Content Based Filtering Techniques in Recommendation System using user preferences

R.Manjula
Research Scholar, Anna University, Chennai, India

A. Chilambuchelvan
Professor, Department of CSE,
R.M.D Engineering College, Chennai, India

Abstract - Recommender systems use several of data mining techniques and algorithms to identify user preferences of items in a system out of available millions of choices. Instead of providing a static experience in which users search for and buy products, recommender systems help to increase interaction to provide a richer experience. Recommender systems can easily identify the recommendations autonomously for individual users based on past purchases and searches, and on other users' behaviour.This technique can predict a user’s preferred items by using the user’s past history data as well as other users’ past history data, and then recommends items to the user. This paper is focused on Recommender systems, and its major challenges for instance cold start problem, data sparsity, scalability and accuracy. *Content-based filtering* constructs a recommendation on the basis of a user's behaviour. As with Collaborative Filtering , the representations of customers’ precedence profile are models which are long-term, and also we can update precedence profile and this work become more available.

Keywords- Recommender systems, Collaborative Filtering, Content based Filtering

I.INTRODUCTION

The tremendous increase in e-commerce and online web services the matter of information search and selection has become increasingly serious and the users are confused for personal evaluation of these alternatives. Recommender systems are a helpful way for online users to deal with loading information and have turn out to be one of the most popular and powerful tools for e-commerce. The recommendation system are the supporting system which is used to help users to find information services, or products such as Books, Music, Movie, Digital Products, Web sites and TV Programs by analyzing the suggestions from other users [1]. Based on the past history the recommender system provide list of items by predicting which item are most suitable to user, preferences and constraints. Recommendation systems use different algorithm and methods to provide personalized recommendations. Recommendation system is also called information filtering system, or recommendation engine used to recommend informational items.

In day to day life, people suggest on recommendation from other people by spoken words, news reports from news media, reference letters, general survey etc. Recommender system assist and augment this natural social process to help people to sieve through available books, articles, web pages, electronic products, grocery items and so forth to find the most interesting and valuable information for users. The recommendation system the Collaborative filtering differs from content based filtering ideas. The idea of content based filtering is that users are interested in items that are similar to item the users previously liked. On the other hand the idea of collaborative filtering is that users like items that the user’s peers liked.

II. RECOMMENDER SYSTEM

Recommender systems generate personalized preferences using information filtering techniques and algorithms with a goal to support decision making of the users[2]. Recommender systems are mainly used in various domains like online booking, online shopping, audio and video recommendations and so on. The concept of recommendation is not new but in use from many years, the difference is due to more number of users asking for recommendations among thousands to millions of choices. It has become a tedious job to recommend someone appropriately without filtering the data for relevant choices. It depends on the factors like users rating given to collection of items based upon the user satisfaction level, their likes and dislikes, age group, gender, occupation, region or locality, community etc.

There are various problems and challenges consequently affect the performance of recommender systems. The cold start problem is one of the challenges and that arrives when either a new user enters into a system or a new product arrives in catalogue. Both situations lead to difficulty in predicting user preference in the absence of availability of sufficient user rating history[3]. The efficiency of recommender systems can be improved by proposed Hybrid recommender system which calculates recommendations for new user on the basis of his or her demographic attributes like age, gender, occupation similarity of existing users in the system. The proposed approach results in generation of more relevant and accurate recommendations as compared to traditional methods of finding recommendations.

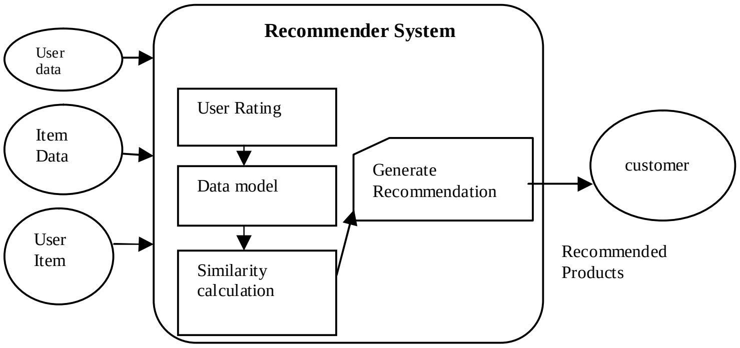

Figure.1 Recommender System

Figure.1 depicts the recommender system and it becomes the one of the most important methods for providing documents, merchandises, and cooperators to response user requirements in providing information, trade, and services that are for society, whether through mobile or on the web. Due to increasing quantity of data and information daily which causes overloading of information and data. So finding the customer's requirements may lead to important problem. The search engines were one of the innovations which helped people a lot for search and they were somewhat as a solution for this problem. But the information could not be personalized by these engines so the recommendation system was introduced[7]. These Recommendation systems utilize users past history from a group to assist for realizing users interestingness and demands in a society from a possibly onerous set of selections. The main aim of recommendation system is creating significant suggestions and recommendations of products or objects for users interestingness like book recommendation on Amazon that use recommendation systems to identify users' tendencies and subsequently, attract users more and more. There are a lot of different methods and algorithms which can assist recommendation systems to create recommendations that are personalized.

In this system the recommendation process first starts by gathering information about the items like author, title, cost etc and uses feature extraction and information indexing. This system uses content based filtering for processing information and data from various sources and try to extract useful features and element about the contents of the items and the constraint based filtering uses features of items to determine their relevance [9]. In this system the feature extraction and representation could be achieved automatically like news from papers but human editors which have to manually insert features from items like movies and songs. Recommender systems help to match users with items and for this various recommender systems have been designed according to availability of exploitable data, implicit and explicit user feedback, domain characteristics etc. Recommender Systems are classified according to approach/paradigm used for predicting preferences.

III. COLLABORATIVE FILTERING

In recommendation, collaborative filtering arrives to model the prior user behaviour and this model can be constructed from a single user's behaviour or also from the other user behaviour that have similar traits. When it takes other users' behaviour into account, collaborative filtering uses group knowledge to form a recommendation based on like users. The recommendations are made from automatic collaboration of multiple users and filtered on those who exhibit similar preferences or behaviours. By using the information from many users who subscribe to and read blogs, you can group those users based on their preferences. From this information, you identify the most popular blogs that are read by that group. Then for a particular user in the group you recommend the most popular blog that he or she neither reads nor subscribes to[4 - 6].

Collaborative filtering (CF) execute to recommend the objects that are suggested to customer based on how customers interests, and categorized in to objects. Collaborative filtering (CF) is used for introducing their filtering system that gives ability to customer for explanation their e-mails and documents[5]. Other customer can ask for documents that elucidated by specific people, but recognition of these people was left to customers. Collaborative filtering (CF) methods are used to identify the close neighbors of customer that is active. Collaborative filtering (CF) algorithms uses patterns which express customers' precedence and interaction for accordance them to customers share similar information and documents. After recognition a match that is possible, the suggestions and recommendations are generated by algorithm. Collaborative filtering (CF) algorithms utilize to anticipate values for vacant cell in matrix.

IV. CONTENT-BASED FILTERING

Content-based methods provide the recommendations by analyzing the description of the items that have been rated by the user and the description of items to be recommended. More number of algorithms has been proposed for analyzing the content of text documents and finding similarities in this content that can serve as the basis for making recommendations. Main goal of classification learners is to learn a function that predicts which class a document belongs to. Other algorithms would use a regression problem in which the goal is to learn a function that predicts a numeric value of the rating of the document. There are two important sub problems in designing a content-based filtering system [8]. The first is finding a representation of documents and next is to recommend for unseen documents.

All of the content-based approaches represent documents by the “important” words in the documents. For example, it represents documents in terms of the 100 words with the highest TF-IDF weights i.e., the words that occur more frequently in those documents than they do on average. The information retrieval can be categorised in two ways like document selection and document ranking methods. In document selection methods, the query is used for specifying constraints for selecting relevant documents. The document and a query both are represented as vectors in high dimensional space corresponding to all the keywords and uses an appropriate similarity measure is used to compute similarity between the query vector and the document vector. The similarity values can be used for ranking documents.

In information retrieval system the first step is to identify keywords for representing the documents, It avoids indexing useless words, a text retrieval system often associates stop list with a set of documents. The irrelevant words are called stop list (the , of, for, with, etc).The information retrieval system needs to identify groups of words where in a group are small syntactic variants of one another and collect only the common word stem per group. A group of different words may share the same word stem. For example apple, apples shares a common word stem. Starting with a set of d documents and a set of t term, we can model each document as a vector v in t dimensional space R^t and this method is called vector space model. The term frequency be the number of occurrence of term t in the document d, that is, freq(d,t).

The term frequency matrix (weight) TF(d.t) measures the associated of a term t with respect to the given document d. It is defined as 0 if the document does not contain the term and non zero otherwise. The relative term frequency is measured using the term frequency versus the total number of occurrence of all the terms in the document. The term frequency is computed

$$\mathrm{TF}(d.t) = \begin{cases} 0 & \text{if } \mathrm{freq}(d,t) = 0 \\ 1+\log(1+\log(\mathrm{freq}(d,t))) & \text{otherwise} \end{cases} \qquad (1)$$

There is other important measure, called inverse document frequency (IDF) in Eq.1 that represents the scaling factor, or the importance of a term t and it will be reduced if a term t occurs in many documents. For example the term information may be less important in many research papers. The Formula for IDF(t) is given in Eq.2.

$$IDF(t) = \frac{\log 1 + |d|}{|dt|} \qquad (2)$$

Where, d is the document collection, and dt is the set of documents containing term t. In complete vector space model, TF and IDF are combined together, which forms the TF-IDF measure is given in Eq.3:

$$\text{TF-IDF}(d,t) = \mathrm{TF}(d,t) * \mathrm{IDF}(t) \qquad (3)$$

Let us examine how to compute similarity among a set of documents based on the term frequency and inverse document frequency. For Example Consider the Table 1.

Table -1 Frequency of terms per document

document /item	t1	t2	t3	t4	t5	t6	t7
d1	0	4	10	8	0	5	0
d2	5	19	7	16	0	0	32
d3	15	0	0	4	9	0	17
d4	22	3	12	0	5	15	0
d5	0	7	0	9	2	4	12

Table 1shows a term frequency matrix where each row represents document, each column represents a term, and each entry registers freq(d_i ,t_j),the number of occurrence of term tj in document di. For example t_6 in d_4 we can calculate

TF(d_{4},t_{6})= 1+log(1+log(15))=1.3377

IDF(t_6) = $\log \frac{1+5}{3}$ = 0.301

Therefore TF –IDF(d_{4},t_{6}) = 1.3377x0.301 = 0.403

To determine the similarity between two documents cosine similarity is used. One of the most obvious advantages of content-based filtering algorithms is these algorithms don not need to domain of knowledge. It is adequate to gather feedback from customers about their precedence. Next advantage of content-based filtering algorithms that we can consider to it is, these algorithms are better than Collaborative Filtering (CF) at finding locally similar objects[10]. Because the explicit focus of content-based filtering algorithms is on similarity of text. However, this item can be a defect in domains where analysis of content in large number is impractical, impossible or difficult, like music and movies. The tendency of algorithms of content-based filtering is get stuck in a "well of similarity", where they suggest objects only from a restrict theme scope. Then the recommendations that are serendipitous can be very difficult to achieve.

A. Similarity Measure

Memory-based CF algorithms check for the complete or a sample of the user-item data to create a prediction. Every user is a part of a group of people with similar interests. By identifying the supposed neighbors of an active user a prediction of tastes on new items for him or her are going to be generated. The neighborhood-based collaborative Filtering rule, a current memory-based CF rule, uses the following steps:

1. Calculate the similarity or weight, w_{ij}, which reflects distance, correlation, or weight, between two users or 2 items, i and j.
2. Generate a prediction for the active user by taking the weighted average of all the ratings of the user or item on a definite item or user, or employing an easy weighted average.

When the task is to build a top-N recommendation, we want to search out k most similar users or items (nearest neighbors) once computing the similarities, so aggregate the neighbors to urge the top-N most frequent items as the recommendation. Similarity computation between items or users could be an essential step in memory-based collaborative filtering algorithms. For item-based CF algorithms, the essential plan of the similarity computation between item i and item j is initial to figure on the users who have rated each of those items so to apply a similarity computation to work out the similarity, w_{ij}, between the two co-rated items of the users [4]. For a user-based CF algorithmic rule, we tend to initial calculate the similarity, w_{uv}, between the users u and v who have each rated a similar items. There are many various ways to work out similarity or weight between users or items.

B. Correlation-Based Similarity

In this case, similarity between two users‟ u and v, or between two items and , is computed by computing the Pearson correlation or different correlation-based similarities. Pearson correlation measures the extent to that two variables linearly relate with one another [4]. For the user based algorithmic rule, the Pearson correlation between user u and v is given in Eq.4.

$$w_{uv}=\frac{\sum_{i\in I}(r_{ui}-\bar{r}_u)(r_{vi}-\bar{r}_v)}{\sqrt{\sum_{i\in I}(r_{ui}-\bar{r}_u)^2}\sqrt{\sum_{i\in I}(r_{vi}-\bar{r}_v)^2}} \qquad (4)$$

Where i ∈I summations are over the items that both the users u and v have rated and is the average rating of the correlated items of the u^{th} user. In item-based algorithm, the set of users denoted by u∈U who rated both items i and j, then the Pearson Correlation is given in Eq.5.

$$w_{ij}=\frac{\sum_{u\in U}(r_{ui}-\bar{r}_i)(r_{uj}-\bar{r}_j)}{\sqrt{\sum_{u\in U}(r_{ui}-\bar{r}_i)^2}\sqrt{\sum_{u\in U}(r_{uj}-\bar{r}_j)^2}} \qquad (5)$$

Where r_{ui} is the rating of user u on item i, is the average rating of the i^{th} item by those users.

C. Vector Cosine-Based Similarity

Vector cosine similarity between items i and j is given by Where "•" denotes the dot-product of the two vectors. an *n* × *n* similarity matrix is computed to get the desired similarity computation, for *n* items. For example, if the

vector *A* = *{x*1, y1} , vector *B* = *{x*2, *y2}*, the vector cosine similarity between *A* and *B* is given in Eq.6.

$$W_{A,B} = \cos(\vec{A}\cdot\vec{B}) = \frac{x1x2 + y1y2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}} \qquad (6)$$

So, the conclusion is, a choice is to be made among all similarity measures. The point to remember at this step is:

i) if the data is subject to grade-inflation i.e.(different users may be using different scales) then use pearson correlation coefficient.

ii) if data is dense i.e. (if almost all attributes have non-zero values) and the magnitude of the attribute value is important, use distance measures such as Euclidean or Manhattan.

iii) if the data is sparse consider using cosine-similarity.

Collaborative filtering uses the model of prior user behaviour for recommendation. The model can be constructed solely from a single user's behaviour or also from the behaviour of other users who have similar behaviour. When it takes other users' traits into account, collaborative filtering uses group knowledge to form a recommendation based on like users. An automatic collaboration of multiple users and filtered on those who exhibit similar preferences or behaviours are basis for the recommendation. Collaborative filtering makes the recommendations by finding correlations among users of a recommendation system. It presents a uniform approach for finding items of potential interest and predicting the rating that the current users would give to an item. To see how such a prediction could be made, consider the example in Table 2. This gives the ratings of 5 items by 5 users. A "+" indicates that the user liked the description of the restaurant and indicates that the user did not like the item.

Table - 2 Rating of User Preferences

Items	**Kiran**	**Arun**	**Chander**	**Mala**	**Janu**
A	-	+	+	+	+
B	+	+	+	+	+
C	+	-	+	-	+
D	-	+	-	+	-
E	+	-	+	-	?

To predict the rating that Janu would give to D, we can look for users that have a similar pattern of ratings with Janu. In this case, Kiran and Janu have identical tastes and one might want to predict that Janu would like D because Kiran does. A more general approach would be to find the degree of correlation between Janu and other users. A weighted average of the recommendations of several users can be found instead of relying on just the most similar user. The weight is given to a user's rating would be found by degree of correlation between the two users. In the most general case, the rating could also be a continuous number rather than just +1 . The Pearson *r* is a measure of correlation that can be used in these circumstances. Let $R_{i,j}$ be the rating of user *i* on document *j*. Then the correlation between user *x* and user *y* is given by:

$$r(x,y) = \frac{\sum_{d\in documents}(R_{x,d} - \overline{R_x})(R_{y,d} - \overline{R_y})}{\sqrt{\sum_{d\in documents}(R_{x,d} - \overline{R_x})^2 \sum_{d\in documents}(R_{y,d} - \overline{R_y})^2}} \qquad (7)$$

Where $\overline{\mathrm{R}_x}$ is the mean value of ratings by user .

In the above example, the correlation between Janu and Kiran is 1.0, between Janu and Arun is – 0.577, between Janu and Chander is 0.577, and between Janu and Mala is –0.577. Therefore, the weight average of the product of each user's rating for D and the correlation between Janu and that user is 0.682. A collaborative algorithm would predict that Janu would like D based on the other users recommendations. Note that in part this recommendation makes use of the fact that Janu and Mala have nearly opposite tastes and that Mala doesn't like D. Thereafter randomly deleted half of each user's ratings and then, for each user, the three items with whose rating had been deleted with the highest recommended rating were found using collaborative filtering. We compared the predicted rating of these three items with the actual rating. We repeated this process of randomly deleting ratings 20 times for each user. On average, 67.9% of the item in the top three items recommended via this collaborative process was actually liked by the user.

Collaborative filtering is most commonly used method to find correlations between user ratings of objects, but it may also be used to find collaborations among the rated objects. For example, there is a perfect correlation between the ratings of D and C in Table 1. As a consequence, one might predict that Janu would like D given that Janu likes C. Similarly, this may be generalized by finding the correlations between item and making predictions based upon the weighted average of ratings for other item. Once again, taking the weighted average of all item in Table 1 would yield the result that Janu would like D. We repeated the experiment described above using correlations among item as the basis of predictions. Under these conditions, 59.8% of the item in the top three item were actually liked by the user. Although basing recommendations on correlations among item does not yield as high a precision as correlations among users in this problem, and that may be combined with other sources of information to provide a better overall recommendation.

These relationships can be viewed on their similarities and differences. The similarities are based on the algorithm used and group of users who have similar interests. If there is a differences then that can be used for recommendation applied through a filter of popularity. It is the process of evaluating or filtering items using the opinions of other users. Collaborative filtering techniques collect user's profiles and the connection among the data are examined according to similarity function. The likely categories of the data in the profiles include user behaviour patterns, user preferences, or item properties. Collaborative filtering technique collects large information about user behaviour, history and then recommends the items based on his similarity with other users communally.

V.CONCLUSION

In World Wide Web, the overload of information leads to the necessity of recommender systems to generate efficient solutions has evolved. Nowadays finding the right recommender for evaluating the reliability of recommender systems is an essential feature. Retrieval of information from huge volumes of data in diversified areas results in a tedious process. Hence, filtering in recommender systems have evolved to make the recommendation process trivial. The aggregate function used for calculating and the quality of recommendation of items depends upon rating distribution and type of aggregate analysis. Moreover, another set of demographic attributes can be exploited for finding clusters and hence recommendation accuracy can be improved in future. For this demographic attribute collection in user profiles can be increased for getting better recommendations.

REFERENCES

[1] A. Bellogí I. Cantador, F. Dí et al., "An empirical comparison of social, collaborative filtering, and hybrid recommenders," ACM Trans. On Intelligent Systems and Technology, vol. 4, no. 1, pp. 1-37, January 2013

[2] A. Yamashita, H. Kawamura, and K. Suzuki, "Adaptive Fusion Method for User-based and Item-based Collaborative Filtering," Advances in Complex Systems, vol. 14, no. 2, pp. 133-149, May 2011.

[3] Hongwu Ye, "A Personalized Collaborative Filtering Recommendation Using Association Rules Mining and Self-Organizing Map ", International Journal of Computer Science & Information Technology (IJCSIT) Vol 6, No 2, April 2014 DOI:10.5121/ijcsit.2014.6204 51

[4] J. Mai, Y. Fan, and Y. Shen, "A Neural Networks-Based Clustering Collaborative Filtering Algorithm in E-Commerce Recommendation System," in Proc. 2009 Int'l Conf. on Web Information Systems and Mining, pp. 616-619, June 2009.

[5] M. C. Pham, Y. Cao, R. Klamma, et al., "A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis," Journal of Universal Computer Science, vol. 17, no. 4, pp. 583-604, April 2011.

[6] N. Mittal, R. Nayak, M. C. Govil, et al., "Recommender System Framework using Clustering and Collaborative Filtering," in Proc. 3rd Int'l Conf. on Emerging Trends in Engineering and Technology, pp. 555-558, November 2010.

[7] Rong Hu, Wanchun Dou,Jianxun Liu, "ClubCF: A Clustering-based Collaborative Filtering Approach for Big Data Application , " IEEE Trans. On Emerging topics in Computing, vol. 20, no. 11, pp. 1519- 534,October2014.

[8] R.D.Simon, X. Tengke, and W. Shengrui, "Combining collaborative filtering and clustering for implicit recommender system," in Proc. 2013 IEEE 27th Int'l Conf. on. Advanced Information Networking and Applications, pp. 748-755, March 2013.

[9] X. Li, and T. Murata. "Using Multidimensional Clustering Based Collaborative Filtering Approach Improving Recommendation Diversity," in Proc. 2012 IEEE/WIC/ACM Int'l Joint Conf. on Web Intelligence and Intelligent Agent Technology, pp. 169-174, December 2012.

[10] Z.Zheng, H. Ma, M. R. Lyu, et al., "QoS-aware Web service recommendation by collaborative filtering," IEEE Trans. on Services Computing, vol. 4, no. 2, pp. 140-152, February 2011.