Proposed Approach for Book Recommendation Based on User k-NN

Rohit, Sai Sabitha and Tanupriya Choudhury

Abstract Large data repositories helped us in support systems but created a huge problem for meaningful information retrieval. Filtering of data based on user requirements solved this problem. This process of data filtering when combined with prediction developed recommendation systems. Initial work in recommendation systems can be listed in the areas of cognitive science, approximation theory, marketing models, and automatic text processing. This paper focuses on recommendation system for books. In this paper, training and testing models are designed to predict user ratings for new users. The predicted user ratings are used to propose three types of recommendations based on three different user attributes.

Keywords Recommendation system · Collaborative filtering · Pearson similarity · Cosine similarity · User k-nn

1 Introduction

In present world each person wants quick supplies for his requirements in every field of life including shopping or renting of books. Recommendation systems provide best possible solution to this problem. These are kind of expert systems which help in gathering the related information [1]. Most of recommendation systems work for almost similar purpose that is to recommend items which are most relevant to the users. To fulfill this purpose recommendation systems use different approaches including collaborative, item-based, and hybrid filtering.

Rohit (✉)
Department of CS&E, Amity University, Noida, India
e-mail: rohit.ahlawat@live.in

S. Sabitha · T. Choudhury
Faculty, Department of CS&E, Amity University, Noida, India
e-mail: assabitha@amity.edu

T. Choudhury
e-mail: tchoudhury@amity.edu

S.K. Bhatia et al. (eds.), *Advances in Computer and Computational Sciences*,
Advances in Intelligent Systems and Computing 554,
https://doi.org/10.1007/978-981-10-3773-3_53

Table 1 Output for recommendation

User Id Prediction	Age	Author	Book Title		
4017	48	A. Manette Ansey	Midnight Champagne: A Novel (Mysteries & Horror)	New Orleans, Louisiana, USA	4.102143
4017	48	A. Manette Ansey	Sister (Mysteries & Horror)	New Orleans, Louisiana, USA	3.450125
4017	48	A. Manette Ansey	Vinegar Hill (Oprah's Book Club (Paperback))	New Orleans, Louisiana, USA	3.355193
4228	41	A. Manette Ansey	Unwanted Company	Austin, Texas, USA	3.151324

In this paper we are using collaborative filtering approach to provide recommendations to the users. We are training a book rating data with our training model. This trained data will be sent to testing model. The testing model will predict user ratings for new users. On the basis of these predicted values, a system is proposed to recommend books to new users on their personal attributes which are age, location, and interest. Using these three attributes we are proposing three different models. All models include dataset provided by our training and testing models. To create this training model we used a real-time dataset of books as described in Fig. 5. It has large number of entries which are feasible for our analysis. Main objective of this proposal is to assist new users of any book repository in finding their desired books. Research works have been accomplished by many researchers with similar objective as shown in Table 1. Main purpose of this research work is to design a different approach in the creation of recommendation systems. Our work will provide a base in creation of recommendation systems using User k-NN prediction model.

2 Theoretical Background

2.1 Recommendation System Overview

Lot of work has been done in recommendation systems but interest remains same as it is a problem-rich field and having limitless possibilities both in research and industry. It has large number of practical implementations to solve the problem of information overloading and providing personalized information [2]. Following list of different research works in the field of recommendation systems will support the fact that the recommendation system using user k-NN prediction is least touched and thus have large opportunities for research work (Fig. 1).

Initial work in recommendation system can be listed in the areas of cognitive science [3], approximation theory [4], marketing models [5], and automatic text processing [6]. This work later became the rating estimation for new entries on the basis of different attributes and likes of already present entries similar to them.

S.No.	Author	Year	Area	Based on
1	ZHEN ZHU	2007	Book Recommendation	Apriori Algorithm
2	Binge Cui	2009	Online Book Recommendation	Web Services
3	Yongcheng Luo	2009	Privacy-Preserving Book Recommendation	Multi Agent
4	Maria Soledad Pera	2011	Personalized Book Recommendations	Word Similarity
5	*CHENG Qiao*	2013	Simulation Resource Recommendation	Collaborative Filtering
6	Pijitra Jomsri	2010	Recommendation system for Digital liabrary	Association Rule
7	Salil Kanetkar	2014	Web based recommendation system	Hybrid
8	Anand Shanker Tewari	2014	Opinion based book recommendation	Naïve Bayes Classifier
9	Kumari Priyanka	2015	Personalised book recommendation	Opinion Mining

Fig. 1 Literature survey

Recommendation systems can be categorized based on how recommendations are made [7]:

- Content-based recommendations: Items are recommended on the basis of past preferences of the user.
- Collaborative recommendations: Items are recommended on the basis of past preferences of users with similar taste.
- Hybrid recommendations: These are the combinations of both content-based and collaborative recommendations.

We are using collaborative recommendations and user k-NN method for our system which is explained in Sects. 2.3 and 2.4 respectively.

2.2 Performance Measures

RMSE: Root-mean-squared error is a very good general-purpose error matric for numerical predictions [8]. Its value lies between 0 and ∞, 0 is the best value for any prediction and ∞ is the worst. Hence, this value should be minimized to prove performance of our model better.

MAE: Mean absolute error measures the average of magnitude of errors in a specific prediction [9]. Value of MAE also lies between 0 and ∞, 0 is the best values for any prediction and ∞ is the worst. So, our motive is to minimize this value for the better performance.

2.3 Similarity Measures

There are two main similarity measures which are present in Rapid Miner:

- Cosine-based similarity: This treats the two items as different vectors and the similarity is calculated on the basis of angle between these two vectors. It is also known as vector-based similarity.
- Pearson-based similarity: It checks how much the rating provided by a common user is different from the average rating of that item.

We used Pearson correlation mode because it provided more accurate results than Cosine for our dataset. Value of RMSE in case of Pearson is less than Cosine by a percentage of 10.66 as shown in Figs. 2 and 3.

2.4 Collaborative Recommendation

Collaborative recommendations are provided on the basis preferences of users which are similar in taste to new users [10]. We chose this over content-based because content-based cannot find out the quality of the item [11]. Collaborative recommendations work on collaborative filtering (CF) algorithm which works as follows [12]:

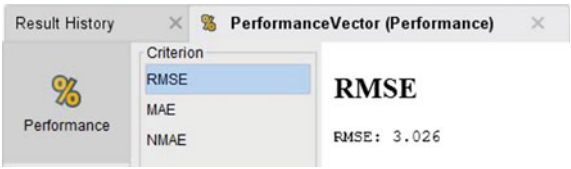

Fig. 2 RMSE of Pearson

Fig. 3 RMSE of Cosine

- Similarity values are calculated between two or more items in a dataset using one of the similarity measures. These measures are explained in Sect. 2.4.
- These similarity values are used to predict ratings for the entries not present in dataset.

In this paper, collaborative filtering is used along with the user k-NN to provide an approach for recommendation system. Collaborative filtering solves most of the shortcomings present in the content-based filtering [13]. Since feedback of other users creates difference between recommendations, there is a possibility of maintaining the effective performance. The approach of this research is as follows.

2.5 k-NN Algorithm

K-nearest neighbors is the method used for both regression and classification [14]. It is a type of instance-based learning and also called lazy learning. Following is the algorithm for k-NN approach.

It is a technique which uses K-instances as represented points in a Euclidean space.

- In K-NN classification, an object is classified by a majority vote of its neighbors, and the object is assigned to the class most common among its *K* nearest neighbors for discrete value.
- For real value, it returns the mean values of the K nearest neighbors (K is a positive integer, typically small). If K = 1, then the object is simply assigned to the class of that single nearest neighbor.

3 Methodology

The methodology to adopt for the research is depicted in Fig. 4:

Datasets from three excel sheets of BX-Book-Ratings, BX-User, and BX-Books details are integrated using data integration techniques.

1. The integrated data is pre-processed.
2. User k-NN algorithm is used for predictive analysis of training samples book ratings.
3. The predictive model is designed using rapid miner.
4. The model is tested using testing samples.
5. Performance of the model will be measured using performance measures named RMSE, MAE, and NMAE.

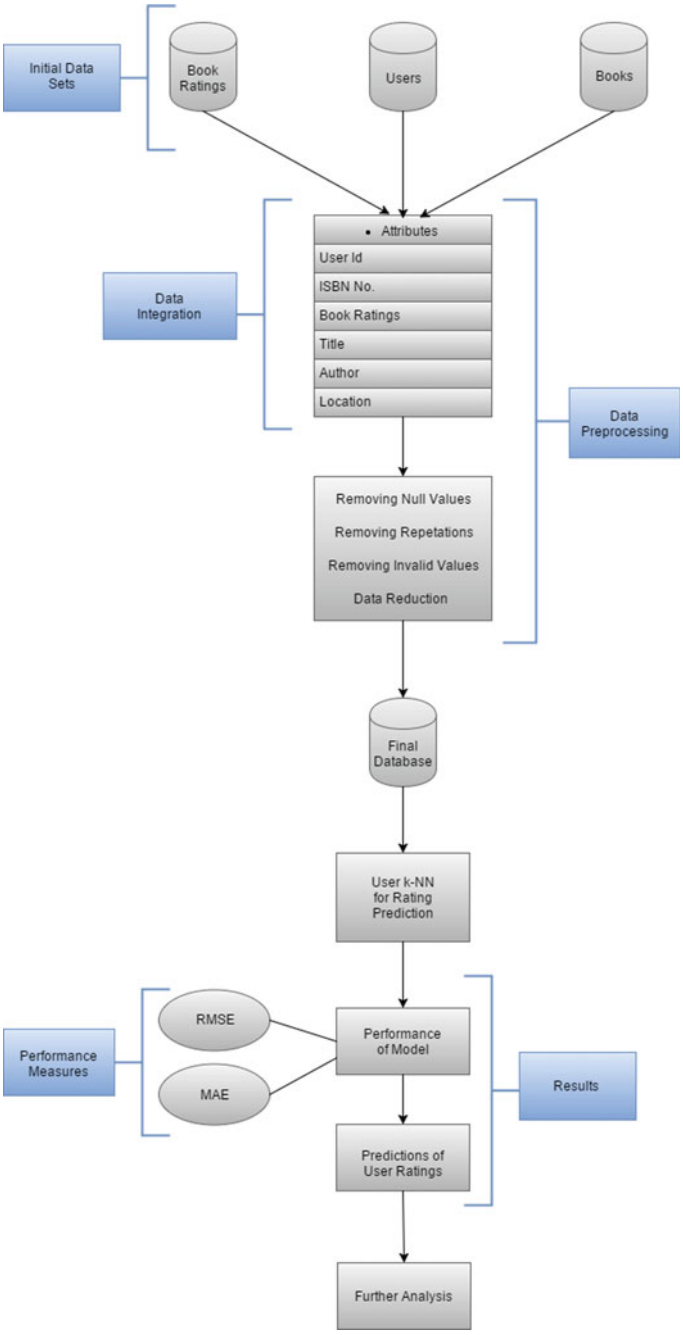

Fig. 4 Methodology

3.1 Data Integration

There were three files in the initial dataset with different attributes in them. Description of those files is provided in Fig. 5. To select most suitable attributes Pearson R Test is performed to calculate the similarity between attributes.

Attributes with high similarity were reflected as single attributes. Formula for Pearson R Test is given below:

$$r = \frac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2 (y-\overline{y})^2}}.$$

Manual integration is also performed to get most suitable attributes. For example, there were image URL in BX-Books excel files which are not usable to this research. Other attributes such as publisher details and year of publication were not relevant to this approach, and hence removed from the attribute list.

3.2 Data Pre-processing

- The dataset of book rating, user details, and book details had 1,149,780 ratings for 271,379 books.

Original Data	
Source	http://www2.informatik.uni-freiburg.de/~cziegler/BX/
No. of Excel Files	3
Names of Files	BX-Users, BX-Books, BX-Book-Ratings
No. of Attributes in BX-Users	3
Entries in BX-User	168097
No. of Attributes in BX-Books	8
Entries in BX-Books	271380
No. of Attributes in BX-Book-Ratings	3
Entries in BX-Book-Rating	1048576

Processed Data		
No. of Excel Sheet	-	1
No. of Attributes	-	6
Names of Attributes	-	User-Id, ISBN, Book-Rating, Book-Title, Author, User-Location
Reduction Range	-	Up to User-Id 5000
Total Entries	-	8660

Fig. 5 Metadata of dataset

- The user ids are made anonymous and mapped to integers.
- Six attributes User Id, ISBN No, Book Ratings, Title, Author, and Location were selected from set of different attributes.
- Data cleaning was performed and repeated; invalid and null values were removed.
- The dataset is reduced till 5000 user ids for better understanding of results.

4 Experimental Setup

4.1 Dataset Used

The dataset was collected in 4-week crawl from the Book-Crossing community. It was downloaded from official website of IIF [15]. The metadata of the original dataset is given and the pre-processed dataset is shown in Fig. 5.

4.2 Tool Used

The Rapid Miner data mining tools are used for the purpose of research and analysis in data mining. It is a tool with integrated environments for data mining, machine learning, predictive analysis, and text mining. It is used for information mining process including results, presentations, validation, and optimization. It provides a large pool of data loading, data transformation, data modeling, and data visualization methods [16].

4.3 Model Construction for Training

Model constructed in Rapid Miner for training of data which will be used to predict user ratings is shown in Fig. 6. Following steps describe the working and flow of the model:

1. "Read Excel" is used to import an excel file in the Rapid Miner process.
2. Set Role method specifies the role of each attribute present in the excel file [17]. In this model Book Ratings are specified as "label", ISBN as "item identification", User Id as "user identification" and all other attributes as "regular".
3. User k-NN is a model for rating prediction and can be used after installing an extension called "Recommender" in your Rapid Miner tool.
4. Apply Model implements the model selected and provides the final result of that model. Here User k-NN model is User k-NN and result is prediction.
5. "Performance" shows the accuracy and validity of your model.

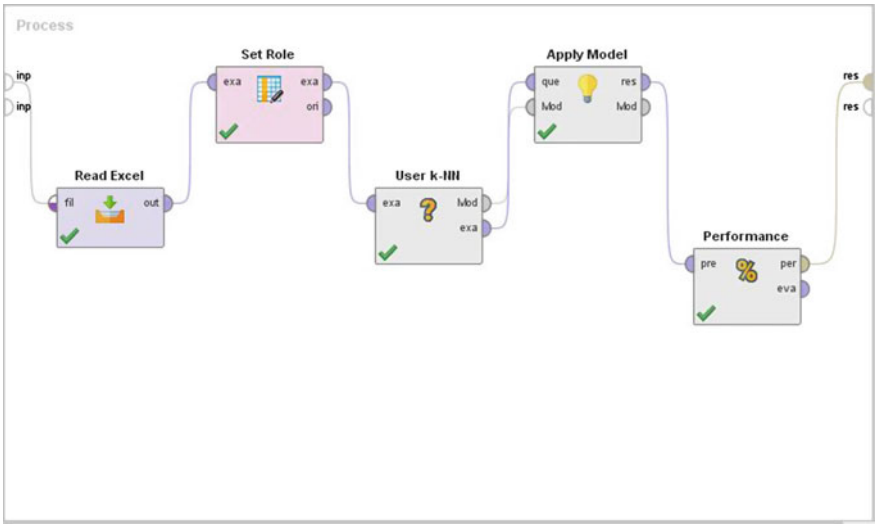

Fig. 6 Model designed for training of data

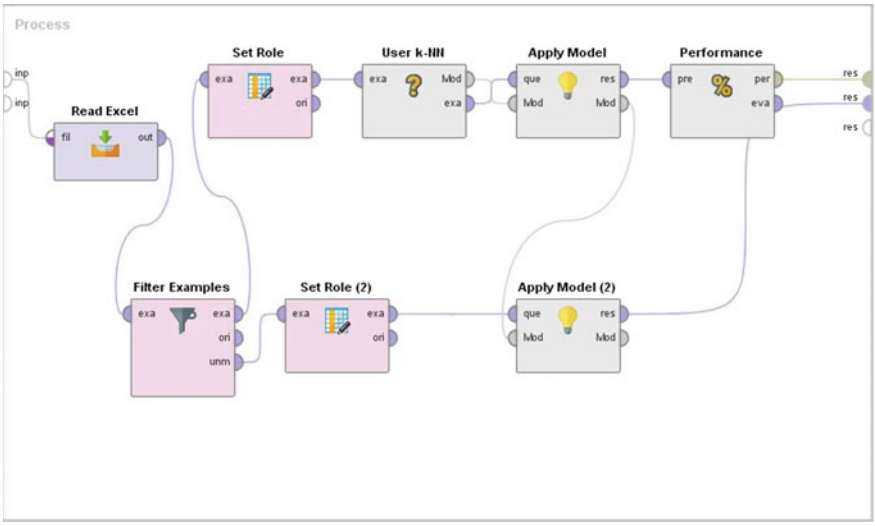

Fig. 7 Model designed for testing of data

4.4 Model Construction for Testing

Model constructed in Rapid Miner for testing of data is shown in Fig. 7. This model tests the prediction of ratings for the new users. Following steps describe the working and flow of the model:

1. "Read Excel", "Set Role", "User k-NN", "Apply Model" and "Performance" work same as in the Training Model.
2. "Filter Example" method separates empty values of user ratings from non-empty values.
3. Empty values are sent to "Apply Model2" which uses the training data and provide prediction for the empty values of user ratings.

5 Result and Analysis

5.1 Output

Outputs of training model and testing model are shown in Figs. 8 and 9, respectively. The model designed for rating prediction trained our dataset on basis of user ratings. Results of the training model are further used in testing of the data. The model designed for testing of data uses output from training model and provides prediction to new users. These results are used in further analysis in the paper.

5.2 Work Flow of Proposed Model

- The new user will enter a search item to the system.
- It can be author's name or a book title.

ExampleSet (8659 examples, 4 special attributes, 4 regular attributes) Filter (8,659 / 8,659 examples): all

Row No.	Book-Rating	Age	ISBN	User-ID	Author	Book_Title	User_location	prediction
1	0	18	195153448	2	Mark P. O. Mo...	Classical Myt...	stockton, calif...	2.649
2	0	26	1841721522	10	Celia Brooks ...	New Vegetari...	albacete, wis...	2.269
3	7	14	375759778	19	ARTHUR PHI...	Prague : A No...	weston, ,	4.506
4	0	19	425163091	20	Stephan Jara...	Chocolate Je...	langhorne, p...	2.649
5	0	24	067176537X	36	Dolores Krie...	The Therape...	montreal, qu...	2.649
6	7	17	553582747	42	Dean Koontz	From the Cor...	appleton, wis...	3.513
7	0	51	425182908	44	Patricia Corn...	Isle of Dogs	black mounta...	2.898
8	0	51	042518630X	44	J.D. Robb	Purity in Death	black mounta...	2.392
9	8	51	440223571	44	Maeve Binchy	This Year It W...	black mounta...	3.992
10	0	51	812523873	44	Laura J. Mixon	Proxies	black mounta...	2.392
11	0	51	842342702	44	Tim Lahaye	Left Behind: A...	black mounta...	1.868
12	9	34	440225701	51	JOHN GRISH...	The Street La...	renton, washi...	4.256
13	7	24	671623249	56	Larry McMurtry	LONESOME ...	cheyenne, wy...	4.576
14	0	24	679810307	56	SUZANNE FI...	Shabanu: Da...	cheyenne, wy...	3.176
15	9	24	679865691	56	SUZANNE FI...	Haveli (Laure...	cheyenne, wy...	4.976
16	7	32	2070423204	64	Michel Tournier	Lieux dits	lyon, rhone, fr...	4.506

Fig. 8 Output of training model

ExampleSet (2660 examples, 3 special attributes, 5 regular attributes) Filter (2,660 / 2,660 examples

Row No.	ISBN	User-ID	Age	Author	Book_Title	User_location	prediction
382	1551665077	3371	25	Nora Roberts	Last Honest ...	groveland, m...	3.454
383	1551665638	3371	25	Jayne Ann Kr...	Call It Destiny...	groveland, m...	3.454
384	1551665794	3371	25	Barbara Deli...	Twelve Across	groveland, m...	4.006
385	1558747109	3371	25	Jack Canfield	Chicken Sou...	groveland, m...	3.454
386	1583486259	3371	25	Tom Maremaa	Imagined	groveland, m...	3.454
387	1854879820	3371	25	Anne Styles	That Cinderel...	groveland, m...	3.454
388	1885983212	3371	25	Jack Kersh	Hotel Sarajevo	groveland, m...	3.454
389	006028871X	3373	30	Louise Renni...	Angus, Thon...	elk grove, cali...	3.454
390	60958022	3373	30	Joanne Harris	Five Quarters...	elk grove, cali...	4.155
391	61013420	3373	30	Stuart Woods	Worst Fears ...	elk grove, cali...	3.454
392	006109157X	3373	30	Stuart Woods	Dead Eyes	elk grove, cali...	4.214
393	61099368	3373	30	Stuart Woods	Palindrome	elk grove, cali...	3.454
394	61099805	3373	30	Stuart Woods	Swimming to ...	elk grove, cali...	3.454
395	312291639	3373	30	Emma McLa...	The Nanny Di...	elk grove, cali...	4.283
396	312957955	3373	30	William Hjort...	Falling Angel ...	elk grove, cali...	3.454
397	312995423	3373	30	Dan Brown	Digital Fortre...	elk grove, cali...	3.454
398	316789089	3373	30	Anita Shreve	The Pilot's Wi...	elk grove, cali...	4.168

Fig. 9 Output of testing model

- Then the user is asked for the required attributes which are age, location, and area of interest.
- Then the dataset which was created by the models will come in picture and will be used for the recommendation.
- Highest rated books of that author will be recommended to the user if he searched by the author.
- If he searched by title, then the books which are categorized in that group are recommended to the user.

Example: New user XYZ asks for following author:
"Manette Ansay"

Then all the books written by A. Manette Ansay will be searched from the dataset created by testing model and following is the sample of that data:

Here we have four books by requested author but the three books with highest rating will be sent as recommendation. The recommendations will be

1. Midnight Champagne by A. Manette Ansay
2. Sister by A. Manette Ansay
3. Vinegar Hill by A. Manette Ansay

Table 2 Values of performance measures

Models	RMSE	MAE	NMAE
Training model	3.025	2.652	0.295
Testing model	2.990	2.631	0.292

5.3 Performance Measures

Performance of prediction model is measured on factors defined in Sect. 2(B). Following table mention performance measures for both models (Table 2):

5.4 Analysis

We are following below-defined procedures for our further analysis and research work. On first access user is asked for following attributes:

- Age
- Location
- Area of Interest

These three possibilities are proposed using above-defined attributes and data created by our training and testing models.

Case study 1: **Recommendation using age**. When recommendations are provided to new user it cannot use ratings as a total base. Suppose new user is 25 years old and recommended item is rated high by persons of more than 60 years old. Then it will not be a fair recommendation for that user. So using output of testing model, new proposal is made which uses age of new user as a main attribute.

In Fig. 10, predictions provided by testing model are put together with users with different age to show the distribution between them.

The model shown in Fig. 11 uses age as an attribute of test data and finds similar objects in data trained by our model.

1. Age groups are created of range 10 using data of Fig. 10.
2. Suppose user lies in Group 1 which is of 0–10, then three books with highest ratings in that age group are fetched from training dataset.
3. These results are provided to the recommender system and will be produced as recommendations to the new user.
4. Next top three books are recommended in case user does not like provided recommendations.

Case study 2: **Recommendation using location**: As stated in case study 1, it is necessary to have an attribute which helps in providing more relevant recommendations. In this case, it is location of new user. On the basis of this, a proposal is made for better recommendations.

In Fig. 12, predictions provided by testing model are put together with users with different locations to show the distribution between them.

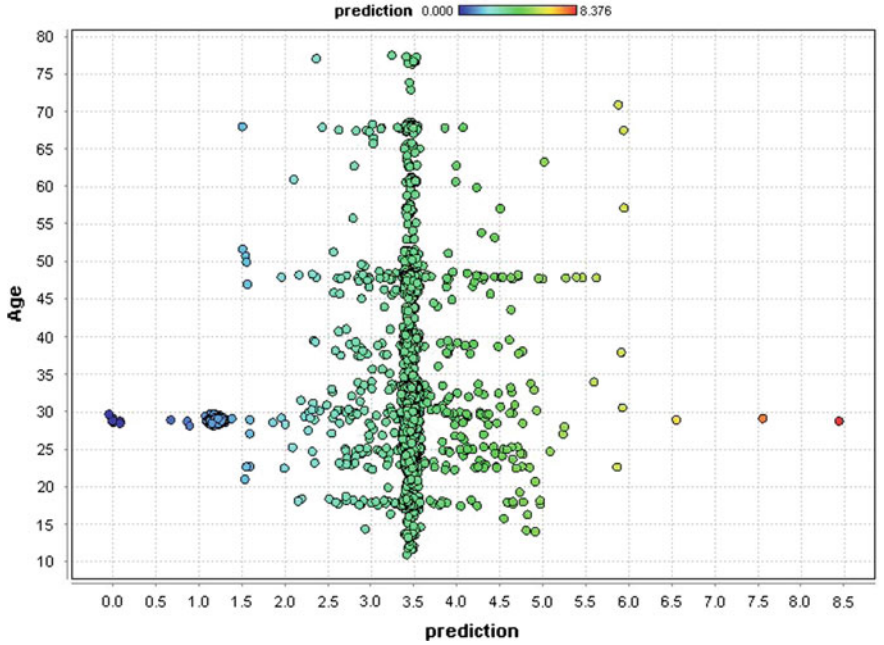

Fig. 10 Age-wise distribution of prediction

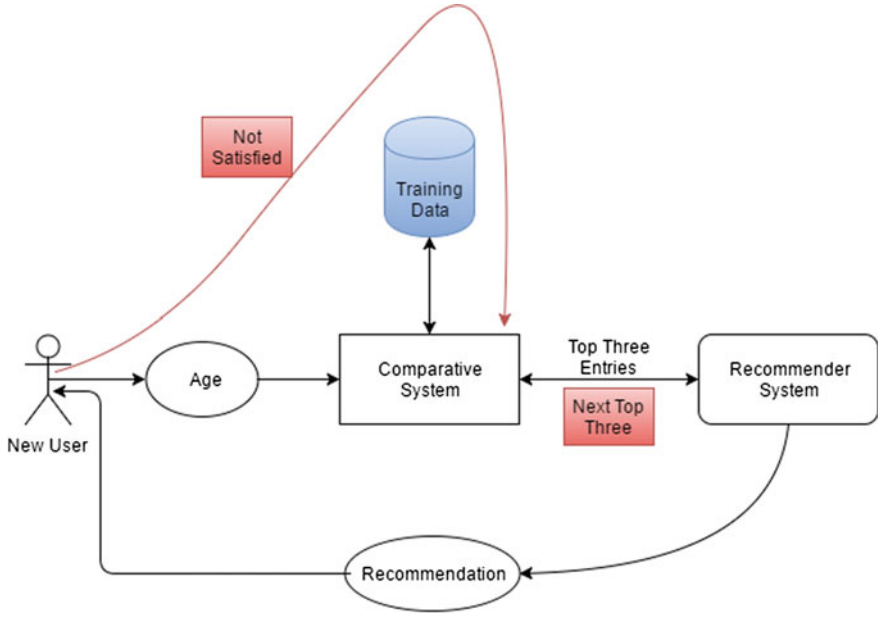

Fig. 11 Recommendation using age

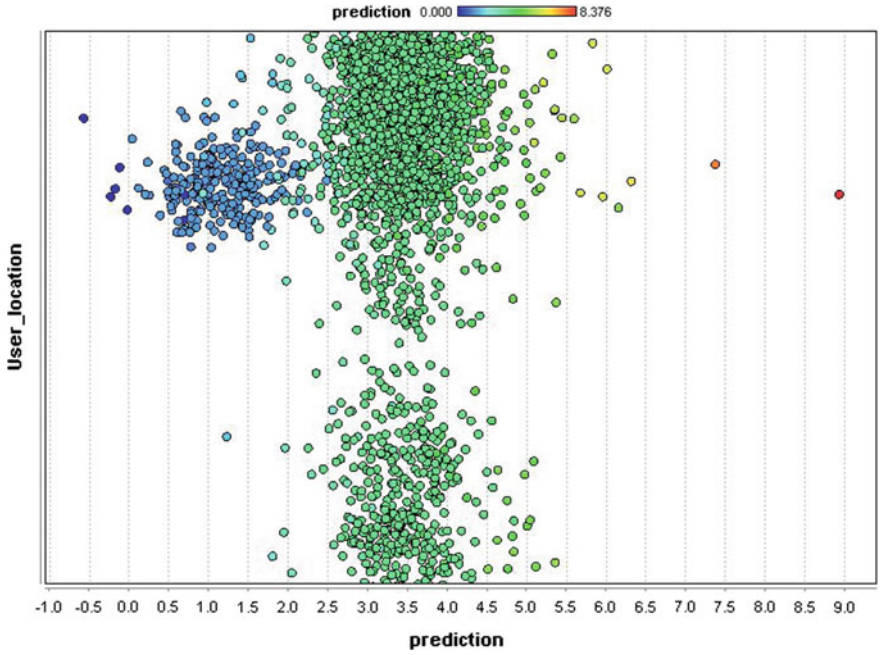

Fig. 12 Location-wise distribution of prediction

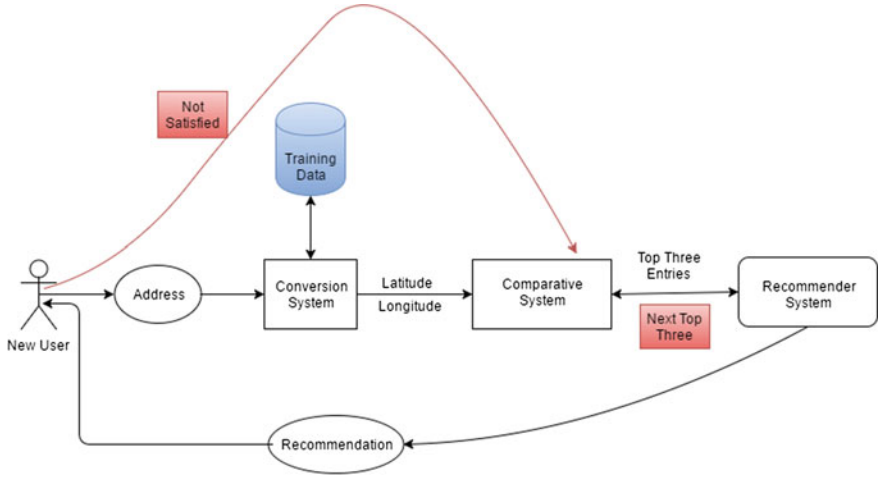

Fig. 13 Recommendation using location

The model in Fig. 13 uses location of users as an attribute of test data and finds similar objects in data trained by our model.

1. Addresses of users in training data and new users are converted to latitude and longitude values using data provided by Fig. 12.

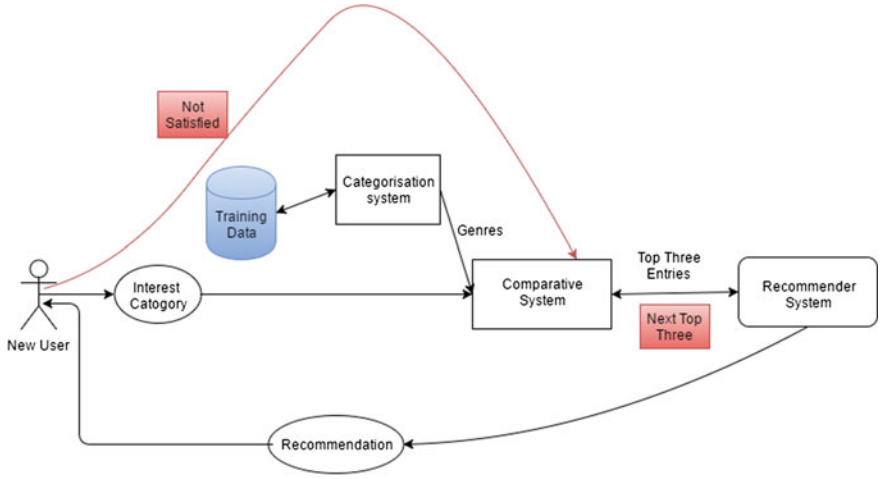

Fig. 14 Recommendation using interest

2. 10 values which are closest to the values of new user are selected.
3. Three books with highest ratings in those 10 entries are selected and sent to recommender system.
4. These results will be produced as recommendations to new user.
5. Next top three books are recommended in case user does not like provided recommendations.

Case study 3: **Recommendation using interest**:
This model uses Area of Interest as an attribute of test data and finds similar objects in data trained by our model (Fig. 14).

1. All books present in training data are categorized in different genres.
2. System provides list of genres and new user selects one of them according to related interest.
3. Three books with highest rating in that genre are selected and sent to recommender system.
4. These results will be produced as recommendations to new user.
5. Next top three books are recommended in case user does not like provided recommendations.

6 Conclusion

Predicted user ratings are well distributed with respect to our three main attributes. All case studies are applicable for development of proposed models except case study 3. It cannot be certified for development as the dataset does not have categorized entries on the basis of area of interest. In future the dataset used can be categorized on the basis of different genres, then it will be used for recommendation on the basis of area of interest.

Acknowledgements We sincerely thank Mr. Cai-Nicolas Ziegler and Book-Crossing community for collection of dataset. This data is freely available for research and we acknowledge the hard work done in the collection of data [18].

References

1. Zhang Haiyan, "Research on the Recommendation System Based on Social Tag (in Chinese)", Information Studies: Theory &Application, vol. 35, no. 5, pp. 103–106, 2012.
2. Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(6), 734–749.
3. Rich, E. (1979). User modeling via stereotypes*. *Cognitive science*, *3*(4), 329–354.
4. Powell, M. J. D. (1981). *Approximation theory and methods*. Cambridge university press.
5. Lilien, G. L., Kotler, P., & Moorthy, K. S. (1992). *Marketing models*. Prentice Hall.
6. Salton, G. (1989). Automatic Text Processing. Addison Welsley. *Reading, Massachusetts*, *4*.
7. Balabanović, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, *40*(3), 66–72.
8. https://www.kaggle.com/wiki/RootMeanSquaredError.
9. http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_cont_var/uos3/uos3_ko1.htm.
10. Tewari, A. S., Kumar, A., & Barman, A. G. (2014, February). Book recommendation system based on combine features of content based filtering, collaborative filtering and association rule mining. In *Advance Computing Conference (IACC), 2014 IEEE International* (pp. 500–503). IEEE.
11. Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web* (pp. 285–295). ACM.
12. Xin, L., Haihong, E., Junde, S., Meina, S., & Junjie, T. (2013, December). Collaborative Book Recommendation Based on Readers' Borrowing Records. In *Advanced Cloud and Big Data (CBD), 2013 International Conference on* (pp. 159–163). IEEE.
13. Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence*, *2009*, 4.
14. Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. Systems, Man and Cybernetics, IEEE Transactions on, (4), 580–585.
15. http://www2.informatik.uni-freiburg.de/~cziegler/BX/.
16. https://RapidMiner.com/products/studio/.
17. http://docs.rapidminer.com/studio/operators/.
18. Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, Georg Lausen; *Proceedings of the 14th International World Wide Web Conference (WWW '05),* May 10–14, 2005, Chiba, Japan.