Big Data in Recommendation System

Yujie Wu
Indiana University Bloomington
Bloomington, Indiana 47401
yujiwu@iu.edu

ABSTRACT

This paper will befocused onhow the company recommend their products and services to their customers based on the dataabout customer's preferences through the case of Netflix and Yahoo.

KEYWORDS

Recommendation system, Netflix, Yahoo

1 INTRODUCTION

With the development of web technology and online business, recommender systems are widely used in e-Commercial business platform such as Amazon, eBay, Monster and Netflix. They process a tremendous number of online commercial activities and provide the personalized online business experience, which relies on the recommender system. The recommender system tells the customers what they are looking for, what they want to buy, and so on. With the recommender system, less popular products can attract peoplefis attention and e-Commercial business model works more efficient and profitable.

The basic problem of recommendation system is personalized matching of items (people, products, services, jobs, etc.) to people[3]. The recommendation system takes the data that produced by peoplefis online activities and some specific criteria such as overall context, user information, community information, properties of items plus the machine learning or data mining algorithms to output some information or suggestion that people may be interested in or related to according to their preferences for some goals[3].

2 NETFLIX RECOMMENDER SYSTEM

Netflix recommender system is an industrial-scale and real-world recommender system. Its recommendation is based on personalization. For customers (household), everything that they could see on the front-end webpage containing rows and columns is recommendation. The columns are sorted by the ranking and diversity. The personalized genre rows focus on user interest which is important for user satisfaction. They are generated based on the usersfi recent activities, ratings, comments, or usersfi preference settings. The recommender system will filter out the movies if they have been watched before and exclude duplicated tags and genres when providing recommendations and suggestions[3].

Netflix recommender system just takes the user preferences and output the immediate recommendations. How is it highly related to Big Data? According to the statistics from Netflix, last two quarters in 2013 have four million new registered subscribers, which leads to total 29.2 million subscribers were actively using Netflix. There are 4 million ratings, 3 million searches, 30 million plays happening in Netflix website every day. At the end of 2013, Netflix reached 44 million members[3]. A large amount of data is collected each day. Therefore, it becomes reality that Netflix recommender system could use Big data which usually beats better algorithms to provide recommendations.

The algorithms that Netflix recommender system is using are Restricted Boltzmann Machines (RBM) and a form of Matrix Factorization. They are developed as part of the Netflix 2007 Progress Prize which worth several million dollars. Restricted Boltzmann Machine is a neural network. The form of Matrix Factorization is an asymmetric form of SVD which can take implicit information into account[1]. Both algorithms consist of a tremendous number of different machine learning techniques. The algorithms consume a large amount of data as their input. Machine learning techniques form an abstract model which is waiting for data stream to shape it. Once the model reaches convergence or it becomes mature enough, the algorithms output the prediction which is used as the recommendation for Netflix users. More data means more precise the outcome is.

The recommendation algorithms are designed based on the hypothesis that the suggestions will increase the member engagement with Netflix service and ultimately attract more users and more profits. To verify whether the algorithm works as expected, Netflix designed a test, named AB test. AB test is an experimental approach to figure out the changes of webpages which maximize an outcome of interest. The test contains two identical versions with only one different variation which possibly affect customer's behavior[3]. For instance, the A version of a website has some webpages that could be accessed through a category list. The version B of that website is modified from version A that the webpages which can be accessed only through a category list now have their own shortcuts listed on the main page of the website. Once executing the AB test, it is obvious whether the modification on that variation increases the user engagement.

To modify the webpage, it should measure or evaluate all related metrics, which is a data-driven process. Metrics could be short-term or long-term. Sometimes, short-term metrics do not fit the long-term goals. For example, larger quantity of clicks does not necessary mean better recommendation. However, long-term metrics such as member retention works better in Netflix[3]. With the choice of metric, Netflix monitors how users interact with different algorithms during the testing.

3 YAHOO RECOMMENDER SYSTEM

The main page of Yahoo contains many modules such as advertising module, search queries recommendation, breaking news recommendation, and application recommendation. All recommendations rely on Yahoo recommender system based on the given context such as user data and user preferences. Yahoo recommender system is not merely an algorithm or a piece of code, it is an environment involves items, context, and metric. Items could be articles, advertisements, movies, songs that users may be interested in. Context could be query keywords, pages, mobile, social media that users provided while surfing online. Metric could be click rate, revenue, engagement that needs to be optimized for achieving some long-term business objectives[4].

Every second, a tremendous amount of data from users and machines is feed to the system. It is a problem that big data matters. Therefore, big data analytics and machine learning algorithms can be applied to improve or optimize the metric and the system while recommendation is on-going.

The data is easy to obtain but its quality is not guaranteed since the nature of data resource. Various factors including the properties of the item, context, feedback, and constraints specifying legitimate matches may affect data quality and eventually the solution. Yahoo recommender system uses collaborative filtering to deal with such problem.

Collaborative filtering assigns each item an individual rating to form a consensus recommendation. To be more specific, collaborative filtering has three branches which are user-based collaborative filtering, item-based collaborative filtering, content-based collaborative filtering. As the name implies, user-based collaborative filtering groups the similar users and find their preferences, then it predicts the interest of current user based on the group of the similar users. Item-based collaborative filtering recommends items to current user based on the rating that is assigned to each individual item. Content based collaborative filtering finds the items with the similar properties that the current user likes[4].

Collaborative filtering is now the most prominent approach to generate recommendations. It presumes that the ratings of the items are given by users. Then it takes a table of data including the users and item ratings to compare the values and return the top-ranked items for the current user[4]. Finally, collaborative filtering outputs a prediction that describes how much the current user likes or dislikes the item.

As mentioned before, the input is a table which has a set of attributes. Each attribute represents an item and each tuple represents a user. Therefore, the value in each cell means the rating of the item given by corresponding user. Collaborative filtering finds some most similar users and their items to the current user, then remove the items that current user have already seen or purchased. Hence, the input data table only includes similar users and items which will be recommended to the current user.

Here remains a problem that how to define the similarity between users. Let A and B be two different users and let I be the set of items that both user A and B rated. Let $r_{a,i}$ be the rating of user A for i^{th} item. Let $\overline{r}_a$ and $\overline{r}_b$ be the average value of all items in set I rated by user A. Therefore, the similarity could be calculated as the following function[4]:

$$sim(a,b) = \frac{\sum_{i \in I}(r_{a,i} - \overline{r}_a)(r_{b,i} - \overline{r}_b)}{\sqrt{\sum_{i \in I}(r_{a,i} - \overline{r}_a)^2}\sqrt{\sum_{i \in I}(r_{b,i} - \overline{r}_b)^2}}$$

The similarity function is called cosine similarity. The function assumes each tuple in the table is a vector. Since similarity function uses cosine value, the possible value of similarity is between -1 and 1. If two vector points to the same direction, cosine similarity value equals 1. If two vector points to the opposite direction, cosine similarity value equals -1.

Once the data table is feed to the algorithm, the prediction of the rating value of some random item i which will be recommended to the current user could be calculated as follows[4]:

$$pred(a,i) = \overline{r}_a + \frac{\sum_{b \in N} sim(a,b)(r_{b,i} - \overline{r}_b)}{\sum_{b \in N} sim(a,b)}$$

where a is the current user and b is a random user in the data table. The set N is group of all users in the data table except the current user. The item with highest rating value will be returned by the algorithm as the ultimate suggestions to the current user.

Yahoo recommender system in advertisement module employs machine learning technologies such as singular value decomposition (SVD) and latent semantic indexing (LSI) to provide recommended keywords. SVD and LSI are also used to recommend music and movies[2]. Like the most machine learning algorithms, SVD and LSI train the model based on the numeric data. Since a tremendous amount of data is gathered in a short period of time, the training time will increment exponentially and leads to a delay in response finally. The solution of Yahoo recommender system is partition the data set and develop new method for certain sets. The training time , as a result, increases log-linearly in practical situations[2].

4 CONCLUSION

Big data is highly involved in recommendation machines. Both Netflix and Yahoo utilize machine learning algorithms such as Restricted Boltzmann Machines, a form of Matrix Factorization, singular value decomposition, and latent semantic indexing. Yahoo also uses collaborative filtering algorithm for item recommendation. Netflix uses AB testing for validating a new recommender algorithm. In the future, more efficient and more elegant algorithms will be invented. Big data will lead to a more precise recommendation.

REFERENCES

[1] Xavier Amatriain. 2014. How does the Netflix movie recommendation algorithm work? Online. (12 2014). https://www.quora.com/How-does-the-Netflix-movie-recommendation-algorithm-work

[2] Dennis Decoste, David Gleich, Tejaswi Kasturi, Sathiya Keerthi, Omid Madani, Seung-Taek Park, David M. Pennock, Corey Porter, Sumit Sanghai, Farial Shahnaz, and Leonid Zhukov. 2005. Recommender Systems Research at Yahoo! Research Labs. Online. (1 2005). https://www.cs.purdue.edu/homes/dgleich/publications/decoste2005%20-%20yahoo%20recommender%20systems.pdf

[3] Geoffrey Fox. 2017. Big Data Applications and Analytics Case Study: e-Commerce and Life Style Infomatics: Recommender Systems I. Online. (9 2017). https://drive.google.com/file/d/0B6wqDMIyK2P7YkIwczVfQlJqVG8/view

[4] Geoffrey Fox. 2017. Big Data Applications and Analytics Case Study: e-Commerce and Life Style Infomatics: Recommender Systems II. Online. (9 2017). https://drive.google.com/file/d/0B6wqDMIyK2P7UVloVElaZ2FXcTg/view

5 BIBTEX ISSUES

6 ISSUES

Have you spellchecked the paper?

Are you using and, a, the properly?