E-book recommendation system using content-based filtering

1 Aarush Gandhi
Department of Computer Science & Engineering. KIET Group of Institutions, Ghaziabad-201206, UP, India
aarushgandhi73@gmail.com

2 Akshat Patwal
Department of Computer Science & Engineering, KIET Group of Institutions, Ghaziabad-201206, UP, India
akshatpatwal31@gmail.com

3 Shaswat Kumar
Department of Computer Science & Information Technology, KIET Group of Institutions, Ghaziabad-201206, UP, India
shaswat.kumar03@gmail.com

4 Sushil Kumar
Department of Computer Science & Engineering, KIET Group of Institutions, Ghaziabad-201206, UP, India
drsushil.cs@gmail.com

5 Shrankhla Saxena
Department of Computer Science & Information Technology, KIET Group of Institutions, Ghaziabad-201206, UP, India
shrankhla.saxena@kiet.edu

Abstract

The system known as the book recommendation system could be a new form of web tool which helps users to get the names of the books of their interest. Using a web recommender is relatively an easy and quicker step to get names of the books and this can be done in very short time. This system directs our user towards those books which can meet their interest through cutting down large databases of books. Best strategy to increase profits and attract customers would be a recommendation system. The prevailing methodologies enable the systems to gather the immaterial information and cause a downfall in attracting the users and finishing there is a fast and reliable method. In this paper we provide the running model of the recommendation systems that's presently used in the web book searching domain. This research paper shows a simple system for book recommendations that help the user to get the best book of their interest. As we all know there are thousands of books of the same genre so all the readers are very confused about which book they should read first. For instance this book recommendation system comes into the picture which suggests you the best book of your interest so that you do not confuse further. In this research paper

we showed you different models to get the best results of books for the customers. Models used are popularity based model (top in whole collection), Popularity based model (top in given place), Same author, and publisher of given book name, books popular yearly, average weighted rating based, correlation based. We also include some filtering techniques like -collaborative filtering, content filtering, nearest neighbor. Firstly we filter our data by using the filtering techniques and then we take intersections of all the models we stated above and then present our records to the customer.

Keywords: Collaborative filtering, Content based filtering, Memory based approach,Recommender system

1. Introduction

A recommender system is basically a taxonomic category of information filtering system that helps users to predict their preferences. Nowadays this system is very popular as every big company is using this system to recommend their product to the consumers. Applications like Netflix, Youtube use this kind of system to recommend the preferred product to the consumer. For specific topics like restaurants and online qualitative analysis, there also are common recommender systems. to look at analysis articles and specialists, partners, money services and insurance, and recommendation systems are developed.

Recommenders usually use some filtering methods like collaborative filtering and content based filtering. Collaborative filtering filters information by using the interactions and data collected by the system from other users. It's based on the idea that people who agreed in their evaluation of certain items are likely to agree again in the future. Collaborative filtering systems focus on the relationship between users and items. The similarity of items is determined by the similarity of the ratings of those items by the users who have rated both items. A content based recommender works with data that the user provides, either explicitly (rating) or implicitly (clicking on a link). Based on that data, a user profile is generated, which is then used to make suggestions to the user. As the user provides more inputs or takes actions on the recommendations, the engine becomes more and more accurate. We use this content based filtering because no data from other users is required to start making recommendations, recommendations are highly relevant to the user. Recommendations are transparent to the user, content-based filtering systems are generally easier to create.

We also used some models like popularity based model(top in whole collection), popularity based model (top in given place), same author and publisher of given book name, books popular yearly, average weighted rating based, correlation based. In the popularity based model (top in whole collection) model we firstly sort the dataset according to the rating in non increasing order and then recommend top n books. In this we take some number of top rating books for the preferred genre. In the popularity based model (top in given place) model firstly we filter out the books on the basis of location then sort the dataset and then recommend top n books. In this model we take some number of books according to the location of the user. In the same author and publisher of given book name model we firstly filtered out the books by the same author or same publisher and then we sorted it according to their rating and recommended top n books. In the popular yearly model we group books which are published yearly and then top books for that year. By this if the user needs a book of the genre which got famous this year will have more priority.

2. Literature Survey

To provide recommendations, there are two approaches that are traditionally used, collaborative filtering (CF) and content-based filtering (CBF). In the CBF approach we work on the content of elements which is based on the user profile like whether the user wants to read action or adventure books and the type of books the user likes [20]. But in CF, it does not use content, it uses past history of the user like what kind of books the user read in past and likes it and by recommending similar books there is a chance that user might recommend books [3]. We builded some recommendation systems which use some models which are demonstrated in Figure 1.

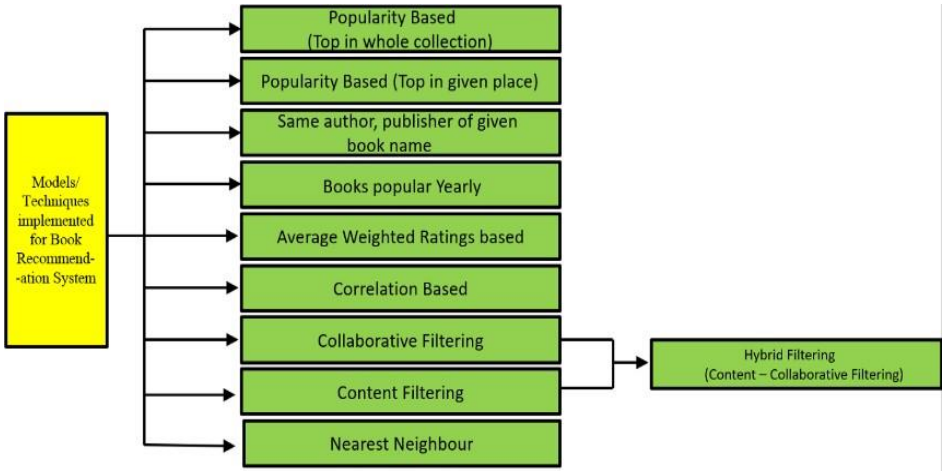

Figure 1: Recommendation models used

Input that is given for the model is:

Enter a book name: Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback)) Enter number of books to recommend: 5

1. Popularity Based (From whole collection):

In this model we first sort the dataset according to the rating in non increasing order and then recommend top books . In the table 1 we can see that

	ISBN	Book Rating	Book-Title	Book-Author	Publication Year	Publisher
408	0016666343	707	The Lovely Bones: A Novel	Alice Sebold	2002	Life Brown
26	0671880107	581	Wild Animus	Rich Shapero	2004	Too Far

748	0385504208	488	The Cd Code	Dan Brown	2003	Doubleday
522	00121955	383	The Red Tent (Best Selling Backst)	Anita Diamant	1998	Paxtor USA
1105	0060828336	320	Divine Secrets of the Ya-Ya Sisterhood: A Novel	Rebecca Wells	1997	Pennai
77384	058035342X	315	Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback)	JK Rowling	1999	Arthur A Line Books
356	0142001740	314	The Secret Life of Bees Sue	Monk Kidd	2000	Penguin Books

Table 1. Top books on the basis of rating

2. Popularity Based (on the basis of location):

In this model firstly we filter out the books on the basis of location then sort the dataset and then recommend top books (refer to table 2).

Enter the name of place: india

	ISBN	**Book Rating**	**Book-Title**	**Book-Author**	**Year-Of-Publication**	**Publisher**
26	09715841	3	Wild Animus	Rich Shapero	2004	Too Far
169	067104761	2	Skin And Bones	Franklin W. Dixon	2000	Aladdin
167	0486264735	2	Pride and	Jane Austen	2000	Dover

			Prejudice (Dover Thrift Editions)			Publications

Table 2. Top 3 books on the basis of location

3. Books by the same author, publisher of the given book name;

In this model, we firstly filtered out the books by the same author or same publisher and then we sorted it according to their rating and recommended top n book [15].

Books by same author:

Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Order of the Phoenix (Book 5)
Harry Potter y el cielo de fuego.
Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Sorcerer's Stone (Book 1)

Books by same publisher:

The Seeing Stone The Slightly True Story of Cedar B. Hartley: Who Planned to Live an unusual Life
Harry Potter and the Chamber of Secrets (Harry Potter)
The Story of the Seagull and the Cat who Taught Her To Fly Book! Book! Book!

4. Books popular yearly

In this model we group books which are published yearly and then top books for that year (refer to table 3).

	ISBN	**Book-Title**	**Book-Author**	**Year-Of-Publications**
253750	964442011X	Tash khun	JamaA' Fash	1378
102496	0373226888	Tommy's Mom	Linda O. Johnston	1902
170971	0404089119	Charlotte Bronte	Clement K.	1906

		and Her Sisters	Shorter	

Table 3: Top books that published yearly

5. Content Based Filtering

In this model we recommend books that have similar types of content and similar titles, basically here we look for similarities in books [3].

Recommended Books:

Harry Potter and the Sorcerer's Stone (Book 1)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Order of the Phoenix (Book 5)

3. DATA SET

The dataset we are using in this work is the Book-Crossing Dataset that comprises of 3 tables:-

1. Books- This table has 8 columns. The names of the columns are ISBN, Book title, Author, Year of publication, Publisher, and URLs representing three different versions of images (large, medium and small)
2. Users- This table has 3 columns. The names of the columns are UserID, Location and age.
3. Ratings- This table also has 3 columns. The names of the columns are UserID, ISBN and Rating.

Books Data: **(271360, 8)**
Users Data: **(278858, 3)**
Books-Ratings: **(1149780, 3)**

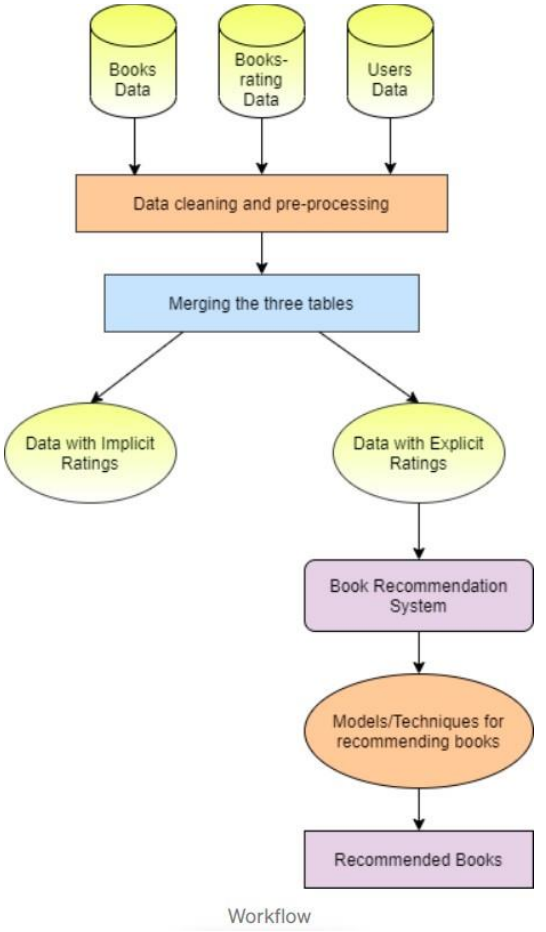

Figure 2: Workflow of the Project

Data cleaning and preprocessing

All the pre-processing and cleaning we have done on the dataset is described below:

Books Table (refer to table 4):
1. Drop all the three image URL features.
2. Check for the number of null values in each column in the table. There are only 3 null values in the table. Replace these three null cells with 'Other'
3. Check for all the unique years of publications in the year of publication column. Two values in this column would appear to be publishers. Also three tuples have author and book name merged in the same column.
4. Manually set the values for these three tuples for each of their features using the ISBN number obtained before.
5. Convert the data-type of the years of publication column to integer.
6. By keeping the range of valid years as less than 2023 and not 0, replace all the invalid years with the mode of publications i.e. 2002.
7. Upper-case all the alphabets in the ISBN column and remove the duplicate rows (tuple) from the table.

ISBN	Book-Title	Book Author	Year of Publication	Publisher
0195153448	Classical Mythology	Mark P.O. Morford	2022	Oxford University Press
0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada
0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial
0393045218	The Mummies of Urumchi	E.J.W. Barber	1999	W.W. Norton & amp; company

Table 4. Books Table

Users Table (refer to table 5)

1. Check for all the null values in the table. The Age column has more than 1 lakh null values.
2. Check for unique values present in the Age column in the Users Table. There are many invalid ages present like 0 or 244.
3. By keeping the valid age range of readers from 10 to 80, replace all the null values and invalid ages in the Age column with the mean of the rest of the valid ages.
4. The location column in the Users Table has 3 values: city, state, and country. These are split into 3 different columns named as City, State, and Country respectively.
5. Removal of duplicate entries(tuples) from the table would be done.

User-ID	Age	City	State	Country
1	35	nyc	new york	usa
2	18	stockton	california	usa
3	35	moscow	yukon territory	russia
4	17	porto	v.n. gaia	portugal
5	35	farnborough	hants	united kingdom

Table 5. Users Table

Ratings Table(refer to table 6)

1. Check for all the null values in the Ratings table.
2. Check for the Rating column and User-ID column to be an integer in the same.

3. Remove punctuation from values of ISBN column and if that resulting ISBN is available in the book dataset only then consider it, else drop that entity.
4. Upper-case all the alphabets present in the ISBN column in the same table.
5. Removal of duplicate entries(tuples) from the table would be done.

User-ID	ISBN	Book-Rating
276725	034545104X	0
276726	0155061224	5
276727	0446520802	0
276729	052165615X	3
276729	0521795028	6

Table 6: Ratings Table

Merged Dataset (refer to table 7)
All three tables are merged and for the final dataset tuples having ratings of 0 are dropped.

ISBN	Book-Title	Book-Author	Year-of-Publication	Publisher	User-ID	Book Rating	Age	City	State	Country
0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	8	5	35	Timmins	Ontario	Canada
074322678x	Where You'll Find me and Other Stories	Ann Beattle	2002	Scribner	8	5	35	Timmins	Ontario	Canada
0887841740	The Middle Stories	Sheila Heti	2004	House of Anansi Press	8	5	35	Timmins	Ontario	Canada
1552041778	Jane Doe	R.J. Kaiser	1999	Mira Books	8	5	35	Timmins	Ontario	Canada

Table 7. Merged Dataset

4. PERFORMANCE AND RESULT ANALYSIS

Here we are presenting the accuracy measures of the proposed approach by constructing a confusion matrix obtained by testing on the test data. The below graphs give a comparison of the recall, precision, Fscore and MAP (Mean Absolute Presidion) of PCC, CPCC, Cosine and Jaccard.

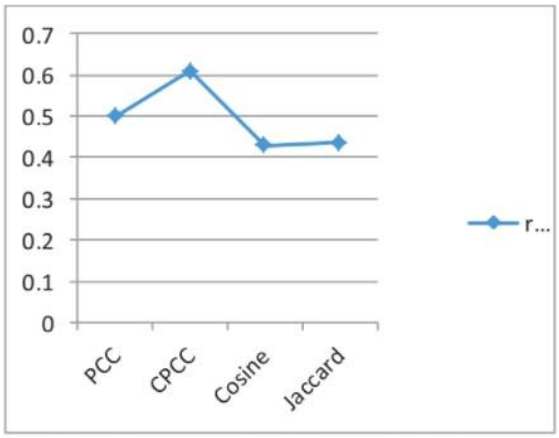

Figure 4. Recall measure for different similarity measures

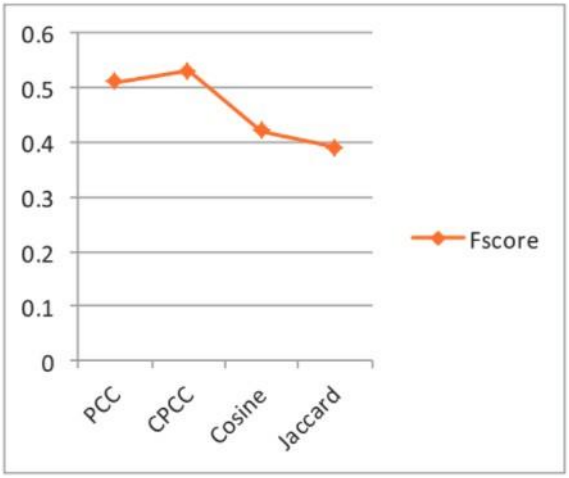

Figure 5. Fscore for different similarity measures

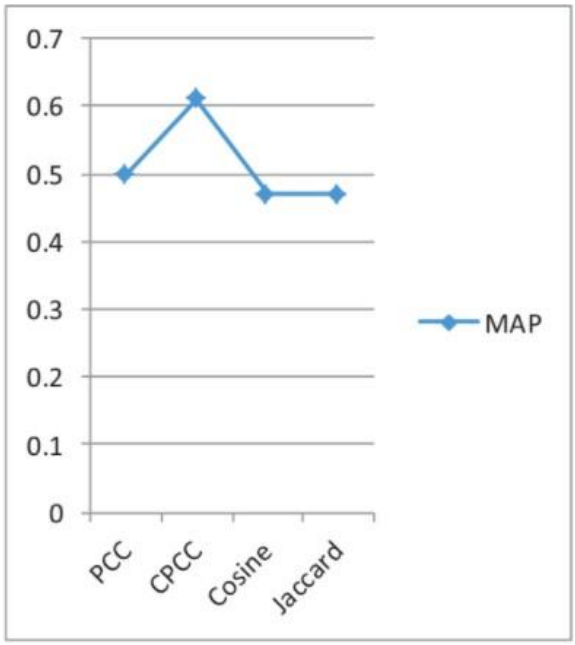

Figure 6. MAP for different similarity measures

5. CONCLUSION

In this paper we have introduced new techniques to recommend books using user based collaborative filtering for assisting the users in an efficient way. We also compared some similarity measures and found the best one. The user based collaborative filtering method builds a user-user similarity matrix by using some similar measures to present the similarity between users in order to utilize it for further processing. We can deduce from the results and visualizations that rating the accuracy followed by a normal distribution implies its consistency and efficiency. Future work should instead target to protect the system data against attacks and develop the different algorithms.

References

[1] Abhilasha Sase, KritikaVarun, SanyuktaRathod, Prof. DeepaliPatil(2015),"A Proposed Book Recommender System", International Journal of Advanced Research in Computer and Communication Engineering.

[2] Murthy ,K.V.S.S.R., N Satyanarayana ,K.V.V (2018),"Intrusion detectioCosineCosinen mechanism with machine learning process " International journal of Engineering and Technology.

[3] Zhongqiu, ZhichengDou, Jianxun Lin, XingXie and XiangYang(2015)," Content-Based Collaborative Filtering for News Topic Recommendation", Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[4] Narsinga rao,M.R.,Venkatesh Prasad ,V.,Sai Teja,P.,Zinda Wali,M.phanindra Reddy,O.(2018),"A survey on Prevention of Overfitting in convolutional neural networks using machine learning techniques' 'International journal of engineering and Technology.
[5] E. UkoOkon, B. O. Eke, P. O. Asagba(2018)," An Improved Online Book Recommender System using Collaborative Filtering Algorithm", International Journal of Computer Applications.
[6] LavanyaReddy,L.Srisai Krishna,J,&vaishnavi,G.V.S.S(2018),"Survey on software reliability prediction using machine learning techniques", Journal of Advanced research in dynamical and Control Systems"
[7] ManiMadhukar(2014)," Challenges & Limitation in Recommender Systems", International Journal of Latest Trends in Engineering and Technology.
[8] Lakshmi ,C.r.,rao, D.T.,&Rao,G.V.S(2018),"Fog detection ad Visibility enhancement under partial machine learning approach"International Conference on Power,control,Signals and instrumentation Engineering
[9] Dharna Patel, Dr. Harish Patidar(2018)," Hybrid Recommendation Solution for Online Book Portal", International Journal for Research in Applied Science & Engineering Technology
[10] Krishna Mohan ,G.,Yoshitha ,N.,Lavanya ,M.L.N,&KrishnaPriya,A. (2018) ,"Assessment and Analysis of software reliability using machine learning techniques"International Journal of Engineering and technology
[11] ParulAggarwal, Vishal Tomar, AdityaKathuria(2017)," Comparing Content Based and Collaborative Filtering in Recommender Systems",International Journal of New Technology and Research.
[12] Surarchita,V.,VenkateswaraRao ,P.,bhattacharyya ,D.,Kim,T.(2016) ,"Classification of penaeid prawn species using radial basis probabilistic neural networks and support vector machines"InternationalJournel of Bio-Science and Bio-technology
[13] Dr. Mohammed Ismail 1 ,Dr. K. BhanuPrakash ,Dr. M. NagabhushanaRao (2018)"Collaborative filtering-based recommendation of online social voting" International Journal of Engineering & Technology.
[14]Vidhyullatha ,p.,Rajeswara Rao,D.(2016),"Machine learning techniques on multi dimensional curve fitting data based on r-square and chi-square methods"International Journal of Electrical and Computer Engineering
[15] Immidi Kali Pradeep, M.JayaBhaskar(2018)"Comparative analysis of recommender systems and its Enhancements" International Journal of Engineering & Technology.
[16] Anila,M.,&Pradeepini ,G.(2017),"Study of prediction algorithms for selecting appropriate classifier in machine learning"Journal of Advanced Research in Dynamical and Control Systems
[17] Item-Based Collaborative Filtering Recommendation Algorithms(2001)" BadrulSarwar, George Karypis, Joseph Konstan, and John Riedl"
[18] Atmakur,V.K.,n SIvaKumar,P.(2018),"A prototype analysis of machine learning methodologies for sentiment analysis of social networks",International Journal of engineering and technology
[19] Dhruv, Ajay, AasthaKamath, Anuja Pawar, and Karan Gaikwad. "Artist Recommendation System Using Hybrid Method: In Emerging research in computing, Information, Communication and Applications, pp. 527 -542. Springer, Singapore, 2019
[20] Raghuwanshi, Sandeep K., and R. K. Patreiya, "Collaborative Filtering Techniques in Recommendation Systems" In Data Engineering and Applications, pp, 11-21, Springer, Singapore, 2019.