International Journal of Multidisciplinary and Current Research ISSN: 2321-3124
Research Article Available at: http://ijmcr.com

Collaborative Filtering: Data Sparsity Challenges

Er.Meenakshi#* and Dr.Satpal^

#Computer Science Department, GRIMT, Radaur, India
^Computer Science Department, Baba Mastnath University, Haryana, India

Received 20 Sept 2018, Accepted 25 Nov 2018, Available online 27 Nov 2018, **Vol.6 (Nov/Dec 2018 issue)**

Abstract

Today internet is a place where the huge amount of data is stored, there is need to sift, which create a problem for the internet user, so recommend system solve the problem. A recommendation system is a system that helps a user found the products and content by forecast the user's rating of each item and showing them the items that they would rate highly. Recommendation systems are everywhere. With online shopping, customer has nearly infinite choices. No one has enough time to try every product for sale. Recommendation systems play an important role to solve the users search the products and content they care about. Recommendation system is a process of filtering the information that deal with information overloaded problems. Recommendation system is important for both user and service provider. It reduces the cost of transaction and selecting item in an online scenario it also improve the quality of decision making process. It is now an effective means for selling their product. So over emphasized of user is not good for recommendation system. To solve the problems of recommendation system like data sparsity we use one of best technique that is collaborative filtering technique.

Keywords: *Internet, web-services, longtail, recommendation system, collaborative filtering. etc.*

Introduction

Before discussing the collaborative filtering we go through with the recommendation system and working process of it. RS is the most popular application of data science today .it is used to predict the "rating" and "preference "that a user given to a item .today almost all company has applied to item or user or other .Amazon used to suggest the product to the user , You tube apply it which video play next, and Face book is used to like and to follow the people. We take recommendation system like this fig. it show how recommend function is rating of user data and new product with the help of RS technique.

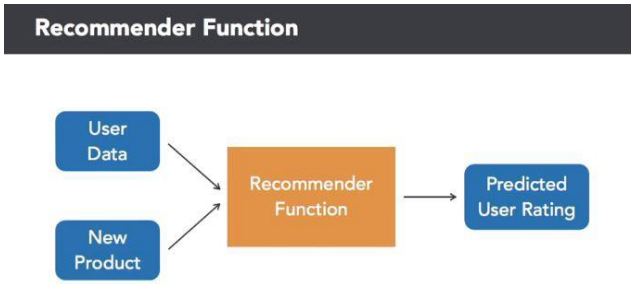

Fig.1 How work a recommend system function

Phases of recommendation system

a) Information collection phase

In this phase information is collects regarding the user so that we make a best user profile because it include the information about the user behavior and all the attribute of user like skill interest, learning style ,choice etc. An accurate model makes a best recommendation for other. Best information is based on feedback.

1) Explicit feedback: it is totally depend on the user rating. Some time it is accurate because user just rating of the product. Its not include other feature of product.

2) Implicit feedback: in this the system automatically cover the action of user like when user purchase product, preference of item and history of user and site searching time or also email content. So it not requires the effort of user .it is less accurate because it is deepened on implicit Preference.

3) Hybrid feedback: the benefit of both implicit and explicit feedback can be joined in a hybrid system to overcome the weakness and to get best result.

*Corresponding author's ORCID ID: 0000-0002-2580-5150
DOI: https://doi.org/10.14741/ijmcr/v.6.6.13

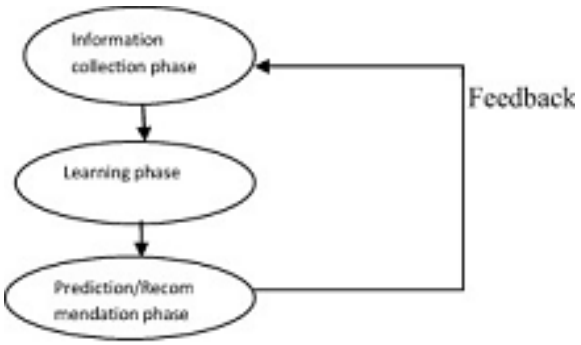

Fig. 2 Recommendation phases

b) Learning phase: in this we apply algorithm to find out and exploit the user information from first phase (information collection) of RS.

c) Prediction /Recommendation phase: in this what type of item user may prefer, recommendation system predict/recommend.

Recommendation filtering technique

Different type of recommendation system provides us to find out the best and useful recommendation to its individual user.

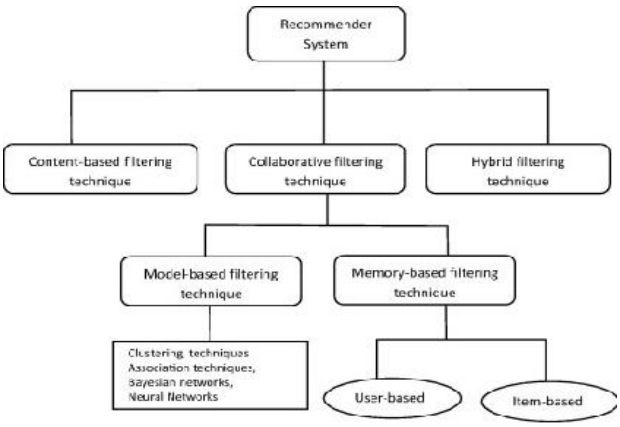

Fig 3 Recommendation system technique

1) Content-based filtering

This technique is domain depended which means main focus on analysis of the attribute of item in order to gerents prediction. It is mainly used where web page, news and publication are recommended .it is mainly used where user profile and past history of content of item is there. to find out the similarities of two item CBF used different type of mode. it could use **Vector Space Model** such **T**erm Frequency inverse documents (TF/IDF)or many more model are there .if the user profile change content based technique has possible to adjust its recommendation within very short period of time.

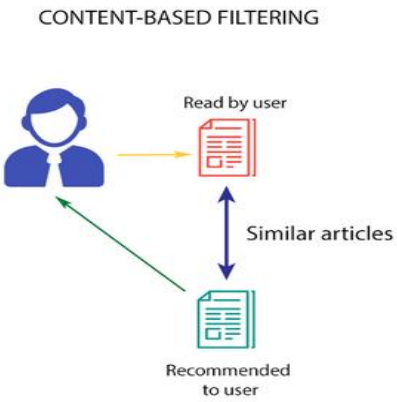

Fig 4 Content based filtering

Major overcome of CB its need to have an in-depth knowledge of the feature of the items in the profile. Content overspecialization is other main problem of CBF. User is restricted for recommendation those item which are already defined in their profile.

2) Collaborative filtering

It is a domain in depended prediction technique for contents which means that is based on assumption "similar user have similar preference ".CF is working on by making a database(user item matrix) of preference for item by the user .By examine their preference ,the RS can (i)active user preference predict for certain item .(ii)A ranked list of item will provide which is mostly like. CF removes the relationship between that item which has not similarities but is linked by default through the group of users accessing them recommendation that are created by CF can be either based on prediction and recommendation. Recommendation is the list of top N that liked by user most. And prediction is a numerical value, Rij which is predicted score of item j the user i. in CF a data base of user's preference is there.CF compare according to their preference. The preference will be gathered implicitly and explicitly.

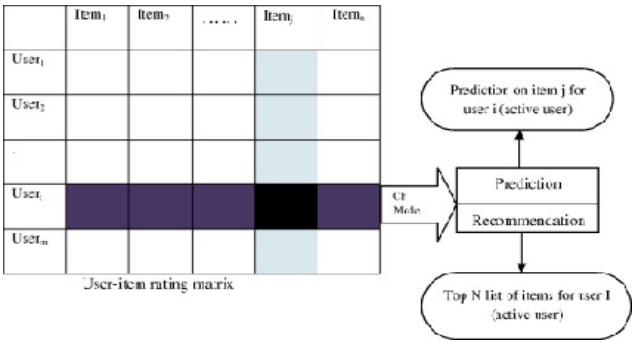

Fig 4 Collaborative filtering process

Collaborative technique has two type: Memory based (neighbor based) and Model based

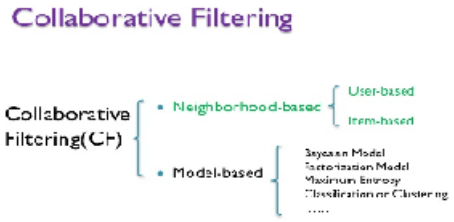

Fig 5 Collaborative Filtering

Memory-based Collaborative Filtering Algorithms

Memory-based collaborative filtering is two type item base or user based in the user based we can compute the rating similarity between the user on the same time.

These models engage statistical techniques to get a set of users called as *neighbors.* Different algorithm is used to find the prediction. The techniques, also known as *nearest-neighbor* or user-based collaborative filtering are more popular and widely used in today. Two most measure method are used one is Pearson correlation coefficient used in two linear variables and another is Cosine similarity used in the area of information retrieval. It use vector.

Model-based Collaborative Filtering Algorithms

During of **t**he process of this model we use mainly machine learning and data mining technique. It gives the recommendation on the bases of user rating. This is totally based o different *machine learning* algorithms such as **Bayesian network, clustering,** and **rule-based** approaches. This model is used the user – matrix technique to identify the relation between item and use this relation to compare the list of top-n recommendation. Some learning algorithms are used in this model like Association rule, Clustering and Artificial Neural network.

Problems in collaborative filtering :1) sparse data

A few item are rated by the user means less information found in result is called sparsity.

- However, it is difficult to find the inequality between customers because the sparsity problem is caused by the insufficient number of the transactions and feedback data, which confined the usability of the

Collaborative filtering

$$\begin{pmatrix} 1.0 & 0 & 5.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.0 & 0 & 0 & 0 & 0 & 11.0 & 0 \\ 0 & 0 & 0 & 0 & 9.0 & 0 & 0 & 0 \\ 0 & 0 & 6.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7.0 & 0 & 0 & 0 & 0 \\ 2.0 & 0 & 0 & 0 & 0 & 10.0 & 0 & 0 \\ 0 & 0 & 0 & 8.0 & 0 & 0 & 0 & 0 \\ 0 & 4.0 & 0 & 0 & 0 & 0 & 0 & 12.0 \end{pmatrix}$$

Fig 6: Sparse data (zero form data is sparse)

- Data sparsity gives a negative effect on the quality of recommendations that is given by traditional Collaborative Filtering algorithms.
- As such, the user simply provided with recommendations for the items most popular with a group of 'randomly' which is selected users, while among members of this group is disagreed.

Cold start problem: a situation where a recommend does not have right information item or user to make relevant prediction

Scalability: when the volume of data is increased it may be cause the bad recommendation because computation may grow with linearly with the no of user and item.

Synonymy: It is a tendency of very similar item that have different names or entries. Like some it is very difficult to find out the difference between closely related items like baby wear and baby cloth.

Data characteristics in Collaborative filtering

Data can be collected either explicitly or implicitly. Firstly we arrange the collected data after arrange we use in user matrix method for further process. In row represent user and column represent item. Matrix element is mentioned by a set of action that is specific user took in the context of a specific item. User does not access every item in many time in the store area. It is fact that sparse matrix cause the sparsitiy problems.

In collaborative filtering, explicitly means data with low sparsity and implicitly collected data with high sparsity. From Web logs(derived data) is heavily sparse data question is here, why would we want to apply collaborative filtering to Web logs? The answer is that collecting data in such manner requires no effort from the users and also, the users are not make to use any kind of specialized Web browsing software.

Experimental setting of data

Programmable Web is the largest web service. Recently this site play a important role in long tail service .so we crawled the data of ProgrammableWeb.com from June 2005 to 2018. Firstly we download the data files from this link:

- www.simflow.net/Team/baibing/DLTSR.rar
- All data files are in .mat files then we read .mat file using python **import scipy.io.**
- By using this method: **scipy.io.loadmat("File path")** we read the mat file.
- Then we print all the keywords from dictionary and By using these key word we get the data from dictionary.

Results: Get Keywords from Dictionary:

Fig 7 Result of keyword from main file

Script to create csv from list

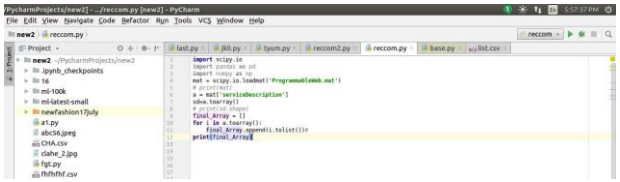

Fig 8 Csv file

- After then we convert into list and and create a csv file of data.

And data shape is (12477*21410) in this fig. in last we can seen no information this data is called our sparse data which we want to find in our objective.

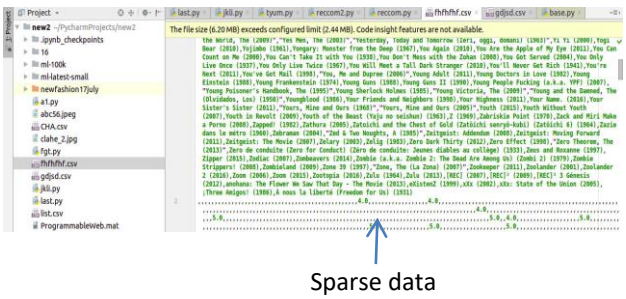

Fig 9 Result of csv file (sparse data)

Conclusion and future work

Because in today recommendation system is a new chance of retrieving personalized information. It's also to help to remove the problem of information over-loading future by using, which is very conman in retrieval system. This paper also discusses recommendation techniques and also cover the advantage and disadvantage of the technique, also discuss how to find out sparse data and its cons. We can remove the sparse data using SDAE(STACK DENOISING ATOENCODERS) is a deep neural network which is used to tackle the problem of unsatisfactory quality of description.

References

[1]. C. C. Aggarwal, J. L. Wolf, K.-L. Wu, and P. S. Yu. Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

[2]. P. Baldi, P. Frasconi, and P. Smyth. Modeling the Internet and the Web: Probabilistic Methods and Algorithms. Wiley, New York, 2003.

[3]. D. Billsus and M. J. Pazzani. Learning collaborative information filers. In Proceedings of the 15th International Conference on Machine Learning, 1998. https://towardsdatascience.com/building-a-recommendation-system-for-fragrance-

[4]. Hu R, Pu P. Potential acceptance issues of personality-ASED recommender systems. In: Proceedings of ACM conference on recommender systems (RecSys'09), New York City, NY, USA; October 2009. p. 22–5.

[5]. Pathak B, Garfinkel R, Gopal R, Venkatesan R, Yin F. Empirical analysis of the impact of recommender systems on sales. J Manage Inform Syst 2010;27(2):159–88.

[6]. Rashid AM, Albert I, Cosley D, Lam SK, McNe

[7]. https://towardsdatascience.com/building-a-recommendation-system-for-fragrance-5b00de3829d

[8]. https://towardsdatascience.com/building-a-recommendation-system-for-fragrance-5b00de3829d

[9]. https://www.researchgate.net/publication/320898302_Research_on_Long_Tail_Recommendation_Algorithm

[10]. https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-recommendation-engine-python/

[11]. https://medium.com/@libreai/a-glimpse-into-deep-learning-for-recommender-systems-d66ae0681775

[12]. https://machinelearningmastery.com/sparse-matrices-for-machine-learning/