A Review on Advanced Algorithms in Recommender Systems

Saranya K., M. Brindha, S. Rajesh, S. J. Savitha
Department of Computer Science and Engineering,
Sri Ramakrishna Institute of Technology,
Coimbatore Tamil Nadu, India

Abstract

Searching in web is becoming a misery for the naive people nowadays because of the information explosion across the internet. Many E-commerce websites are relying on the Recommender Systems (RS) in order to sell and gain from their products to the people (e.g Amazon, Netflix, IMDb, Yahoo). Creation of effective and accurate Recommender Systems are essential for the E-commerce owners to capture the positive impression of the clients by reducing their burden of information overload problem. The main objective of this paper is to give the researchers an outline of the effective algorithms used in developing a stabilized powerful RS tool which suggests not-yet-experienced products but that may suit according to the users current preference. The results of the discussed algorithms below depicts the working process of each algorithm according to the degree of utility of the clients. This review article concludes by leaving the ball in researchers court by selecting the appropriate algorithm for developing their RS depending upon their specific application.

Keywords: Recommender Systems (RS), degree of utility, information overload.

I. INTRODUCTION

Recommender Systems (RS) are the software tools that make use of the data mining techniques, in order to provide useful suggestions (Goldberg D et al. 1992) of items according to the user preference. It is primarily used for the users who lack of personal experience about a particular genre. The recommendation is based on the degree of utility of the user. Over 170 research articles have been published on RS until recent years and it has been gradually increasing day by day due to its wide variety of applications. RS makes use of the hefty amount of information which are scattered across the database, structures and organizes it and provide relevant information as per the request of the active user (information retrieval), thereby pruning the unwanted information (Mahmood.T et al., 2009). In this paper the evaluation of the Recommendation approaches, advanced algorithms used in RS and finally the challenges in RS field are being discussed.

The recommender systems emerged as an individual research area only in the mid 1990’s. Before that it was just a part of information retrieval technique in the data mining domain. The RS development was initiated due to a simple observation that people generally rely on other sources for getting suggestions (Mcshery. F et al., 2009) such as watching a movie, reading a book, collecting compact disks either from their friends or through the critics from magazines and newspapers. Whenever the users have trouble in decision making autonomously due to the lack of knowledge in any domain they seek for the recommendation. Due to the information explosion on the web lot of choices are put in front of the users when they need to get their relevant information. Hence the choices became misery to the individuals (Anand S.S et al., 2005);

this simple idea has led to the emergence of this recent research field RS which has proved to be a solution for information overload problem.

The RS have been broadly classified into two types based on their type of suggestions they are, personalized recommendation systems and non-personalized recommendation systems (Montaner. M t al., 2003). The non-personalized recommendation systems are very common ways of predicting items to generic users who does not feed any information about them to the system. Hence the system will provide all popular items that have been widely viewed by other users; there is no need for research in such type of recommendation. The personalized recommendation system came into existence where the user needs to get the suggestion according to their specific preference and constraints. Here the system needs to get the information from user either explicitly (i.e. providing ratings or asking for feedbacks) or implicitly (i.e. Observing the user’s pattern of browsing or navigation of pages and selecting items).

The RS field has gained drastic attention by researchers in recent years due to the following facts;

- Usage of RS in popular websites such as Amazon, Netflix, Yahoo, YouTube etc.
- Wide range of conferences and workshops are being held on RS with reference to the ACM Recommender Systems.
- It has been implemented in the educational field as a subject and also as a course offered to both UG and PG students.
- There have been several dedicated issues of RS in reputed academic journals.
The RS plays two roles, one as a service provider and the other as a user (Montaner, M. & al., 2003). The service provider creates RS in order to provide functions such as:

- To make his products get sold out.
- To sell a variety of products other than the popular ones.
- To make the user satisfied.
- To create the trust for the user and attract more clients.
- To improve themselves by knowing the actual needs of the users.

The user makes use of the RS if and only if it effectively satisfies their needs for utilization (Bridge D & al., 2006). The big part here is that the RS should balance the goals of both the players and it is determined by the technique used for its development.

The RS is an information processing system which collects a variety of data from various sources and provides the suggestions (Burke, R & al., 2007). Since the data and knowledge source is vast and very diverse it is the technique that decides the effective usage of that information. The data used by the RS can be classified into three types, namely items, users, and transaction (Schafer & al., 2007).

The items are the products or objects requested by the active user for his utility it can be about anything like movies, songs, news, magazines, books etc (Mahmood, T & al., 2007). The items are featured by their value and complexity (Ricci F & al., 2006). The value is calculated by positive or negative points and complexity is derived from the urge to use of the active user (Arazy, O & al., 2009). The user is the individual who interacts with the system to get what he needs. The individual may have various needs and preferences and the RS should take his goals into consideration and should provide the relevant information using an effective recommendation technique (Benshimon D, 2007). And finally the transaction is the recorded interaction between the system and the user; it is a lot like data that stores the information regarding the item and the context of those items. This information is very much helpful for the RS to make its accurate prediction.

Ratings are the most famous example of transaction data which are obtained from the users either explicitly or implicitly (Schafer & al., 2007). Some types of ratings are as follows;

- Numerical ratings, which contains 1-5 stars.
- Ordinal ratings provide alphabetical options to be marked by users.
- Unary ratings just indicate if the item has been purchased or not.
- Binary ratings get the good or bad comment from the active user.

The remainder of this paper is organized as follows. In section 2, evaluation of recommender systems is being discussed. Section 3 depicts the advanced algorithms in RS. Section 4 narrates the challenges in RS followed by the conclusion in section 5.

II. EVALUATION OF RECOMMENDER SYSTEMS

As the recommendation systems are being widely spread among the research communities as well as the commercial communities, a proper guidance of design and implementation is required to enhance the qualities of the RS, since many approaches have been suggested on employing this system a system designer is responsible for selecting an appropriate algorithm to decide which property of the application to be focused. In earlier days the effectiveness of recommendation systems is evaluated using the accuracy property. But now the users are more interested in discovering new items, exploring diverse packages and simultaneously maintaining their privacy. In order to do so, some comparison has to be done among the various approaches using different properties using some evaluation metrics. This performance comparison is carried out in three stages, they are: offline experimentation, user studies and online experimentation.

Initially the experimental setup should be done with the following guidelines:

- First a strict hypothesis has to be formed, and the experiment has to be designed in such a way that it tests that hypothesis.
- The experimentation should be carried out with same data sets on different algorithm or different data sets with the same algorithm in order to understand the cause of good performance.
- Generalization has to be made by training diverse data sets with the particular algorithm to make better predictions.

First of all the offline experiments are performed by pre-collected data sets. The data sets are the rating of items or choosing the users. It is one of the evaluating techniques performed at low cost without using any real world users. Simulation of the behavior of users is calculated using these datasets. One of the advantages of using offline experiments are they allowed us to compare many candidate algorithms, resulting in the filtration of unwanted approaches.

The user studies are another way of simulating the user’s behavior before deploying the recommendation system for real time usage. Generally the recommendation systems rely on the interaction of the users for evaluating its better performance. The user study is conducted by recruiting a set of test users and asking them to perform some particular set of tasks. By doing so their activities, behavior and the accuracy of results with the system are recorded. This is the direct way of evaluating the system performance. Finally the online evaluation technique is developed in order to overcome the limitations of above evaluating techniques. Here the recommendation systems are measured using online testing. Hence, if the system should be evaluated in the best way, then the above three steps should be followed in order to minimize the negative effects in the online evaluation experiment.

III. ADVANCED ALGORITHMS IN RECOMMENDER SYSTEM

Various types of advanced algorithms have followed in recent years in the RS domain; some of them are given below;
A. Aggregation of Preference in Recommender Systems

As the name itself depicts this algorithm aggregates or combines all the possible suggestions using aggregation functions and properties (Burke R et al., 2002). Generally this process is done by using mathematical functions such as arithmetic mean or max/min functions, but the result is not accurate and justified. So, this algorithm forms, some reasonable aggregate functions such as,

- Collaborative filtering (CF) methods.
- Content based filtering (CB) methods.
- Demographic filtering (DF) methods.
- Utility based (UF) methods.
- Knowledge based (KB) methods.
- Hybrid recommender (HR) methods.

In the CF method the justification used to provide the suggestion for the particular item depends on the ratings of similar users. Likewise, the justification that is provided with the recommendation that is given referring to the similarity in the item’s features comes under the CB method. The DF method delivers the correctness of the preference by providing suggestion depending on similar profile of individuals. The items are suggested based on the current needs, preferences and tastes of the clients in the UF method. Whereas the KB method analyses the complete background knowledge about the user and justifies its prediction. Finally, the HR method combines all the above justification techniques and eliminates its drawbacks. The main role of the aggregation function is to combine multiple inputs merge them as a single result and provide it to the user.

The aggregate functions are constructed by the following sequential ways, i.e. data are collected and pre-processed, followed by constructing the syntax semantics, interpreting it and understanding its functional behaviour. Finally the weight and parameters are determined.

Say if there is a new user \( u \) and neighbourhood of similar taste, users \( U_n = \{ u_1, ..., u_i \} \) the probability of the user \( u \) to choose a new item \( d_i \) is the summation of the scores or ratings of all the neighbourhood users of similar tastes \( R(u,d_i) \).

Aggregation of preferences is done by the formula,

\[
R(u, d_i) = \sum_{j=1}^{m} \text{sim}(u, u_j)R(u_j, d_i) \quad (1)
\]

Similarity between the user and similar users can be calculated using Weighted arithmetic mean \( \text{sim}(u, u_j) = w_j \) which gives the weights and \( R(u_j, d_i) \) which are the inputs for aggregation.

Aggregation of features is given by the equation,

\[
R(u, d_i) = f(x_1, ..., x_n) \quad (2)
\]

where \( R = (u, d_i) \) is the overall rating.

And the measure of similarity is calculated by using,

\[
R(u, d_i) = \sum_{j=1, j\neq i}^{q} \text{sim}(d_i, d_j)R(u, d_j) \quad (3)
\]

Where \( \text{sim}(d_i, d_j) \) are the previously rated items by the user.

B. Active Learning In Recommender Systems

The motivation of RS is to present items which are highly preferable for the user. Apart from this myth there is also another dimension of the system which is ultimately getting knowledge about the user, finding their preference so as to increase the profitability of the RS. In order to do this, an algorithm is formed which is known as Active Learning (AL) RS (Schain A.I 2002). The process of integrating the methodology of finding the user’s likes and dislikes in addition to the predictive nature of RS ultimately produces some better way of personalizing the recommendations to the users. This is called as AL in RS.

If a new user is using the RS, the system does not have enough information about the new user to predict the item which satisfies the customer, in such case the system can ask the user to rate some new items from the system catalogue and capture the intention and the interest of the user; this process is called as the training points. The selection of the items to put forth in front of the user should be from different genres and should be limited so as to reduce the user’s inconvenience (Swearingen.K et al., 2001). The calculation of training points is very important in the active learning mechanism. Using these training points a model is built to nail the user’s intention. There is a notion that AL is so irritating or boring concept, but still it helps in finding and exploring new data which helps in improving the prediction accuracy (Settles et al.,).

An item is the input which is denoted by the variable \( x \) \( \in X^{\text{train}} \). Set of all the items is given by \( X \). Preference of user \( u \) is given by the function \( f_u \) where \( u \) is the target user; rating of an item \( x \) is the output function denoted by \( y = f(x) \) whereas \( Y = \{1, 2, ..., 5\} \) are the possible outcomes which are the ratings. In active learning items with their user ratings are partitioned into two types training point and testing point. Items that are belonging to training set is denoted by \( X^{\text{train}} \) and the training set is given by the equation \( T = \{(x_1, y_1) \} \) where \( x_1 \in X^{\text{train}} \).

\( X^{\text{test}} \) refers to items in test set where prediction error is given by \( L \) which is loss function; for eg. mean absolute error (MAE) or mean squared error (MSE).

\[
L_{\text{MAE}} \left( f(x), f'(x) \right) = |f(x) - f'(x)| \quad (4)
\]

\[
L_{\text{MSE}} \left( f(x), f'(x) \right) = (f(x) - f'(x))^2 \quad (5)
\]

Our objective is to minimize the generalization error with respect to training set. Now active learning is denoted by,

\[
\hat{G}(x^{\text{train}} \cup \{x\}), \text{ Or } \hat{G}(x) \quad (6)
\]

Where \( \hat{f} \) = function approximation from training points, \( \hat{G} = \text{generalization error} \).

C. Multi-Criteria Recommender Systems

The generic process of recommender system is used to produce the needful items to the requested user. By the way the traditional RS makes use of the ratings of the target user or the ratings of the user community (neighbours) to predict the item to the user’s interest otherwise known as
the utility function. The utility function is calculated using the items (i), users (u) and delivers R(u, i) where R is the recommendation result [19]. The utility function here takes only one aspect, i.e. the ratings of the user from their history and predicts the result. This method produces results with less accuracy; in order to improve the accuracy of the results Multi-Criteria Recommender System (MCRS)( Roy et al., 1996) is introduced.

The MCRS takes various aspects of the utility function such as the requested user’s ratings, his community or friend’s ratings (neighbour) into consideration and predicts the items (Adomavicius G et al., 2005). The MCRS makes use of multi criteria decision making (MCDM) methods to run the system i.e. for example if two users has watched three movies commonly and has rated the three movies as 4 out of 5 we could only grasp the outer layer of the users interest this is the single aspect prediction in utility function; suppose if the three movies are split into various genre such as cinematography, music, director etc then the first user may give rating as (4,3,2) and second user may give his rating as (5,5,1) so the utility function is split into multi criteria to make better judgment of the user’s preference and better prediction is made. Generally the utility function is R can be written as;

\[ R: \text{Users} \times \text{Items} \rightarrow R_{ij} \]  

Extending the above equation to multi-criteria recommender rating with and without using overall ratings is given as follows;

\[ R: \text{Users} \times \text{Items} \rightarrow R_{ij} \times R_{ij} \times \ldots \times R_{ij} \]  

\[ R: \text{Users} \times \text{Items} \rightarrow R_{ij} \times R_{ij} \times \ldots \times R_{ij} \]  

Bernard Roy one of the pioneers of MCDM method proposed certain methodologies (Resnick P et al., 1994) for implementing this method into the generic recommendation system they are as follows;

- Identifying the set of items to be considered while making the decision.
- Identifying all the parameters and functions that declare the preferences of the users.
- Developing a public preference model using the functions declared in above step i.e. comprising all the preference of the user.
- Selection of the items from the alternatives by decision making process of the constructed software system.

D. Robust Collaborative Recommender Systems

Collaborative algorithm works on the basis of user to user recommendation i.e. users having similar tastes are captured similarity with the active user is measured and the prediction is made. When a system is launched it is also exposed to vulnerabilities and attacks similarly the RS also suffers from attacks by the attackers. In order to protect it from the attacks robust collaborative systems has to be designed. Researchers started to research on these vulnerabilities on collaborative systems on 2002 and laid foundation to investigate, detect and avoid such attacks by building robust collaborative systems (Herlocker J et al., 1999).

The aim of attackers is to create fake profiles inject into the RS and makes the system to provide irrelevant items to the user i.e. the adversary makes an item look alike a good choice for recommending but it is not the good choice and vice versa. There are two types of attack one is efficient attack, here the attacker creates low cost manipulation to the system which causes a large impact on the system (o Mahony et al., 2002). This type of attack is easily detectable finding the group of fake profiles and eliminating them from the database can be done. Second one is the inefficient attack where it is costly way of inducing modification to the RS database and has a very large impact. This type is not easily detectable.

The high knowledge attack is one that needs the complete understanding of the processing of the RS, i.e. the mean and the standard deviation of the items is well known to the attacker (Lam S.K et al., 2004). In the low knowledge attack the attacker knows only the public knowledge about the RS. The informed attack is done by the attacker who has the in depth knowledge about the mathematical properties attributes and the working of the RS.

IV. CHALLENGES IN RECOMMENDER SYSTEM

Some of the challenges prevailing in the recommender system area where the researchers and practitioners focus on are given below,

- The spatial problem is one of the most discussed issues in the RS field. The algorithms that are designed in offline method are not capable of handling huge amount of data (Sarwar B et al.,2002).
- The annoying nature of RS i.e. the recommendation is proactively given to the user even before the user request for it. This becomes disturbing for the users (Sae-Ueng,S et al., 2008).
- The privacy of the user is affected since the system collects too much information about the user for the better performance of the system (Ramakrishnan et al., 2001).
- Combining the long term and short term needs of the users and providing results according to their interest is still lacking in the RS (Aimeur et al., 2000).
- Efficient recommender systems integrated to mobile devices is still an open challenge since mobility has been an emerging trend for storing personalized data.
- Optimization of multiple recommendations.
- Distributed RS using one to one conversation between client and server creates a centralized system effect (Han P et al.,2004).
V. CONCLUSION

This study provides the review of the various algorithms used in the RS field. Each algorithm has its own way of working process which has been reviewed in the above section, and the common challenges that is prevailing in the RS field are discussed. Each type of algorithm is used according to specific application. The future of recommender systems lies in doing justice to the serendipity and developing a strong RS that supports the user needs without information overload problem. Despite of all of its advantage “cold start problem” is one of the major barrier in developing a robust RS which is yet to be addressed and its in the hands of researchers to give best solution to this limitation depending upon their own application.

REFERENCES


