Survey on Data Mining Techniques for Recommendation Systems

[1]Susan Thomas, [2]Jayalekshmi S.
[1]M.Tech Student, [2]Associate Professor
[1]Department of Computer Science and Engineering,
[1]LBSITW, Poojappura, Trivandrum, Kerala, India

***Abstract*— Recommender or recommendation systems are nowadays widely used. Recommendation system is a system which provides recommendations to users according to their tastes. Also it is a system that can guide the user in a personalized way. Recommender systems can be classified into: content based collaborative filtering and hybrid approach. There are number of data mining algorithms used in recommendation system. But each of it has its advantages and disadvantages. In this paper a comparison between content based and collaborative filtering techniques and the advantages and disadvantages of the algorithms used in those techniques are discussed.**

***Index Terms*— Recommendation system, Content based approach, Collaborative filtering, Model based approach, Memory based approach**

I. INTRODUCTION

Data mining is one of the most attractive interdisciplinary Data mining is one of the important research area in computer science nowadays. Data mining is the process of extracting useful information from large amount of data. It is of two types, directed data mining and undirected data mining. In directed data mining, a model is build based on the available data and then it will describe rest of the data. Whereas in undirected data mining some relationship is established among the variables.

Recommendation systems are systems which provide recommendations to users according to their tastes and it guides the user in a personalized way. Also this system will predict the 'rating' or 'preference' that the user would give to an item. They are applied in variety of applications such as movies, books, music, products, restaurants etc.

Recommendation systems must have sufficient amount of data in the systems for providing recommendations. Data collected can be implicit and explicit data. The main goal of collecting data is to know the preferences of the users, so that accurate recommendations will be obtained. When a user rates an item via some interface then it is explicit feedback. When a user buys an item, it means the user likes the item and when the user returns the item then it is implicit feedback.

Recommendation system can be classified into content based approach, collaborative filtering approach and hybrid approach. Content based recommendation systems will recommend items based on the description of the items and profile of the user. Collaborative filtering recommendation system will recommend items based on the similarity between the users who have rated the same item before. Hybrid is a combination of content based and collaborative filtering approaches [5].

Recommendation systems are used in many applications such as news recommendations, product recommendations, movie recommendations etc.

The advantages of recommendation systems are:

- Recommend items according to the users preference
- Removal of unnecessary information
- System is always up-to-date

The disadvantages of recommendation systems are:

- Need lots of data for giving accurate recommendations
- Changing the preferences of the users
- Performance and scalability problems in dealing large amount of data
- Privacy is a problem

Here in this paper, section 2 gives a brief history of recommendation systems, section 3 gives an overview of data mining algorithms used in recommendation systems, section 4 gives comparison between different techniques in content based and collaborative filtering approaches and section 5 gives the conclusion.

II. BACKGROUND

In early days recommendations was through word of mouth. As years passed by the amount of data grew, this results in recommendation systems. Usenet was the origin of recommendation system. It is a discussion system which is distributed worldwide. Usenet was newsgroups [6]. As the data increases it results in data overload. So because of this other solutions for data overload were developed. Tapestry is a recommendation system developed by Xerox in 1992. It coined the term "collaborative filtering" [6]. It was a manual collaborative filtering system. Large amount of emails and messages in newsgroup are handled by Tapestry. Then came Grouplens developed by John Riedl and Paul Resnick. It collected ratings from Usenet readers and used those ratings. It was an automated collaborative filtering system. It collects the human judgements known as rating for items and then the system will automatically collaborate with the users of similar interest. With such systems each user does not know about the

preferences of other users. Also such system does not want to know who the other users are or what items are present in the system in order to obtain recommendations.

With the advent of www came new recommendation systems such as Ringo, Firefly, Alexa Internet, PHOAKS, YAHOO! etc. Ringo was developed by Patte Maes in 1994 and it was a music recommendation system. Firefly is for both music and movie recommendations, which was developed by group of engineers in 1995. In 1997 another recommender system PHOAKS was developed. This system will search in the Usenet group for URLs and then it will post the important URLs to the website, as it indicates the most popular sites [11].

The most commonly used recommendation systems today are Amazon and TripAdvisor. Amazon is a recommendation system which will recommend items such as DVDs, food, toys etc. TripAdvisor is a travel recommendation system which will recommend the best hotels in different parts of the world.

III. LITERATURE SURVEY

Recommendation system is a system that will produce recommendations as output to users. It can be classified into two ways: content based approach and collaborative filtering approach. These two approaches can be further classified into memory based and model based approaches.

In memory based approach, data is stored in memory. It operates on data which is the users, items and ratings. In model based approach, a model is build. That is, it is based on the ratings of the dataset that a model is build. Using that model and some information's extracted from the dataset, recommendations can be done without using the entire dataset every time.

Content Based Approaches

Every item has its own description and each user will have a profile which has the user's preferences [7]. In content based recommender system, items are described using keywords. Various candidate items are compared with items previously rated by the user and the best matching items are recommended. Content based approaches are classified into memory based and model based approaches. In memory based approaches, some of the techniques in content based methods are Term frequency-inverse document frequency and K- Nearest neighbor. Whereas in model based approaches, the technique is decision tree.

Term frequency-inverse document frequency

Data can be in structured or unstructured format. Structured data is represented in database that is, it consists of rows and columns. News articles are example for unstructured data. In such cases there is no attribute name with values for it. Using this method free texts are converted to structured representations. Here each word is treated as an attribute, with a value whose type is Boolean denotes whether the word is present in the article or with a value whose type is integer denotes the number of times the word is present in the article [7].

Term frequency is the number of times the term appears in the document. Inverse document frequency is the number of documents in a collection divided by the number of documents that contain the term. The term frequency-inverse document frequency (tf*idf) is a value which is associated with a term, which gives the importance or relevance of that term in the document. The tf*idf weights are computed for every term. The word with the highest weight is more frequent in the document and that is the most central topic. The disadvantages of this method are that words can be used in different context, but this method does not take this into account.

$$w(t,d) = \frac{tf_{t,d}\log(\frac{N}{df_t})}{\sqrt{\sum_i (tf_{t,d})^2 \log(\frac{N}{df_{t_i}})^2}} \quad (1)$$

The tf*idf weight, w(t,d), of a term t in a document d is a function of the frequency of term in the document $(t_{f_{t,d}}, d)$, the number of documents that contain the term (d_{f_t}) and the number of documents in the collection (N).

The advantage of this method is that it can convert unstructured data into structured data and the disadvantage is that it does not consider the context in which the term is used.

K- Nearest Neighbor

K-nearest neighbor algorithm is one of the simplest methods which stores textual descriptions of items in memory. Textual descriptions include explicitly or implicitly labeled items [7]. In this method a new unlabelled item is classified by comparing that item with all other items in memory using similarity function. Most commonly used similarity function is Euclidean distance metric.

$$d(A,B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2} \quad (2)$$

where d(A,B) is the Euclidean distance between user A and user B, x_A is the rating of user x on item A, y_A is the rating of user y on item A, z_A is the rating of user z on item A.

Consider the following example shown below in Table 1. The rows represent the names of hotels in a city and the columns represent the properties of those hotels and the last column gives whether he/she likes or dislikes that particular hotel.

Table 1 Nearest Neighbor For Content Based Approach

Name	Food	Service	Cost	Like(L)/ Dislike (D)
Horizon	5	4	4	L
Sealord	4	3	3	L
Pearl	1	4	4	D
Blue Sky	5	4	4	L
Hail	1	3	2	D
Residency	4	5	5	L
New	5	4	4	?

Here 'New' is a hotel where we can predict whether the user likes/dislikes that particular hotel based on the similarity of likes/dislikes by the user on other hotels he has rated. Euclidean distance metric is used to compute the similarity between the data in nearest neighbor approach. The table 2 shows the calculations.

Table 2 Similarity Computation Using Euclidean Distance Metric

d(new,horizon)	0
d(new,sealord)	1.73
d(new,pearl)	4.89
d(new,blue sky)	0
d(new,hail)	4.58
d(new,residency)	1.732

From table2 it can be seen that the similarity between hotels 'new' and 'horizon' is 0, 'new' and 'sealord' is 1.73, 'new' and 'blue sky' is 0 and 'new' and 'residency' is 1.732. Similarity between them are very close to each other and since the similarity of those hotels that are similar to new are liked by the users in the past, from this computation one can predict that the user likes the hotel 'new'.

The advantages of this method are simplicity and efficiency. The disadvantages are it is expensive to find the k nearest neighbors when the dataset is large and the performance depends on the number of attributes selected.

Decision Tree

Decision tree is a model based approach used in recommendation systems. With the extracted information of the dataset, a model is build using decision tree. It has several benefits such as efficiency and flexibility. Decision tree is a predictive model. Based on the input attributes, this model maps the input to predicted output. In decision tree, attributes are represented by the interior nodes. The set of values of the attributes are represented as arc from parent to child node.

In the construction of decision tree, it always starts at the root node with training set. Item attribute at each node is chosen as the split attribute. To classify an item, we start at the root, and apply the predicate at the root to the item. If the predicate is true, go to the left child, and if it is false, go to the right child. Then repeat the same process at the node visited, until a leaf is reached. That leaf node classifies the item as liked or not [1].

That is item attribute at each node is chosen as the split attribute. The input set is divided by the values and so each child node receives only a subset of the input set that matches the values of the attribute specified by the arc to child node. This process repeats recursively until there are no more split attributes [7].

The table 3 below shows the previous history of a user's likes and dislikes to movies. Based on this information that is training data a model is build using the decision tree approach. From this model one can predict whether the user likes or dislikes the movie based on the content features of the movie.

Table 3 Example For Building A Decision Tree

Staring	Genre	Year	Like(L)/ Dislike(D)
Dileep	Comedy	1998	L
Mohanlal	Action	2000	D
Gopi	Science Fiction	1989	D
Gopi	Action	2010	L
Dileep	Comedy	1988	L
Dileep	Science Fiction	2010	L
Mohanlal	Comedy	2011	D
Mohanlal	Action	1985	D
Dileep	Action	2005	L

Suppose if the test data has content features such as staring by Dileep, genre is science fiction and year is 2000. Based on the model in the Figure 1, it is clear that the user likes the movie.

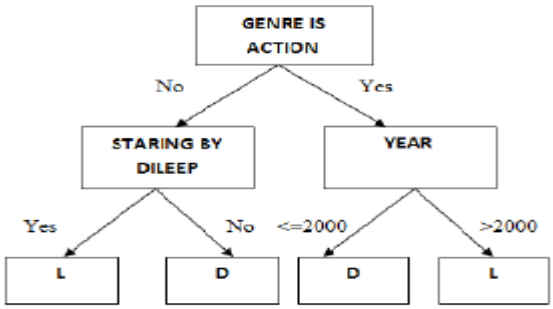

Fig. 1 Decision Tree of a user

The advantages are fast at classifying data and it deals with both strings and number. The disadvantages are large number of trees is to be build for each user or for each item, rating of one item can only be predicted at a time and it is not used for unstructured data

Collaborative filtering approaches

Collaborative filtering approach is mostly used in recommendation systems. This approach will compare the behavior of one user with other users. Then at the next step the interest and preferences of the nearest neighbors to the current user is taken into consideration. In memory based approaches, some of the techniques in collaborative filtering methods are User-based nearest neighbor algorithm and Item-based nearest neighbor algorithm. Whereas in model based approaches, the technique is clustering.

User-based Nearest Neighbor algorithm

This algorithm will compare the similarity between the users based on the ratings provided by them and then generate predictions according to it. The steps are as follows [3]:

1. Compute the similarity between current users and all other users using Pearson correlation coefficient.

$$sim(a,b) = \frac{\sum_{p \in P}(r_{a,p} - \bar{r_a})(r_{b,p} - \bar{r_b})}{\sqrt{\sum_{p \in P}(r_{a,p} - \bar{r_a})^2}\sqrt{\sum_{p \in P}(r_{b,p} - \bar{r_b})^2}} \quad (3)$$

   where, a, b – users, P - total number of items, $r_{a,p}$ - rating of user a on item p, $r_{b,p}$ - rating of user b on item p, r_a is the average rating of user a on all.
2. Selection of K most similar users is done by taking similarity weights between users which is greater than a certain threshold.
3. Prediction is then calculated as follows:

$$pred(a,p) = \bar{r_a} + \frac{\sum_{b \in N} sim(a,b) * (r_{b,p} - \bar{r_b})}{\sum_{b \in N} sim(a,b)} \quad (4)$$

   where pred(a,p) is the prediction of user a on item p, sim(a,b) is the similarity of users a and b, r_b is the average rating of user b on all items, N is the total number of users.
4. Items with highest rating are recommended.

The advantages are no content analysis, quality improves and serendipity. The disadvantages are new user problem, new item problem and lack of scalability. [8]

Item-based Nearest Neighbor algorithm

This algorithm will compare the items the target user has rated in the past with the target item and then K most similar items are selected. Then next the predictions are done according to it. The steps are as follows [10, 2]:

1. Compute the similarity between target items with all other items the user has rated in the past.

$$sim(a,b) = \frac{\sum_{u \in U}(r_{u,a} - \bar{r_u})(r_{u,b} - \bar{r_u})}{\sqrt{\sum_{u \in U}(r_{u,a} - \bar{r_u})^2}\sqrt{\sum_{u \in U}(r_{u,b} - \bar{r_u})^2}} \quad (5)$$

   where sim(a,b) is the similarity between the items a and b, $r_{u,a}$ is the rating of user u on item a, $r_{u,b}$ is the rating of user u on item b, U is the users who have rated both the items i and j.
2. K most similar items are selected.
3. Prediction is then calculated as follows:

$$pred(u,p) = \frac{\sum_{ierateditem(u)} sim(i,p) * r_{u,i}}{\sum_{ierateditem(u)} sim(i,p)} \quad (6)$$

where pred(u,p) is the prediction of user u on item p, sim(i,p) is the similarity of items i and p, $r_{u,i}$ is the rating of user u on item i.

The advantages are no content analysis, quality improves and serendipity and the disadvantages are new item problem, cold start problem and sparsity.

Clustering Neighborhood approach

In clustering technique, clusters are created in such a way that each cluster will have group of users who have similar tastes and preferences. The collection of objects within the same cluster will have same features than the objects in different clusters. The clustering neighborhood algorithm is as follows [9]:

1. Clustering algorithm divides the user-item database A into p partitions, that is A_1, A_2, A_3.....A_p.
2. Then the neighborhood of the active user u is the users in the cluster in which the active user belongs. If u belongs A_i then the entire partition A is used as the neighborhood.
3. Then that particular cluster in which the active user belongs is selected. Then prediction is calculated using basic collaborative filtering technique.

$$R_{a,j} = \overline{P_{C_a}} + \frac{\sum_{i\,rates\,j}\left(P_{C_{i,j}} - \overline{P_{C_i}}\right) r_{a,i}}{\sum_i |r_{a,i}|} \quad (7)$$

where $r_{a,i}$ denotes the correlation between the active user C_a and its neighbor C_i who have rated the product P_j, $\overline{P_{C_a}}$ denotes the average of customer C_a and $P_{C_{i,j}}$ denotes the rating given by the customer C_i on the product P_j.

The advantages are it solves the scalability, sparsity problem to an extent and also improves the prediction performance.
The disadvantages are tradeoff between performance and scalability and also it has an expensive model building [8].

IV. COMPARISON BETWEEN CONTENT BASED AND COLLABORATIVE FILTERING TECHNIQUES

Content based recommendation system will recommend items based on the similarity between the items the user has rated in the past. Collaborative filtering recommendation system will recommend items based on the similarity between users. Table 4 below gives the advantages and disadvantages [5].

Table 4 Advantages And Disadvantages

	Content based techniques	Collaborative filtering techniques
New user problem	**Yes**	**Yes**
New item problem	**No**	**Yes**
Sparsity	**No**	**Yes**
User independence	**Yes**	**No**
Scalability	**No**	**No**

- **New user problem:** Addition of a new user will cause problem for accurate recommendation because, the system should first learn about the user preferences.
- **New item problem:** Addition of a new item will cause problem for accurate recommendation because that item will not have been rated by users before.
- **Sparsity:** This is a problem in collaborative filtering recommendation system because there will be some items rated high but by only few number of users. In that case recommendations will be poor.
- **User independence:** In collaborative filtering recommendation systems, the recommendations are done based on the similarity between users. So collaborative filtering recommendation systems depend on users whereas content based recommendation system is independent of users.

V. CONCLUSION

Recommender or recommendation systems are a system which provides recommendations to users according to their tastes. Also it is a system that can guide the user in a personalized way. It has been applied to variety of applications. Collaborative

filtering techniques are most commonly used. Even though recommendations are obtained, it does not consider the sentiment of the reviews. In future work more accurate recommendation can be done by doing sentimental analysis using NLP techniques on the reviews of items. Also instead of finding similarity between users based on their ratings, similarity can be found by considering the similarity between the text reviews. In this way more accurate recommendations can be obtained.

VI. ACKNOWLEDGMENT

I am thankful to my guide Mrs. Jayalekshmi S., Associate Professor of Computer Science and Engineering, for her guidance and encouragement for this paper work.

REFERENCES

[1] Rasoul Karimi, Alexandros Nanopoulos, and Lars Schmidt-Thieme,."A supervised active learning framework for recommender systems based on decision trees",Springer, 2014, pp 1-25.
[2] W. H. Hu, F. Yang and Z. W. Feng, "Item based collaborative filtering recommendation algorithm based on MapReduce",Multimedia Communication and Computng Application , 2015, pp. 95-100.
[3] Zhi-Dan Zhao, Ming-Sheng Shang, "User-based Collaborative Filtering Recommendation Algorithms on Hadoop", IEEE International Conference on Knowledge Discovery and Data Mining, 2010, pp. 478-481.
[4] Joseph A. Konstan, John Riedl, "Recommender systems: from algorithms to user experience",Springer,2012, pp. 101-123.
[5] G. Adomavicius, and A. Tuzhilin, "Towards next generation of recommender systems: A Survey of the State of-the-Art and Possible Extensions", IEEE Transactions on Knowledge and Data Engineering, 2005, pp-734-749.
[6] Paul Resnick, Hal R. Varian, "Recommender Systems", Communications of the ACM , 1997, Vol. 40, pp. 56-58.
[7] Michael J. Pazzani and Daniel Billsus, "A Content Based Recommendation Systems", Springer-Verlag Berlin Heidelberg, 2007, pp 325-341.
[8] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey on Collaborative filtering Techniques", Advances in Artificial intelligence, 2009, pp. 1-19.
[9] Gui-Rong Xue, Chenxi Lin, Qiang Yang, WenSi Xi, Hua-Jun Zeng, Yong Yu,and Zheng Chen, "Scalable Collaborative Filtering Using Cluster-based Smoothing", In: Proceedings of the ACM SIGIR Conference,2005, pp. 114-121.
[10] Badrul Sarwar, George Karypis, Joseph Konstan and John Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms", Communications of the ACM 2001, pp. 285-294.
[11] Loren Terveen, Will Hill, Brian Amento, David McDonald and Josh Creter, "PHOAKS: a system for sharing recommendations",Communications of the ACM, 1997, pp. 1-3.