# (E-AggNN) Implementation of Effective and Efficient Algorithm Collaborative Filtering and Content Based similarity Search on Recommender System

Ms. Sonali Patil[1], Ms. Swati Joshi[2]

*M.E. in Computer Science & Engineering Pune University, Pune [1]*
*Pursuing M.E in Computer Science & Engineering Department BSIOTR, Pune University[2]*

## ABSTRACT

*In this work, we used sum variant of FANN, and other for the max variant. Here investigates a novel technique to solve a "cold start" problem. This problem occurs when new item added in the list means there is null score assign to that particular item our aim is to provide solution that is manufacturer and customer oriented. Our result generates recommendations as per the manufacture and Customer oriented using collaborative filtering and ranking based mechanism.For better and fast performance nowadays all the search engines are enhanced with recommendation system. Most of the time Recommendation process depends on the past entities, so collaborative filtering technique uses this past data for recommending process. Collaborative filtering alone is not able to provide fine grained recommendation for all the users' queries. Hybrid Recommendation system is required which uses both past and present data which are aggregated for the high dimensional parameters.*

**Keyword :** *Collaborativefiltering,HybridRecommendationsystem,Aggregation*

## 1. Introduction

In this work, we used sum variant of FANN, and other for the max variant. Here investigates a novel technique to solve a "cold start" problem. This problem occurs when new item added in the list means there is null score assign to that particular item our aim is to provide solution that is manufacturer and customer oriented. Our result generates recommendations as per the manufacture and Customer oriented using collaborative filtering and ranking based mechanism.

In late years, the utilization of various positioning criteria has been investigated, in which the last rankings of objects are got by joining the individual rankings as indicated by some aggregation function (for example, min, max or sum). The ranking criteria utilized as a part of utilizations of this type of similitude pursuit have varied incredibly from region to region. In media applications, ranking criteria have been characterized as far as a few separation capacities processed over various arrangements of discriminative elements, for example, shading highlights and surface elements. In keyword based search, criteria have been characterized as for every individual keyword utilized as a part of the query; while in subspace comparability look, ranking criteria have been characterized on every individual measurement of the focused on subspace measurements.

So as to boost benefit, the engineers might wish to decide potential destinations for the development of new outlets that would minimize the aggregate (or maximum) separation to target sets of clients. On the off chance that there are more components to be viewed as, for example, the area value, then the k most suitable locales might be recovered from the database as hopefuls. Total similitude look strategies could likewise advantage applications that make utilization of pertinence criticism, a type of query refinement in which the client is given the chance to choose a few objects from a past query result to serve as the premise for an ensuing query. Razente et al. proposed total closeness inquiries as a pertinence criticism system for content based picture recovery. Besides, as pointed out in total similitude inquiries can be used in bunching and anomaly discovery. For instance, the nature of an answer for bunching or anomaly discovery can be assessed by the aggregate (or most extreme) separation between the focuses and their closest group centroid. Conglomeration of similitude is an essential operation in recommender framework: in substance based proposal, collections are by and large utilized as a part of deciding thing to-thing closeness, while in community sifting; total assumes a part in the determination of client closeness.

As a situation delineating the requirement for total similitude search, we could envision a spatial database comprising of an arrangement of potential areas to build a retail outlet, inside of which land engineers can posture queries on accumulations of gatherings of potential clients at various areas.

Similarity search is used as ranking criteria in applications of this form of have differed greatly from area to area. In multimedia applications [1], ranking criteria have been defined in terms of several distance functions computed over different sets of discriminative features, such as colour features and texture features. In keyword based search [2], criteria have been defined with respect to each individual keyword used in the search; while in subspace similarity search [3], ranking criteria have been defined on each individual dimension of the targeted subspace dimensions Given a group of query objects Q, aggregate similarity or aggregate nearest neighbour (AggNN) search aims to retrieve the k objects from the database S that are most highly ranked with respect to Q, where the ranking criterion similarity measure) is defined as an aggregation (usually sum or max) of distances between the retrieved objects and every object in Q. As a scenario illustrating the need for aggregate similarity search, we could imagine a spatial database consisting of a set of potential locations to construct a retail outlet, within which real estate developers can pose queries on collections of groups of potential customers at different locations. In order to maximize profit, the developers may wish to determine potential sites for the construction of new outlets that would minimize the total (or maximum) distance to target sets of customers. If there are more factors to be considered, such as the land price, then the k most suitable sites may be retrieved from the database as candidates. Aggregate similarity search techniques could also benefit applications that make use of relevance feedback, a form of query refinement in which the user is given the opportunity to select several objects from a previous query result to serve as the basis for a subsequent query. In [4], Razente et al. proposed aggregate similarity queries as a relevance feedback mechanism for content based image retrieval. [6], aggregate similarity queries can be utilized in clustering and outlier detection. For example, the quality of a solution for clustering or outlier detection can be evaluated by the total (or maximum) distance between the points and their nearest cluster centroid [6]. Aggregation of similarities is a fundamental operation in recommender systems [7]: in content-based recommendation, aggregations are generally used in determining item-to-item similarity, whereas in collaborative filtering, aggregation plays a role in the determination of user similarity.

Because of its significance and simplification, AggNN has gotten a lot of consideration in the writing. In particular, it has been tended to in the settings of street systems, Euclidean vector spaces and metric spaces. All in all, in any case, AggNN routines tend to support just those protests that are like all query objects in Q, which by and by might extraordinarily restrain its execution when the qualities of the objects of Q differ significantly. For instance, in our land improvement situation, rather than endeavouring to draw in all focused on clients, a designer might be content with pulling in a huge extent of the potential clients, by deciding areas minimizing the aggregate (or maximum) separations to at any rate this extent of potential clients. Another conceivable situation is that of substance based picture recovery, in which pictures connected with a given idea are looked for. On the off chance that for instance the client wishes to acquire pictures of a car, he or she may give various such pictures as query cases. In any case, it is improbable to accept that any one database picture could coordinate numerous query examples at the same time.

We outline our calculations by receiving a multi-step look procedure together with tests for conceivable early end taking into account a measure of characteristic dimensionality.

The main contributions of this paper are:

- Two approximation algorithms for the FANN issue, one for the addition variant and the other for the max variant.
- A hypothetical investigation of our strategies, indicating conditions under which a careful result can be ensured. We likewise indicate conditions under which rough result can be ensured for a variable separation estimate proportion.
- A broad exploratory assessment, appearing that our calculations can create query results with great precision.

## I. . RELATED WORK

Aggregate similarity search was first studied by Papadias et al. [1] for the sum variant of the AggNN problem in Euclidean space. Their solution was soon improved upon by using the R-tree index together with branch-and-bound pruning of the search space. For general metric spaces, a similar approach was adopted by Razente et al. and Michael E. Houle, Xiguo Ma, and Vincent Oria [2] using distance-based indexes such as the M-tree. To improve the efficiency of the search, several approximation methods were also proposed.

Gediminas Adomavicius, and YoungOk Kwon[2] suggest Recommender systems are becoming increasingly important to individual users and businesses for providing personalized recommendations. However, while the majority of algorithms proposed in recommender systems literature have focused on improving recommendation accuracy (as exemplified by the recent Netflix Prize competition), other important aspects of recommendation quality, such as the diversity of recommendations,

## I. LITERATURE SURVEY

The performance of these algorithms degrades quickly as dimensionality increases. To get around the curse-of dimensionality problem of the MBM method, approximation methods have been proposed, but only for max Ann queries in the Euclidean space.The basic idea is to find the centre of the minimum enclosing ball (MEB) of the Q, and then simply return the nearest neighbour of this center from P . Li et al. Gediminas Adomavicius, and YoungOk Kwon[2] suggest Recommender systems are becoming increasingly important to individual users and businesses for providing personalized recommendations. However, while the majority of algorithms proposed in recommender systems literature have focused on improving recommendation accuracy (as exemplified by the recent Netflix Prize competition), other important aspects of recommendation quality, such as the diversity of recommendations.

K. Bradley and B. Smyth [7] proposed a new class of diversity-preserving algorithm capable of addressing this without compromising similarity or efficiency.

Saúl Vargas and Pablo Castells [9] presented a formal framework for the definition of novelty and diversity metrics that unifies and generalizes several state of the art metrics. Then identify three essential ground concepts at the roots of novelty and diversity: choice, discovery and relevance, upon which the framework is built. Item rank and relevance are introduced through a probabilistic recommendation browsing model, building upon the same three basic concepts.

It is more appropriate to view the problem of generating recommendations as a sequential optimization problem and, consequently, that Markov decision processes (MDPs) provide a more appropriate model for recommender systems.

[10] Showed that this simple method gives a √2-approximate answer to the max Ann query in any dimensions, and its query cost is essentially the same as one standard NN query also proposed a few heuristic for approximating Ann queries, but with no provable approximation ratios.

## II. PROPOSED SYSTEM

The main contributions of this paper are:

- Two approximation algorithms for the FANN issue, one for the addition variant and the other for the max variant.
- A hypothetical investigation of our strategies, indicating conditions under which a careful result can be ensured. We likewise indicate conditions under which rough result can be ensured for a variable separation estimate proportion.

- A broad exploratory assessment, appearing that our calculations can create query results with great precision.

- Therefore, using the item ratings and user profiles, recommender system has been proposed to provide diverse recommendations i.e. highly personalized items with only a minimal accuracy loss as well as suggest a sequence of items instead of a single recommendation to improve the quality of recommendations and use consumer-oriented or manufacturer oriented ranking mechanisms so both consumer and manufacturer will get benefit from recommendations.

A. Getting Dataset details

User can see the information for dataset with different categories are provided. User can select any item from the list.

B. Rating generation

Ratings can be generated explicitly by the user or the system can predict implicitly. So for explicit ratings, user can give rating to any travels. This information is stored in rating table with user id, travel id and rating. Ratings are made on a 5- star scale. 1-star indicates very poor rating. 2-star indicates poor rating. 3-star indicates ok rating. 4-star indicates good rating. 5-star indicates very good rating.

C. Similarity computation module

This is a first module in item based collaborative filtering approach. When user requests for any travel, that travel is considered as target item. Now system has to find similarity between target travel and other travels for that active user has rated.

The proposed system has used adjusted cosine similarity method.

D. Prediction computation module

- Using the similarity computation results in previous step, system selects N most similar items. Here N=10 has been taken.

For prediction calculation, weighted sum approach is used.

To compare the efficiency of our algorithm with previous approaches like content-based recommendation algorithms were implemented and applied in this study.

## III. SYSTEM DESIGN

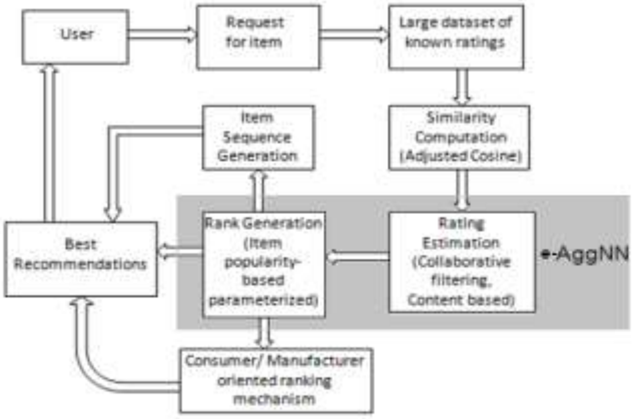

User
Request for item
Large dataset of known ratings
Item Sequence Generation
Similarity Computation (Adjusted Cosine)
Best Recommendations
Rank Generation (Item popularity-based parameterized)
Rating Estimation (Collaborative filtering, Content based)
e-AggNN
Consumer/ Manufacturer oriented ranking mechanism

Fig 1: System Architecture

The proposed recommendation ranking approaches were tested with several travels rating datasets, including Travel (data file available at yatra.com), travels (data file available at makemytrip.com), and Yahoo! Travels (individual ratings collected from travel pages at travels.yahoo.com). We pre-processed each dataset to include users and travels with significant rating history, which makes it possible to have sufficient number of highly-predicted items for recommendations to each user (in the test data).

The data collection phase where three main recommender algorithms (collaborative, content-based, and ECSN) were applied. The next three columns (new items, new users, and existing users) present the updated situation of MyExpert social network for each week of experiments based on the number of users and items.

For each user of MyExpert, the item categories get ordered based on PS value and stored in a stack data structure (ItemStack) such that the category with the biggest PS is accessible at the top of the stack. To produce the recommendation list for user u, the highest ranked category at the top of stack is moved to TopItemCat using POP (ItemStack).

Then, the new submitted items in MyExpert (100 items per week) are searched to find items using category ID of TopItemCat. The recommendation list that is supposed to be suggested to each user u includes the 10 most relevant items. To have more items with highest PS value in this list, the PrioritizedCount array has been considered to identify the number of items for each top scored item category:

PrioritizedCount={3,2,2,1,1,1,1,1,1,1}

Based on this identified priority, the highest scored category can have up to 3 items while the two next highest ones come with at most 2 items in recommendation list. The others have the same value of one item. In the body of while loop, the ordered recommendation list (TopItemsList) is generated for each user u € U

To compare the efficiency of our algorithm with previous approaches like content-based recommendation algorithms were implemented and applied in this study.

In situations where content-based methods suffer from the cold-start problem when new items or new users are involved , the our recommender algorithm utilized social networking features to solve this issue. Collected feedback for experiments and analyzed to show how the proposed recommender algorithm in present research mitigated this issue.

### A. *Experimental Setup*

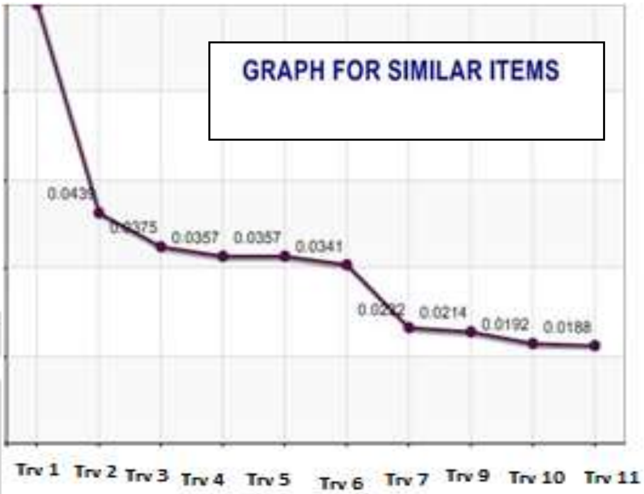
The system is built using .NET framework (3.5) on Windows platform. The Visual Studio 2008 is used as a development tool. The system doesn't require any specific hardware to run, any standard machine is capable of running the application

## IV. RESULT ANALYSIS

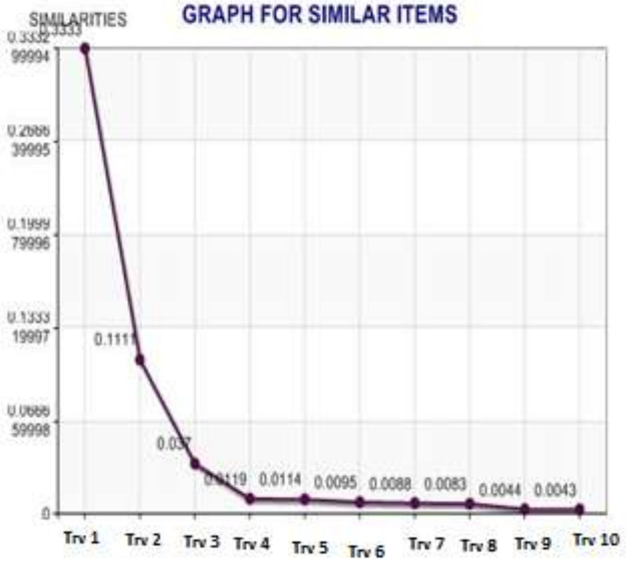
In database creation module, information related to user, items and ratings has been stored in different tables. Thus system can retrieve the data properly from database and also get travel ratings explicitly from the users. Item similarity computation module in collaborative filtering technique has been implemented with travel's known ratings as an input. Similarities are obtained between target travel and other travels. It uses most similar items as an input and system predicted rating of travel is obtained as an output. Output value lies between 1 to 5 as per the 5-star scale of rating. Graphs show the system predicted ratings with respective travel items.

Finally prediction result shows the accuracy with respect to rating .higher rating items have the maximum chances to select in ranking list.

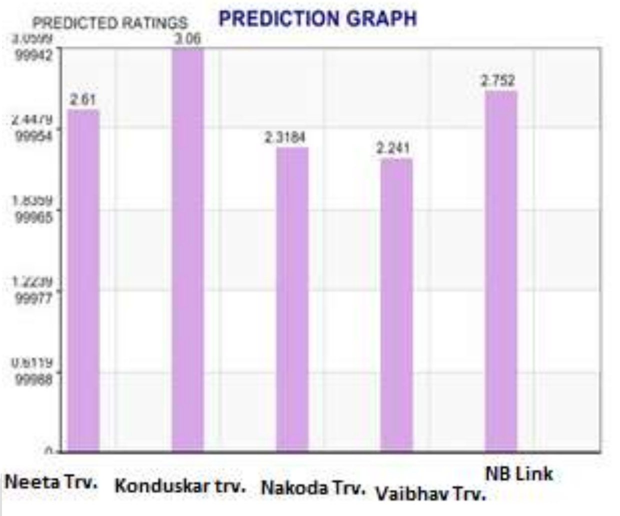
```
For each u € U
{
(1)Generating the ordered stack of item categories (ItemStack) based on PS value computed by Definitions 2 and 3.
(2) SIC ← 0// Initializing the Selected Items Count (SIC) by 0
(3) SC ← 0  // Initializing the selected category (SC) by 0
(4) While(SelectedItems< 10)
(5)   {
(6)TopItemCat←POP (ItemStack),
(7)TopItemList←FindNewItems(TopItemCat, PrioritizedCount(SC))
(8) SIC −= count(TopItemsList)
(9)  TopItemsList to RecommendationList
(10) }
}
```



Graph1: Similariy item set Retrieved by e-AggNN (Dataset 1)



Graph2 .Similar items graph for the travel dataset-2



Graph3. System predicted ratings graph with respective items

## 4. CONCLUSIONS

Recommender systems provide valuable suggestions to users with the help of user rating databases. Therefore user rating database creation is one important step in the proposed system. Database creation module is implemented by accepting explicit ratings from the different users. Similarity computation module computes similarity between target item and other items while prediction estimation module predicts the rating for the target item. Accuracy of predictions can be measured with statistical accuracy metrics.

Further, the system proposed recommendation technique which is based on content i.e. user preferences and item profiles. Item popularity based parameterized ranking technique will ranks the items such that recommendation accuracy will be maintained and the diversity will be increased. Quality of recommendations will be improved using consumer/ manufacturer oriented ranking and item sequence generation techniques

## 5. ACKNOWLEDGEMENT

I take this special opportunity to express my sincere gratitude towards professor and all the people who supported me during my entire project work. I would like to express my gratitude to my guide and also the project coordinator Prof. Sonali A.Patil for providing special guidance. I would also like to thank our HOD Prof. G. M. Bhandari who always has enough time to solve student's problems at any time. Finally thanks to all teachers who are always supportive at us.

## 6. REFERENCES

[1.] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. Journal of ACM,45(6):891–923, 1998.

[2.] S. Berchtold, C. B ohm, D. A. Keim, and H. P. Kriegel. A cost model for nearest neighbor search in high-dimensional data space. In PODS , 1997.

[3.] M. Berg, M. Kreveld, M. Overmars, and O. Schwarzkopf. Computational geometry: algorithms and applications Springer, 1997.

[4.] SeokKee Lee a,1, Yoon Ho Cho b,*, SoungHie Kim a,2" Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations" Information Sciences 180 (2010) 2142–2155

[5.] C.C. Aggarwal, P.S. Yu, Data mining technique for personalization, Data Engineering 23 (1) (2000).

[6.] B. Sarwar, G. Karypis, J.A. Konstan, J. Riedl, Application of dimensionality reduction in recommender system – a case study, in: Proceedings of the ACM WebKDD Workshop, 2000.

[7.] K. Bradley and B. Smyth, "Improving Recommendation Diversity," Proc. of the 12th Irish Conf. on Artificial Intelligence and Cognitive Science, 2001.

[8.] S.T. Park and D.M. Pennock, "Applying collaborative filtering techniques to travel search for better ranking and browsing," Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 550-559, 2007.

[9.] Saúl Vargas and Pablo Castells, "Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems", RecSys'11, October 23–27, 2011.

[10.] H. Li, H. Lu, B. Huang, and Z. Huang. Two ellipse-based pruning methods for group nearest neighbor queries. In GIS , 2005.

## BIOGRAPHIES

| | |
|---|---|
| | Ms.Swati Shripad Joshi.<br>*Pursuing M.E in Computer Science & Engineering Department BSIOTR, Pune University* |
| | Prof. Sonali Appasaheb Patil<br>Mtech CSE, Phd pursuing from BSAU, Chennai.<br>Asst.prof.Department of Computer Engineering<br>JSPM'S BSIOTR,WAGHOLI,PUNE |