International Journal of Advanced Science and Technology
Vol. 29, No. 8, (2020), pp.4213-4220

Solvıng The Cold-Start Problem Usıng User Profıle Based Bagged Ensemble Model For Movıe Recommendatıon

Udayabalan Balasingam[1], Prof.Dr.N.P.Gopalan[2]
[1]*Research Scholar, Department of Computer Applications, NIT, Trichy.*
[2]*Professor, Department of Computer Applications,Trichy*

Abstract

Effective and accurate recommendations is a key-challenge for today's ecommerce dominated world. Accurate recommendations can not only improve sales for the organizations, but also provides easy selection opportunities for consumers. This ultimately reduces the choice overload for consumers. Cold start is a major challenge for any recommendation system. This work presents a user-profile based model, UPBEM, to handle the cold start issue in recommendation systems. Correlation between users is identified based on their profile and predictions are made based on the current user's information, along with the information of similar users. A bagging ensemble is used for the recommendation process. Experiments were conducted on the MovieLens data. Comparisons and results show reduced MAE and RMSE, exhibiting improved performances of the UPBEM model

Keywords: *Recommendation System, Ensemble, Bagging, User Profile, Categorical Encoding*

1 Introduction:

Recommendation models are highly useful in this technology filled world [1]. They enable easier selection for users and also provide better future requirements for organizations. Growth of ecommerce has resulted in most of the purchases happening online. The availability of huge number of choices has resulted in the selection fatigue, where the users are not able to decide their choice [1,2]. This mandates for a system that can filter the required products effectively based on the users. A personalized recommendation, rather than generalized recommendation can result in a huge reduction in the decision making process for customers. Organizations based on e-commerce also have an interest on recommendation systems,

as they can provide effective identification of requirements, which in turn improves customer loyalty [3]. Applications requiring recommendation systems include, e-commerce sites, movie recommendations, song recommendations, real-estate sites, matrimonial sites etc. Recommendation systems work by analyzing the user's likes and dislikes and performs predictions based on their past data. The complexity involved in building effective recommendation systems are manifold, of which cold-start is an important issue that has to be addressed [4]. The domain is not laden with old customers. It is also a major responsibility of organizations to retain first time users. Recommendations for such customers is complicated, as the model does not have any data

for training [5, 6]. It also applies to customers with low profile information. The generally adopted solution is to provide the overall best products as recommendations [7]. This, however, is not the perfect solution. Personalization even to the smallest extent can provide more correlated results [8]. This work presents a solution for the cold-start problem in recommendation systems by using user profile matching techniques. Profiles that are similar to the current user'sCertain software quality traits, for example, maintainability, ease of use, dependability can't be actually indicated and estimated. At the beginning periods of software process it is hard to characterize a total software specification. Thus, despite the fact that product may adjust to its specification, clients don't live up to their quality desires.

2 Related Works

This section deals with some of the more recent and significant researches in the domain of recommendation models. A genre clustering based model that signifies the importance levels based on weights is presented by Fremal et al. [9]. Weight assignment is based on the metadata. A temporal based model that considers the changing user behavior was proposed by Liu et al. [10]. User preference is not only identified by their past data, but also using their evolution data. Other similar temporal dynamics based models include methods by Lathia et al. [14] and Zheng et al. [15]. Markov state computing model is used to perform recommendations. A linear Gaussian Regression based model that aims to provide effective recommendations was presented y Zhang et al. [11]. A bipartite network based model was presented by Daminelli et al. [12].

A life stage based correlation model for recommender systems was presented by Jiang et al. [13]. A technique to handle data sparsity and

temporal change in user behavior was presented by Li et al. [16]. The technique works by combining features of movies and user interest details. It is a hybrid model and considers both short-term and long-term interests of users for analysis. Other works in the domain include works by Stanescu et al. [17], Leng et al. [18] and Geng et al. [19].

A Katarya et.al suggested the cuckoo search based recommendation model [20]. This is based on collaborative filtering method and uses cuckoo search for the prediction process. Feedback based collaborative filtering model was presented by Hu et al. [21]. This technique considers both implicit and explicit feedbacks for analysis. Goal programming based recommendation model was presented by Inan et al. [22]. This techniques combines multiple information like similarity scores of movies and the goal programming model to perform recommendations. A neural network based collaborative filtering model was presented by Yu et al. [23]. It uses a Contextual-boosted Deep Neural Network model for recommendation. A similar model was presented by Xiao et al. [24].

3 III. MOVIE RECOMMENDATION USING USER PROFILE BASED BAGGED ENSEMBLE MODEL (UPBEM)

Cold start is a major issue encountered by recommendation systems. Cold start problem occurs due to the unavailability of past customer data. It becomes complex to provide predictions without proper training data. This work presents a User Profile Based Bagged Ensemble Model (UPBEM) to solve the cold start problem. The model proposed consists of four modules; the user profile creation module, movie data preprocessing module, training data creation and tree based bagging model for recommendation. This Research is based on a suggestion for films but the model produced is generic and can be extended to any domain. The proposed model architecture is illustrated in Figure 1.

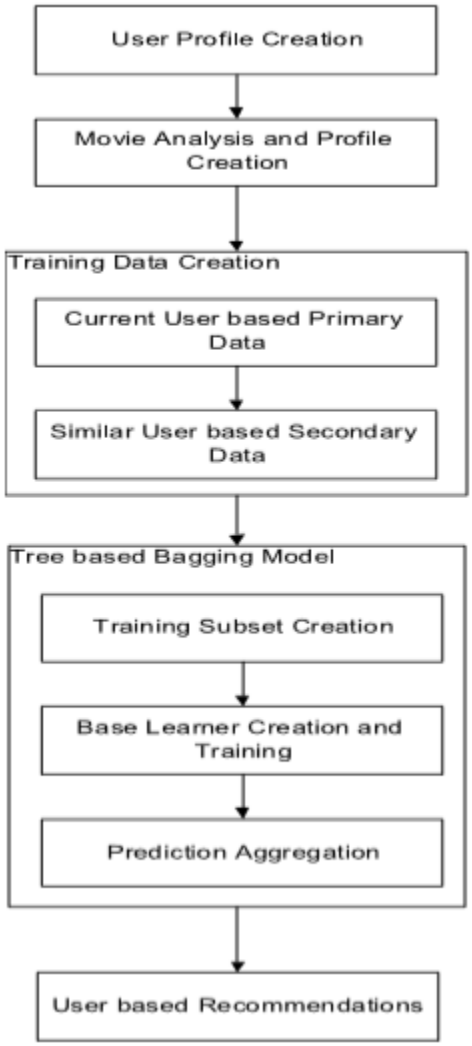

Fig 1. UPBEM Architecture

A.User Profile Creation User profile is the process of creating a numerical vector for each user based on their properties like gender, occupation, demography etc. Any categorical information provided by the user can be used to build their profile. The major crux of the profile creation process is that, the background of a user provides an overview of the user's personal interests. This can aid in identifying users with similar interests. The major difficulty in creating user profiles is that the data is usually rich in categorical and string features. These are required to be carefully analyzed and categorized for effective profile building. Distinct identifier values like User ID, string values like name and address and categorical

values like age and occupation should be distinguished. Identifiers and string-based values are eliminated, as they are usually distinct for each user and hence cannot be used for analysis. Categorical data, although being textual in nature, contain information that can be used for analysis. These values, however, cannot be used directly for analysis. Machine learning models can only operate on numerical data. Hence, they are converted to numerical data using one-hot-encoding techniques. One-Hot-Encoding identifies distinct entities contained in the attribute and creates an additional feature column for each of these entities. If the instance contains the specific entity, the column is marked with a one else a zero. This results in the creation of several features, however, it greatly improves the prediction levels of the machine learning

models.

B.Movie Data Preprocessing Similar to building a user profile, movie based profiles are also to be created for effective analysis. Movie profiles contain the genre-based information of movies. For effective usage, genres are represented as single attribute textual entities. A single movie can be categorized under multiple genres. Hence it is mandatory to convert the genre information into features. Every genre corresponds to a single attribute and if a movie corresponds to a genre, it is marked as one else zero. Creating the user profile and the movie profile usually results in a large number of attributes and the resultant data is sparse. Hence real-time recommendation requires a fast and an effective model for predictions. C.Training Data Creation Training data for recommendation models usually corresponds to data pertaining to a single user. Training data also pertains to that user. All further processing is also performed on a per-user basis. Training data for a recommendation model is composed of movie genre information, along with the rating provided by the user under consideration.

Rating information is usually available as an independent entity. It contains rating information of all users. Rating information pertaining to the single selected user is extracted and combined with the genre matrix created in the previous phase. After the completion of the join operation, several movies remain unrated. These movies are considered for prediction. The movies with rating information forms the primary training data. The primary training data might contain sufficient number of instances for a long time user. However, for a new user this might be very low, or might be empty. This issue is termed as the cold start problem. The proposed model overcomes this issue by generating the secondary training data. Even for users without cold start problem, this additional data proves to be an added advantage during the training process. The secondary training data is generated by identifying similar users and utilizing their rating detains for the training process. This work performs similar identification of users based on user profiles. The initial module discussed the process of creating profile of the user.This data is used to identify similarity between users. Correlation is used to determine user similarity. Users with 80% or more correlation with the selected user are considered by the proposed approach. The value of 80% is a threshold indicator and can be modified based on the data being used and the user's status. The value can be increased for old users and can be reduced substantially for new users with low rating instances. Highest rated movies by these selected users are identified. The genre matrix for these movies and the ratings are combined to form the secondary training data. The primary and the secondary training data are combined to form the final training data. this process is performed for each user and every time a recommendation is required for a user. D.Tree based Bagging for Rating Prediction The training data is usually a large and sparse matrix. Hence it requires an effective processing model to handle the sparsity and also provide faster

predictions. This work proposes a tree-based bagging technique that can be used for training and prediction. Bagging is a type of ensemble that uses subsets of data to train multiple machine learning models. Final predictions are based on an aggregate of the predictions from each of the learned models. Bagging method consists of three phases; the subset creation phase, the training phase and the aggregation phase. The subset creation phase, also called the bag creation phase is used to filter a random subset of the input data for the training process. This work uses 60% of the training data for subset creation. Unlike general bagging models, every subset contains all the primary training data and randomly sampled secondary training data. This ensures that all the primary data are used for rule building by all the base learners. Base learner training forms the next phase of the process. Multiple instances of base learners are created and trained using the created data subsets. This work uses Decision Tree as the base learner. This results in multiple trained base learners. The data prepared for prediction is passed to the trained base learners. Each base learner provides predictions for each of the instances. Hence multiple predictions are obtained for each instance. These predictions are passed to the aggregation phase. A mean based aggregation is performed in this phase, resulting in the prediction of rating for each of the movies in the prediction list. The movies rated as highest are filtered and provided to the user as recommendation.

4 Results And Discussion

The proposed model uses MovieLens data [25, 26] for experimentation. The MovieLens data consists of 3 components. The user detail component, movie detail component and the rating component. The first two components are independent of each other. They are connected by the final rating component, which represents the rating provided by users for movies. The proposed

model has been implemented using Python and Scikit library. Performance evaluation is performed based on the two standard metrics for regression; Root-Mean-Square-Error (RMSE) and Mean Absolute Error [27, 28].

$MAE =$

$1\ N$

$\sum|yi - y'i|\ n\ i=1$

$RMSE = \sqrt{}$

$1\ N$

$\sum(yi - y'i)2\ n\ i=1$

Where N is the total number of test reviews, yi and yi' represent the actual ratings for the test reviews and those predicted.. MAE measures predictive effectiveness. Strong forecasts show smaller MAE values. RMSE represents predictive stability. This represents the rate of uncertainty in the forecasts. Low MAE values are a good predictor whilst high RMSE values indicate high predictive variability. Low MAE values represent a good predictor, while high RMSE values indicate high variability in predictions. The proposed model has been compared with SW I, CM II and MLR models [12] and Cuckoo Search based model [23].

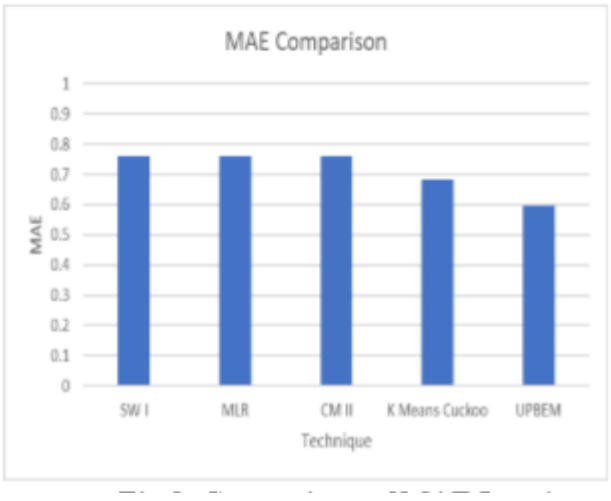

Fig 2. Comparison of MAE Levels

MAE levels of the proposed model is compared with the existing models and shown in Figure 2. The MAE levels of the proposed UPBEM model exhibits the lowest levels, showing the effective predictive capabilities of the proposed model. RMSE levels of the proposed model is compared with existing

models and shown in figure 3. The proposed model exhibits significant reduction in the RMSE levels compared to the existing models.

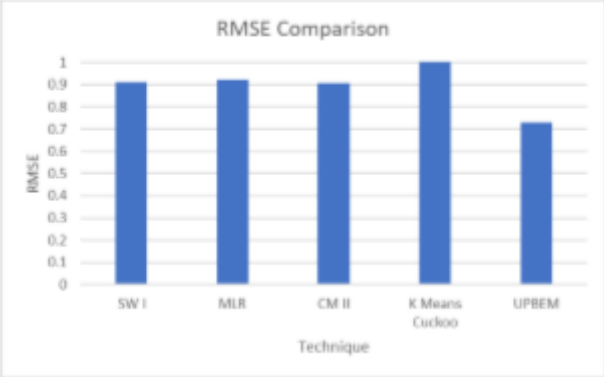

Fig 3. Comparison of RMSE Levels

A tabulated comparison of the performance of the UPBEM model with the exisitng models is shown in Table I. The best results are shown in bold form.. The UPBEM model is shown to display the lowest MAE values, with a decrease from 0.08 to 0.16.Similarly, a reduction in the RMSE levels between 0.2 and 0.8 has been observed. This proves the ability of the UPBEM model to exhibit effective recommendations.

Table I. Performance Comparison of UPBEM

Model	MAE	RMSE
SW I	0.7616	0.9096
MLR	0.7611	0.9212
CM II	0.7615	0.9086
K Means Cuckoo	0.6842	1.231
UPBEM	**0.5950**	**0.7310**

5 Conclusion

This work proposes a user-profile-based recommendation model that can be used to solve a cold start problem. Cold start problem results in the inability of the recommendation model to provide prediction for a new user. The objective of this work is to solve the issue by identifying similar users and their likes and providing recommendations to the new user based on this data. The model achieves higher correlation in recommendations due to the user specific nature of recommendations. The model has been evaluated using the MovieLens data, however, the architecture is generic and hence recommendations can be performed in any domain. The model was observed to achieve effective predictions for new users and also was observed to enhance the predictions for existing users.

The major limitation of the model is that the user and the movie profile building phases results in the

creation of huge number of features, which are skewed in nature. The model is static in nature and cannot handle behavioral changes of users. Future enhancements will be towards standardizing the process and to enable effective feature building, and also towards handling the temporal behavioral changes of users.

References

1. .Schwartz, Barry. "The paradox of choice: Why more is less." New York: Ecco, 2004.
2. Bobadilla, Jesús, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. "Recommender systems survey." Knowledge-based systems, Volume 46, Pages: 109-132, 2013.
3. Wei, Jian, He, Jianhua, Chen, Kai, Zhou, Yi, and Tang, Zuoyin. "Collaborative filtering and deep learning based recommendation system for cold start items". Expert Systems with Applications, Volume 69, Pages: 29–39, 2017.
4. Silva, Nícollas, Diego Carvalho, Adriano CM Pereira, Fernando Mourão, and Leonardo Rocha. "The pure cold-start problem: A deep study about how to conquer first-time users in recommendations domains." Information Systems, Volume 80, Pages: 1-12, 2019.
5. Zhang, Desheng, He, Tian, Liu, Yunhuai, Lin, Shan, and Stankovic, John A. "A Carpooling recommendation system for taxicab services". IEEE Transactions on Emerging Topics in Computing, Volume 2, Issue 3, Pages: 254–266, 2017.
6. Zhang, Daqiang, Hsu, Ching Hsien, Chen, Min, Chen, Quan, Xiong, Naixue, and Lloret, Jaime. "Cold-start recommendation using bi-clustering and fusion for large-scale social recommender systems". IEEE Transactions on Emerging Topics in Computing, Volume 2, Issue 2, Pages: 239–250, 2017.
7. W. C. McDowell, R. C. Wilson and C. O. Kile Jr, "An examination of retail website design and conversion rate", Journal of Business Research, Volume 69, Issue 11, Pages: 4837–4842, 2016.
8. Conversion rate of online shoppers in the united states as of 4th quarter 2017, by device, accessed:201802-14, 2017.
9. S. Frémal and F. Lecron, "Weighting strategies for a recommender system using item clustering based on genres," Expert Systems with Applications, Volume 77, Pages: 105-113, 2017.
10. Liu, Zhen, Ke Tan, Xiao-Qing Wang, and Shi-Hao Tang. "A learning framework for temporal recommendation without explicit iterative optimization." Applied Soft Computing, Volume 67, Pages: 529-539, 2018.
11. Y. Zhang, Y. Zhuang, J. Wu, and L. Zhang, "Applying probabilistic latent semantic analysis to multi-criteria recommender system", AI Commun. Volume 22, Issue 2, Pages: 97–107, 2009.
12. S. Daminelli, J.M. Thomas, C. Durán, and C.V. Cannistraci, "Common neighbours and the localcommunity-paradigm for topological link prediction in bipartite networks", New J. Phys. Volume 17, Issue 11, Pages: 1-11, 2015.
13. P. Jiang, Y. Zhu, Y. Zhang, and Q. Yuan, "Life-stage prediction for product recommendation in ecommerce", in: Proceedings of the 21st ACM SIG KDD International Conference on Knowledge Discovery and Data Mining, ACM, Pages: 1879–1888, 2015.
14. N. Lathia, S. Hailes, L. Capra, and X. Amatriain, "Temporal diversity in recommender systems", in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Pages: 210–217, 2010.
15. C.L. Zheng, K.R. Hao, and Y.S. Ding, "A collaborative filtering recommendationalgorithm incorporated with life cycle", Advanced Materials Research, Volume 765, Pages: 630–633, 2013.
16. Li, Jing, Wentao Xu, Wenbo Wan, and Jiande Sun. "Movie recommendation based on bridging movie feature and user interest." Journal of computational science, Volume 26, Pages: 128-134, 2018.
17. A. Stanescu, S. Nagar, and D. Caragea, "A Hybrid Recommender System: User Profiling from Keywords and Ratings", Proceedings of 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, Pages: 73-80, 2013.
18. Y. Leng, C. Liang, Y. Ding, and Q. Lu, "Method of Neighborhood Formation in Collaborative Filtering", Pattern Recognition and Artificial Intelligence, Volume 26, Issue 10, Pages: 965-974, 2013.

19. X. Geng, H. Zhang, J. Bian, and T. Chua, “Learning Image and User Features for Recommendation in Social Networks”, Proceedings of 2015 IEEE International Conference on Computer Vision, Pages: 4274-4282, 2015.
20. R. Katarya and O.P Verma, “An effective collaborative movie recommender system with cuckoo search,” Egyptian Informatics Journal, Volume18, Issue 2, Pages: 105-112, 2107.
21. Hu, Yutian, Fei Xiong, Dongyuan Lu, Ximeng Wang, Xi Xiong, and Hongshu Chen. "Movie collaborative filtering with multiplex implicit feedbacks." Neurocomputing, Pages: 1-10, 2019.
22. .Inan, Emrah, Fatih Tekbacak, and Cemalettin Ozturk. "Moreopt: A goal programming based movie recommender system." Journal of computational science, Volume 28, Pages: 43-50, 2018.
23. Yu, Shuai, Min Yang, Qiang Qu, and Ying Shen. "Contextual-Boosted Deep Neural Collaborative Filtering Model for Interpretable Recommendation." Expert Systems with Applications, 2019.
24. Xiao, Han, Yidong Chen, Xiaodong Shi, and Ge Xu. "Multi-perspective neural architecture for recommendation system." Neural Networks, 2019.
25. .https://grouplens.org/datasets/movielens/
26. F.M. Harper and J.A. Konstan, “The movie lens datasets: History and context,” ACM Transactions on Interactive Intelligent Systems (TiiS), Volume 5, Issue 4, Pages: 1-20, 2016.
27. . S. Dooms, T. De Pessemier and L. Martens, “Offline optimization for user-specific hybrid recommender systems,” Multimedia Tools and Applications, Volume 74, Issue 9, Pages: 3053-3076, 2015.
28. . X. Ge, J. Liu, Q. Qi and Z. Chen, “A new prediction approach based on linear regression for collaborative filtering,” IEEE Eighth International Conference In Fuzzy Systems and Knowledge Discovery (FSKD), Volume 4, Pages: 2586-2590, 2011