I.J. Modern Education and Computer Science, **2018, 11, 36-42**
Published Online November 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2018.11.05

Cross-Domain Recommendation Model based on Hybrid Approach

Smriti Ayushi
PES University/Department of Computer Science, Bangalore, 560085, India
Email: smriti.ayushi@gmail.com

V R Badri Prasad
PES University/Department of Computer Science, Bangalore, 560085, India

Received: 21 August 2018; Accepted: 08 September 2018; Published: 08 November 2018

***Abstract*—**The demand of recommendation has aroused severely since there are huge number of choices available and the end user desires to extract information in least time and with high accuracy. The traditional recommendation systems generate recommendations in the same domain but now cross domain recommendations are gaining importance. The cross domain recommendations address well the limitations of single domain analysis such as data sparsity and cold start problem. Under this research work cross domain recommendation model is designed based on the study of various supervised classification algorithms. 3 domains are under consideration music, movie and book. Model is capable of generating one to many cross domain recommendations exploiting movie domain knowledge to generate recommendations for books and music. Data is collected through survey and data pre-processing has been performed. Study is carried out over K-Nearest Neighbor, Decision Tree, Gaussian Naïve Bayes and Support Vector Machine classifiers and also over majority voting Ensembling, cross validation and data sampling by applying these classifiers to choose the best classifier to form the base of content-based recommendation. Recommendation model uses a hybrid approach of combination of content-based recommendation, user to user collaborative filtering and personalized recommendation techniques. The model perform Twitter sentiment analysis over the recommended entities generated by the model to help the user in decision making by knowing the positive, negative and neutral polarity percentage based on tweets done by people. The designed model achieved good accuracy on testing.

***Index Terms*—**K-Nearest Neighbor (KNN), Decision Trees (DT), Support Vector Machines (SVM), Gaussian Naïve Bayes (GNB), Content-based Filtering, Collaborative Filtering, Personalized Recommendation, Cross-Domain Recommendation, Sentiment Analysis.

I. Introduction

Recommender system is a machine learning technique that facilitates learning of users' choice of products and services which vary with the users' characteristics and hence recommends them a product or service accordingly. Today the major issue faced by e-industry is availability of huge data for anything. It has become difficult for the user to come across all of the items he/she might like. In addition to this customers have busy life and desire to spend least time on searching vaguely and also they want quick and useful recommendations at the same time based on their interests. So there comes an efficient requirement of a medium which keeps suggesting user on his/her interest basis so that task of customers becomes easy and remain interesting. Thus by performing various schemes of data analysis and user's analysis, the required information specific to an individual can be extracted from huge number of available choices and user can come across all what he/she might like. Cross-domain recommendation is an approach to do knowledge transfer. Using information from domains where there is sparse information in a domain where knowledge is scarce. Under this research an efficient cross domain recommendation model is tried to be achieved using hybrid approach leveraging all the three recommendation techniques: Content-based filtering, collaborative filtering and personalized recommendation. For achieving accurate results an exhaustive study over supervised classifiers has been done to choose the best one amongst all.

Data is collected from the surveys undertaken, data cleaning, all the required pre- processing steps has been done along with converting the string values into numeric data and further various classifier algorithms are applied. Bollywood movies and songs data are taken into consideration under the study and books in both Hindi and English language. Data has been collected on a large scale having many features and big data set is considered

for better analysis, then noise and unwanted fields were removed for better outcomes. Classifiers studied include: K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machines (SVM) and Gaussian Naïve Bayes (GNB). Accuracy score outcomes from all the algorithms are compared and with the Decision Tree classifier accuracy is highest. So decision tree algorithm was fixed for performing cross recommendations for better outcomes. Majority voting Ensembling technique is also performed with all the four classifiers. 4-fold Cross-validation technique is also performed in which dataset is divided into 4 partitions. Out of the 4 partitions, over 3 partitions training is performed and remaining one partition is kept for validation. As the result of experimental study decision tree gave best accuracy over dataset.

Cross domain recommender is designed to deal with data sparsity and cold start problem which are mostly faced by single domain recommenders. Using movie as input domain, recommendations are given for output domains: book or music as per user choice. Generated output from model consist of recommendations in the selected output domain. The output recommendation list is the integrated result from all three techniques: content-based, collaborative and personalized. Over the recommended list in output domain received from model, further tweets are extracted and analyzed to get user polarity over the recommended output entities in terms of negative, positive and neutral polarity to help user in decision making.

II. Literature Review

The undergone research under this area illustrates that since interests reflects a wide variety so for providing user specific recommendations demands leveraging their preferences from all the available domains or systems. This idea of using diverse data can lead to more accurate and efficient recommendations and even user personalized recommendations can be given as multiple domains are getting involved and cold start problem and data sparsity problem will also be dealt. Thus cross domain recommenders works on the principle of generating or enhancing recommendations in a domain which has sparse knowledge by utilizing knowledge from the domains where user preferences and interests are available along with user information. Source domain can be single as well as multiple. The popular recommender systems are studied like amazon, Facebook, YouTube, Netflix, etc. Existing recommenders mostly use the techniques of item-to-item or user-to-user collaborative filtering, personalized recommendation or the combination of both. The proposed recommender system is designed using hybrid approach based on the content based recommendation technique, collaborative filtering technique and personalized technique. Recommendations will be based on the integrated learning from all the three techniques with the aim of achieving efficient and higher accuracy recommendations. Cross recommendation is least researched with supervised learning algorithms, mostly it has been achieved by clustering (unsupervised learning) models or tensor decomposition technique or rank matrix. This research work uses supervised learning algorithms for analysis.

III. Motivation of this Research

The research gained its motivation from the fact that in today's era e-industry, other commercials and customers require recommendations systems to save time and resource. The requirements of e-industry, in enhancing their business they want to expose most of the items to the customers as per their interest areas and likings so that there sale and hence revenue increase. Similarly customers also face challenge due to the expansion of industries and business and wide scope and options available to them. In addition to this customers have busy life and desire to spend least time on searching vaguely and also they want quick and useful recommendations at the same time based on their interests. So there comes an efficient requirement of a medium which keeps suggesting user on his/her interest basis so that task of customers becomes easy and remain interesting. Thus by performing various schemes of data analysis and user's analysis, the required information specific to an individual can be extracted from huge number of available choices and user can come across all what he/she might like. So that user can choose well. Under this work an efficient cross domain recommendation model is tried to be achieved using hybrid combination of content-based, collaborative filtering and personalized recommendation technique for achieving accurate results an exhaustive study over supervised classifiers has been done to choose the best among Decision Tree, K-Nearest Neighbor, Gaussian Naïve Bayes and Support Vector Machines classifier algorithms .

IV. Methodology

The framework steps used for this analysis is depicted in below figure. Different steps had their own important role. Fig.1. depicts the order of steps followed to achieve the cross-recommendation model and lastly sentiment analysis over tweets made by people over the generated recommendations from the model.

A. Data Collection

Bollywood movies, Bollywood music data and books is collected through the survey of up to 35-40 users and data size is 4500+ with the following features: UserID , Name, Age, Gender, Occupation and Ratings provided by user and EntityName (Movie, Book, Music) , EntityID, Genre, Owner (movie director, music artist, book author). Age attribute has wide range and data is collected from people over various occupations such as student, doctor, housewife, IT professional, PhD scholar, business, finance, pilot, service, professor, associate analyst and retired. When it comes to entity Genre, many

different genres are considered so that user choice can be properly taken care of and ratings can be given accordingly which includes Romantic, Comedy, Action, Horror, Social Drama, Kids, Suspense, and Autobiography under movie genre and Thriller, Suspense/Mystery, Romantic, Classics, Autobiography under book and Romantic, Ghazal, Devotional, Rock/Disco, Classical, Folk under music genre.

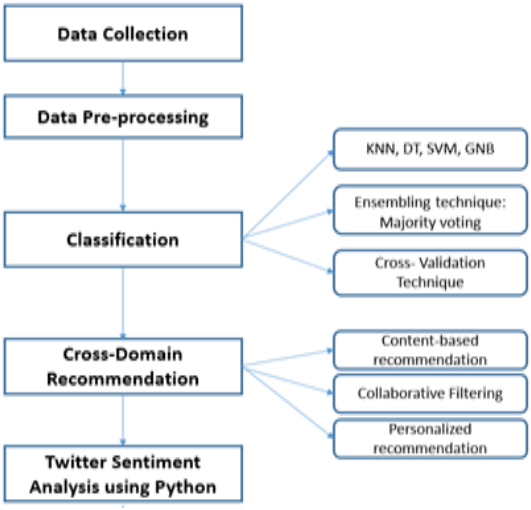

Fig.1. Framework depicted with flow diagram

B. Data Pre-Processing

Data is formatted by removing all the unnecessary characters followed by converting the data to numeric from string format for applying classifier algorithm. Then there is a need of converting .xlsx file into .csv file because the classifier functions can take csv files as parameters. Several python scripts are written to format the data for further analysis.

C. Classification

In the analysis following four classifiers are studied: K-Nearest Neighbor, Gaussian Naïve Bayes, Support Vector Machines, and Decision Tree. These classifier algorithms are supervised learning methods in which labels are assigned and target class is predicted for new data points. All the classifier algorithm implementation codes work on the concept of dividing the dataset into test and training set based on the given split ratio and then model fitting, prediction and calculating accuracy score by finding the ratio of the predicted to the actual. All the algorithms are applied on the dataset to determine which classifier will give best accuracy over the dataset as classification is the base of content-based recommendation.

D. Cross-Domain Recommendation

Python scripts are written to perform cross domain-recommendation like if we enter movies name as input to the system to reflect our choices and interests. We can get list of recommendation for books and music. Here movie will be treated as input domain and exploiting information from this domain we can get recommendations in output domain: book and music whatever user wish to. Three recommendation techniques are performed under this project: content-based filtering, collaborative filtering, and personalized recommendation.

Content-based Recommendation

Content based filtering uses the content similarity concept. Similarity is found over the dataset. User has enter movie choices and his ratings is also available in the data record and genre of movie is also there in the dataset. Also data is available about music and books along with other attributes like genre, age, etc. Hence correlation is found and based on movie inputs recommendations are given in books and music.

Content-based filtering Algorithm Implementation

Step 1: Classifier model is formed for classification.
Step 2: Fit data onto model.
Step 3: As per user's interest choice entered, classes are predicted on similarity of content basis.
Step 4: Output of (3) can be single as well as multiple classes.
Step 5: Recommendation ( )
- i. Load dataset and scan.
- ii. For all the predicted class code will search for other instances within the same class and with highest rating. Hence mapping in same user interest space found.
- iii. Repeat ii and get entities for output domains
- iv. Output recommended list in cross domain.

Content-Based & Collaborative Filtering Recommendation Technique

User will enter friends name at the time of login then all the friends' data will be analyzed from the dataset and on the basis of friends liking and interest user will get recommended list in the output domain. This analysis is based on the fact that friends are like-minded people and probably have same interest in movies, books or music. User should enter friend names from the dataset itself so that his friend's data can be fetched and analyzed. Recommendation results are the outcome of both content based and collaborative filtering applied.

Collaborative Filtering Recommendation Algorithm Implementation

Step 1: Friend List is fetched from configuration file.
Step 2: Friend name is checked in the dataset and the highest rated entities by the friends are considered.
Step 3: Repeat (2) and entities are appended in the list.
Step 4: Output list of recommendations through CF is received.
Step 5: Intersection of the output recommendations from content based and Collaborative Filtering.

Content-Based, Collaborative Filtering & Personalized Recommendation Technique

Personalized recommendation is achieved by using users past interest in database. For the user who had

already interacted with the cross domain recommender model can make use of personalized recommendations as well. Based on the interest in store new recommendations are made in the output domain chosen. Personalized recommendation is helpful since recommendations are given as per specific user based on his own earlier choices. Recommendation results received from the model are the outcome of: content based, collaborative filtering and personalized recommendation technique applied.

Personalized Recommendation Algorithm Implementation

Step 1: Database is loaded.
Step 2: Get users past interest in output domain from the database.
Step 3: Search for genre in entities of users past interest amongst output domain.
Step 4: From database fetch entities names with similar Genre.
Step 5: Get union of the recommendations given from content + Collaborative Filtering and recommendations from personalized technique.
Step 6: Output the recommended list after integration of Content + CF+ personalized.

Integration of Recommendation Techniques

Output from all the recommendation techniques is combined by taking the union of recommendations received after the implementation of each recommendation algorithm: content-based, collaborative filtering, and personalized. Output recommended list obtained for the output domain selected is more accurate since all the technique are utilized. Taking union of recommendations from content, collaborative filtering and personalized also helps to avoid making the recommendation localized to a user, union is taken so that content and user-to-user collaborative filtering variety will also be provided.

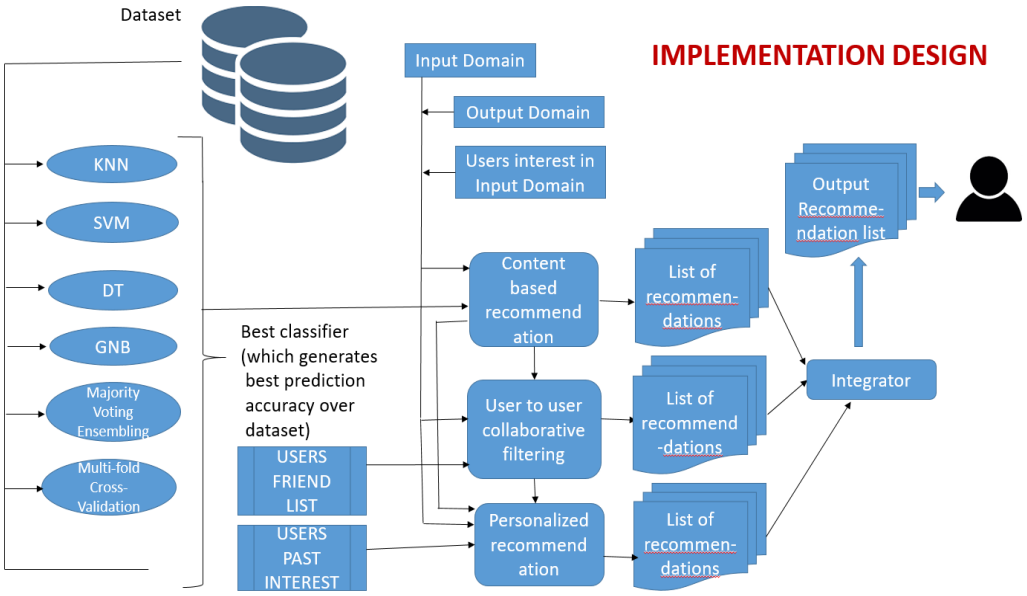

Fig.2. Cross Domain Framework Implementation

E. Twitter Sentiment Analysis Using Python

On the recommendations which are given as an output of cross-recommender model in the desired output domain i.e. can be books/music/movie, this is done to give the proposed model a real-time interaction with the social site such as twitter. Twitter sentiment analysis is performed using python. The aim is to give the user insight for the recommendations he has got from recommender model that what polarity that entity has got from real-time sentiment analysis of twitter. Out of recommended list entities will be picked and its polarity will be given from sentiment analysis over tweets that how people are liking, disliking or neutral about that entity. Polarity will be in terms of percentage of positive, negative and neutral votes or poles. This will help the user who is using designed cross domain recommender model to reach at a conclusion and helps user in decision

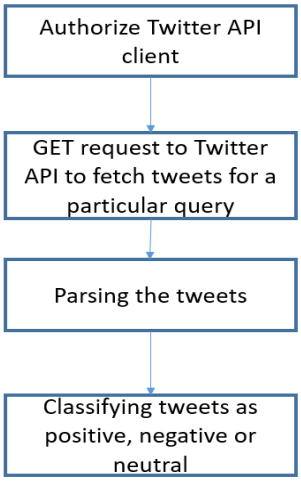

Fig.3. Steps followed in Twitter sentiment analysis using python

making whether he/she should go for that entity or not. Twitter sentiment analysis is performed through python scripts using Tweepy, Textblob and re libraries.

V. Experiment and Results

A. Experiment with Classifiers: K-NN, DT SVM, and GNB

In the KNeighborsClassifier( )function, for KNN classier algorithm implementation which takes as input 3 parameters n_neighbors = k_value, weights and algorithm. All the parameters are varied and the accuracy output is tested and compared to select best suited parameters and the comparative analysis values are listed in the table format. K-value is fixed as 3 because it is tested practically and theoretically, at k=3 maximum accuracy is achieved. Weights can be uniform, distance or even any customized user defined function can also be used here. Algorithm can take following values: auto, ball_tree and brute.

Table 1. Accuracy score on variation of 3 parameters that is passed to the KNeighborsClassifier( ) function for K-NN algorithm implementation.

N-neighbor (K-value)	Weights	Algorithm	Accuracy score in %
3	Uniform	auto	79.10
3	Uniform	ball_tree	78.95
3	Uniform	brute	79.39
3	Distance	auto	81.53
3	Distance	ball_tree	81.38
3	Distance	brute	81.82

In the Decision Tree classifier algorithm, classifier function, model fitting, model prediction and accuracy score is determined using the two criterion: Gini Index and Entropy (Information Gain) for selecting the best out of the two attribute selection criterion and keeping it fixed for further cross domain recommendation procedure. Though almost same accuracy in percentage is achieved from both of the criterion.

Table 2. Accuracy score determined by applying different attribute selection criterion on dataset in Decision Tree algorithm implementation

Criterion	Accuracy Score in %
Entropy	93.52
Gini Index	93.52

In the SVM algorithm three classifier model types is tested SVC with linear kernel, Linear SVC and SVC with RBF kernel to determine best model out of the three. The percentage accuracy score determined by each model is listed in the table below. SVC with linear kernel gave better accuracy score on our dataset.

Table 3. Output result of different SVM Classifier algorithm model over dataset.

SVM model	Accuracy Score in %
SVC with linear kernel	91.68
Linear SVC	84.03
SVC with RBF kernel	51.58

Gaussian Naïve Bayes is chosen because of the attributes in the dataset are real-valued and have normal distribution. Accuracy determined from Gaussian Naïve Bayes classifier algorithm is listed in the table below.

Table 4. Output result of GNB classifier algorithm over dataset.

Classifier	Accuracy Score in %
Gaussian Naïve Bayes	77.92

After performing analysis over models and criterion within individual classifiers. A comparative analysis is done over the accuracy determined from all the 4 classifiers under study. Graph is created using Matplotlib package in python to represent analysis graphically. On applying SVM classifier over the dataset 91% accuracy was achieved, 93% with Decision Tree, 77% from Naïve Bayes and 79% on the application of K-NN algorithm choosing k value as 3.

Decision Tree is found to give the best accuracy on our dataset. Hence further for content-based recommendation decision tree classifier is used for classification learning model.

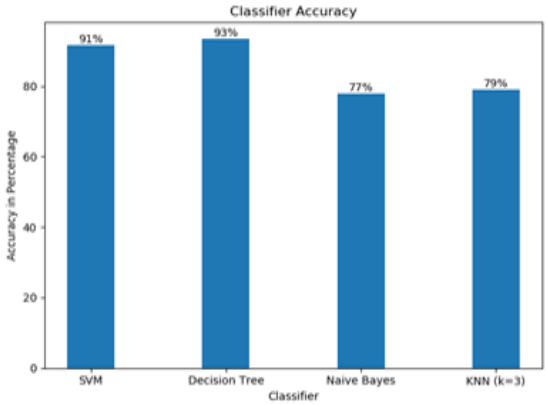

Fig.4. Accuracy score comparison of different classifier algorithms experimented

B. Experiment with Majority Voting Ensemble Learning

Firstly entire dataset is divided into test and train split by giving split ratio of 0.3 which means (70% is training set and 30% is testing set). On training data all 4 classifiers: K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Gaussian Naïve Bayes (GNB) are applied and then for the test set classes are predicted from all classifiers and for each instance (tuple) count is kept that how many classifiers have predicted class1, how many predicted class2 and so

on. Finally the highest count of whichever class comes that will be assigned at last. This application is termed as majority voting because votes from all the classifier is considered and then class is assigned. Ensembling because learning from all the classifiers is aggregated to determine the results.

Table 5. Output result of Majority Voting Ensembling technique over dataset

Technique	Accuracy Score in %
Majority Voting Ensembling	69.16

C. Experiment with Cross-Validation Technique

Dataset is divided into 4 partitions. Out of 4 partitions: Over 3 partitions training is performed. One partition is kept for validation. Above step is repeated in rotation (for 4 rotations). Every time changing the partition to be validated and remaining 3 partitions are trained. Average of accuracy is taken at last, for all the 4 rotations. The mentioned procedure is performed for each of the classifier: DT, KNN, SVM and GNB to compare the accuracy achieved over dataset.

Table 6. Output result of Cross-Validation technique with 4 classifier algorithms over dataset

Cross-Validation	Accuracy Score in %
K-Nearest Neighbor	69.27
Decision Trees	94.06
Support Vector Machines	92.62
Gaussian Na ïve Bayes	78.80

D. Experiments with Recommendation Techniques

Accuracy of designed cross-domain recommender framework is calculated by keeping some part of the dataset aside as testing data and giving it as an input to the system (proposed recommender model) without exposing the rating given to the entities by the user, then from the designed model recommendations are given. Further comparing the predicted ratings and actual ratings given from the database and thus hit and miss can be evaluated. This can be performed over all the instances of the test data and average can be taken hence finally accuracy score of the system is calculated. The recommendations can be considered as above 3 rated entities and hence output recommended list from model can be compared with actual rating in the dataset on the lines that if the same recommended entity has got actual rating in the database above 3 it is a hit whereas below 3 can be considered a miss. Performing this process helped in determining accuracy of the model and understand the results and further improve upon or explore as a future work of this research. Accuracy is improved in the hybrid of all the 3 techniques as integrated learning is performed their leveraging content similarity, like-minded friends choices and personal interest of user.

Table 7. Comparison of the accuracy results achieved for output domains from designed model for different combinations of recommendation techniques

Recommendation Techniques	Books	Music
Content Based Filtering	80.6 %	76.6 %
Content + Collaborative Filtering	81.3 %	81.6 %
Content + Collaborative +Personalized	84.3 %	84.3 %

VI. Conclusion

Exhaustive classifier study and experimentation done on dataset and best is chosen out of all. Direct recommendation systems are readily available but cross recommendations have better judgement and greater scope of research. Integration of several domains are further capable of generating higher accuracy in suggestions. Single domain recommenders usually suffer from data sparsity problem and cold start problem. If any new user comes who wants to get recommendations there is not enough data previously available to recommend him over any domain but in the case of multiple domains input can be taken from multiple domain thus recommendation is possible in domain with scarce information. Framework uses content based, collaborative filtering and personalized technique to yield high accuracy in recommendations with the integrated learning done from all the methodologies being used. Twitter sentiment analysis helps in enhancing application by providing help to user in decision making over the recommendations received. Model achieves good percentage accuracy score on testing.

References

[1] Meng Jiang, Peng Cui, Xumin Chen, Fei Wang, Wenwu Zhu and Shiqiang Yang,"Social Recommendation with Cross-Domain Transferable Knowledge", *IEEE*, 2015.

[2] SharuVinayak, Richa Sharma and Rahul Singh, "Cross Domain Recommender Systems: A Review" *IJRCCT*, Vol 5, Issue- 6, June 2016.

[3] Enkh-AmgalanBaatarjav, JedsadaChartree, and ThiraphatMeesumrarn, "Group Recommendation System for Facebook", *ACM*, November 2008.

[4] Greg Linden, Brent Smith, and Jeremy York, "Amazon.com Recommendations Item-to- Item Collaborative filtering", *IEEE Computer Society*, 2003.

[5] James Davidson, Benjamin Liebald, Junning Liu, PalashNandy and Taylor Van Vleet, "The YouTube Video Recommendation System", ACM Journal, 2010.

[6] S.EphinaThendral and C.Valliyammai, "Clustering Based Transfer Learning in Cross Domain Recommender System", *IEEE Eighth International Conference on Advanced Computing (ICoAC)*, 2016.

[7] Hongxing MA, JianpingGou, "Sparse Coefficient-Based k-Nearest Neighbor Classification", IEEE, July, 2017.

[8] Zhen-Yu Chen, Zhi-Ping Fan and Minghe Sun, "A SVM Ensemble Learning Method Using Tensor Data: An Application to Cross Selling Recommendation", IEEE, 2015.

[9] Vivek kumar, Krishna Mohan Shrivastva and Shailendra Singh, "Cross Domain Recommendation Using Semantic Similarity and Tensor Decomposition", *ELSEVIER in International Conference on Computational Modeling and Security* (CMS 2016)2016.
[10] Shulong Tan, Jiajun Bu, Xuzhen Qin, Chun Chen, Deng Cai, "Cross domain recommendation based on multi-type media fusion", *ELSEVIER*, 2014.
[11] SnehaKhatwani and Dr. M.B. Chandak, "Building Personalized and Non Personalized Recommendation Systems", *International Conference on Automatic Control and Dynamic Optimization Techniques, IEEE,* Pages: 623 – 628, 2016.
[12] Sun Lin, "E-Commerce Personalized Recommendation System Based on Web Mining Technology Design and Implementation", *Intelligent Transportation, Big Data and Smart City (ICITBS), International Conference, IEEE,* 2015.
[13] Julia Hoxha, Peter Mika and Roi Blanco, "Learning Relevance of Web Resources across Domains to make Recommendations", *12th International Conference on Machine Learning and Applications*, IEEE, 2013.
[14] Zihan Zhang, Xiaoming Jin, Lianghao Li, Guiguang Ding and Qiang Yang, "Multi-Domain Active Learning for Recommendation", *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (AAAI-16).
[15] Rich Caruana and Alexandru Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms", *ICML 06 Proceedings of the 23rd international conference on Machine learning, ACM,* Pages 161-168 Pittsburgh, Pennsylvania, USA- June 25-29, 2006.
[16] Erfan Ahmed, Md. Asad Uzzman Sazzad, Md. Tanzim Islam, Muhitun Azad, Samiul Islam and Dr. Mohammad Haider Ali, "Challenges, Comparative Analysis and a Proposed Methodology to Predict Sentiment from Movie Reviews Using Machine Learning", *International Conference On Big Data Analytics and computational Intelligence (ICBDACI), IEEE,* 19 October 2017.
[17] Prerna Mishra, Ranjana Rajnish and Pankaj Kumar, "Sentiment Analysis of Twitter Data: Case Study on Digital India", *IEEE,* 16 February 2017.
[18] F.O. Isinkaye, Y.O. Folajimi and B.A. Ojokoh, "Recommendation systems: Principles, methods and evaluation", *ELSEVIER*, Volume 16, Issue 3, Pages 261-273, November 2015.
[19] Carlos A. Gomez Uribe and Neil Hunt "The Netflix Recommender System: Algorithms, Business Value, and Innovation", *ACM Transactions on Management Information Systems*, Vol. 6, No. 4, Article 13, December 2015.

Authors' Profiles

Smriti Ayushi received a B.Tech. degree from Sharda University, Gr. Noida and and M.tech. degree from PES University, Bangalore. Her area of interest includes machine learning, artificial intelligence and Internet of Things.

V R Badri Prasad holds a MS (1997) in Software Systems from BITS Pilani and is currently pursuing his Ph.D from Jain University. His interest areas are Parallel and Distributed Computing, Computer Vision, Microprocessors, Computer Architecture, Data Structures and Algorithms. He He has professional society membership (lifetime) in ISTE and IET. He has secured a State Rank of 110 during his SSLC Examination. He likes being creative and executes tasks in parallel.

How to cite this paper: Smriti Ayushi, V R Badri Prasad, " Cross-Domain Recommendation Model based on Hybrid Approach", International Journal of Modern Education and Computer Science(IJMECS), Vol.10, No.11, pp. 36-42, 2018.DOI: 10.5815/ijmecs.2018.11.05

