A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems

Sri Hari Nallamala¹, Usha Rani Bajjuri², Sarvani Anandarao³, Dr. D. DurgaPrasad⁴ and Dr. Pragnaban Mishra⁵

¹Assistant Professor, CSE, VasireddyVenkatadri Institute of Technology, Guntur, A.P.
²Assistant Professor, CSE, Lakireddy Bali Reddy College of Engineering, Mylavaram,A.P.
³Assistant Professor, IT, Lakireddy Bali Reddy College of Engineering, Mylavaram, A.P.
⁴Professor, CSE, PSCMR College of Engineering & Technology, Vijayawada, A.P.
⁵Associate Professor, CSE, K L E F (Deemed to be University), Guntur, A.P.

E-mail: nallamala.srihari@gmail.com

Abstract. In the modern age and many prestigious applications use the recommendation method to play an important role. The system of recommendations collected apps, built a global village and provided enough information for development. This paper presents an overview of the approaches and techniques produced in the recommendation framework for collaborative filtering. Collaborative filtering, material and hybrid methods were the method of recommendation. In producing personalised recommendation the technique of collaborative filtering is particularly effective. There have been several algorithms over ten years of study, but no distinctions have been made between the various strategies. Indeed, there is not yet a widely agreed way to test a collaborative filtering algorithm. In this work we compare various literature techniques and review each one’s characteristics to emphasise their key strengths and weaknesses.

Keywords–Collaborative Filtering, Content Based Filtering, Data Mining, Machine Learning, Recommender Systems.

1. Introduction

Because of the vast amount of information on the internet, material of interest has been challenging for users. The main objective of research on recommended systems (RS) was this issue called knowledge overload. The main approaches to the construction of such a framework are content-based filtering (CB) and collaborative filtering (CF). In both methods, however, some writers point to constraints. The user is not shocked by the most commonly cited content-driven approach and is not filtered basing over subjective issues viz. consistency and design. And it is a constraint for collective methods to produce a broad variety of positive and negative estimates. These studies do not, however, provide comparable findings to support the idea that the overload issue in view of a certain context is more sufficient or not to be resolved.

Recommendation is an information philtre system that recommends products for users; philtres the data and recommends objects. Recommendation system. It is typically used for film lenses, book crossings, jesters and wiki lenses using mutual filtering to display information about goods and objects that may be of interest to consumers. For suggestion of all things, user concern in the past is seen and evaluated. The recommender framework uses descriptions of the profile, the behaviour, the expectations and the behaviours of the registered user for their entire group of users in order to present the recommendations. It depends a lot on the estimation of similarities.
1.1 Types of Recommender System

The Recommendation Framework can be categorised according to technique used into three categories, i) content-based, ii) collaborative and iii) hybrid.

Filtering based on content suggests elements for users which are practically identical to those that the user had previously chosen or wished. First the relationship between the object and its properties are established in the term of the matrix, and then machine similarity based on the features of the contrasted items using different mathematical functions selects the most related items to the target item. The most common feature of similarity is the Modified Coefficient of Cosine, Cosine or Pearson. A high level of prediction can result in strong similarity steps.

Collaborative filtration is a way to classify the like customers and suggest what popular customers like. This framework recommends products with previously identical tastes for the active user or for the target user. The similarity in the users' preferences are measured (or) determined on the basis of the similarity of the different users' rating background. The most popular method of CF is a similarity computation of a list of consumer expectations, which e-commerce websites typically gather from the rating of goods. This is why it's often called "people-to - people relationships." In the content-based recommendation, the equation used is the same methodology but concentrates on peer feedback. The most effective, trendy and broadly deployed technique in the recommendation systems is this filtering.

Hybrid filtration means combination of content-based and collaborative filtering that means classification of customers and giving suggestions by using both methods of filtering.

2. Review of Literature

Shivhare Hirdesh and al., 2015 [1] offers a combinatory approach to the creation of a recommending method (RS) through a combination of fuzzy-c-means clustering technique and also a genetic algorithm. The proposed recommendation system of the FCMGENSM offers better similarity and consistency measures than the current system recommended by the GENSM, but more time for computation of the planned RS than the current RS system is taken. Zan Wang al., 2014 [2] projected a hybrid model-based film recommendation method that uses an enhanced clustering of K-means coupled to a transformed user area by genetic algorithm (GA). Thus the original user space turns into much denser, trusted and is not searched in the entire user space but used for neighbourhood collection. By this method it is possible using conventional film recommendation systems to effectively estimate film ratings for new users.

Yahya H. Al Shamriet al., 2014 [3] suggested fluctuating weightings for the most frequent memory dependent CRS similarity steps. Fuzzy weighting can be interpreted as a learning method for evaluating user expectations. Fuzzy weighting is quick, efficient, and needs more room than genetic algorithm learning. Experimental findings suggest that fluctuating weighting increases the CRS efficiency independently of the fluid weighting component, where fuzzy-weighted likeness measures outperform the PPC, cover and medium absolute errors of their standard counterparts.

Darvishi-mirsolekarluet al., 2013 [4] presented a study of the Partnership Filtering Recommender Schemes. It is the most common and effective way for the target consumer to suggest the object. These users are involved and have the same interests. The biggest problem in collaborative filtering is scalability. In view of the incremental expansion of consumers and items, the time needed to locate the next neighbour of the target consumer or item increases.

Forsati and Mohammadi. al., 2013 [5] through the use of a mixture of collaborative and content driven methods to address the downside of each approach and greatly increase the range of recommendations in contrast with each individual approach, implementing a novel hybrid method recommendation. The new approach to clustering is applied based on GAs and generates a 2-layer graph. Using this algorithm on both graph layers, the comparison of web pages and users measured and the use of collaborative, content-based and hybrid approaches suggested recommendations.

Shiva Nadiet al., 2011 [6] suggested fuzzy recommendation framework based on ants co-operation (FARS). It operates in two levels. First, user behaviour is offline modelled and results for online suggestion are used for second process. The output is measured by means of log files. The findings are encouraging and give us extra comprehensive and usable suggestions.
Subhash K. Shinde et.al., 2011 [7] proposed a new updated C-fuzzy means clustering algorithm, which works in two phases, with users collecting opinions in the form of a username matrix. Subhash K. Shinde et.al, 2011 [8] recommendations for active users with similarity measures are created online in the second level. Kwoting Fang et.al. 2003 [9] suggests a method that advises the use of a fuzzy clustering system with two approaches. It indicates the same preference for the users in question and suggests what they want. The reminder was used to test the suggested accuracy with 2 (item by item) & (user by user) algorithms.

3. Modern And Latest

3.1 Content Based Filtering vs. Collaborative

Filtering Material was used for the first filtering approximations. Based on their contents, these systems choose which things to recommend. The client summary is therefore a reflection of the content of the client. This method of filtering is particularly successful when the documents having text are being retrieved where the keywords reflect each document. But there are many restrictions in these structures. Second, a computer should be able to evaluate the objects. This is hard for multimedia knowledge to be found, in which the computer perception is very different from the user perception of the material (colours, textures). Although the attributes assignment by an individual (annotated multimedia material) solves that issue, it is at least partially inadequate to sieve content to handle most of the current information. The lack of assessment of a product’s consistency is another big problem with contents-based filtering. For instance, if both articles use similar terms, a good article cannot be separated from a bad. Independent on each person's preferences, ideas, culture etc., the eminence of an object is an extremely subjective aspect that is difficult for a computer to evaluate. Finally there is no way to locate inaccurate items that are of interest to the customer, i.e., very nice items that are not actually connected to the client profile. Such problems are less susceptible to collaborative filtering systems, as they don't depend on the contents of the items, but on the views of other users. Here, the system suggests products with other users with related preferences (or) interests which have earned high ratings. The objects are actually rated by individuals in these techniques. The method therefore doesn’t have to evaluate content (therefore it extends to any item category together with non-annotated multimedia content), and also takes into account the consistency or subjective assessment of objects. The client profile is the collection of ratings given to various objects in collaborative filtering systems. This can be directly recorded, that is, by requesting the client (or) by examining their device dealings. So, the rating is usually expressed as a uniform value (who displays only the corresponding items), binary value or, most often, numerical value on a finite scale. In the table branded as the rating-matrix, the user ratings are kept. In order to create recommendations, this table is processed 2 types of algorithms, memory-based & also model-based, can be separated, based on how rating-matrix data are processed. The entire table is used by memory algorithms to measure their prediction. Generally, the user selects users (or items) identical to the active user using similarity steps. Then, from these neighbours' scores, the prediction is determined. (This is why the neighbour is also named). Although the process for getting neighbours focus on discovering similar users (or) items, majority part of these algorithms could be classified as user-based (or) item-base algorithms. Model-Based algorithms initially build a model to represent users' behaviours & consequently to envisage their ratings. Later, the model parameters are calculated offline by using the rating-matrix data. There are various techniques, often related to ML in literature using linear algebra methods SVD, factor analytics, PCAs, MMMF, clusters, & neural networks, as well as graphical diagrams. There are several literary approaches. Memory-based algorithms are usually simpler; however, they yield fairly exact results. However, there are significant issues with scalability as all the data have to be processed to determine a single prediction. These are not suitable for on-line systems that suggest products in real time with a huge number of users (or) objects. In addition, they are much more reliant than the model-based problems of some recommendation systems, which we emphasise.

Sarwar et al. 2001 [10]; Huang et al. 2004 [11] &B.B.Fernandes et al., 2017 [12] discusses about sparsity in the ranking matrix. Every user only rates a little portion of the products accessible in most recommended systems and then majority of the cells from the rating-matrix are vacant. In those situations, it is difficult to find similarities between different users or objects.

Schein et al. 2002, [13] suggests with regard to the previous problem, the problem is that the users who were recently adds to the system have trouble making recommendations. In such situations, users have not yet rated enough objects so that their preferences cannot be imagined by the recommending method. Other
systems address this problem by asking the consumer to rate a number of things first. These initial evaluations might therefore lead to prejudices within the framework. Finally, remember that the issue of cold start affects new products because it is not recommended until appropriate users have evaluated it.

Spam attacks, specifically by users fascinated in confusing the system for recommending an assured product may occur in recommended systems. A variety of techniques were studied in the past affecting both the neighbour-based & model-based algorithms.

4. Evaluation of Collaborative Filtering Algorithms

Within three key topics, their evaluation raises many difficulties and problems after nearly 2 decades of research was doing on collaborative filtering algorithms.

4.1 What in the recommendation framework do we evaluate?

There has still not been a consensus on the features to be measured. The most frequent patterns are to determine whether the algorithm is correct or correct. The three metric forms are as given below to calculate the algorithm’s efficiency.

**Prediction Accuracy**: It analyses the diversity between the rating expected by the system and the actual rating. The MAE (mean absolute error) is the most common type of metric. Additional related metrics like MSE, squared mean root error (RMSE), or absolute normalisation mean error are used too. After its submission to the Netflix Prize competition, RMSE has become very successful in modern times.

**Classification Accuracy**: This tests the system's distinction between good and poor. For instance, the precision, recall and ROC are known metrics of this kind. These metrics are ideal for finding good objects, in particular if users prefer binary. By comparison these metrics don’t determine the exact order of things on the recommendation list when users communicate their preferences in a digital range. You only test whether good items are recommended, regardless of which item is better. This is not always appropriate, of course. For instance, in the recommendation list, users often concentrate on the first items, so that we often want the best items in those positions. Such metrics are not the best choice in such situations.

**Rank Accuracy**: This tests the system's ability to organise the things the user had suggested. In certain cases, such calculations are too sensitive as they order the method to recommend the best outcomes (items), where in fact decent items are recommended rather than actually the best. Such metrics assess the association among the prediction & real classification (Pearson, Spearman's & Kendall's), the half-life utility and a regular distance calculation (NDPM). Such are the metrics of prediction and real classification.

The literature is also able to find measures which are not allied to the algorithm's accuracy. A widely used metric is the coverage that calculates the percentage of things that can be predicted by the device. Algorithm coverage, when you have to find all the good elements and not only suggest a few, is particularly important. Some studies have also measured the satisfaction of the customer with the suggestions in recent years. The thought is to test ideas such as the ability to suggest new and unforeseen objects, the way a user communicates (for example, how user reviews have been implemented), or how much a user can believe a recommendation by the system. During the increase of the number of users or objects, other measures assess the performance from the machine perspective or its scalability. The actions or the ability to generate successful suggestions in sparsity contexts of the algorithm with new users or objects was also assessed.

4.2 How should an evaluation can be performed?

The offline assessment in collaborative filtering studies is the most frequent approach. It consists of a dataset alienated into 2 sub-sets: training and assessment. The training subset is the data recognised by the algorithm, which are the data used by the algorithm to determine the recommendation (or) estimates. These are contrasted to the data in the assessment subset and this approach will commonly used, there are substantial variation between crams, the dataset used, the technique used for the division of the data into the training & assessment subsets and so forth. EachFilm, MovieLens, Jester and later days, Netflix, while some works are not accessible to the public with proprietary data. The synthetic data sets have also been used, created specifically for a specific property. There are some alternatives in terms of the methods used to construct the training subset. It requires creating a training collection to take each user's N ratings.
4.3 Which one is the best algorithm (or algorithms) in a particular context?
On the basis of the works contained in literature, it is difficult to address this question. Many either
determine the algorithm proposed using the best measures or methodologies, or they only research the best
circumstances for this particular algorithm. Since the suggestion algorithm is usually contrast to other
algorithms, a simple methodology, like the user-oriented one, is usually selected instead of more modern or
context-related algorithms for which the algorithm is built. Taking into account the decades of algorithms
built over the past decade, the lack of work comparing the various technology is shocking. The most
noteworthy are just the evaluation of a specific family of algorithms, their behaviour in different
conditions, or the comparison of many algorithms.

4.4 Tendencies-based collaborative filtering
Many, but not all, this type of filtering algorithms (collaborative) are focused on the relationship between
users and items presented so far. While many different methods for processing the data were used, the most
important thing was to find more or less secret relationships. The theory is that if two users have the same
pattern of ranking, the missing ratings possibly would coincide. But most methods need a lot of knowledge
to locate these relationships because it is a very complicated process. These similar type algorithms
therefore face severe problems with sparse datasets. We have created an algorithm, which looks at the
discrepancies between users or objects rather than searching for relationships between them. Users will rate
objects in various ways and their variations in views and preferences are related. That's not the only
explanation, however, users with similar penchants may rate items differently; some users appear to give
+ve ratings, and leaving –ve ratings for very bad items; however, other users save their top ratings for the
most advanced items and usually give negative ratings. Moreover, these differences are not limited to
specifically graded schemes. For example, the time that a consumer interacts with the system or the money
he or she will spend on buying our goods will affect the rating received by an implied system. It seems
obvious that the ratings of users will not only depend on the actual quality of the object, but also on many
factors. These variances had been previously found, but their integration into algorithms had a secondary
function, typically restricted to a pre-standardized rating to prevent the prediction from affecting them
negatively. That is to say, in the measurement of similarities between users or objects the differences were
disregarded. However, this is the basis of our algorithm. In our view, the consistency and functionality of a
specific item for the consumer is a more reliable predictor. Indeed, in previous work such variations were
also modelled using machine learning [14-21]. As stated that there are two hidden variables in their
probable model which distinguish user rate and actual user preferences. Recently, G. Potter [22] took a
psychological view, which attributes these differences to social or emotional users' causes, based on
behavioural economics research. For instance, if a user is looking at a good movie, and he is looking at one
better, the latter will be valued more highly than if the user had only looked at the second film. However,
our algorithm infers these distinctions more simply: the user patterns and objects. Instead of complicated
estimates, the tendencies can be easily estimated, and even less than the data necessary to find relationships
can be reliably estimated.

\[
\tau_u = \frac{\sum_{i \in L_u} (v_{ui} - \bar{v}_j)}{|L_u|},
\]

\[
\tau_j = \frac{\sum_{u \in U_j} (v_{ui} - \bar{v}_u)}{|U_j|}.
\]

The design of precise algorithms with very limited computation-time and memory requirements is a
very important function. The trend definition refers to whether a consumer uses the positive or negative
rating of products. This pattern should not be confused with the user average value. For instance, a
consumer who only rate good items is likely to have a high mean. However, each item could indeed have a
higher mean rating than the rating provided by the user. If many users also liked those items, the consumer
appears to rate the products negatively, while their mean is high in this case. In this case, the Thus, we
define the user's tendency as the average distinction between its ratings & the mean object.
5. Conclusion

This article provides a summary analysis of literature studies in accordance with the film recommendation system. In this analysis, we compared the behaviour of the algorithms in multiple contexts and not just under the most favourable conditions, with an optimistic comparison of different collaborative filtering algorithms. Similarly, using a clearly defined approach would make it easier in the future to compare other approaches to solve one of the major problems when evaluating collaborative filtering algorithms. It is interested that new algorithms and other datasets are used in future studies in various fields. It can also be especially important to apply the trend-based algorithm to contexts like knowledge recuperation on the web. This technique looks like ideal for huge quantities of data in this area. Another idea is to explore other solutions based on the similarities between users or products to conventional techniques. The good results of our experiments with the trend-based algorithm suggest the path to be taken in future works.

References

[1] Hirdesh Shivhare et al., 2015, Recommender system using fuzzy c-means clustering and genetic algorithm based weighted similarity measure, IEEE International Conference on Computer, Communication and Control (IC4-2015)

[2] Zan Wang, Xue Yu, Nan Feng, Zhenhua Wang, 2014, An Improved Collaborative Movie Recommendation System using Computational Intelligence Journal of Visual Languages & Computing 25(6).

[3] Al-Shamri, Mohammad Yahya H. & Al-Ashwal, Naji H., 2014, Fuzzy- Weighted Similarity Measures for Memory-based Collaborative Recommender System, Journal of Intelligent Learning Systems and Applications, (6) 1–10, Scientific Research.

[4] F.Darvishi-mirshekarlou, SH.Akbarpour, M.Feizi-Derakhshi, 2013, Reviewing Cluster Based Collaborative Filtering Approaches, International Journal of Computer Applications Technology and Research, ISSN: 2319–8656, Volume 2– Issue 6, 650 - 659.

[5] R. Forsati et al., 2013, A fuzzy co-clustering approach for hybrid recommender systems, International Journal of Hybrid Intelligent Systems 00, 1–11, DOI 10.3233/HIS-130166.

[6] Shiva Nadiet al., 2011, A Hybrid Recommender System for Dynamic Web Users, International Journal Multimedia and Image Processing (IJMIP), Volume 1, Issues 1/2, March/June.

[7] Subhash K. Shinde et al., 2011, Hybrid Personalized Recommender System Using Modified Fuzzy C-Means Clustering Algorithm, International Journal of Artificial Intelligence And Expert Systems (IJAE), Vol.1 Iss.4.

[8] Subhash K. Shinde et al., 2011, Hybrid Personalized Recommender System Using Fast K-medoids Clustering Algorithm, Journal of Advances In Information Technology, Vol. 2, No. 3, doi:10.4304/jait.2.3.152-158.

[9] Fang K, Liu C-Y, 2003, Recommendation system using fuzzy c-means clustering. In: Book of information technology and organizations: trends, issues, challenges and solutions, vol 1. Idea group publishing, p 137–139.

[10] Sarwar B, Karypis G, Konstan J A, Riedl J T, 2001, Item-based collaborative filtering recommendation algorithms, 10th International World Wide Web Conference. 285–295.

[11] Huang Z, Chen H, Zeng D, 2004, Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Trans. Information Systems. 22(1): 116–142.

[12] Bruno BandeiraFernandes, Juarez AP Sacenti, and Roberto Willrich, 2017, Using Implicit Feedback for Neighbors Selection: Alleviating the Sparsity Problem in Collaborative Recommendation Systems.In Proceedings of WEBMEDIA. ACM, 341–348.

[13] Andrew I. Schein, AlexandrinPopescul, Lyle H. Ungar, and David M. Pennock, 2002, "Methods and Metrics for Cold-Start Recommendations".

[14] S H Nallamala, et al., 2018, “An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records”, (IJET) (UAE), Vol. 7, No 2.7, SI 7, ISSN: 2227 – 524X.

[15] S H Nallamala, et al., 2018, “A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment”, (IJET) (UAE), Vol. 7, No 2.7, SI 7, ISSN: 2227 – 524X.

[16] S H Nallamala, et al., 2019, IAATCSE, “Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems$, ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2.

[17] S H Nallamala, et al., 2019, IRTE “Breast Cancer Detection using Machine Learning Way”, ISSN: 2277-3878, Vol-8, Issue-2S3.

[18] S H Nallamala, et al., 2019, “Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in
the Breast Cancer Treatment”, IJSTR, ISSN: 2277-8616, Vol. 8, Issue 11.

[19] S H Nallamala, et al., 2020, A Review on “Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management”, TEST Engineering & Management, ISSN: 0193-4120, Vol.83, No.3.

[20] S C Polavarapu et al, 2018, “RFID based automatic tollgate collection”, IJET (UAE), Vol. 7, No 2.1, 2018, ISSN: 2227 – 524X.

[21] KollaBhanu Prakash, et al., 2020, “Accurate Hand Gesture Recognition using CNN and RNN Approaches”, International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3216 – 3222.

[22] Potter, G., 2008, “Putting the collaborator back into collaborative filtering”, the 2nd-Netflix-KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition. ACM Press, New York.