Binary Journal of Data Mining & Networking 5 (2015) 36-40
http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

A Hybrid Based Recommendation System based on Clustering and Association

Jaimeel M. Shah[1] , Lokesh Sahu[1]

*Corresponding author:

Jaimeel M. Shah

[1]Parul Institute of Eng. & Tech.
ME Computer Engineering

Abstract

Recommendation systems play an important role in filtering and customizing the desired information. Recommender system are divided into 3 categories i.e collaborative filtering , content-based filtering, and hybrid filtering are the most adopted techniques being utilized in recommender systems. The main aim of this paper is to recommend the best suitable items to the user. In this paper the approach is to cluster the data and applying the association mining over clustering. The paper describes about different hybridization methods and discuss various limitations of current recommendation methods such as cold-start problem ,Graysheep problem,how to find the similarity between users and items and discuss possible extensions that can improve recommendation capabilities in range of applications extensions such as , improvement of understanding of users and items incorporation of the contextual information into the recommendation process, support for multi-criteria ratings.

Keywords: Recommendation system, Collaborative filtering (CBF) , Content filtering(CF), Hybrid filtering , Coldstart , Graysheep , association mining, clustering .

Introduction

In everyday life, people rely on recommendations from other people by spoken words, reference letters, news reports from news media, general surveys, travel guides etc so recommendations plays an important role in finding the best items. A recommender system is the information filtering that applies data analysis techniques to the problem of helping customers find the products they would like to purchase by producing a predicted likeness score or a list of recommended products for a given customer. Recommender systems work from a specific type of information filtering system technique that attempts to recommend information items (movies, TV program/show/episode, music, books, news, images, web pages, scientific literature etc.) or social elements (e.g. people, events or groups) that are likely to be of interest to the user[1].

The recommender system also compares the user profiles and seeks to predict the ratings. With the help of Recommender systems are filtering and sorting data can be easily done. Moreover the Recommender system use opinions about the community of users and to determine content of interest using certain rules extractions. Recommendation systems are classified into 3 approaches i.e collaborative, content based or knowledge-based method to have a better recommendation. Each recommendation systems have some strategy to recommend better, the most useful strategies are listed below

A. Collaborative based Recommendation systems

Collaborative filtering Algorithm recommender system became one of the most researched techniques of recommender systems .if users shared the same interests in the past, they will also have similar tastes in the future. So, for example, if user A and user B have a purchase history that overlaps strongly and user A has recently bought an item that B has not yet been, the basic rationale is to propose this item also to B. The collaborative filtering technique recommends items based on user-based approach and item-based approach [2].

User-Based Approach

In the User-based approach the user plays an important role. If certain majority of the customer have same taste then they join into the one group .Recommendations are given to user based on evaluation of items by other users form the same group, with whom he/she shares common preferences. If the item was positively rated by the community, it will be recommended to the user.

Item-Based Approach

Here in Item-Based Approach the items play an important role Recommendations is based on evaluation of items. The system generates recommendations with items in the neighborhood that a user would prefer.

B. Content Based Recommendation System

Here in Content-based recommender systems deals with profiles of users that are created at the beginning. A profile has information about a user and his taste which is based on how user rates the items. In the recommendation process, the engine compares items that were already rated by user with items he did not rate and looks for similarities. Those items that are mostly similar to the positively rated ones, and the one which are positively rated by the users are recommended to the users[2].

In the context of content-based recommendation, the following questions that arise:

- How to determine which items match, or are at least similar to or compatible with, a user's interests?

- Which techniques can be used to automatically extract or learn the item descriptions?

C. Hybrid Based Recommendation System

Here in Hybrid Recommendation it is a combination of both collaborative approach and content based approach .With the help of Hybrid Recommendation different types of problems can be easily overcome the problem such as Cold-Start problem can be handled using the hybrid recommendations[2].

The combination of approaches can proceed in different ways:
How implementations of algorithms take place?
How to utilize some rules of content filtering in collaborative filtering??.
How to extract rules in the Hybrid recommendations?

Challenges And Issues

Cold-Start Problem

Its difficult to give recommendations to new users because his profile is almost empty and he has not rated any items so the taste of user remains unknown to the system so this type of the problem is called as coldstart problem. In some recommender systems this problem can be solve with the survey at the time of creating a profile. Another problem is when user has not rated before when new to the system. Both of these problems can be overcome with the help of hybrid approaches [2].

Data-Sparsity

Sparsely is the problem of lack of information. Suppose we have a huge amount of users and items but user have rated only few items. If a user has evaluated only few items then its difficult to determine the taste of the user.so to overcome this we use collaborative and hybrid approach to create neighborhoods of users based on their profiles.

Scalability

Due to increase of numbers of users and items, the system needs more resources for processing information and forming recommendations. Majority of resources is consumed with the purpose of determining users with similar tastes, and goods with similar descriptions. This problem is also solved by the combination of various
types of filters and physical improvement of systems.

Gray Sheep

Gray sheep problem means where user doesnot consistently agree or disagree to the group of the people and due to this reason for such user recommendation seems to be difficult[7] .

Explainability

Explainability is another important aspect of recommender systems. An incomplete reasoning such as "you will like this item because you liked those items" due to this recommendations of items will be difficult .

Different Hybridization Methods

Hybrid recommender systems combine two or more recommendation techniques for better performance with some drawbacks of any individual one[7].

Weighted Method

Here in Weighted method scores of several recommendations are combined together and it help to produce the single recommendation.The example of weighted method is P-TANGO system that uses hybrid Recommendations . Here first of all equal weight is assigned to both content and collaborative recommenders but gradually adjust the weights as the prediction of ratings are confirmed. Pazzani's combination hybrid does not use numeric scores, but rather use the output of each recommender as a set of votes, which are then combined in a consensus scheme.

Switching Method

Here in Switching method system uses some criterion to switch between recommendation techniques. The DailyLearner system uses a content/collaborative hybrid in which a content- based recommendation method is applied first. If the content-based system cannot make a recommendation with sufficient confidence, then a collaborative recommendation is attempted. this switching hybrid doesnot completely avoid problem.

Mixed Method

When large recommendations take place the Mixed method come into the action.here inthis method is used in Television System used. First of all content based method is used for textual description of tv-shows and use of collaborative method for finding the preferences of the user and Recommendations from the two techniques lead to suggest a final program. with the help of this mixed method new item -start up problem can be overcome : the content-based component can be relied on to recommend new shows on the basis of their descriptions even if they have not been rated by anyone. It does not get around the "new user" start-up problem, since both the content and collaborative methods need some data about user preferences to get off the ground, but if such a system is integrated into a digital television, it can track what shows are watched (and for how long) and build its profiles accordingly.

Feature combination

In Feature combination the features from different recommendation data sources are used together into a single recommendation algorithm. Feature combination hybrid lets the system consider collaborative data without relying on it exclusively, so it reduces the sensitivity of the system to the number of users who have rated an item. Conversely, it lets the system have information about the inherent similarity of items that are otherwise opaque to a collaborative system.

Cascade

Here in Cascade method it involves the stage process In this technique, one recommendation technique is employed to produce a ranking of candidates and a second technique refines the recommendation from the candidate set. The restaurant recommender EntreeC, described below, is a cascaded knowledge-based and collaborative recommender. Like Entree, it uses its knowledge of restaurants to make recommendations based on the user's stated interests. The recommendations are placed in buckets of equal preference, and the collaborative technique is employed to break ties, further ranking the suggestions in each bucket.

Feature Augmentation

To improve the performance of a core system Feature Augmentation is used. For example Libra System makes content-based recommendations of books based on data found in Amazon.com, using a naïve-bayes text classifier and this help in finding the quality of books. In Feature Augmentation one technique is used to produce a rating of an item and that information is then incorporated into the processing of the next recommendation technique. So the difference between the Cascade and augmentation are as follows: in feature augmentation the feature used by second recommendation is the one which is the output of the first one where as in cascading second recommender doesnot use the output of first one but the results of the two recommenders are combined in a prioritized manner.

Meta-level

Here two recommendation techniques can be merged by using the model generated by one as the input for another. The difference between the meta-level and augmentation is that in augmentation output of first recommender is used as input for second one where as in meta-level the entire model will be consider as a input for the second one . The first meta-level hybrid was the web filtering system Fab .

Different methods to find Similarity

Similarity is define by data analysis in term of distance function. The distance function can be calculated using Euclidean distance or Manhattan Distance.

$$\begin{aligned} d(i,j) &= \sqrt{(x_{i1}-x_{j1})^2+\ldots+(x_{in}-x_{jn})^2} \\ d(i,j) &= |x_{i1}-x_{j1}|+\ldots+|x_{in}-x_{jn}| \end{aligned}$$

Different methods that are used for calculating the similarity are as follows[5]

Cosine Similarity

In this approach, items are thought of as vectors in the m dimensional user-space where the dimension is the attribute by which the item are rated. The cosine of the angle between the vectors that represent two items is their similarity. We know from calculus the dotproduct formula:

$$\vec{i}\cdot\vec{j} = ||\vec{i}||\cdot||\vec{j}||\cdot\cos\Theta$$

$$\Longrightarrow sim(i,j) = \cos\Theta = \frac{\vec{i}\cdot\vec{j}}{||\vec{i}||\cdot||\vec{j}||}$$

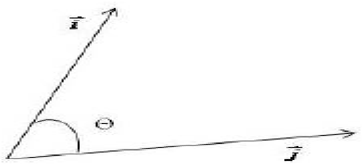

Figure 1: Vector representation of items. The more is small, the more similar are the items

Conditional based Probability

Another way to compute the similarity is to use a measure that is based on the conditional probability of liking (or rating) an item given that the user already showed his interest for another item. If an item i has a good chance of being purchased after an item j was purchased then i and j are similar. The similarity is given by $sim(i, j) = P(i|j) \times \alpha$
where is a factor dependent on the probem

Pearson Corelation Similarity

The similarity is given by amount of corelations between items or usersIf the set of users who both rated i and j are denoted by U then the correlation similarity is given by:

$$\begin{aligned} sim(i,j) &= corr_{ij} \qquad (3) \\ &= \frac{\sum_{u \in U}(R_{u,i} - \overline{R_i})(R_{u,j} - \overline{R_j})}{\sqrt{\sum_{u \in U}(R_{u,i} - \overline{R_i})^2}\sqrt{\sum_{u \in U}(R_{u,j} - \overline{R_j})^2}} \end{aligned}$$

Prediction based on explicit rating

In this case, users are required to express their ratings on items. e.g the user1 has given an movie 1. Let $I^0 = I\, x : x = 1, 2, \ldots, n^0 \wedge n^0 \leq n$ where

Our Approach and Basics

In this paper, a hybrid recommendation system is used to combine the content based and collaborative methods to capitalize their strengths and to achieve a good performance. Here the main purpose is to recommend the best suitable items to the user so to recommend the best suitable items to the user we require a strong rules to generate it. First of all the data gets cluster based on the certain characteristics. The main aim of the clustering is to generate the several cluster based on certain characteristics so it can be easily known which item belong to which cluster. Secondly applying the association mining algorithm on the clusters that are generated and to generate the frequent itemsets. The main purpose of applying the association rules is to generate the strong rules from the frequent itemsets. Now let us understand the strong rules that how it can be generated. When the rules satisfy the minimum support and minimum confidence than it is said to be called as strong rules and if it fails to satisfy this condition that this type of rules is said to be called as weak rules.The calculation of support and confidence are as follows.

Support

Every association rule has a support and a confidence.
"The support is the percentage of transactions that demonstrate the rule."

n is the total number of items in the database the set of items that users ux and uy have both rated. The similarity between ux and uy is given by a 5 ratings so to compute this we have formula such as

$$\begin{aligned} \kappa_{x,y} &= sim(u_x, u_y) \\ &= \frac{\sum_{h=1}^{n'}(r_{u_x,i_h} - \overline{r_{u_x}})(r_{u_y,i_h} - \overline{r_{u_y}})}{\sqrt{\sum_{h=1}^{n'}(r_{u_x,i_h} - \overline{r_{u_x}})^2}\sqrt{\sum_{h=1}^{m'}(r_{u_y,i_h} - \overline{r_{u_y}})^2}} \end{aligned}$$

Prediction based on Implicit rating

Implicit rating does not mean that a user will not show his appreciation toward an item, it simply means that he does not do it directly or explicitly as with the preceding approach. The rating of each item is captured implicitly. For example, if a user spend more time looking on an item, The item gets a higher rating.

$$\begin{aligned} \lambda_{x,y} &= sim(u_x, u_y) \\ &= \frac{\sum_{h=1}^{n'}(r_{u_x,c_h} - \overline{r_{u_x}})(r_{u_y,c_h} - \overline{r_{u_y}})}{\sqrt{\sum_{h=1}^{n'}(r_{u_x,c_h} - \overline{r_{u_x}})^2}\sqrt{\sum_{h=1}^{m'}(r_{u_y,c_h} - \overline{r_{u_y}})^2}} \end{aligned}$$

<u>Example</u>: Database with transactions (customer_#: item_a1, item_a2,)
1: 1, 3, 5.
2: 1, 8, 14, 17, 12.
3: 4, 6, 8, 12, 9, 104.
4: 2, 1, 8.
support {8,12} = 2 (,or 50% ~ 2 of 4 customers)
support {1, 5} = 1 (,or 25% ~ 1 of 4 customers)
support {1} = 3 (,or 75% ~ 3 of 4 customers)

The confidence is the conditional probability that,
given X present in a transition , Y will also be present.
Confidence measure, by definition:

Confidence(X=>Y) equals support(X,Y) / support(X)

Thus based on the confidence and support that are decided the strong rules will be generated Third step is to divide the items into favorite and non-favorite itemset i.e if ratings less than 3 it is consider as non favorite itemset and if rating is more than 3 so it is consider as favorite itemset. As we get the favorite itemset we will see the rules wether which rules belong to the favorite itemset and the items that is derived from rules was not rated then we will recommend that derived items to the user and similarly for non-favorite itemset we will check the similarity among the items. Those who have the highest similarity will be recommended to the user.

Recommendation Approach	Recommendation Technique	
	Heuristic-based	Model-based
Content-based	Commonly used techniques: • TF-IDF (information retrieval) • Clustering Representative research examples: • Lang 1995 • Balabanovic & Shoham 1997 • Pazzani & Billsus 1997	Commonly used techniques: • Bayesian classifiers • Clustering • Decision trees • Artificial neural networks Representative research examples: • Pazzani & Billsus 1997
Collaborative	Commonly used techniques: • Nearest neighbor (cosine, correlation) • Clustering • Graph theory Representative research examples: • Resnick et al. 1994 • Hill et al. 1995 • Shardanand & Maes 1995 • Breese et al. 1998 • Nakamura & Abe 1998	Commonly used techniques: • Bayesian networks • Clustering • Artificial neural networks • Linear regression • Probablistic models Representative research examples: • Billsus & Pazzani 1998 • Breese et al. 1998 • Ungar & Foster 1998
Hybrid	Combining content-based and collaborative components using: • Linear combination of predicted ratings • Various voting schemes • Incorporating one component as a part of the heuristic of each other	Combining content-based and collaborative components by: • Incorporating one component as a part of the model for the other • Building one unifying model

Conclusion

Recommendation System plays an important role over recent years where various method such as content ,collaborative ,and hybrid method are proposed .This paper mainly describes various limitations in recommendation system. The paper mainly consist of approach where it help us to recommend the best suitable items to the user by applying association mining on clustering Moreover it also deals with various hybridization methods like weighted method, which is used to overcome the certain limitations Moreover, paper also describes about the various approaches regarding the hybridization and it will help to know whether which approach to be used for hybridization

References

[1]. Diego Campo, Miquel Sonsona, Jose-Miguel Pulido “A hybrid recommender combining user, item and interaction data”.

[2]. Mohammad Hamidi Esfahani, Farid Khosh Alhan “New Hybrid Recommendation System Based On C-Means Clustering Method”.

[3]. Prof. Vipul Vekariya, DR.G.R.Kulkarni “Hybrid Recommender systems: survey and experiments”.

[4]. Ujwala H.Wanaskar, Sheetal R.Vij, Debajyoti Mukhopadhyay “A Hybrid Web Recommendation System based on the Improved Association Rule Mining Algorithm”.

[5]. Gediminas Adomavicius and Alexander Tuzhilin “Towards the Next Generation of Recommender Systems:A Survey of the State-of-the-Art and Possible Extensions” .

[6]. Gediminas Adomavicius and Alexander Tuzhilin “Towards the Next Generation of Recommender Systems:A Survey of the State-of-the-Art and Possible Extensions” .

[7]. R.Srinivasa Raju I.Kali Pradeep I.Bhagyasri “Recommender Systems for E-commerce: Novel Parameters and Issues” .

[8]. Robin Burke “Hybrid Recommender Systems: Survey and Experiments” .