Indian Journal of Science and Technology, *Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/99896, November 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Online Product Recommendation using Relationships and Demographic Data on Social Networks

R. Satish Srinivas, C. S. Anish Balaji and P. Saravanan*

School of Computing, SASTRA University, Thanjavur, India; rsatish57@gmail.com, anish.cs24@gmail.com, sharan.doit@gmail.com

Abstract

Objectives: There have been many traditional product recommender systems in the past, but they were mostly based on collaborative or content filtering, historical transaction records or website browsing history of the users. This approach leads to sparsity and cold-start issues. **Methods/Statistical Analysis:** The proposed system is a hybrid one and can improve the suitability and the accuracy of recommender systems with the help of users' and products' demographic info and the ratings of a brand name, constantly updated in the e-commerce and social media websites. **Findings:** Experiments on the actual dataset reveal that friends with similar tendencies select the similar items and this approach can solve the relevant data sparsity and cold start problems. Initially, the system loads the data about the friends who have given product ratings and selects a target cold start user. Then it predicts his/her tech level and cost level using BMART algorithm. With the help of these, it finds the user-friends and user-items similarity using the cosine and Euclidean cluster similarity methods respectively. It then feeds these into the proposed ranking algorithm to find the relevance scores of the products to the target user and based on this the recommended product links are displayed. **Application/Improvement**: The outputs reveal that the proposed system gives very accurate and personalized product recommendations to the user.

Keywords: Hybrid Ranking, Recommender System, Regression Ranking, Relevance Score, Social Media

1. Introduction

Many e-commerce systems have used traditional product recommender systems in the past, but they were mostly based on collaborative filtering, historical transaction records or website browsing history of the users. Moreover, they also rely upon the relevant items rated only by random users of the website to provide the product recommendations. Demographic information of users has been used from social media and it gives personalized product recommendations to the users using string search and bagged multiple additive regression tree algorithm[1]. This is done by extracting purchase intent tweets, matching user-user demographics, user item demographics and recommending items. The results have indeed shown that this system is most optimal in generating recommendation results in the best way to match users' choices.

Analyzing the social media message is the state of art procedure in the recommender system. Based on competitive[2] advantage a systematic approach to analyze method can be used. CART algorithm[3] has been used to extract the tweets from twitter to classify them into health and non-health messages. This classification outperforms with other classification interms of precision, error rate, and accuracy.

With the extensive use of social media websites, it is now possible to extract the social relationships among users and product demographics from online e-commerce websites. These can include attributes like the user's age, gender, education, friendship in the network and also the product's rating records (tags). Further experiments on real dataset reveal that friends with similar tendencies select the similar items. Thus the proposed system can improve the suitability and the accuracy of recommender

*Author for correspondence

systems, solving the cold start and data sparsity problems. Many social media websites contain abundant information about users' demographic info and the ratings of a brand name, constantly updated in their pages. Considering the issues and challenges in the recommender system, a regression based hybrid recommender system has been proposed in the manuscript.

The system proposed can be broken down into the following working components:

- In this first module initially the task is to classify the target variables based on both gender and career. Then using the friends' age, tech level and cost level in that category the method predicts the tech level and cost level of the target user.
- In this second module, p1 is the tech level of the target user and q1 is the cost level of the target user. p2…p50 are the tech level of products. Similarly q2…q50 are the cost levels of products. Here the system considered the target user to be the centre of the cluster and the products to be other elements in the cluster. Then it finds out the distances between them and calculated the similarities based on the fact that similarity is inversely proportional to the distances. Next the cosine similarity method is used to find out the similarity between u1-u2,…,u1-u50 using the users' attributes.
- The above results are fed into the ranking methodology using a weighted aggregate of the three elements for finding relevance scores of the products to the target user. Finally they are sorted and ranked in the decreasing order and shown as recommendation links to the user. Experiments on the large datasets show that this approach is more efficient than the other popular conventional methods and hence this system can provide very personalized and effective product recommendations to the users.

The remaining sections in this manuscript has organised as follows. Section 2 focuses on the work related to recommender systems and its challenging. The proposed work and algorithms are discussed in Section 3. Section 4 elaborates the experimental setup, results and discussion. The performance of the proposed method is evaluated in Section 5. Finally, Section 6 concludes the article with future work.

An approach of social manipulation method[4] has been proposed for recommender system based social network data. Users' relationships and rating records are used to judge the values that are missing in the user item matrix and applying the bi-clustering algorithm on them. Relationships between users can improve the prediction accuracy and from this there can be proposed a social manipulation approach which implements social media information to relate and improve recommender systems.

The structural and personal destruction have been analyzed[5] by hurricane Katrina, characterizing the geography of exposure. To understand health issues and challenges to the poor populations during recovery can be done. Multiple Additive Regression Trees are used as the methodology to classify and predict flood levels, damage and 911 calls from various areas affected. It has been successful in interpreting meaningful and useful information with regards to measures influencing the possible future health hazards and disaster recovery, especially the impact of all the measures because of exposure to storm. To provide a working guide to the boosted regression trees, an optimal method for fitting data mining models that fit a single clear cut model has been developed. The default bagging ratio used is 1/2, which means that, always half of the data are chosen just at random, without replacement, from the full trained data set. They have provided a useful basis for evaluating the models[6].

Classification and regression trees[7] present a complicated abstract representation of the relationship among the variables. The data set has been utilized as the first step in the building of an information providing model or a final visualization of the important relationships. For a response variable which is binary, we group the data into groups by the variable which is known as classification. Instead when the response variable is instead numeric or continuous we predict the output using regression and provide a relevant model.

It has been found useful to discover[8] how any two collaborative filtering algorithms can be optimised by the calculation of the demographic correlations among the members of user or product domain space. That algorithm calculates a predicted value for the already rated items rather than generating a top recommended list of the movie lens data set and the testing done shows the improvised approach can change from becoming worse than the base of the filtering algorithms, to outperform them, based on the roles of the demographic correlations in the predicted results. A method for constructing a regression tree called GUIDE[9] has been proposed. It is particularly developed to remove variable selection bias, a disadvantage that can undermine the reliability of the

interpretations that can arise from a tree structure. Thus that system solves the problems that can very much affect the ability to interpret the regression tree: being wrong in variable choosing, being insensitive to local interactions, and complex tree structures.

It has been found highly beneficial to give personal recommendations to the users if the user preferences are unavailable and the cold start problem thus needs to be solved[10]. Two algorithms are used which are word-space similarity and topic-genre similarity on the movies space of IMDB website. The recommendations to the target user were finally obtained using the hierarchical topic clustering algorithm and sorted based on the genres preferred. A flexible frame work[11] has been recommended to scrap similar videos using semi automatic web scrapers. To recommend a movie for a new user has been proposed using prediction[12]. It is based on available user's ratings on a given list of movies is the task. A combination of model and memory based approaches where rating to be given by the new user is predicted, by taking a weighted average, on forming a cluster and finding available user's similarity using correlation coefficients.

Several approaches have been proposed for recommender system, but most of them have the cold start problem[13], which means the user has no ratings about items in that product space. Usage of selected attributes of the target user to recommend products solved the same. Combined usage of demographic info of user and products from neighbourhood of the target user product recommendations by avoiding invalid data on analysis of attributes was done.

Recommendation of products based on cluster distance and regression is a problem[14]. Even though CF algorithms for recommender systems were easily portable, they suffered from data sparsity and cold start problems. Usage of various neighbourhood models to find the similarity between users and items in a cluster and combination of this using neighbourhood-aware matrix factorization algorithm is efficient. The predictions of one model are used to estimate unknown variables in the other ones thereby enhancing the prediction accuracy. A neighbourhood based methodology[15] that predicts similarity measures between users, items and between users and items can be used for this. Cold start problem was resolved with minimum error thereby ensuring quality recommendations to a new user.

A ranking algorithm[16] for effective recommendations is the key challenge for recommenders that were far better in performance and memory usage from other traditional collaborative filtering based algorithms. Usage of a training set as input, a performance measure function and the number of iterations as parameters by running 'T' rounds and at each round creating a weak ranker. Finally, the algorithm outputs a ranking model 'f' by linearly combining the weak rankers.

A collaborative algorithm for recommenders[17] has been proposed to selected criteria of users for effective recommendations can be beneficial. Some of the selected criteria that matched the requirements of recommenders for recommendation were the users, the entities, the value dimensions, the belief system and the ideal candidate. A weighted sum approach that used some parameters determined the type of recommendation to be made to the target user. Usage of models to combine predictions on selected parameters from user-product demographics can be efficient[18]. Regression on the same and combining (blending) predictions using some models for output can improve accuracy. Improved performance from other methods which uses linear combination of predicted results can provide optimal results.

But none of the above systems provides a hybrid recommender system which can use the existing users' demographic data in social media combining it with the product ratings from an e-commerce website for the cold start users.

2. Proposed Methodology

2.1 User Tech and Cost Level Prediction using Multiple Additive Regression Trees

A regressive approach[19] that follows the method of regression and takes a weighted average of the predicted values bought a key solution for the problem to be solved. Findings showed significantly lower MAE with enhanced prediction quality. Dramatically better performance and high quality recommendations have been achieved better than user-based algorithm.

Regression is the concept in data mining which is used to predict numerical values of variables. The difference in it from classification is that it is used for continuous variables prediction instead of categorising variables into classes. Many ranking model based on regression tree have been proposed including social and cloud data recommendation[20]. In regression trees the initial levels are categorical and the leaf nodes to give a predicted value for the dependent variable. The following Figure 1 is the module flow diagram of the implementation of the system:

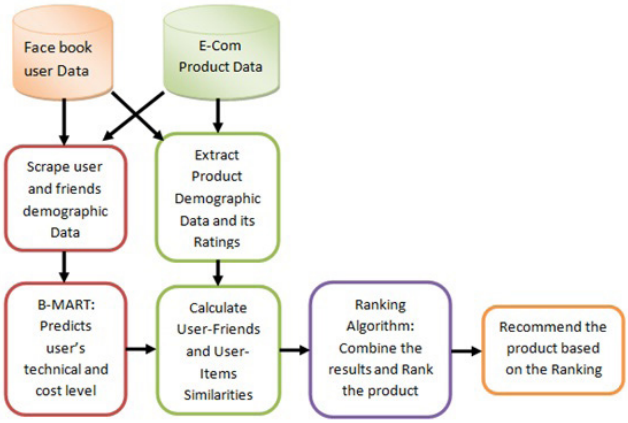

Figure 1. Architecture and implementation of system.

In this first module initially the task is to classify the target variables based on both gender and career. Then using the friends' age, tech level and cost level in that category the method predicts the tech level and cost level of the target user.

Regression Equation,

$$y = a + b * x$$

Slope,

$$b = (N \acute{O}XY - (\acute{O}X)(\acute{O}Y)) / (N \acute{O}X^2 - (\acute{O}X)^2)$$

Intercept,

$$a = (\acute{O}Y - b (\acute{O}X)) / N$$

where X is the ages of users in the category of that gender and Y is their tech levels. With these the equation is formed and giving target user's age (x) the tech level (y) is predicted. Then the same is done for the regression based on the classification of career. After this equal weight is given to the above two predictions and are bagged with a factor of 0.5 and combined to give the final tech level prediction. Similarly the cost level of the target user is also predicted. Here bagging in its simplest terms has been done by giving equal weight to the predictions and combining them to give a more accurate prediction and using them in the upcoming hybrid similarities finding and ranking algorithm.

The usage of user's similarity on their attributes and ratings on products for recommendations is another big challenge for the recommenders[21]. Combination of the same by a hybrid methodology that indicates the type of recommendation to the target user solved the problem by computation of users rating similarity and the user attribute similarity.

2.2 User-Item Similarity using Euclidean Distance

The Euclidean distance is the measure of the straight-line distance between two points in the Euclidean space. It is found useful to provide an outlook of the similarity algorithms and also compare their performance in recommender systems[22]. Finding the pair-wise similarity between movies by using Pearson correlation, Euclidean and cosine similarity measures. Pearson correlation comes out to be the most accurate measure giving the least mean squared error from the actual value expected for recommender systems.

With this the Euclidean distance can be used for n number of points. The formula used here is:

$$d(p, q) = \sqrt{(q1 - p1)^2 + (q2 - p2)^2}$$

In this second module, p1 is the tech level of the target user and q1 is the cost level of the target user. p2…p20 are the tech level of products. Similarly q2…q20 are the cost levels of products. In the proposed system we have used this distance to calculate the user-items similarity for the 20 products with the formula 1/(1+d). Here the system considered the target user to be the centre of the cluster and the products to be other elements in the cluster.

Then it found out the distances between them and calculated the similarities based on the fact that similarity is inversely proportional to the distances. The advantage is it is reportedly faster than most other means of determining correlation. This is done because the Euclidean distance is a fair measure of how similar values are for specific preferences or items.

2.3 User-User Similarity using Cosine Similarity

Cosine similarity is the measure of the similarity between vectors of a product space that calculates the cosine of the angle between them. To predict news articles customers are likely to view/read next the cosine similarity function can be used, which is the measure of similarity between two vector spaces derived from the cosine of the angle between them, here the news articles. The recommendation strength for each article for the target user is obtained and the cosine similarity function considers the attributes in the vector space to find the similarity.

The cosine of angle 0° is 1, and 90 degree is 0. Hence it is a representation of orientation and not magnitude: Two vectors with equal orientation can have a cosine similarity of 1, two vectors at 90° have zero similarity, and two vectors if diametrically opposite can have a similarity of -1, independent of their magnitude. Cosine similarity can have a range of [0,1]. The formula used here is as follows.

$$\text{Similarity} = \cos(\grave{e}) = \frac{A.\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i \mathbf{B_i}}{\sqrt{\sum_{i=1}^{n} \mathbf{A}_i^2}\sqrt{\sum_{i=1}^{n} \mathbf{B}_i^2}}$$

The third module has used this formula with gender, age, career, tech level and cost level as a vector of attributes to find out the similarities between u1-u2…u1-u50. The advantage of cosine similarity is that it is very efficient to be evaluated, especially for data sparse vectors, as only the non-zero dimensions are to be considered.

2.4 Finding Relevance Score from Proposed Ranking Algorithm

The fourth module now feed the user-user similarities, user item similarities and ratings of products from friends into our proposed ranking algorithm to get the final product recommendations.

There are three major elements in the ranking algorithm:

- First if products p1 and p2 have the same ratings from friend's u2 and u3, but the user-user similarity between target user u1-u2 is greater than that between u1-u3, then the product p1 will get a higher ranking in the recommendation.
- Second if friend's u2 and u3 have same user-user similarities with u1 and have rated products p1 and p2 respectively but user-item similarity between u1-p1 is greater than that between u1-p2 then p1 will get a higher ranking than p2 in the recommendation.
- Finally if both user-user similarities and user-item similarities are same then if the number and cumulative ratings for product p1 is greater than those for product p2 then p1 will get a higher ranking than p2 in the recommendation.

Using the above the relevance scores of the products to the target users are obtained and ranked by sorting in the decreasing order and displayed as recommendations to the user.

3. Experimental Evaluation

3.1 Experiential Setup Data Sets

By combining demographic[23] info of social media users and collaborative filtering approaches for efficient recommendation, user based collaborative filtering by heuristics and memory based algorithms and user similarities can be found using Pearson correlation coefficient and this enhanced accuracy than traditional rating based recommender system.

The following data tabulated in Table 1 is the sample product demographic dataset taken from the amazon e-commerce website and actually contains from p1 to p20.

Table 2 consists of the user demographic sample dataset scraped from the Facebook social media website using the Graph API. Here the target user is u1 and the u2 to u50 are his friends who have rated some products in their respective pages:

Table 3 shows the sample dataset of the ratings of the friends of the target user:

Table 1. Product data set

PID	PRODUCT	PRICE	TECH
P1	GOOGLE NEXUS 6P	35000	A
P2	SAMSUNG S6 EDGE+	54000	A
P3	APPLE IPHONE 6S	42000	A
P4	LG NEXUS 5X	36000	A
P5	SONY XPERIA Z5	44000	A

Table 2. User demographic data set

USER	GENDER	AGE	CAREER	MUTFRIEN
U1	M	21	ENG	0
U2	M	22	ENG	45
U3	M	32	ENG	22
U4	M	28	ENG	20
U5	F	27	CAC	11

Table 3. User-product rating data set

rate_id	USER	PRODUCT	RATINGS
1	U1		0
2	U2	P6	5
3	U3	P8	4
4	U4	P1	5
5	U5	P12	3

3.2 Results and Discussion

The new user has no idea about the products he/she wants to buy. Hence initially only the random users' ratings are considered and it may be completely irrelevant to the target user as it is not based on any trust relationships or relevance scores. The only transaction going on in this output is the registering/checking of the user credentials from the database and verifying them.

Figure 2 shows graphical user interface where the user has to give permission if the system can use the insensitive personal data like demographic info from the social media in the e-commerce recommender system. It is necessary to give personalized recommendations to the user, and assumes that the user has at least fifty friends who have given ratings to some models in the product domain. The transaction happening in this output is connecting the e-commerce user profile to the user profile in the social media. Thus the product demographics and user-friends demographics are loaded into the database for further processing and hence can sole the cold start and data sparsity problems.

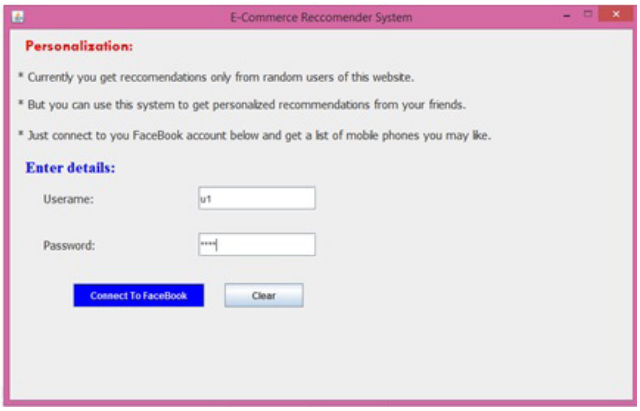

Figure 2. Social media connection window.

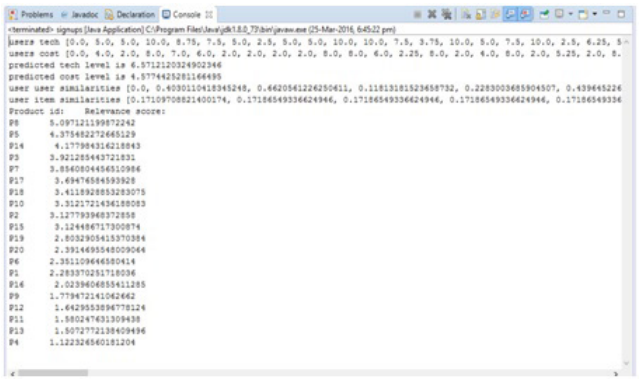

Figure 3. Output of proposed algorithms.

Figure 3 gives the various ranking data generated by the proposed system. The first line show the tech/cost levels of the friend of the users which is based on the aggregate of the products they've rated. Then using bagged multiple additive regression trees the tech and cost levels of the target user is predicted. After this the user-friends similarities are calculated using the cosine similarity and the user-item similarities are calculated using the Euclidean cluster similarity algorithms. Based on these and the product ratings of the friends the proposed ranking algorithm finds the relevance score of the products to the target user and sorts them in descending order to give the final recommendations to the user.

The product recommendations to the target user have showed in Figure 4. It is based on the descending order of the relevance scores of the products to the target user. It displays the mobile phone models which is the chosen product domain and gives links to the products in the e-commerce website.

4. Performance Evaluation

Variance is a measure of how far the given numbers are spread out and hence a small variance means that the data points are very close to the mean (expected value) and hence to each other, while a high variance can mean that the data points are farther away from the mean and from each other. Thus they can be used as a valid comparison metric for the recommender system.

The formula for the variance is:

$$S^2 = \frac{\Sigma_{i=1}^{N}\left(x_i - \bar{x}\right)^2}{N-1}$$

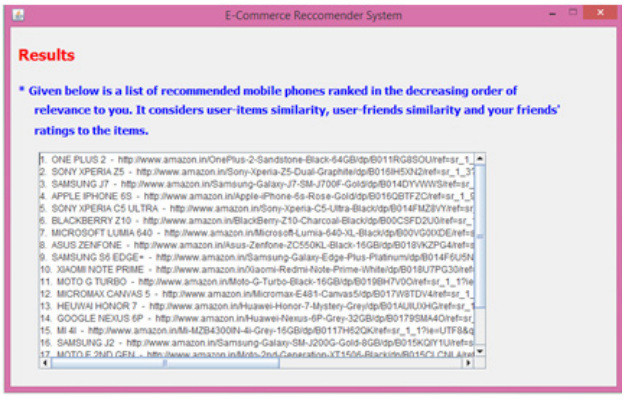

Figure 4. Final recommendations by the proposed system.

while the formula for the weighted variance is:

$$S_w^2 = \frac{\Sigma_{i=1}^{N}(x_i - \overline{x_w})^2}{\frac{(N' - 1)\Sigma_{i=1}^{N} w_i}{N'}}$$

where w[i] is the weight for the i^{th} observation, N' is the number of non-zero weights, and $\overline{x_w}$ is the weighted mean of the observations. The weighted variance is adopted here because the rank is assigned to the products and hence weight 20 is given to the top ranked result (hence weight 1 for the 20^{th} ranked product).

From the comparison graph (Figure 5), the proposed system has proved the following.

- The regression is the first basic one in which only the general regression algorithm is used and hence for a user all the users irrespective of their demographic information will be considered and recommended for the target user and has variance of 4.16.
- The second one is the Multiple Additive Regression Tree in which multiple predictions are just added separately without bagging factors and has variance of 3.58.
- The third one is the Bagged Multiple Additive Regression Tree in which bagging factors are considered like hybrid similarities and has variance of 2.98.
- The fourth one is the proposed ranking algorithm which takes the results from the above methods and with the product ratings gives a weighted aggregate as the final relevance scores for the final product recommendations. Here the predicted target user's tech and cost levels and the recommended products' rank tech and cost levels and used to give the least variance value of 2.44 and hence can be considered to be a better recommender system.

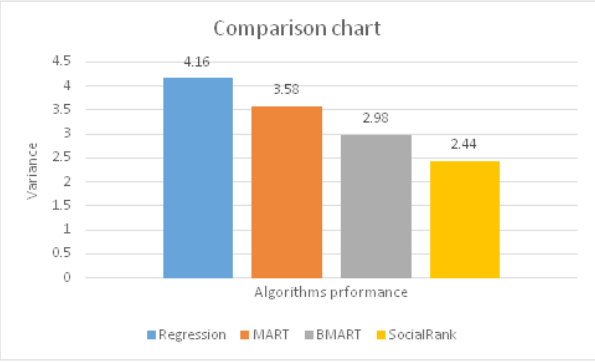

Figure 5. Performance comparisons of regression, MART, BMART and social rank methods.

5. Conclusion and Future Work

The main focus of this project has been to solve the cold start and data sparsity problems and provide recommendations resolving them. In the proposed system, based on the observational fact that relationships among users, can optimize the prediction accuracy and combine the predictions from various attributes to improve accuracy. The bagged multiple additive regression tree algorithm's prediction is based on the target user's attributes and gives a value that has a very less variance rate from the actual value. The cosine and Euclidean cluster similarity measures also provide suitable methods to find out the user-user and user-item similarities. Finally, the devised ranking algorithm also uses the above results and provides a unique method to combine them as a weighted aggregate with the product ratings to calculate a relevance score that gives accurate and personalized product recommendations to the user. Thus the system solves the cold start and data sparsity problems that exist in the current recommender systems.

As the user gets familiar with the product domain, there needs to be some improvements in the algorithm to provide updated recommendations to the user. Moreover, an automatic data scraper would be more ideal if the users in the social media give permissions and participate in the system. This analysis and design of the above ways can lead to a better deployment of the system in the real world.

6. References

1. Zhao WX, Li S, He Y, Wang L, Wen JR, Li X. Exploring demographic information in social media for product recommendation. Knowledge and Information Systems. 2015; 49(1):1–29.
2. Singla ML, Apoorv D. How social media gives you competitive advantage. Indian Journal of Science and Technology. 2015 Feb; 8(S4):1–6.
3. Nivedha R, Sairam N. A machine learning based classification for social media messages. Indian Journal of Science and Technology. 2015 Jul; 8(16):1–4.
4. Sun Z, Han L, Huang W, Wang X, Zeng X, Wang M, Yan H. Recommender systems based on social networks. Journal of Systems and Software. 2015 Jan 31; 99:109–19.
5. Curtis A, Li B, Marx BD, Mills JW, Pine J. A multiple additive regression tree analysis of three exposure measures during Hurricane Katrina. Disasters. 2011 Jan 1; 35(1):19–35.
6. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. Journal of Animal Ecology. 2008 Jul 1; 77(4):802–13.

7. Morgan J. Classification and regression tree analysis. Technical Report; 2014.
8. Vozalis M, Margaritis KG. Collaborative filtering enhanced by demographic correlation. AIAI Symposium on Professional Practice in AI, of the 18th world Computer Congress; 2004 Aug.
9. Loh WY. Regression tress with unbiased variable selection and interaction detection. Statistica Sinica. 2002 Apr 1; 12(2):361–86.
10. Fleischman M, Hovy E. Recommendations without user preferences: A natural language processing approach. Proceedings of the 8th International Conference on Intelligent user Interfaces ACM; 2003 Jan 12. p. 242–4.
11. Mathew A, Balakrishnan H, Palani S. Scrapple: A flexible framework to develop semi-automatic web scrapers. IRECOS. 2015 May 31; 10(5):475–80.
12. Gong S. A collaborative recommender based on user information and item information. Proceedings of the International Symposium on Information Processes; Academy Publisher. 2009 Aug 21. p. 1–4.
13. Safoury L, Salah A. Exploiting user demographic attributes for solving cold-start problem in recommender system. Lecture Notes on Software Engineering. 2013 Aug 1; 1(3):303.
14. Töscher A, Jahrer M, Legenstein R. Improved neighborhood-based algorithms for large-scale recommender systems. Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition ACM; 2008 Aug 24. p. 4.
15. Kumar S, Kumar S. An approach for recommender system by combining collaborative filtering with user demographics and items genres. International Journal of Computer Applications. 2015 Oct; 128(13):16–24.
16. Xu J, Li H. Adarank: A boosting algorithm for information retrieval. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 2007 Jul 23. p. 391–8.
17. Akhtarzada A, Calude CS, Hosking J. A multi-criteria metric algorithm for recommender systems. Fundamenta Informaticae. 2011 Jan 1; 110(1-4):1–1.
18. Jahrer M, Toscher A, Legenstein R. Combining predictions for accurate recommender systems. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2010 Jul 25. p. 693–702.
19. Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th International Conference on World Wide Web ACM; 2001 Apr 1. p. 285–95.
20. Mourougan S, Aramudhan M. Regression tree based ranking model in federated cloud. Indian Journal of Science and Technology. 2016 Jun; 9(22):1–7.
21. Gong S. A collaborative recommender based on user information and item information. Proceedings of the International Symposium on Information Processes; Academy Publisher. 2009 Aug 21. p. 1–4.
22. Rajendra, Qing W, Raj JD. Recommending news articles using cosine similarity function. Warwick Business School Journal. 2015; 1–8.
23. Hu R, Pu P. Using personality information in collaborative filtering for new users. Recommender Systems and the Social Web; 2010 Sep 26. p. 17.