Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/3328
DOI URL: http://dx.doi.org/10.21474/IJAR01/3328

RESEARCH ARTICLE

INTELLIGENT RECOMMENDATION SYSTEM.

Shailendra Singh Kathait[1], Shubhrita Tiwari[2] and Piyush Kumar Singh[3].

1. Co-Founder & Head of Analytics, Valiance Solutions.
2. Data Scientist, Valiance Solutions.
3. Indian Institute of Technology, Kharagpur.

Manuscript Info

Manuscript History

Received: 23 December 2016
Final Accepted: 29 January 2017
Published: February 2017

Key words:-
Hybrid Recommendation · Publication sites · Content-Based Filtering · Collaborative Filtering · Item-Item Collaborative Filtering · User-User Collaborative Filtering · Trending Articles · User persona · Novel combination · Tags

Abstract

Searching for articles of interest on publication sites can be difficult and time-consuming. Sometimes it takes lot of efforts to find the most relevant article because of which the reader looses interest completely. Recommender systems help the users find articles of their interest with personalized suggestions. In this paper, Hybrid Recommender System is implemented which is a novel combination of content-based filtering, collaborative filtering, trending article algorithm and user persona and recommend articles considering all the possible factors. User short-term interest is catered by suggesting trending articles while long-term interest is catered by observing what kind of content the user prefers to read and by finding out similar users and recommend what they are reading. The model makes the recommendation based on tags assigned to each article and knowledge of articles read by each user. This model doesn't require ratings of articles by each user as generally users usually don't rate article after reading them as compared to giving rating to movies after watching. The model built takes into consideration many aspects including the trend emerging at current time as well the interest of the user, the time period, geographical location, browsing history etc. then make recommendations accordingly.

Introduction:-

Articles in huge numbers are published every day across different categories. The information portal sites include articles of categories like stock markets, finance, banking, insurance, entertainment, social feeds etc.

Web sites are deploying recommendation systems for suggesting articles to users according to their taste. A key part of the news is that user has a long-term interest in certain categories and short-term interest in recent happenings. The short-term interest of user about some recent event can be dealt by recommending the trending article of that time on the basis of view counts of articles within a particular timeframe. As far as the long-term interest of the user is concerned, the recommendation can be done on the basis of user behavior and preferences. For this purpose, content-based filtering and collaborative filtering techniques are used to generate the recommendation.

Corresponding Author:- Shailendra Singh Kathait.
Address:- Valiance Solutions, A-75, Sector-58, Noida, Uttar Pradesh, 201301, India

Content-based filtering finds articles which are similar on the basis of tags assigned to each article. Each article is assigned weights on the basis of term frequency and inverse document frequency of each tag. After which user probability of reading an article is calculated. On the other hand, collaborative filtering uses the correlation between the articles on the basis of the ratings given to article by different users.

The disadvantage of content-based filtering is that it leads to over-specialization that is the recommended article is similar to already read article and may not be useful for the user. This method does not use the interaction information between users to generate recommendations.

Collaborative filtering relies on past preferences or rating correlation between users. However, this technique can lead to bad prediction if the article is unpopular and very few users have given feedback about them.

To overcome these difficulties, a hybrid model is proposed that takes into account all the possible aspects that contribute towards making the most relevant recommendation to the user.

In this paper, we have developed a hybrid intelligent model, in which users are suggested articles on the basis of following factors:-

Directional Variables:-

1. Trending Articles for a particular time period
2. Reader's interest (based on browsing history)
3. Geographical Location
4. Time Period
5. Reader's specifications (gender, age etc.)
6. Reader's behavior.
7. Other Parameters like time of log-in etc.

Considering all the above mentioned aspects and applying appropriate filtering, the most relevant recommendation is generated.

Related work:-

The related work done in this either uses collaborative filtering or content-based filtering for recommendations. The other factors that contribute to the recommendations is ignored, as a result of which irrelevant recommendations are made in many cases. For example if the geographical location of the reader is not considered, then he'll be recommended articles of a different nation, to which he has nothing to do. In our project, we made use of all the possible information of the user as well as of the articles to make recommendations. The most appropriate and relevant tags were assigned to the articles that considered all the data that can be extracted from the articles and summarized it. All the user attributes are considered to make the most suitable recommendation of his interest.

In Item-Based Collaborative Filtering Recommendation Algorithms [1], only collaborative filtering approach is used for recommendations that only considers the article's similarity irrespective of trending articles and other parameters.

In Personalized News Recommendation Based on Click Behavior [2], again the user behavior is considered and on the basis of past browsing history, the recommendations are made.

In Content based recommender systems [3], the ratings given by the reader is considered as an important parameter in-order to make appropriate recommendations. The interests of the reader that do not give rating to the article are left unconsidered in this case.

In order to build an intelligent system, all the possible parameters should be considered along with their weights (impact factors), that generates the most relevant recommendation for the user.

Information Retrieval:-

For the purpose of the recommendation of articles, a lot of data on user's reading habit and clicks counts of the article have to be extracted beforehand. In this paper, article attributes such as click counts (number of times

article is read), time-stamp (time at which article is read), article tags (different genres which tells what the article is talking about, for e.g. an article can have tags such as Stocks, Finance, Markets, Demonetization, Sports etc.) are known.
The Intelligent System developed by us was for a leading publishing house of Asia. The publishing house provided data that consisted of the following information:

1. Articles' Content
2. Articles' Ratings
3. User browsing history
4. User details
5. Article's Tags

The tags were assigned to the articles by implementing an unsupervised algorithm that assigned the most relevant tag to the articles. The algorithm implemented segmentation of Chinese characters, removal of stop words, preparation of dictionary, implementing TF-IDF concept and finally multi-stage tagging that generated the relevant tags for articles.

The information on user reading habit was provided in the form of articles read by different users. The sample dataset containing attributes of articles is shown in Table 1, dataset containing attributes of users is shown in Table 2, and dataset containing details of user is shown in Table 3.

User-ID	Login time	Articles read
U1	Male	A1,A7
U2	Female	A100
U3	Male	A12,A3

Table 1:- Article Attributes

Article-ID	Tags	Time of publishing	Click-count
A1	Finance, Stocks, Economy	3rd January 2017	678
A2	Wine, Finance, Trading	8th March 2012	8753
A3	Sports, Basketball	6th April 1992	5300

Table 2:- User Attributes.

User-ID	Gender	Age	Location
U1	Male	30	G1
U2	Female	31	G2
U3	Male	45	G3

Table 3:- User Specifications

Proposed Methodology:-

Hybrid Recommender System (HRS) suggests articles to user considering the short-term as well as the long-term interests. The different techniques that Hybrid Recommender System uses are as follows [4]:

1. For short-term interest of the user, trending articles are recommended.
2. For long-term interest of user, articles are recommended on the basis of articles read by user.
3. Other important parameters (user attributes) are considered to make recommendations.

The following different approaches are used for the purpose of recommendation:-

a. Content-based filtering.
b. Item-Item Collaborative filtering.
c. User-User Collaborative filtering.

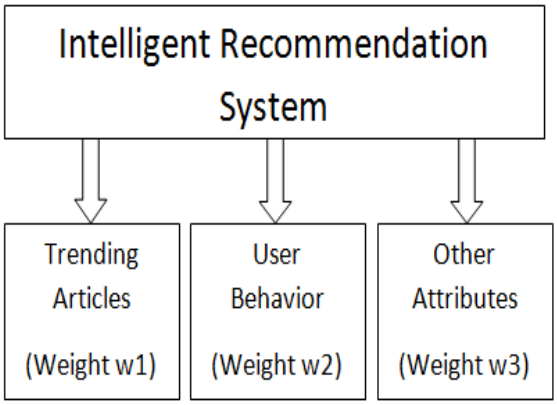

Figure 1:- Recommender System Block Diagram

Recommending Trending Articles:-
Recommending trending articles imply recommending the latest trending articles that has got maximum views in a given time-frame and are the most trending topics of interest. The following methods were used for recommending trending articles to the user:

Firstly, on the basis of the views (click-count) and time of publishing of the article, the trending article of last months, 30 days, 7 days, and 1 day was selected for recommendation. The first step in this is the categorization of the articles. The articles are categorized in to different categories like Business, Art, Entertainment, Education, History etc. This categorization is done on the basis of tags of articles. After the categorization is completed, the time-frame for which the trending articles will be selected is decided, that varies from category to category. For example, for the "News" category, the trending article of last 1 or 2 day is selected, for entertainment, the trending article of last 30 days can be selected, for stock market, the trending article of last 1 or 2 hour will be a relevant recommendation. The final recommendation is built after sorting of the selected categorized articles for a given time-frame, on the basis of click-counts (number of views of article) and determining article which has highest click counts in each category in given time-frame. The time-frame chosen can be last 3 months, last 30 days, last 15 days, last 7 days, last 1 day, last 1 hour etc. This technique recommends articles to users which can cater to their short-term interest.

The trending articles in Twitter as well as in other networking sites was determined for recommendation.

The trending article was further determined on the basis of user behavior. This is done by observing user 's hourly, daily and weekly interests, that is, the type of article user reads at a particular time of the day, on weekdays and weekends etc. For example, if user watches news at night on weekdays, and entertainment articles on weekends, then the recommendation at these instances (that is at night on weekdays and on weekends) is built around news and entertainment. Similarly other behavior of users is observed and recommendations are built on the basis of this. In the same manner, user-user matching can also be implemented, that is recommending articles read by 1 user at a particular instant to other user that has read one or more articles in common with the 1st user but has not read other articles read by 1st user.

The trending articles obtained in step-1, step-2 and step-3 are taken and are assigned some weights (w1 to articles from step-1, w2 to articles from step-2 and w3 to articles from step-3), w2 chosen to be maximum and w3 minimum. On the basis of these weight-ages, number of recommendations and the order of these recommendations is developed.

Content-based filtering of Articles:
A content-based filtering system selects items based on the correlation between the content of the items and the user's preferences as opposed to a collaborative filtering system that chooses items based on the correlation between people with similar preferences [4][5]. A genre matrix is built in which each row tells which tags a specific article contains. It can be compared as specific row of the genre matrix signifying an article 'A001' is a

k-dimensional vector, where each dimension corresponds to distinct tags and 'k' is the total number of tags in document. The dimension of matrix is m ∗ k, where m is number of articles and k is number of distinct tags in all the m articles. A sample Article-Genre matrix is shown in Table 4.

Table 4:- Article Genre Matrix

Article	Chemicals	Equipment	Wine	Trading	Finance
A1	1	1	0	0	0
A2	0	0	0	0	1
A3	0	0	0	0	1
A4	0	0	0	0	1
A5	0	0	0	0	1

Another matrix 'IsRead ' is built which tells whether a particular user has read the particular article or not. This matrix is given in Table 5. The dimension of the matrix is n ∗ m, where n is the number of users and m is the number of articles. The 'User Profile ' is generated from 'Genre ' and 'IsRead ' matrix by taking their matrix multiplication. This generated user profile gives the information of inclination of user towards particular tags. Thus each row of user profile signifies specific user's interest in various tags. The user profile matrix is n ∗ k dimensional matrix, where n is number of users and k is number of distinct tags. The greater score of any tag against specific user signifies that the user has interest in that tag or article containing that tag as compared to article containing different tag. A sample user profile is shown in Table 6.

Table 5:- User Article 'Is-Read' Sample.

User-Article ID	A1	A2	A3
U1	0	1	1
U2	1	0	0

Table 6:- User Profile

User-ID	Automobile	Share	Wine	Finance
U1	9	5	6	10
U2	16	1	12	9
U3	12	0	0	7

For recommending articles to test user, a dot product of k-dimensional vector in user profile U_i and k-dimensional vector in genre matrix M_i is taken for test user and each article. This assigns a score to all article for particular user entirely on the basis of the tags of the article. The higher score is generated for articles which contains tags towards which test user is more inclined. The articles are sorted on the basis of score and the top articles are recommended.

Score: The score of every article m_i for given test user u_i is given by (1):

$$\text{Score}(U_i, M_i) = u_i.m_i \quad (1)$$

Where u_i and m_i are the k-dimensional vectors from user profile and article-genre matrix and k is the total number of distinct tags.

Item-Item Collaborative Filtering:-

Item-based collaborative filtering uses interaction between users to recommend articles. In the algorithm, the similarities between different items in the data-set is calculated by using one of the similarity measures, and then these similarity values are used to generate recommendation for users [6][7]. The similarity value between any two articles is measured by observing all the users who have read both the article. The similarity measure used in this paper is Pearson (correlation) based similarity. This measure is based upon how much the articles read by common user for pair of articles deviate from average times each article is read. The Pearson similarity score between article M_i and article M_j is given by (2).

$$sim(i,j) = \frac{\sum_{uEU}(M_{u,i}-M_i)(M_{u,j}-M_j)}{\sqrt{\sum_{uEU}(M_{u,i}-M_i)^2}\sqrt{\sum_{uEU}(M_{u,j}-M_j)^2}} \quad (2)$$

The similarity score between any two pair of articles is calculated by calculating Pearson correlation score between the respective columns of article in user-article 'IsRead ' matrix[8].

For a test user for whom we want article recommendation, similarity score of all the articles corresponding to every article read by user is added up and top-N article after sorting is recommended to the test user. This method is iterated for each user to generate the recommendation for each user. For user U1, similarity score of read articles A1, A2 and A3 given in Table 7 is added up column wise to get the score of each article. The following score is given to each article after adding up A1: 2.5, A2: 2, A3: 2, A4: -2.5, A5: -2.5. As the unread articles A004 and A005 get negative score imply that they are not similar and hence not recommended.

Table 7:- Similarity Scores

	A1	A2	A3	A4	A5
A1	1	0.5	1	-1	-1
A2	0.5	1	0.5	-0.5	-0.5
A3	1	0.5	1	-1	-1
A4	-1	-0.5	-1	1	1
A5	-1	-0.5	-1	1	1

User-User Collaborative Filtering:-
User-User collaborative approach applies the same idea as that of item-item collaborative filtering, but it calculates the similarity between users rather than calculating similarity score between articles [9].]. A Pearsons correlation matrix of users is built based on 'IsRead ' matrix i.e. on the basis of m-dimensional vector of two users in IsRead matrix, where m is total number of articles. The similarity is calculated between rows of User-Article 'Is read' matrix in Table 5. The similarity score obtained is shown in Table 8.
It is calculated using the same Pearson similarity equation (2).

Table 8:- User-User Correlation Matrix.

User	U1	U2	U3
U1	1	0.667	-1
U2	0.667	1	0.667
U3	-1	-0.667	1

On the basis of similarity score of users, top-N similar users of test user is generated. The articles read by all these similar user is used for building recommendation to test user. The articles read by similar user of test user for the purpose of recommendation to test user can be sorted in order by popularity using click counts of article.

Other Attributes Filtering:-
The recommendations obtained from the aforementioned methods are further filtered on the basis of following factors:-
1. Geographical location
2. Date and Time
3. User Attributes (Age, Gender etc.)

Observations and Results:-

The different recommendations generated using A,B,C,D are further assigned different weights and then the overall recommendation is generated considering all the possible factors. The weights to each article obtained through different steps are assigned according to "Design of Experiments" principle.

Table 9:- Weighted Recommendation Table

Recommendations	Weight
Step-A recommendations	0.42
Step-B recommendations	0.42
Step-C recommendations	0.08
Step-D recommendations	0.08

The final recommendation is made on the basis of the obtained weights for each recommendation obtained using the proposed algorithm.

The response rate obtained in case of Hybrid system is much better as compared to that obtained in case of collaborative and content-based filtering individually.

Response rate is given by (3):

$$Response\ Rate = \frac{No.of\ relevant\ recommendations}{Total\ no.of\ recommendations} \quad (3)$$

The number of relevant recommendations included the number of articles (from the set of recommended articles) that were read (clicked) by the user.

In order to test our hybrid model and make comparison, the experiment was performed for 100 users. A set of these 100 users was taken, and all the data related to the user was considered. The user attributes were collected as well as the articles read by user was found out. After this we made 5 recommendations each using content-based filtering, collaborative filtering and the using the hybrid model. Then we observed the clicks by user within next 24 hours span of time.

In order to obtain a comparison a design of experiments model was constructed and following steps were implemented:

1. Firstly the recommendations were generated by applying content-based filtering on the articles and then the response rate was calculated. This was done by observing the number of articles (from the set of recommended articles), that was read (clicked on) by the user.
2. Secondly the response rate was calculated by generating recommendations on the basis of collaborative filtering.
3. Finally the response rate was calculated for the recommendations generated on the basis of the Hybrid model and the result obtained in all the 3 cases was tabulated in Table 10.
4. For the 5 recommendations made in case of content-based filtering, the user clicked on 2 recommendations giving response rate of 40%.
5. For the 5 recommendations made in collaborative filtering, the user clicked on 2 recommendations again giving response rate of 40%.
6. In case of hybrid model, the user clicked on 4 recommendations out of 5 made, giving the highest response rate of 80%.
7. The similar procedure was followed for all the 100 users and average response rate was calculated for each category, the results are tabulated in Table 10.

Table 10:- Comparison Table.

Approach	Response Rate
Content-based	0.44
Collaborative	0.48
Hybrid Model	0.85

Conclusion:-

The model proposed in this paper is a hybrid one, that has made improvements in different sections for generating overall relevant recommendations, taking into account all the factors that can contribute to the same. In this paper, implicit feedback of user for an article is used, which observes action of users whether or not user has read that article. If user gives an explicit feedback such as rating an article, liking an article or sharing an article, it would generate better recommendation as compared to implicit feedback. Another way of tracking user behavior is by storing the time spent by user on every article that he read.

Also, if there exist term frequency of every tags, TF-IDF (Tag Frequency-Inverse Document Frequency) method can be used. In this method, weights of each article according to frequency a particular tag in that article and IDF (inverse document frequency) which tells how frequent that tag has occurred in all articles. The product of k-dimensional vector of user profile and k-dimensional weights of each article is calculated to generate similar article based on content .In the method of collaborative filtering, Pearson correlation matrix is used for calculating similarity between articles and between users. As the variable is dichotomous, other methods of similarity such as Jaccard-Needham, Yule, and Kulzinsky give a better result as compared to Pearson correlation coefficient [9].

The exclusion of the articles for recommendation that has already been read by user also varies according to the category of the article. There can be some articles that the user will want to read more than once for example if it is a lengthy article, or some historical one or else of deep interest or Education related and need to read more than once for grasping completely.

References:-

1. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms" in GroupLens Research Group/Army HPC Research Center Department of Computer Science and Engineering University of Minnesota, Minneapolis, MN 55455.
2. Jiahui Liu, Peter Dolan, Elin Rønby Pedersen, "Personalized News Recommendation Based on Click Behavior", 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA.
3. De Gemmis M., Lops P., Semeraro G., "Content Based Recommender System: State of the Arts and Trends, In Recommender systems handbook", Springer US pp.73 -85 (2011).
4. Burke R., "Hybrid recommender systems: Survey and experiments- User modeling and user-adapted interaction", vol.12(4), pp.331-370 (2002).
5. Pelanek R., "Recommender System: Content Based, Knowledge Based, Hybrid Recommendation" (2016).
6. D., Sowell B., Steinberg L. E., Tuladadh A. S., "Recommender Systems", Carleton College (2007).
7. Burke R., Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction, vol.12(4), pp.331-370 (2002)
8. Ricci F, Rokach L, Shapira B, "introduction to Recommender Systems handbook", Springer, US (2011).
9. A. P., Wang J., "Unifying User-based and Item-based Collaborative Filtering Approaches by SimilarityFusion", Delft University of Technology (2006)