Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 9; April-June, 2015 pp. 1-5

http://www.krishisanskriti.org/acsit.html

# An Elaborative Survey on Recommendation Techniques

**Sangeeta[1] and Neelam Duhan[2]**

*[1,2]Computer Engg. Department, YMCA University of Science & Technology, Faridabad*
*E-mail: [1]sangeeta2yadav@gmail.com, [2]neelam.duhan@gmail.com*

**Abstract—***Recommender systems are type of information filtering systems which consider user preferences for the purpose of recommending or suggesting items (it may be documents, products, news articles or any service). Researchers use various techniques for implementing recommender systems in different domains. This paper provides an overview of recommender systems and detailed survey of current methodologies used for recommendation. Comparison of various techniques i.e. collaborative filtering, content-based filtering and hybrid recommendation is also performed in terms of their pros and cons.*

**Keywords***: Collaborative filtering, Item-based technique, User-based technique, Content-Based technique, Hybrid technique.*

## 1. INTRODUCTION

World Wide Web (WWW) is an information system which is tremendously increasing. Advertising on popular web pages can be lucrative, and e-commerce or the sale of products and services via the Web continues to grow. In such a large information system it is difficult to find what an individual wants. To solve this issue, searching (search engine like google, yahoo etc) came into account. But, if a user doesn't know what information he wants to get, then searching is not a solution. This problem is solved by recommender systems. Recommendation uses data mining techniques to filter information wherein "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques"[1].

Now what really needed is new technologies that can make recommender system most successful. In the past, various techniques have been used by developers for recommendation [2], out of which collaborative filtering is the most successful method. Personalized recommendation is also popular in these days. One of the algorithms to implement personalized recommender system is hybrid collaborative filtering based on user preferences and item features [14].This article describes possible techniques of recommendation and analyzes their hybridization.

In the rest of the paper, Section 2 describes history of recommendation techniques and Section 3 describes collaborative filtering technique in detail. In the end of Section 2, comparison of two collaborative filtering techniques i.e. Item-based and User-based is provided. Next section 4 explains Content-Based filtering technique. Section 5 describes Hybrid technique (Combination of Collaborative and Content-based technique) and then comparison between Collaborative filtering and Content-based filtering is provided. Finally, the whole survey is concluded in the last section.

## 2. RECOMMENDATION TECHNIQUES

Recommendation approaches can be broadly classified in three categories as follows:

***Collaborative filtering:*** It is most familiar and widely used method for recommendation. Collaborative techniques work well for complex objects such as movies, music, online products etc. These techniques make use of user-item rating matrix to compute predictions. These systems aggregate rating on items by a number of users and find the similarity based on these ratings.

For example, *Javawock* is a recommender system that helps to provide useful Java components (library class files) to a developer based on collaborative filtering. If a developer gives an unfinished Java program to the system, it tries to find Java library class files used in the provided program. Then, the system recommends to the developer Java library class files that were used in the similar programs but were not used in the developer's program [3].

***Content-based filtering:*** Content based algorithms make use of item or user profiles and features for recommendation. These systems find the related items based on item features. Content based algorithms analyze items description or user profile to find the items of individual's interest.

For example, *Content REcommendation System* is based on content-based filtering technique. "This system collects and mines the private data of user at the client side. Then it discovers, stores and updates private Dynamic User Profile (DUP) at the client side. The system fetches preferred message

from the content server according to DUP. An important usage of this technology is a personalized advertising system in the RSS (Rich Site Summary, or RDF Site Summary) reader application" [4].

***Hybrid recommendation methods:*** Hybrid recommenders use a combination of collaborative and content-based filtering methods. This removes shortcomings of both methods and give best results. Collaborative recommender systems suffer from problems like data sparsity, cold start and scalability. Content-based methods may have problems like over-specialization and new user. To overcome most of these limitations, hybrid methods are used. For example, "*Fab* is a recommendation system designed to help users sift through the enormous amount of information available in the WWW. Operational since Dec. 1994, this system combines the content-based and collaborative methods of recommendation in a way that exploits the advantages of the two approaches while avoiding their shortcomings. Fab's hybrid structure allows for automatic recognition of emergent issues relevant to various groups of users. It also enables two scaling problems pertaining to the rising number of users and documents [5].

## 3. COLLABORATIVE FILTERING

Collaborative technique is most general and famous for recommender systems because it is very simple and easy to implement; and gives best results. This technique uses ratings of products provided by the user for prediction. Collaborative Filtering works under three basic steps as follows:

- Similarity Calculation
- Prediction
- Recommendation

**Similarity Calculation:** Similarity between two users/items can be determined using any mathematical formula which gives best results. One of the methods to compute of similarity is cosine method [6] given in (1).

$$\text{Sim(i,j)} = \cos(i,j) = \frac{i \times j}{\|i\| \|j\|} \quad (1)$$

where i and j are items represented as vectors and Sim(i,j) gives similarity between i and j.

**Prediction:** Prediction can be computed with the help of similarity calculated in (1). Predicted values are list of recommended items to a particular user. One of the prevalent is prediction formula [6] given in (2).

$$\text{P(u,i)} = \overline{R}_i + \frac{\sum_{j \in NI} sim(i,j) * (R_{u,j} - \overline{R_j})}{\sum_{j \in NI} |sim(i,j)|} \quad (2)$$

where sim(i,j) represents similarity between i and j calculated from (1). $\overline{R}_i$ and $\overline{R}_j$ represents average rating of item i and j respectively. $R_{u,j}$ represents rating of any item j given by any user u. P(u,i) gives prediction of item i for a target user u.

**Recommendation:** Recommendation is a process of selecting top N recommendations from resulting prediction. These top N items are final recommendations.

Basically Collaborative Filtering can be implemented using two approaches as follows:

1. Item based method
2. User based method

### 2.1 Item based method

Item based method is a collaborative filtering technique which analyzes the set of items a target user has rated and calculates how similar they are. After computing similarity of all rated items by target user, k most similar items are selected for prediction. Similarity and prediction computation is given below in detail.

#### 3.1.1 Item Similarity

This is most crucial step in item based collaborative technique. Similarity computation largely effects quality of recommendation. Here are three methods for item similarity computation.

- Cosine based similarity
- Co-relation based similarity
- Adjusted cosine based similarity

In this paper a movie database is considered, wherein movies are regarded as items. This similarity computation is illustrated below using this database.

**a) Cosine based similarity:**

In this method, two movies are thought of as different vectors. The similarity between these two vectors can be determined by calculating cosine angle between these vectors [7]. Let us assume two movies i and j as vectors then their similarity is given by (1).

**b) Co-relation based similarity:**

In this case, similarity between two movies i and j is calculated by using Pearson Correlation [6]. Firstly isolate the co-rated cases and then compute the Pearson correlation based similarity as in (3).

$$\text{Sim(i,j)} = \frac{\sum_{a \in U_{ij}} (R_{a,i} - \overline{R_i})(R_{a,i} - \overline{R_j})}{\sqrt{\sum_{a \in U_{ij}} (R_{a,i} - \overline{R_i})^2} \sqrt{\sum_{a \in U_{ij}} (R_{a,i} - \overline{R_j})^2}} \quad (3)$$

where $R_{a,i}$ denotes the rating of movie i by user a. $\overline{R}_i$ and $\overline{R}_j$ are average rating of movies i and j.

**c) Adjusted cosine based similarity:**

Adjusted cosine based similarity provides better quality similarity because it considers user's average rating for each co-rated pair [6]. This will overcome the difference between

rating scales of individual users. Mathematically, similarity is given as in (4).

$$\text{Sim(i,j)}= \frac{\sum_{a \in U_{ij}} (R_{a,i} - \overline{R_a})(R_{a,j} - \overline{R_a})}{\sqrt{\sum_{a \in U_i} (R_{a,i} - \overline{R_a})^2} \sqrt{\sum_{a \in U_i} (R_{a,j} - \overline{R_a})^2}} \quad (4)$$

where $R_{a,i}$ and $R_{a,j}$ denotes the rating of movie i and j by user a. $\overline{R_a}$ is average rating given by user a.

### 3.1.2 Item Prediction

After computing similarity between two items, prediction can be calculated to provide top N recommendations. Prediction can be computed using weighted sum method given in (5).

$$\text{P(u,i)}= \frac{\sum_{j \in NI} sim(i,j) * R_{u,j}}{\sum_{j \in NI} |sim(i,j)|} \quad (5)$$

where sim(i,j) represents similarity between i and j calculated from (1). $R_{u,j}$ represents rating of item j given by user u. P(u,i) gives prediction of item i for target user u.

## 2.2 User based method

User based collaborative filtering is similar to the nearest neighbor method. This technique finds the nearest neighbor of a target user. Alternatively, it finds users with similar taste of target user. In this method, first step is to find nearest neighbor for which we need to obtain the users history profile. By analyzing the history a rating matrix can be prepared in which each entry represents the rating of the user given to an item [8]. Each row in a matrix represents individual user and column represents an item, and the number at the intersection of a row and a column represents the user's rating value. If a user has not yet rated the item, intersection of that row and column is empty. The second step is to compute the similarity between target users and find their nearest neighbors. The Pearson correlation coefficient method is the most widely used for similarity computation.

Generally, user-based methods utilize entire database to find recommendations. This approach is very popular in past but now there are other alternatives also which provide better recommendation. Main challenge of the approach is that user item matrix is very sparse and recommendations based on the matrix are of poor quality.

Assume that, according to user's psychological, emotional and trend change, user's rating on the similar item is unstable. This unstable evaluation affects the selection of nearest neighbor and most similar item for recommendation. To solve this issue, researchers give Stability Degree definition based on user s[9]. Stability Degree (SD) use target user's neighbors as the reference object. SD=0 means that user b is completely stable relative to user a. SD=l means user b is completely unstable relative to user a.

## 2.3 Hybrid method

User-based and Item based methods have their own advantages and disadvantages. To minimize limitations of both, hybrid techniques are presented. In literature, there are various ways to combine the User-based and Item based techniques to make hybrid algorithm. One way can be the *dividing algorithm* which works in two steps.

First step implements User-Based technique to calculate the user similarity of target user u by using the characteristics of the user rating data. Second step calculates the item similarity through the modified Pearson-r correlation method to search the nearest neighbors of the item. Then, two algorithms are fused with the control factor α to predict the items as in (6) [14].

$$P_{u,i} = \alpha * P_{user}(u,i) + (1-\alpha) P_{item}(u,i) \quad (6)$$

where $P_{u,i}$ is predictions of user u and item i. $P_{user}(u,i)$ is result calculated by User-based method and $P_{item}(u,i)$ prediction calculated by item-based method.

Comparison between Item-based and User-based method is given as follows in Table 1.

**Table 1: Comparison between Item based and User based Collaborative Filtering**

| | Item based | User based |
|---|---|---|
| Strengths | • No user cold start<br>• Improves scalability<br>• Accuracy is good | • Easy implementation<br>• User similarity is dynamic |
| Shortcomings | • Item cold start<br>• Serendipity-Unusual interest<br>• Item similarity is static | • Cold start<br>• Sparsity<br>• Scalability<br>• Accuracy is less |

From the table, it can be observed that item based techniques offer quality whereas user based offer ease. Item based techniques are better than user based techniques in terms of accuracy which is very important factor. On the other hand, both techniques have their disadvantages so hybrid approach is also used for recommendation.

# 4. CONTENT-BASED FILTERING

Content based recommender systems analyze item feature to find the predictions for an individual user. These systems also use user profiles and histories to find interest of the target user. A profile may contain various type of information. For example, user's age, gender, marital status, profession etc. One may be the details about user interest on different items and second is history of user's interaction with any recommender system. Items viewed by the user in past are useful for prediction. Content based filtering provides more personalized recommendation. Content-based methods are usually implemented using following methods.

- Heuristic based methods (Classification Techniques)
- Model-based methods (Probabilistic)

### 2.4 Classification Techniques

To implement content based recommenders, classification is most widely used technique. Traditional algorithms like item based methods give good performances for enough data sets but scalability is a problem there. To minimize the scalability problem, classification or clustering techniques are used. Similar items are grouped in a cluster and whole users are divided into n clusters. Whenever a new user comes, algorithm matches the similarity of the user to all the clusters and adds the new user to a most similar cluster [15]. There are several methods for classification like tf-idf, decision tree method, nearest-neighbor method and linear classifiers. Classification learning algorithms make use of training data set to make a model and use that model for recommendation. Training data can be collected by tracking user's activities on recommendation system or by taking a feedback from target user. TF-IDF is popularly used algorithm in Classification techniques. In this algorithm, TF represents the term frequency in document d and IDF represents inverse document frequency or appearance frequency of word.TF can be calculated as given in (7).

$$TF_{i,j} = \frac{f_{i,j}}{max_z f_{z,j}} \quad (7)$$

where $TF_{i,j}$ represent term frequency of keyword i in document j. $f_{i,j}$ is the number of times keyword i appears in document j. $max_z f_{z,j}$ calculates maximum of all keywords z. But, a large number of occurrences in a document do not mean that it is also important to the document. To overcome this effect IDF is calculated as in (8).

$$IDF_i = \log \frac{N}{n_i} \quad (8)$$

where N represents total number of documents and $n_i$ represents number of documents in which keyword i occurred. Weight of keyword i in document j is given in (9).

$$W_{i,j} = TF_{i,j} \times IDF_i \quad (9)$$

The content of document can be given as in (10).

$$\text{Content}(D_j) = (W_{1j}, \ldots \ldots W_{kj}) \quad (10)$$

Finally content of document is given by collection of weighted keywords.

Decision tree method builds the decision tree by partitioning training data into subgroups until those subgroups contain instances of single class. Decision tree can give best results for structured data [10] and it gives poor performance for unstructured data.

### 2.5 Probabilistic Techniques

Naïve Bayes is a probabilistic approach and forms a model based on previously observed data. Naïve Bayesian classifier implements content based recommender systems. This algorithm is most popular and used in text classification. Naïve Bayesian classifier is used in many recommendation systems [12]. The probabilistic model computes the a posteriori probability,

P(c|d), of document d belonging to class c. This result is based on the a priori probability, P(c), the probability of scrutinizing the document d in class c, P(d|c), the probability of observing the document d given c, and P(d), the probability of observing the instance d. Using these probabilities, the Baye's theorem is applied to compute P(c|d) as in (11).

$$P(c/d) = \frac{P(c)P(d/c)}{P(d)} \quad (11)$$

## 5. HYBRID RECOMMENDATION METHOD

Recommender systems find the items of user requirement and interest in a very large database worldwide. So, this requirement of recommender system is increasing day by day with an increase in number of internet users and its database. To implement recommender systems effectively and efficiently, we need new algorithms or techniques. Above mention techniques i.e. collaborative filtering and content-based method have their own advantages and disadvantages. To minimize their shortcomings and to take advantages of both methods, hybridization is done. There can be several methods to combine both the techniques.

One of them is to collect data from user profiles and apply clustering on the gathered information to make users group. After this content-based step, use collaborative filtering method to calculate similarity between two clusters and within a cluster [11]. This similarity is used to compute prediction.

One practical example that implemented hybrid approach is FAB[5], a recommender system that allows for automatic recognition of emergent issues in various users group. FAB mainly solves two problems: One is scaling problem (increasing number of users and increasing number of items) and the second one is to automatically identify emerging communities of interest in user population. Comparison between collaborative filtering and content-based filtering is as follows in Table 2.

From the table, it can be observed that collaborative filtering is easy and popularly used for recommendation where as content based filtering gives more personalized recommendations. Collaborative filtering needs more computation for recommendations and content based methods need to access a user's private profile which creates privacy and security issues. So, to overcome limitations of both the techniques, hybrid methods are more commonly used.

**Table 2: Comparison between Collaborative Filtering and Content-Based Filtering**

| | Collaborative filtering | Content-based filtering |
|---|---|---|
| Strengths | • It does not need domain knowledge.<br>• Quality improves with increasing ratings.<br>• It is easy to implement. | • It provides more personal recommendation.<br>• It does not need other users data.<br>• It can recommend to people of unique taste.<br>• It is not popularity biased.<br>• It can give Explanation for each recommendation. |
| Shortcomings | • New item can not be recommended with this technique.<br>• It is less scalable.<br>• Data matrix is very sparse which give poor results.<br>• User's have multi-criteria ratings which effect accuracy.<br>• Shilling attack: a user can give high rating to its own product and negative ratings to its competitor.<br>• Popularity biased-it recommends popular items. | • It is less scalable.<br>• Recommendation to a new user is not possible with this technique.<br>• Security & Privacy issues.<br>• It is over specialized to the user's interest. It does not recommend out of interest which a user may like. |

## 6. CONCLUSION

This paper represented general approaches of recommendations. Every approach has its own pros and cons. Collaborative Filtering is most widely used and provide accurate results but it is not scalable and suffers from cold start problem. Content based Filtering is another popular method which gives best results and these algorithms are scalable also but it suffers from over-specialization (means it cannot predict outside of user's profile). Security and privacy issue of user's profile is also there which can not be ignored. So, to minimize these challenges hybrid techniques have been proposed in literature by various researchers.

In conclusion, recommender systems are very popular in today's e-business era and there is need to improve quality of recommendation and performance by introducing new technologies.

## REFERENCES


[1] Larose, D. Discovering knowledge in Data. An Introduction to Data Mining. Wiley Interscience, 2008.

[2] Dhoha Almazro, Ghadeer Shahatah, William Nzoukou, Mona Kherees, Romy Martinez and William Nzoukou. Survey paper on recommender systems. *arXiv:1006.5278v4 [cs.IR] 24 Dec 2010.*

[3] Masateru Tsunoda, Takeshi Kakimoto and Naoki Ohsugi.Javawock: A Java Class Recommender System Based on Collaborative Filtering.

[4] Ting chen, Wei-Li Han, Hai-Dong Wang, Yi-Xun Zhou, Binxu and Bin-Yu Zang.

[5] Content Recommendation System Based On Private Dynamic User Profile. *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007*

[6] Marko Balabanovic and Yoav Shoham. Fab: content-based, collaborative recommendation.(Special Section: Recommender Systems). *Communications of the ACM, March 1997 v40 n3 p66(7).*

[7] Tieli Sun Lijun Wang, *Qinghe Guo.* A Collaborative Filtering Recommendation Algorithm Based on Item Similarity of User Preference. *Second International Workshop on Knowledge Discovery and Data Mining*

[8] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-Based Collaborative Filtering Recommendation Algorithms. *WWW10, May 1-5, 2001, Hong Kong.ACM 1-58113-348-0/01/0005.*

[9] Zhi-Dan Zhao, Ming-Sheng Shang. User-based Collaborative-Filtering Recommendation Algorithms on Hadoop. *2010 Third International Conference on Knowledge Discovery and Data Mining.*

[10] Xiangwei Mu, Yan Chen and Taoying Li. User-Based Collaborative Filtering Based on Improved Similarity Algorithm. 978-*1-4244-5540-9/10/$26.00 ©2010 IEEE.*

[11] Michael J. Pazzani1and Daniel Billsus.Content-based Recommendation Systems. *The Adaptive Web Lecture Notes in Computer Science Volume 4321, 2007, pp 325-341.*

[12] Byeong Man Kim & Qing Li & Chang Seok Park & Gwan Kim & Ju Yeon Kim. A new approach for combining content-based and collaborative filters. *J Intell Inf Syst (2006) 27: 79–91.*

[13] Pasquale Lops, Marco de Gemmis and Giovanni Semeraro. Content-based Recommender Systems: State of the Art and Trends. *F. Ricci et al. (eds.), Recommender Systems Handbook, DOI 10.1007/978-0-387-85820-3_3, © Springer Science+Business Media, LLC 2011.*

[14] Yan-ni Chen and Min Yu . A Hybrid Collaborative Filtering Algorithm Based on User-Item. *2010 International Conference on Computational and Information Sciences.*

[15] Liang Hu, Guohang Song, Zhenzhen Xie, and Kuo Zhao. Personalized Recommendation Algorithm Based on Preference Features. *TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007-0214ll08/11llpp293-299 Volume 19, Number 3, June 2014.*

[16] Junhao WEN and Wei ZHOU. An Improved Item-based Collaborative Filtering Algorithm Based on Clustering Method. *Journal of Computational Information Systems 8: 2 (2012) 571-578.*