iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VIII **Month of publication:** August 2020

DOI: https://doi.org/10.22214/ijraset.2020.31095

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

A Switching Hybrid Approach to Improve Sparse Data Problem of Collaborative Filtering Recommender System

Tuyet-Van Tran Thi[1,2], Thanh-Nhan Huynh-Ly[1,2]
[1]*Department of Information Technology, An Giang University, An Giang, Vietnam*
[2]*Vietnam National University Ho Chi Minh City, Vietnam*

Abstract: Collaborative filtering is a powerful technique that has been used in many recommender systems with considerable success. This technique uses the interested databases of users with items to predict what products they might like. Nevertheless, these interested databases collected very few under 10%, greatly affect the efficiency of the recommender system. Many research projects have given different solutions to solve this problem. Initially, they have made significant efficiency, improved somewhat sparse data problem, but they still have existed their shortcomings. In this paper, we research and propose a new method for improving sparse data problem, which offers higher efficiency of recommendation than the approaching traditional collaborative filtering and some other methods.
Keywords: Collaborative filtering, recommender system, sparse data, improve sparse data problem, hybrid recommender system.

I. INTRODUCTION

Recently, the recommender systems (RS) has been widely developed in many fields, especially e-commerce. However, the RS is continuously evolving and is taken an interest in many researchers because there are many issues to be researched, such as the Spare Data Problem (SDP), the Cold Start Problem (CSP), ... These issues directly affect the accuracy of the prediction. Therefore, this is one of the focus research of RS [1].

There are two main approaches in RS, which are content-based filtering (CB) and collaborative filtering (CF). In this paper, we focus only on the collaborative filtering approach of the RS, also known as the collaborative filtering system. An RS consists of *N* user U = {u_1, u_2, ..., u_n}, M products I = {i_1, i_2, ..., i_m} with evaluation matrix R = (r_ij). In the collaborative filtering system, the number of users | U | and the number of products | I | are huge. The mission of the collaborative filtering system is to predict current users based on the R matrix with most $r_{ij} = Ø$ values. However, each user gave only a few reviews for the product set, which makes the R-rating matrix with a much smaller number of $r_{ij} \neq Ø$ evaluations than $r_{ij} = Ø$. The percentage of evaluation data is meager, specifically in MovieLens (http://grouplens.org/datasets/movielens/) evaluation data, only accounting for 4.3% or the rate in EachMovie dataset accounting for 2.4%. In RS, people consider this problem as SDP.

Table 1: Example of user-item evaluation matrix

	i_1	i_2	i_3	i_4	i_5
u_1	5	Ø	Ø	4	4
u_2	Ø	4	Ø	3	5
u_3	Ø	4	5	2	3
u_4	5	Ø	5	Ø	Ø

SDP has hindered the same calculation process. For example, we need to determine the similarity between user u_4 and u_2 in Table 1. As the number of products of both u_4 and u_2 is evaluated neither overlap nor intersect, the similarity between u_4 and u_2 is calculated according to the same measurement of 0. This directly affects the training method and the predictive results because the vacant user reviews u_2 are never considered during the training process and contributed to predictions for u_4 users. Products i_1 and i_2 are the same cases. The similarity between these two products is also zero because no user reviews them on both two products.

Besides, SDP makes it less reliable to identify neighborhoods for existing users. For example, we need to predict products for user u_4 in Table 1, based on similarities, and we will calculate u_4 similar to u_1 because r [u_1, i_1] = r [u_4, i_1] = 5.

As a result, products i_4 and i_5 will be notified to u_4 because u_4 is similar to u1 that u1 "likes" i_4, i_5. However, we can calculate u_4 similar to u_3 because r [u_3, i_3] = r [u_4, i_3] = 5, i_4, i_5 will be removed from the list of products allocated to u_4 because of u_4 similar to u_3 that u_3 "dislikes" i_4, i_5. Therefore, if either u1 or u_3 are considered to be neighbors of u_4, the predicted result will become less reliable. If both u1 and u_3 are considered to be neighbors of u_4, it will cause a conflict because u1 and u_3 are not completely the same.

When the system adds new users, the users need to make some initial evaluation of a few products, and then the system will predict them for the next product. There is a similarity in new products that have not been evaluated by any user; they will not be notified to any user until a few users evaluate them. In RS, people call it a slow start problem (CSP).

II. RELATED STUDIES

There have been many researchers focusing on solving this problem to improve the efficiency of the consulting system. It can be summarized: reducing the number of dimensions, clustering, graphical representation method, hybrid method between collaborative and content-based filtering.

The most straightforward strategy to reduce the number of dimensions of the evaluation matrix is to create product clusters of user clusters and then use these clusters as basic units to generate predictions. The author Ungar & Foster [2] uses the K-median technique to cluster users and products independently, then user clusters and products are assembled to create clusters of similarly high levels following both user and product.

Xun Zhou et al. [3] proposed the use of hidden semantic detection (LSM) based on the specific value decomposition technique (SVD). However, in many cases, useful information can be lost in the process of reducing the number of matrix sizes, which makes a limitation on the result of the predictions.

Another approach makes limitation to the problem of sparse data based on the exploitation of indirect relationships on the evaluation matrix. The article [4] represents users and products as a two-sided graph (Bipart Graph Model), one side is the user set, the other is the product set, each side says from the top of the user to the top of the product, which is set if the user has purchased or appreciated for the corresponding product. It depends on the user and product relationship representation, and data is filled in the blank cells in the evaluation matrix performed by weighted spread on the two-sided graph.

Some authors believe that the information on the sparse matrix is not effective enough to provide consulting results, so it should be combined with other data sources such as more information about products and user information by hybridizing filter consulting system collaborates with the consulting system collaborating based on the content [5] or integrated with social network connections into a predictive model [6].

Some other reseachers who are even more unique base on book users' preferences to advise about the film [7]. Combining with other data sources is also a good solution to solve the sparse matrix problem, but this method also has many difficulties when analyzing and selecting the appropriate data source.

We propose to combine two traditional methods of collaborative filtering, which are user-based and product-based (Item-based), according to the conversion hybrid methods described in item 3.

III.RESEARCH METHODS

In the process of researching methods to improve the system efficiency, we found that the information from the assessment matrix is simple but essential and needed to be thoroughly explored. Considering the two approaches of collaborative filtering, which are user-based and product-based, each of them has advantages and disadvantages that can complement each other if they know how to combine them in the right stages.

For example, a new user is added to the system. If a user-based collaborative filtering method is used, he or she has absolutely no evaluation yet; as a result, this method cannot make a prediction. If we combine with the product-based method at this stage, it can completely overcome the situation of the new user because now we base on similar products to make predictions. In contrast, if a new product is added to the system, we can still make predictions based on the user. With the switch hybrid method, we can choose the method according to each specific case to promote the maximum effectiveness of the methods. Based on the research of existing methods, we propose a new method, which is the switch hybrid consulting method between user-based consultations and product-based consultations. According to the architecture, as shown in Figure 1, the essence of the method is thoroughly utilized user-item matrix evaluation data to find out the conversion conditions between user-based and item-based methods to bring the most optimal effect.

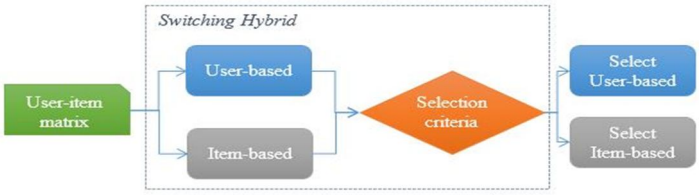


Figure 1: A proposed hybrid model

Calling φ_u and φ_i respectively to represent the same threshold of two users and two products. φ_u and φ_i are values between -1 and 1, which represents the similarity of two users or two products. For example, φ_u = -1 means the lowest similarity, and two users have completely opposite interests. Another example, φ_u = 0, means that two users are not similar, and the interests are not the same. The last instance is φ_u = 1, which means that two users have completed the same interests. The value of φ helps us determine the nearest neighbor set of users or products.

Calling S (u) is a greater set of similarities than the threshold φ_u of user u and neighbor u, we have:

$$S(u) = \{v | sim(u, v) > \varphi, \forall v \neq u\}, \quad (1)$$

Calling S (i) is a greater set of similarities than the threshold φ_i of product i and similar products i, we have:

$$S(i) = \{j | sim(i, j) > \varphi, \forall j \neq i\}, \quad (2)$$

It is now possible to determine the conditions to convert the user-based method or item-based method.

It is considered to the first case is that both two sets S (u) and S (i) are empty, which means that the neighbors of both u and i are not enough to make predictions, so we will make predictions based on the whole evaluation matrix by the value average of all the evaluations on the matrix symbolized $\bar{r}$.

$$P_{u,i} = \bar{r}, \quad (3)$$

The second case is the set S (u) is empty, and the set S (i) is not empty, which means that the number of user u 's neighbors is not enough to give the result of the prediction, so we will base on the neighbor set of the product i to give the predictive results following formula (3.4) but select only neighbors in S (i) set.

$$P_{u,i} = \bar{r}_i + \frac{\sum_{j \in I}(r_{u,j} - \bar{r}_j) * sim(i,j)}{\sum_{i \in I} |sim(i,j)|} \quad where\ sim(i,j) \in S(i), \quad (4)$$

In contrast, when the set S (i) is empty, and the set S (u) is not empty, the number of neighbors of product i is not enough to give a prediction result, so we have to base on user u's neighbor set to give the results of the prediction following formula (3.2).

$$P_{u,i} = \bar{r}_u + \frac{\sum_{v \in U}(r_{v,i} - \bar{r}_v) * sim(u,v)}{\sum_{v \in U} |sim(u,v)|} \quad where\ sim(u,v) \in S(u), \quad (5)$$

The last case is that both sets S (u) and S (i) are not empty. At this time, we base on both two neighboring sets to give prediction results. The result is calculated following formula:

$$P_{u,i} = \left(\frac{|S(u)|}{|S(u)|+|S(i)|}\right)\left(\bar{r}_u + \frac{\sum_{v \in U}(r_{v,i} - \bar{r}_v) * sim(u,v)}{\sum_{v \in U} |sim(u,v)|}\right) + \left(\frac{|S(i)|}{|S(u)|+|S(i)|}\right)\left(\bar{r}_i + \frac{\sum_{j \in I}(r_{u,j} - \bar{r}_j) * sim(i,j)}{\sum_{i \in I} |sim(i,j)|}\right) \quad (6)$$

Therefore, we can summarize the conditions to choose two methods, which are user-based and item-based, shown in Figure 2.

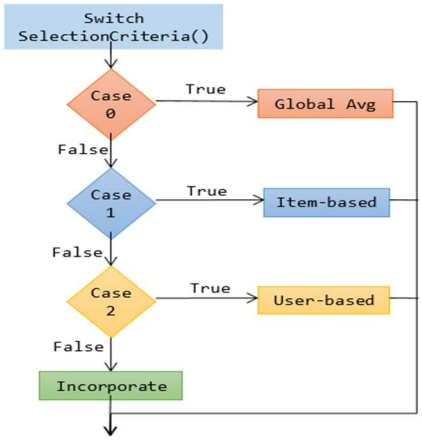


Figure 2: Switch hybrid algorithm model

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

IV. RESULTS AND DISCUSSION

A. Experimental Method

The proposed algorithm has experimented on the MovieLens data set. This is a data set that is researched by the community about consulting systems used popularly [8, 9]. MovieLens is a database built by the GroupLens research group of Minnesota University. The GroupLens group has gathered many data sets with different sizes to serve for researching consulting systems such as MovieLens 100k (ML_100k), MovieLens 1M (ML_1M), MovieLens 10M, MovieLens 20M, ... In the experimental scope of the project, only two sets of MovieLens 100k and MovieLens 1M are used for testing. The two data sets are described in detail in Table 2.

Table 2: MovieLens data description

	ML_100K	ML_1M
Users	943	6040
Movies	1682	3900
Ratings	100.000	1.000.000
Sparsity	93.7%	95.8%
Value of ratings	1-5	1-5

We have used MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) to evaluate the efficiency of the proposed algorithm. These two measures are commonly used in research on consulting systems generally and collaborative filtering particularly. MAE calculates the deviation of the evaluation prediction of the algorithm and the actual evaluation (evaluation of the test set).

$$MAE = \frac{\sum_{u \in U} \sum_{i \in testset_u} |rec(u,i) - r_{u,i}|}{\sum_{u \in U} |testset_u|}, \qquad (7)$$

In formula 7, MAE calculates the absolute average deviation between the consulting prediction rec (u, i) and the actual assessment value r_ (u, i) for all users u ∈ U and all the product i, which belong to the test set $testset_u$.

Like MAE, RMSE is also used to measure the accuracy of the prediction, but it emphasizes more about larger deviations, which is calculated by formula 8.

$$RMSE = \sqrt{\frac{\sum_{u \in U} \sum_{i \in testset_u} (rec(u,i) - r_{u,i})^2}{\sum_{u \in U} |testset_u|}}, \qquad (8)$$

The values of MAE and RMSE are inversely proportional to the accuracy of the prediction, which means the smaller MAE and RMSE are, the higher accuracy of the prediction is. We have used the LibRec library in the installation process of comparison algorithms and proposed algorithms to ensure the correctness and science in the experimental method. LibRec is a library that has been developed by Java to serve the research of the consulting system in the library. There have been many algorithms of the authors in the consulting system field published. To experiment, we have installed three algorithms, including user-based and item-based, which are the two algorithms of traditional collaborative filtering and the switch hybrid method.

B. Experimental Results

We conducted experiments on 5-fold of 100k MovieLens data sets with two measurements MAE and RMSE, and then calculated the value average. Experiments were compared by three methods: user-based, item-based, and the switch hybrid.

After experimenting, we found that the results of the switch hybrid algorithm have MAE and RMSE values that are much smaller than the two traditional algorithms (user-based and product-based filtering) (Figure 3 and Figure 4). This shows the proposed algorithm has higher predictive accuracy.

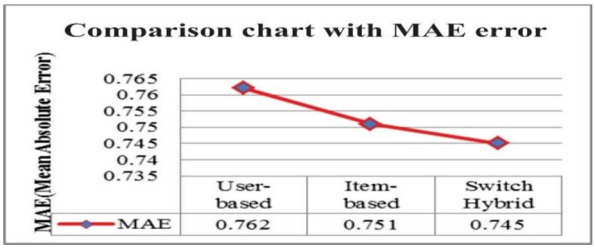


Figure 3: Comparison chart with MAE error

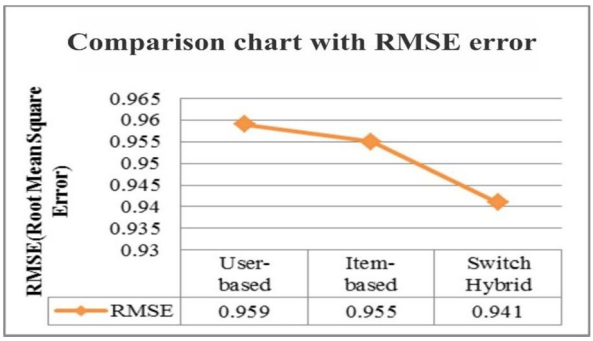


Figure 4: Comparison chart with RMSE error

As our experiments are installed and developed on Librec.net library, it is easy to compare with some algorithms installed in this library. The results in Table 3 show that the predictive results of the switch hybrid method, which we propose are quite better than the UserCluster, ItemCluster, URP, BUCM, GPLSA, etc. methods. This experimental result is carried out on the same MovieLens 100k data set.

Table 3: Comparison of other methods

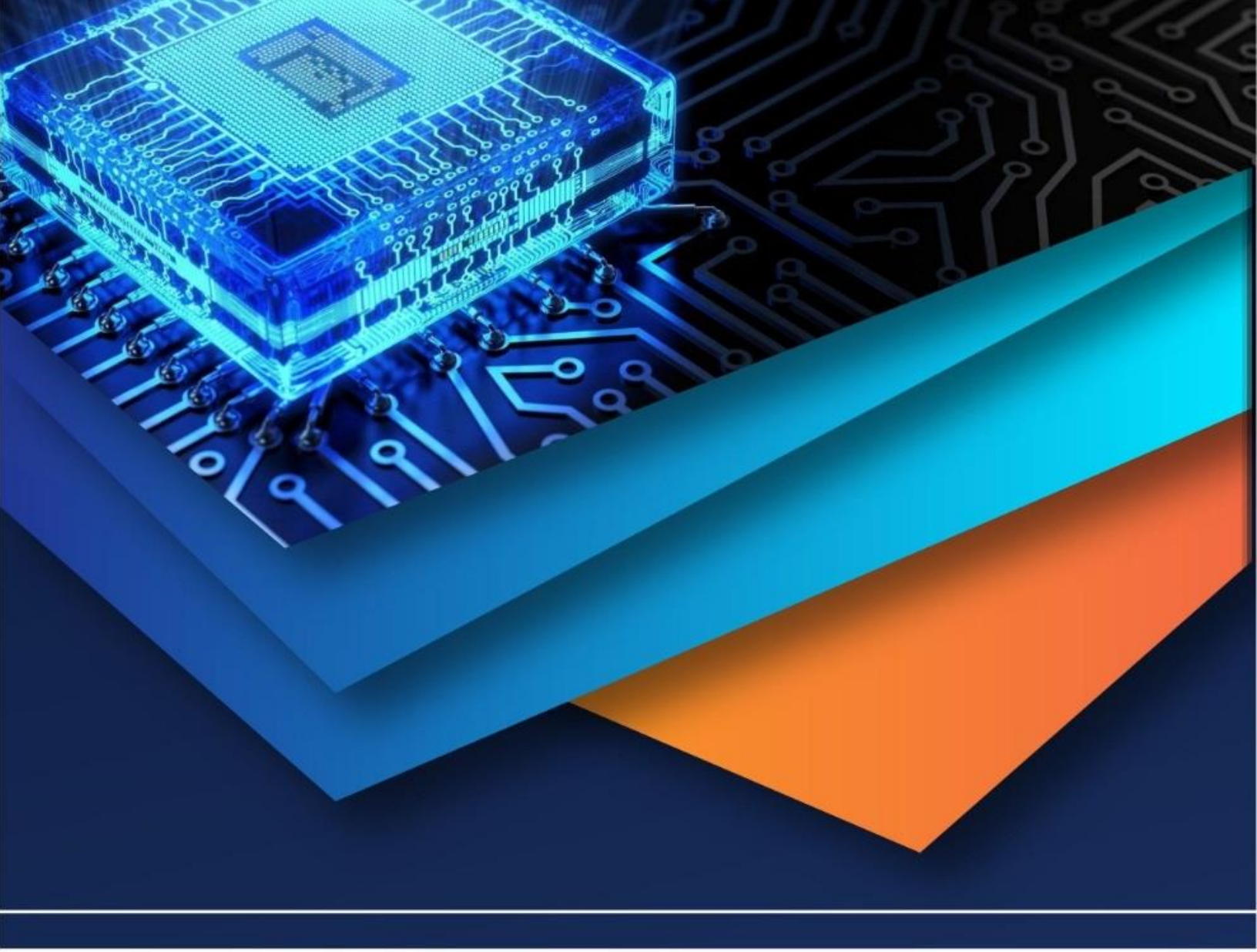
Methods	MAE	RMSE
UserCluster	0.839	1.048
ItemCluster	0.820	1.023
URP	0.792	0.984
BUCM	0.847	1.038
LDCC	0.743	0.937
Switch Hybrid	0.738	0.938
RegSVD	0.739	0.936

V. CONCLUSIONS

We have presented a switch hybrid method between user-based collaborative filtering and product-based collaborative filtering in a new way. Experimental results on the MovieLens dataset show that the proposed method gives more accurate predictions than the two traditional filtering methods and some other methods in the Librec library. The proposed method sought to improve the accuracy of the prediction on sparse matrices to provide more accurate predictive results, enhance the efficiency of the consulting system rather than "thicken" the sparse matrix. In the next study, we will develop an extra "thickening" step by adding a "retraining" stage. It means when giving user evaluation prediction to a product, we can choose the "reliable" values to enter the training that supports the next prediction with the hope of improving the predictive results of this switch hybrid method.

REFERENCES

[1] C. Desrosiers and G. Karypis, "Solving the sparsity problem: Collaborative filtering via indirect similarities," in Technical Report. Department of Computer Science and Engineering University of Minnesota 4-192 EECS Building 200 Union Street SE Minneapolis, MN 55455-0159 USA, 2008.
[2] L. H. Ungar and D. P. Foster, "Clustering methods for collaborative filtering," in AAAI workshop on recommendation systems, 1998, pp. 114-129
[3] Zhou, Xun & He, Jing & Huang, Guangyan & Zhang, Yanchun. (2014). SVD-based incremental approaches for recommender systems. Journal of Computer and System Sciences. 81. 10.1016/j.jcss.2014.11.016.
[4] N. D. Phuong and T. M. Phuong, "A graph-based method for combining collaborative and content-based filtering," in PRICAI 2008: Trends in Artificial Intelligence, Springer, 2008, pp. 859-869.
[5] L. Pasquale, G. d. Marco and S. Giovanni, "Content-based Recommender Systems: State of the Art and Trends," in Recommender Systems Handbook, Springer, 2011, pp. 73-105.
[6] Huynh-Ly Thanh-Nhan, Le Huy-Thap and Nguyen Thai-Nghe. 2017. Toward integrating social networks into Intelligent Tutoring Systems. In proceedings of the 2017 International Conference on Knowledge and Systems Engineering (KSE 2017), pp.112-117, ISBN 978-1-5386-3576-6
[7] B. Li, Q. Yang and X. Xue, "Can Movies and Books Collaborate? Cross-Domain Collaborative Filtering for Sparsity Reduction," in IJCAI, 2009, pp. 2052-2057.
[8] A. Gunawardana and C. Meek, "Tied boltzmann machines for cold start recommendations," in Proceedings of the 2008 ACM conference on Recommender systems, ACM, 2008, pp. 19-26.
[9] D. Jannach, M. Zanker, A. Felfernig and G. Friedrich, "Hybrid recommendation approaches," in Recommender systems: an introduction, Cambridge University Press, 2014, pp. 124-142.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY