram G. Sanjeevi
Department of Computer Science & Engg.
NIT Warangal
Warangal, India
Email: sgs [AT] nitw.ac.in

Abstract— **Recommendation Systems finds the user preferences based on the purchase history of an individual using data mining and machine learning techniques. To reduce the time taken for computation Recommendation systems generally use a pre-processing technique which in turn helps to increase high low performance and over comes over-fitting of data. In this paper, we propose a hybrid collaborative filtering algorithm using firefly and agglomerative hierarchical clustering technique with priority queue and Principle Component Analysis (PCA). We applied our hybrid algorithm on movielens dataset and used Pearson Correlation to obtain Top N recommendations. Experimental results show that the our algorithm delivers accurate and reliable recommendations showing high performance when compared with existing algorithms.**

Keywords-Machine learning; Recommendation systems; Meta-heuristic optimization

I. INTRODUCTION

Recommendation Systems (RS) are very efficient techniques that provide valuable suggestions to the customers and helps the websites to earn customers. Websites such as Amazon, Facebook, YouTube, Netflix, online News articles and many e-commerce websites to help their customers with the trends, useful products, new products demand w.r.t the recommendations or suggestions provided by the like minded customers purchase history or similar items in a customers purchase history or trust worthiness between users. RS are software tools which suggests books in Amazon, Videos in YouTube, Friends suggestions Facebook, helps in finding news articles on a online news websites.

The main motive of recommender systems is to recommend interesting items to the online users or customers among the potentially huge set of items present in their website reducing their search time and to help them making their decisions. A RS recommends the products taking historical purchase information of user, product information given by other users, Product information compared to other products, trust between the user w.r.t other users then tries to build more accurate and personalized recommendations.

Now a days many e-commerce websites, digital media channels, social media, soon., are evolving. Even the large varieties of products are available in the websites. Choosing a product or following the interesting updates on the internet became a hard task for a user. This eventually lead to exponential growth in data. Processing huge volumes of data and analyzing patterns then providing the recommendations became tough task for the systems. In recent years, many algorithms result in-feasible for huge datasets due to (i) low accuracy (ii) high execution times and (iii) complexity. This can be due to data is sparse, cold start problem and poorly scalable. In order to increase the credibility of the sparse data pre-processing techniques such as PCA[17], LDA are used these techniques reduce the dimensions of the data without losing important information. PCA finds the patterns in the high dimensional data set known as 'features'.

Another aspect where the research is concentrating is the Data analyzing or information processing which has a prominent role in areas such as machine learning, statistical inference and data mining. Machine learning deals with the building a model and automates the data analysis. Statistical inference deduces the properties of a probability distribution using data analysis. Pattern recognition[6] in data mining deals with the efficient extraction of knowledge from the data. Many meta- heuristic algorithm such as GA[1], PSO[3], Firefly[13] deals with the optimization of the algorithms which helps in reducing the search space and helps data analysis process to speed up.

Recommendation systems mainly are of three types (i) Content based filtering, (ii) Collaborative filtering (CF), (iii) Hybrid RS. CF provides a less complex architecture to find relevant recommendations and known as dominant techniques in RS. CF[10, 11, 12] is further split into user based CF, item based CF and trust based. User based CF[9, 10] methods tends to recommend items from the like minded customers which are obtained by the similar historical ratings provided by the customers on previous purchases made. That means, customers who think similar in past will agree on future selections. While processing User based CF on huge data set may suffer from (i) sparse data (ii) Excess execution time (iii) Poor performance in terms of accuracy. Data sparsity in recommendation systems mainly occur due to (i) less feed-backs provided by the users (ii) exponential data growth (iii) very few items being purchased by the customers (iv) listed item may be new. The data sparsity can be handled efficiently by adding pre-processing techniques such as LDA, PCA.

To analyse the patterns in recommendations one can follow one of the two methods[19] (i) model based recommendations: the model is trained before-a-head with the historical information of the customers in a website then recommendations are calculated when the customer becomes active on the website using the similarity matrix (ii) memory based recommendations: recommendations are calculated with the historical data and stored in the memory allocated to the user.

In our hybrid technique, we have addressed data sparsity, cold start and time constraint issues and reduced them efficiently. The algorithm have given more accurate recommendations compared to existing techniques. The proposed algorithm follows model based cluster based CF. The proposed system trains the RS using nature inspired stochastic firefly algorithm coupled with agglomerative hierarchical clustering to find the like minded customers. For accelerating the clustering process priority queues are implemented using binary heap. The entire phase done on training data and known as offline mode of the system. Top N recommendations are obtained using Pearson Correlation method which is done on test data and this phase is known as online mode. The performance of our proposed algorithm is measured using Movielense dataset which shows our proposed system have high accuracy compared to existing cluster based collaborative filtering techniques.

The rest of this paper is assembled as follows: In section 2, review on model driven cluster based collaborative filtering is presented. Section 3, Hybrid Cluster based Collaborative Filtering using Firefly and Agglomerative Hierarchical clustering approach. Section 4 deals with experimental results. Eventually, we driven some conclusions and presented some directions to the future improvements in section 6.

II. PRELIMINARIES

Recommendation Systems (RS) main aim is to provide a preferred personalized suggestions to customers and helping them with recommended lists and reducing their work which couldn't have happened with the huge variety of items. Recommendations are most successful intelligent systems in e-Commerce, NetFlix, Spotify, YouTube, Facebook, etc., Recommendation can be mainly divided into content based filtering, collaborative filtering and hybrid filtering. Among which collaborative filtering RS(CF) takes more credit and most successful RS. CF is further sub-divided into user based CF, item based CF and trust based. User based CF takes user profiles and their historical ratings into consideration and provides the recommendations. Item based CF[7, 18] takes items list and item ratings into consideration and provides the recommendations based on similar items. Item based CF gives more performance compared to user based CF but this systems suffers when the new items are added to the dataset and which is not preferable when the new type of items keeps on evolving. Trust based recommendations depends on the trust between the users which varies from 0 to 5.

User based CF can be achieved either by model based CF model or by memory based CF. Both models perform well. When memory is a constrain then its supported to use model based CF like cluster based, neural networks, bayesian networks soon. Cluster based CF models give more preference to making clusters which can represent like minded customers. Existing algorithms such as k-means[3], SOM[5], GAKM[1], PCA-GAKM[2] soon uses Cluster based CF techniques.

Firefly algorithm (FA) is an meta heuristic which is nonlinear and stochastic in nature proposed by yang. FA[13, 14, 15, 20] have many advantages and have been in usage of application like energy efficient applications, travelling sales person problems, soon. Previous studies proved that the FA is well known for its speed and gives best optimal solutions comparatively.

III. METHODOLOGY

A. Pre-processing

Data pre-processing is very important when data is so huge , yet very sparse. With such data calculating distance matrix, clustering methods convergence will take more time which may result in low accuracy of the entire process. There are many pre-processing techniques like LDA, PCA, clustering sometimes considered as a pre-processing technique even. PCA employees orthogonal projection of highly correlated variables to a set of uncorrelated variables. It allows us to keep the features that have high variance. The main intent of PCA is to cut down the no.of dimensions using the variance. The first principal component have the most possible variance which considers highly correlated features. PCA linearly transforms input space of m- dimensions to some output space of n- dimensions(n<<m). The loss of information due to reduced dimensional space many times leads to inaccurate performance. In general a model should contain the Expected Variance ratio" 60% or more to expect good performance. (i)Load the Sparse data set. (ii)Subtract the mean. (iii)Calculating the covariance matrix (iv)Calculating Eigen vectors and Eigen values of the covariance matrix (v)Choosing the components and forming features.

B. Firefly Algorithm

Firefly algorithm (FA) is an meta heuristic which is nonlinear and stochastic in nature proposed by yang. FA is stochastic as FA have one or more random components known as stochastic components tends to form global optimal solutions fast compared to deterministic methods. The entire FA depends on two basic points degree of attractiveness between fireflies and light intensity. The light intensity of firefly hang on light intensity emitted at distance =0 and the distance between the fireflies. The attractiveness depends on

light intensity seen and the distance between them. The movement of firefly is extremely based on the attraction and distance between the fireflies.

$$I_i = I_0 e^{-\gamma r_{ij}} \quad (1)$$

I(i) is light intensity at 'i', I(0) is light intensity at r(ij)=0

$$\beta_{ij} = \beta_0 e^{-\gamma r_{ij}} \quad (2)$$

$$x_i = x_i + \beta_{ij}(x_j - x_i) + \alpha\varepsilon_{ij} \quad (3)$$

C. Clustering

In Clustering based CF clustering plays a crucial role as its name suggests. Clustering divides the objects into clusters where the inter cluster distances among the objects are very less compared to intra-cluster distances between the objects. Clustering algorithms are of typically partition based, density based and hierarchical. Partiotional based clustering[3] suffers from (i)its convergence to a local minima (ii)unknown and unpredictable cluster size and initial K values. These are the main reasons which led us go to hierarchical clustering for our proposed system. Hierarchical clustering analysis[3,6] is further classified into bottom-up (Agglomerative) and top-down approach(divisive) methods. Hierarchical clustering have a very simple processing structure and can produce very much acceptable level of performance. For speeding up the clustering process priority queues are implemented using binary heap.

D. Pearson Correlation

Pearson correlation gives the interdependence between two variables using the divergence between the variables and product of the standard deviations of the variables.

$$\text{Pearson coefficient} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \quad (4)$$

Here, 'x' and 'y' are variables, cov(x,y) is the covariance of 'x' and 'y' and 'σ' is their standard deviations.

IV. IMPLEMENTATION

In this section, our proposed hybrid algorithm is discussed and given step to step insights of the algorithm guiding to the enhancement of performance in terms of accuracy of the recommendation system. This approach is a model based CF , which creates clusters of like minded users offline. Then provides the recommendations list for an active user directly from the clusters formed in offline finding user based recommendations for the customers. The offline phase starts with the pre-processing step where the current dimensions are reduced without loosing important data using PCA. PCA reduces the inaccuracy caused due to data sparsity. Later, we develop FAH(wP) (Firefly Agglomerate Hierarchical with Priority Queue) on the training data set to form clusters which have like minded users. Using clusters produced in offline phase to find the Top-N Recommendations. For finding Top-N Recommendations Pearson Correlation (PCC)[18] is used. Algorithm goes as follows:

The dataset consists of users with id 'U' and list of items 'I' and their corresponding rankings 'R'

1) Data pre-processing is done using Principal Component Analysis(PCA). PCA finds the correlation between the independent values qualified by finding new principal axes. Suppose we have 'm' users and 'n' items gives 'm×n' dimensions of ratings. Where 'm' and 'n' are independent variables. PCA finds the correlation between independent attributes 'm' and 'n' then reduces 'n' to 'd' where d is much less than 'n' (d<<n) and 'd' is called the principal components. For retain the accuracy we have taken the expected variance ratio of 90% is considered so that not much information is lost.
2) Assuming there are 'm×d' dimensions after step 1, using Firefly algorithm[15] we rank the users(fireflies). Here ranking of the fireflies means grouping of the fireflies w.r.t the movement of fireflies. The objective function for the fireflies is calculated and update. This process continues until all the users are grouped. As firefly have the capability to reach the solution in comparatively less time, handles the outliers efficiently the grouping of the users will be done fastly.

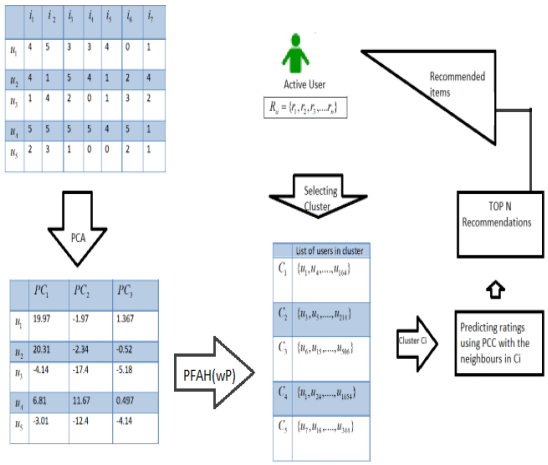

Figure 1: PFAH(wP) algorithm for collaborative recommendation system

3) Agglomerative Clustering[3, 6] is used to cluster the ranked users into limited 'k' groups so that applying the Pearson correlation in the online phase will get accurate recommendations prediction. Initially each firefly is considered to have a cluster of its own (Each row 'm' is considered as a cluster initially) later on two similar clusters are merged until K clusters remains where (K<d). In the priority queue users who are nearest neighbors of the clusters are cached which accelerates the clustering step.
4) Once the clusters are formed PFAH(wP) finishes finding the clusters and offline phase is successfully achieved.

When an active user arrives the online phase starts and TOP N recommendations are listed based on which cluster the user belongs to instead of getting recommendations from the whole user space. The prediction of ratings can be achieved from Pearson correlation[18].

TABLE I. PFAH(wP) APPROACH

Algorithm: PFAH(wP)

Initialization:
Maximum iteration T=200, t=0
α = 0.5
No.of Clusters k;
Objective function

$$f(x) = \sum_{x_i \in X} \min_{i \le j \le k} (dist(C_i, C_j))$$

Initialize each firefly as a cluster

$$C = \{x_i\} \forall i = 1,2,...,m$$

Calculate distance between clusters

$$d_{C_i, C_j} = d_{ij} \forall i, j$$

Randomly select k fireflies

Light intensity at I_i at x_i is determined by f(x)

```
While (t<T)
  For i=1 to m
    For j=1 to m
      if (I_i < I_j)
        Move firefly i to j using equation(2)
      End If
    End For
  End For
  Update the light intensity f(x) using equation(1)
  t= t+1
End While
//output: sorting of fireflies according to the light intensity
Cl<- []
len[x] <-1
For x in C
  nn[x] <- f(xi)
  mdist <- d[x, nn[x]]
End For
Q <- Priority queue with indexes in C, keys in mdist
For i in C
  a <- min element in Q
  b <- nn[a]
  l <- mdist[a]
  While l ≠ d[a,b]
    nn[a] ← arg min_{x>a} d[a,x]
    Update mdist, Q with (a, d[a,nn[a]])
    a <- min element from Q
    b <- nn[a]
    l <- mdist[a]
  End While
  Remove a from from Q
  Append (a,b,l) to Cl
  len[b] <- len[a] + len[b]
  C <- C \ {a}
  For x in C such that x<a
    If nn[a]=a then
      nn[x] <- b
    End If
  End For
  For x in C such that x<b
    If d[x,b] < mdist[x] then
      nn[a] <- b
      Update mdist, Q with (x, d[x,b])
    End If
  End For
  nn[b] ← arg min_{x>b} d[x,b]
  Update mdist, Q with (b, d[a,nn[b]])
End For
Return Cl
```

V. RESULTS AND DISCUSSION

A. Time Complexity

The time complexity of PFAH(wP) is divided into two parts: (i) Cluster Formation (ii)Collaborative Filtering Step;

i. *Cluster Formation complexity:* It is also known as complexity for offline mode. For the pre-processing step the $O(\min(d^3, m^3))$. For firefly Algorithm the average case complexity in terms of objective function evaluation is $O(m \log m)$ which is very optimal compared to Genetic algorithm average case $O(m^{3/2} \log m)$. Finally for constructing agglomerative hierarchical clustering it usually takes $O(n^3)$ by implementing the priority queues using binary heap[4] we reduced the complexity to $O(n^2 \log n)$.

ii. *Collaborative Filtering Step:* it is also known as Online Phase. The general case scenario time complexity of PCC is $O(nm^2)$. The worst case scenario of user based prediction is $O(m^2)$.

Through the complexity analysis above for the two offline and online phases it is proved the constructed PFAH(wP) algorithm is reliable and comparable in real time recommendations reach.

B. Dataset Description

The performance evaluation for FAH(wP) is done on very popular dataset “MOVIE LENSE” which is available in Kaggle with ratings ranging from 1Lakh to 10Million. In order to compare with the existing system we used 1Lakh ratings provided by 943 users on 1682 movies on a distinct scale of 1-5 and ‘0’ in the ratings indicate ratings are not provided or movie is not watched. The data sparsity level of dataset is 0.9309. At first data set is split randomly into 80% for training set and 20% for test set where training set is utilized to form the

clusters using FAH(wP) then test set is utilized for prediction using PCC and providing TOP N recommendations. To measure the quality of the approach similarly as in existing approach we have employed MAE, Precision and Recall measures. By MAE calculating Given5 ratings, Given20 rating and ALLBUT10 ratings we tried to check if the algorithm suffices the cold start problem.

We compared our algorithm with PCA-GAKM which used genetic algorithm having higher time complexity compared to firefly algorithm[13, 15] and our approach takes comparatively lesser iterations. Along with PCA-GAKM we I tried to compare our algorithm with FAH which is Firefly Agglomerative Hierarchical clustering without the prior pre-processing and the priority queues and PFAH(woP) which is Firefly Agglomerative Hierarchical clustering without the priority queues The number of iterations T=200. we repeatedly done the process by changing the K-values (Number of clusters) and Top N (Top recommendations list) values from 3 to 20. our approach outer performs PCA-GAKM when the K- values are 9 to 19 where as PCA-GAKM have good performance when the K-values are in between 12-18 this may be resulted due to the rankings of fireflies done by Firefly algorithm. Later on we set K value to 16 so as to compare our approach with existing approach for numerical experiments. For this experiments we hide the 10 ratings from the test data Prediction is applied on this hidden data using our clustering algorithm to compare the accuracy of the recommendations.

C. MAE evaluation

MAE is defined as how much deviation in recommendations compared to user specified ratings.

$$\text{MAE}= \frac{\sum_{i,j=1}^{n} \left| r_{i,j} - P(u_i, i_j) \right|}{n} \qquad (5)$$

Here, 'n' is total number of ratings in predicted values $r_{i,j}$ is actual rating given by user 'i' on item 'j', $P(u_i, i_j)$ is predicted rating of user 'i' on item 'j'.

MAE is calculated using formula (5). PCA-GAKM, FAH, PFAH(woP) and PFAH(wP). PFAH(wP) gets the good predictions when the cluster size ranges in between 15 to 20, and it will stay approximately stable till cluster size is 60. PFAH(wP) shows distinct improvement when considered with other approaches w.r.t MAE. Fig.3 shows the MAE for the different cluster sizes. FAH(wP) produces mean value as 0.7639, standard deviation as 5.0465e-03 compared to the mean value 0.7821, standard deviation 4.747e-03 in PCA-GAKM. And the t-test with the statistical significance at 1% results in comparison to PCA-GAKM shows 14.0920.

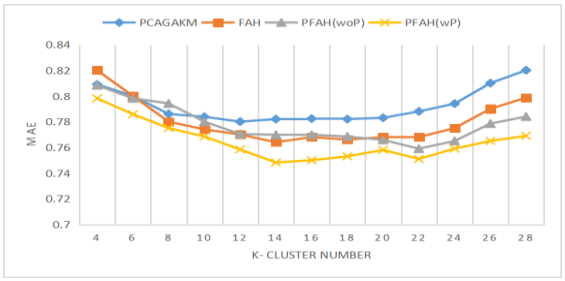

Figure 2: MAE w.r.t to cluster sizes ranging from 4-28

D. Precision

Precision is defined as how relevant the produced recommended ratings are classified.

$$\text{Precision}= \frac{|\text{interesting} \cap TopN|}{N} \qquad (6)$$

It is the ratio of the number of to the point recommendations recovered to total number of irrelevant and relevant recommendations retrieved. Precision is calculated using equation(6). Here we are fixing the cluster size as 20. Fig.4 shows the comparison between all approaches. And it is proved that FAH(wP) outer performance PCA-GAKM it implies FAH(wP) can recommend reliable interesting recommendations.

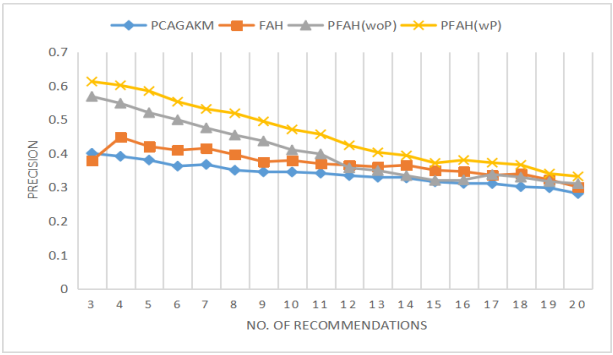

Figure 3: Precision values w.r.t Number of Recommendations

E. Recall

Recall is defined as how truly the relevant ratings are classified.

$$\text{Recall}= \frac{|\text{interesting} \cap TopN|}{|\text{interesting}|} \qquad (7)$$

Recall is the ratio of the number of to the point recommendations recovered to total number of relevant recommendations in the test set. Recall can be calculated using equation(7). By fixing the cluster size as 20. Fig.5 shows the comparative analysis between the 2 approaches. And it is proved that PFAH(wP) outer performance PCA-GAKM it implies FAH(wP) can recommend reliable interesting recommendations.

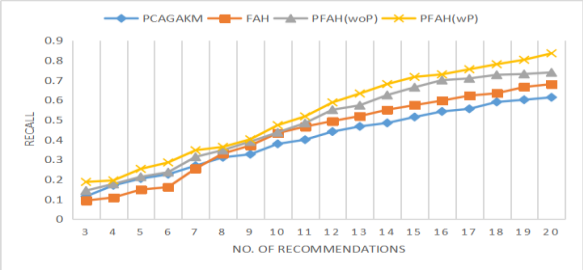

Figure 4: Recall values w.r.t Number of Recommendations

F. Cold start issue for new users with fewer ratings

Many algorithm suffers from cold start problem when the data is sparse. The user with fewer ratings may not get the recommendations based on his/ her interest. In order examine the way PCA-GAKM, FAH, PFAH(woP) and PFAH(wP). PFAH(wP) behaves when only Given 5, 20 ratings and when ALLBUT 10 ratings are given. Our algorithm outer performs in comparison with existing algorithms.

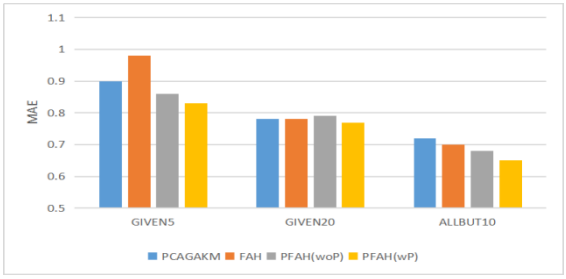

Figure 5: MAE comparison when fewer ratings are given

VI. Conclusion and Future Scope

The contribution of this work is the proposed hybrid firefly agglomerative cluster algorithm for finding recommendations. The proposed hybrid PFAH(wP) algorithm clusters the like minded users very efficiently and with in less iterations compared to PCA-GAKM. Our proposed hybrid PFAH(wP) algorithm is found to have shown better results to those obtained by the PCA-GAKM algorithm in compared w.r.t accuracy. In future one can consider including the advanced methods to increase the accuracy.

References

[1] K. Kim, H. Ahn, "A recommender system using GA K-means clustering in an online shopping market", Expert Syst. Appl. 34 (2) (2008) 1200–1209.

[2] Zan Wang, Xue Yu , Nan Feng, Zhenhua Wang, "An improved collaborative movie recommendation system using computational intelligence", Journal of Visual Languages and Computing 25 (2014) 667–675.

[3] Daniel Müllner "Modern hierarchical, agglomerative clustering algorithms".

[4] Takio Kurita "An efficient agglomerative clustering algorithm using a heap", Pattern Recognition, Volume 24, Issue 3, 1991 205-209.

[5] G. Adomavicius, A. Tuzhilin, "Toward the next generation of recommender system: a survey of the state-of-the-art and possible extensions", IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 734–749.

[6] G.Thilagavathi D.Srivaishnavi N.Aparna "A Survey on Efficient Hierarchical Algorithm used in Clustering", IJERT Vol. 2 Issue 9, September - 2013.

[7] G. Linden, B. Smith, J. York, "Amazon.com recommendations: item to item collaborative filtering", IEEE Internet Comput. 7 (1) (2003) 76–80.

[8] F. Gao, C. Xing, Y. Zhao, "An Effective Algorithm for Dimensional Reduction in Collaborative Filtering", in LNCS 4822, Springer, Berlin, 2007, 75–84.

[9] K.Q. Truong, F. Ishikawa, S. Honiden, "Improving accuracy of recommender system by item clustering", IEICE Trans. Inf. Syst. E90-D (9) (2007) 1363–1373.

[10] J. Wang, N.-Y. Zhang, J. Yin, J, "Collaborative filtering recommendation based on fuzzy clustering of user preferences, in: Proceedings of the Seventh International Conference on Fuzzy Systems and Knowledge Discovery", Yantai, Shandong, 2010, pp. 1946–1950.

[11] G. Pitsilis, X.L. Zhang, W. Wang, "Clustering recommenders in collaborative filtering using explicit trust information", in: Proceedings of the Fifth International Conference on Trust Management IFIPTM, Denmark, Copenhagen, 2011, pp. 82–97.

[12] B.M. Sarwar, G. Karypis, J. Konstan, J. Riedl, "Recommender systems for large-scale e-commerce: scalable neighborhood formation using clustering", in: Proceedings of International Conference on Computer and Information Technology, Dhaka, Bangladesh, 2002.

[13] Rogério B. Francisco, M. Fernanda P. Costa , Ana Maria A. C. Rocha, "Experiments with Firefly Algorithm",Computational Science and its Applications -ICCSA 2014, pp.227-236.

[14] Yang, X-S.: "Firefly algorithms for multimodal optimization. In: Watanabe O, Zeugmann T, (eds.) Stochastic algorithms: foundations and applications", SAGA 2009, LNCS, vol. 5792, pp. 169–78. Springer-Verlag (2009).

[15] Yang, X-S.: "Nature-Inspired Metaheuristic Algorithms", Luniver Press, Beckington, UK, 2nd edition, 2010.

[16] Lukasik, S., Zak, S.: "Firefly algorithm for continuous constrained optimization tasks". In: Chen, S.M., Ngugen, N.T., Kowalczyk, R. (eds.), ICCC 2009, Lecture notes in Artificial Intelligence, vol. 5796, pp. 97-100. Springer (2009).

[17] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, "Recommendation systems: Principles, methods and evaluation", Egyptian Informatics Journal (2015) 16, 261–273.

[18] Suryakant,Tripti Mahara, "A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment", International Multi-Conference on Information Processing-2016 (IMCIP-2016).

[19] Xiao Ma, Hongwei Lu, Zaobin Gan, Qian Zhao, "An exploration of improving prediction accuracy by constructing a multi-type clustering

based recommendation framework", Neurocomputing 191 (2016) 388–397.

[20] Yang, Xin-She, " Firefly Algorithm, Stochastic Test Functions and Design Optimisation", International Journal of Bio-inspired Computation, 2010.