ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 2, pp: (191-198), Month: April - June 2016, Available at: www.researchpublish.com

Recommendation System Using Collaborative Filtering and Content Analysis in Web Usage Mining

Prof. Ujwal U.J [1], Prof. Dr. Antony P.J [2], Vijay D R[3]

KVG College of Engineering, Sullia, 574327, India

Abstract: **Recommender frameworks and systems are presently famous both industrially and in the research group, where numerous methodologies have been proposed for giving recommendations. This paper talks about the collaborative filtering and conventional method for measuring their execution against client rating information sets. The paper will then proceed onward to talk about building dependable, precise information sets; understanding recommender frameworks in the more extensive connection of client data needs and undertaking support and the collaboration amongst clients and recommender frameworks. A cooperative separating methodology is going to channel and perceive the comparable administrations under same group and took after by those assessments proposals are made. Pre-calculation and truncation is crucial to conveying cooperative separating by and by, as it places an upper bound on the quantity of things which must be considered to create a suggestion and wipes out the inquiry time expense of similarity calculation. It accompanies the little cost of diminishing the quantity of things for which expectations can be created. This paper proposes a joined collaborative filtering technique which utilizes cosine closeness strategy to register the likenesses which results in better recommendations.**

Keywords: **Recommendation Systems, Collaborative Filtering, Content Analysis, Web Usage Mining, Rating Systems, User Review Systems, Movie recommendation system.**

I. INTRODUCTION

Recommendation systems have turned into an essential exploration range subsequent to the presence of the main papers on shared separating in the mid-1990s. There has been much work done both in the business and the educated collaborative on growing new ways to deal with recommender frameworks in the course of the most recent decade. The enthusiasm for this zone still stays high since it constitutes an issue rich exploration territory and due to the plenitude of down to earth applications that help clients to manage data over-burden and give customized suggestions, substance, and administrations to them. Case of such applications incorporate suggesting books, CDs, and different items at Amazon.com, films by MovieLens , and news at VERSIFI Technologies. Also, a portion of the sellers have fused proposal abilities into their trade servers . Be that as it may, notwithstanding these advances, the present era of recommender frameworks still requires further changes to make proposal strategies more viable and material to a significantly more extensive scope of genuine applications, including suggesting get-aways, certain sorts of monetary administrations to financial specialists, and items to buy in a store made by a "shrewd" shopping basket. These changes incorporate better techniques for speaking to client conduct and the data about the things to be prescribed, more propelled proposal demonstrating strategies, consolidation of different relevant data into the suggestion procedure, usage of multcriteria evaluations, improvement of not so much meddling but rather more adaptable suggestion strategies that likewise depend on the measures that all the more adequately decide execution of recommender systems[1].

Proposal frameworks in e-business [2] turn out to be progressively critical because of the extensive number of decisions shoppers face. All in all, suggestion frameworks first take an arrangement of information which could be client profiles, an arrangement of thing evaluations, and so on, recognize similarities among the information, and pass the comparable

sets for forecast count. Among the procedures utilized as a part of building proposal frameworks, collaborative filtering is a standout amongst the most encouraging methodologies. LikeMinds, a financially accessible suggestion framework makes utilization of cooperative filtering, and our methodology is propelled by that framework.

The measure of data in Internet is developing so quick that it overpowers web clients. This is additionally a noteworthy issue for e-trade in light of the fact that online customers can't just investigate and think about each conceivable item. To reduce the issue, suggestion frameworks are acquainted with e-trade. Among all the current strategies, Collaborative Filtering is a standout amongst the most encouraging ways to deal with assemble suggestion frameworks. Collaborative Filtering gathers client's inclinations for things, searches for an arrangement of neighbors having comparable inclinations, and induces rating on a specific thing in light of the data gathered from neighbors. With the anticipated rating, the framework will prescribe items which have high anticipated appraisals to clients. Suggestion framework is especially valuable when online traders are offering vast number of items that are in comparative areas, for example, Music CDs, Movies, and so forth. The proposal depends on the appraisals on items that clients have already evaluated. A helpful proposal framework is productive since practically speaking the framework handles a huge number of evaluations and forecast is figured in genuine times. A helpful suggestion framework likewise creates precise forecast.

II. RELATEDWORK

Collaborative filtering recommender framework is a very much considered point. Such frameworks attempt to anticipate the evaluations of specific things for an objective client taking into account the things already appraised by different clients [3]. Utilizing cooperative separating method, a suggestion framework tries to recognize neighbors who have the minimum rating difference to the objective clients' evaluations. The neighbors can be deciphered as having comparative taste to the objective clients', so the suggestion is figured taking into account data from those neighbors as it were. LikeMinds [4] is a standout amongst the most well known applications utilizing shared separating. LikeMinds basically attempt to recognize neighbors in view of the rating distinction of every evaluating sets between the objective clients and the applicants. LikeMinds take an objective client and an arrangement of hopeful clients as info, and registers the closeness taking into account the rating contrasts between the objective client and the applicant clients. The hopeful with most astounding closeness score are considered as a tutor to the objective client. The objective client's forecast is then processed in light of the guide's evaluating.

A Collaborative Filtering Recommendation Algorithm [5] in view of User Clustering and Item Clustering was executed to take care of the issues of versatility and sparsity in the collaborative filtering, and it proposed a customized suggestion approach joins the client grouping innovation and thing bunching innovation. Clients are grouped in view of client's appraisals on things, and every clients bunch has a bunch focus. An arrangement of comparability measures are introduced and a metric of significance between two vectors. At the point when the estimations of these vectors are connected with a client's model then the closeness is called client based similarity, while when they are connected with a thing's model then it is called thing based likeness. The likeness measure can be successfully used to adjust the evaluations essentialness in an expectation calculation and in this manner to enhance exactness.

An Efficient Recommender System utilizing Collaborative Filtering Methods with K-distinctness Approach [6] was executed utilizing shared separating component with k-detachability approach for online advertising. The framework take after the synergistic recommender strategy in which a client rating is conglomeration of different characters utilizing network yet dataset turns out to be extremely uproarious and hard to isolated. In this way, the K-Separability approach broadens direct detachability of information groups into k>2 fragments on the separating hyperplane. It can be executed by single layer or 2-layer perceptron. K-Separability can reveal complex factual conditions i.e. positive or negative. At long last, the framework doesn't have to channel the neighbor-hood of the objective channel as different frameworks. All neighbor-hoods are viewed as and to a great degree valuable if there should arise an occurrence of inadequate dataset.

Another half and half colonialist focused calculation on information bunching was one of the unsupervised learning branches where an arrangement of examples, more often than not vectors in a multi-dimensional space, are gathered into groups in a manner that examples in the same group are comparable in some sense and examples in various bunches are disparate in the same sense. Group investigation is a troublesome issue because of an assortment of methods for measuring the likeness and difference ideas, which don't have a widespread definition[7].

III. PROPOSED SYSTEM

The system utilizes the center calculations for collaborative oriented filtering and customary method for measuring their

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 2, pp: (191-198), Month: April - June 2016, Available at: www.researchpublish.com

execution against client rating information sets. The framework will then proceed onward to talk about building dependable, precise information sets; understanding recommender frameworks in the more extensive setting of client data needs and assignment support and the collaboration amongst clients and recommender frameworks.

A. Collaborative Filtering:

Collaborative filtering (CF) is a prominent suggestion calculation that constructs its forecasts and proposals with respect to the evaluations or conduct of different clients in the framework. The crucial presumption behind this technique is that other clients' conclusions can be chosen and collected so as to give a sensible forecast of the dynamic client's inclination. Collaborative oriented separating, likewise alluded to as social filtering, channels data by utilizing the proposals of other individuals. It depends on the possibility that individuals who concurred in their assessment of specific things in the past are liable to concur again later on. A man who needs to see a motion picture for instance, may request suggestions from companions. The suggestions of a few companions who have comparative interests are trusted more than proposals from others. This data is utilized as a part of the choice on which motion picture to see.

B. Proposed Combined Collaborative Filtering:

Algorithm: Combined CF algorithm

Step 1: Calculate similarity between each two items and construct item similarity matrix, and calculate similarity between each two users and construct user similarity matrix.

Step 2: For the active item, k items that have highest similarity are selected to represent the neighborhood of the active item. Meantime, For the active user, k users that have highest similarity are selected to represent the neighborhood of the active user.

Step 3: The missing rate is predicted by a weighted sum of the combination of item neighbor's ratings and of user neighbor's ratings.

Proposed combined CF model integrates item-oriented CF algorithm and user-oriented CF algorithm into an unified framework. Both item similarity matrix and user similarity matrix contribute to the final rating prediction. Two neighbourhood relationships, user-user relationship and item-item relationship, are retained in the same time. This integration of item-oriented CF algorithm and user-oriented CF algorithm not only reward a improved prediction accuracy, but also bring greater robustness of the recommendation system to sparseness problem. The above algorithm summarizes the whole algorithm for Combined CF model.

Proposed combined CF display incorporates thing focused CF calculation and client arranged CF calculation into a bound together structure. Both thing similarity network and client comparability lattice add to the last appraising expectation. Two neighborhood connections, client relationship and thing relationship, are held in the same time. This incorporation of thing focused CF calculation and client situated CF calculation compensate an enhanced forecast precision, as well as bring more noteworthy vigor of the suggestion framework to inadequacy issue. The above calculation condenses the entire calculation for Combined CF show. In step 1, the thing comparability weight between things can be figured through equation 1 and client similarity weight between clients can be ascertained through equation 3. Item similarity matrix and user similarity matrix can be constructed based on each two items similarity weight and each two users similarity weight respectively. In step 2, we assume the rating of client an on thing m is missing, so thing m is the dynamic thing and client an is the dynamic client. K things that have the most astounding similarity to the dynamic thing m are situated from the thing closeness lattice. Interim, k clients that have the most astounding likeness to the dynamic client an are likewise situated from the client similarity grid. In step 3, the rating of client an on thing m can be anticipated utilizing a weight total of k client weighted normal and k thing weighted normal, as follows:

$$\begin{aligned} p_{a,m}(hybrid) &= \alpha \times p_{a,m}(user) + (1-\alpha) \times p_{a,m}(item) \\ &= \alpha \times (\overline{r_a} + \frac{\sum_{k \in K}(r_{k,m} - \overline{r_a}) \times w_{a,k}}{\sum_{k \in K} w_{a,k}}) + \\ &\quad (1-\alpha) \times \frac{\sum_{k \in K'} r_{a,k} w_{m,k}}{\sum_{k \in K'} w_{m,k}} \end{aligned} \quad (1)$$

where α is adjusting parameter, $P_{a,m}$(*user*) means the rating of user weighted average, $P_{a,m}$(*item*) means the rating of item weighted average, K is the neighborhood set of the k users that are most similar to user a, K' is the neighborhood set of the k items that are most similar to item m.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 2, pp: (191-198), Month: April - June 2016, Available at: www.researchpublish.com

The item similarity weight $w_{i,j}$ between items *i* and *j in item based CF* can be calculated as follows:

$$w_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \overline{r_i})(r_{u,j} - \overline{r_j})}{\sqrt{\sum_{u \in U}(r_{u,i} - \overline{r_i})^2}\sqrt{\sum_{u \in U}(r_{u,j} - \overline{r_j})^2}} \quad (2)$$

Where U is the set of users who have rated both items *i* and *j*, $r_{u,i}$ is the rating of user *u* on item *i*, $r_{u,j}$ is the rating of user *u* on item *j*, *ri* is the average rating of item *i* across users set *U*, and *rj* is the average rating of item *j* across users set *U*

The rating of user *a* on item *m* can be predicted using a user weighted average as follows:

$$p_{a,m} = \frac{\sum_{k \in K} r_{a,k} w_{m,k}}{\sum_{k \in K} w_{m,k}} \quad (3)$$

Where $p_{a,m}$ means the rating of user *a* on item *m*, *K* is the neighborhood set of the k items that are most similar to item *m*.

The user similarity weight $w_{a,u}$ between users *a* and *u in user oriented CF* can be calculated as follows:

$$w_{a,u} = \frac{\sum_{i \in I}(r_{a,i} - \overline{r_a})(r_{u,i} - \overline{r_u})}{\sqrt{\sum_{i \in I}(r_{a,i} - \overline{r_a})^2}\sqrt{\sum_{i \in I}(r_{u,i} - \overline{r_u})^2}} \quad (4)$$

Where *I* is the set of items rated by both users *a* and *u*, $r_{a,i}$ is the rating of user *a* on item *i*, $r_{u,i}$ is the rating of user *u* on item *i*, r_a is the average rating of user *a* across items set *I*, and *ru* is the average rating of user *u* across items set *I*.

The rating of user *a* on item *m in user-oriented CF* can be predicted as follows:

$$p_{a,m} = \overline{r_a} + \frac{\sum_{k \in K}(r_{k,m} - \overline{r_a}) \times w_{a,k}}{\sum_{k \in K} w_{a,k}} \quad (5)$$

Where $p_{a,m}$ means the rating of user *a* on item *m*, *K* is the neighborhood set of the k users that are most similar to user *a*.

C. Content Based Filtering:

In a content based recommender framework, catchphrases or credits are utilized to depict things. A client profile is worked with these traits (maybe as a client upvotes or "likes" something). Things are positioned by how nearly they coordinate the client characteristic profile, and the best matches are prescribed.

D. Cosine Similarity Calculation Method:

The evaluations of client a and client u can be viewed as the two vectors in the n-dimensional vector space, so similarity between the two clients can be figured by the cosine of the angle between the two n-dimensional vectors. The computing equation of cosine similarity is as follow:

$$\mathrm{w(a,i)} = \cos(\vec{a}, \vec{i}) = \frac{\vec{a} * \vec{i}}{\|\vec{a}\| . \|\vec{i}\|} \quad (6)$$

The steps of user-rating item type similarity calculation are as follow: Firstly, register the numbers of every client rating every sort; furthermore, apply cosine similarity strategy to process the likeness between sets of clients with in regards to the checks of clients rating thing sorts as n-dimensional vectors. In this way, the comparability amongst client and client represents as follows:

$$\mathrm{w(a,i)} = \cos(\vec{a}, \vec{i}) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}^2}} \quad (7)$$

Where *j* is the item type that both user *a* and user *i* have rated, is the count of user *a* rating type *j*, and is the item type set which users *a* have rated. Above similarity technique utilizes client rating items sort data rather than client rating data to process clients' closeness, and for the number of item types is far not as much as that of items, so this strategy can lighten rating information sparsity issue and enhance versatility of calculation.

E. Data Representation:

The implementation of our approach was based on the movielens data set that was collected by the grouplens Research Project at the University of Minnesota through the movielens web site[8]. The characteristics of this set are:

- 100,000 evaluations (in a scale from 1 to 5) of 1682 films by 943 users.
- Each user has evaluated at least 20 films.

The types to which movies have a place are found with be as per the types in the Internet Movies Data Base (www.imdb.com). Basic demographic data for the clients (sex, calling, age) is accessible. Be that as it may, we didn't utilize this data, since its commitment is not sufficiently huge to legitimize the expansion in framework many-sided quality.

The client assesses movies that he/she has found in a size of five degrees (5: perfect work of art to 1: terrible film). Movies worth proposal are viewed as those that get grades 4-5, while those that get 1-3 are rejected. This is fundamental with a specific end goal to take in the inclinations of the client and develop the client's profile. We mull over two components: the content of films that the people have as of now seen and the movies that persons with comparable inclinations have delighted in.

F. Workflow of the proposed system:

The proposed system implements the below workflow as shown in figure 1 for both user and item data sets to determine the predictions .

The user and item data sets are taken as inputs in the initial stage and they are preprocessed .Preprocessing involved reading the data from the data set which will be usually delimited using various special characters. The huge volume of data is parsed and stored into database for further processing. Once the data is preprocessed the similarity is calculated using the cosine similarity method which is very much efficient compared to pearson and spearman correlation methods. The average weights are calculated for the similar items and the nearest neighbor is determined using the k-nearest neighbor method.

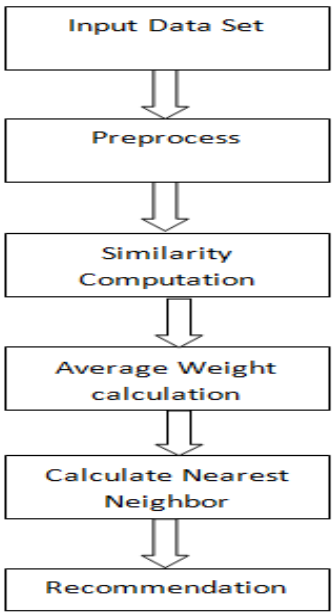

Fig 1. Workflow of proposed system

Once the predictions are generated for both item and user data set both the results are merged as shown in figure 2 to obtain the final movie recommendations.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 2, pp: (191-198), Month: April - June 2016, Available at: www.researchpublish.com

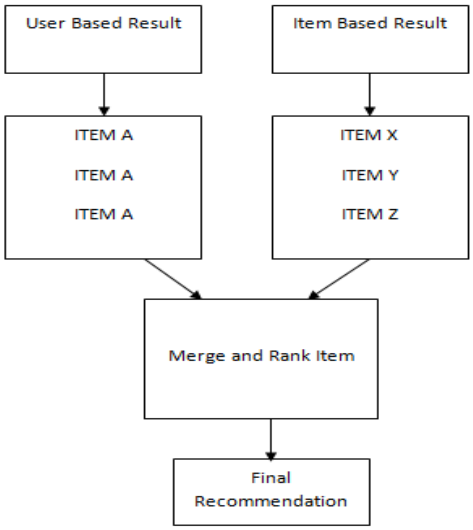

Fig 2. Merge the results and generate final recommendation

IV. PERFORMANCE ANALYSIS

Performance investigation is essentially the procedure of assessing how a specific programming project is working. This procedure typically starts with how the system loads and what happens when every progression in utilizing the project is executed. The object of execution examination is to guarantee the product project is working at ideal productivity and to recognize and revise any issues that may adversely affect that effectiveness.

Mean absolute Error (MAE) is a common used metric to measure the predictive accuracy, defined as the average absolute difference between the ground truth and predicted rating value, as follows:

$$MAE = \frac{\sum_{a,m} |p_{a,m} - g_{a,m}|}{N} \quad (8)$$

Where $p_{a,m}$ is the prediction rating of user a on item m, $g_{a,m}$ is the ground truth rating of user a on item m, N is the test number.

To begin with, we alter the neighborhood number k to 25, then test the impact of parameter α to the half and half CF show. As outlined in Fig.3, we can see that the execution is entirely steady with parameter α traversed from 0.1 to 0.6, the execution disintegrate with parameter α spread over from 0.7 to 0.9. That implies client situated model assume a superior part than item focused model in the combined CF show.

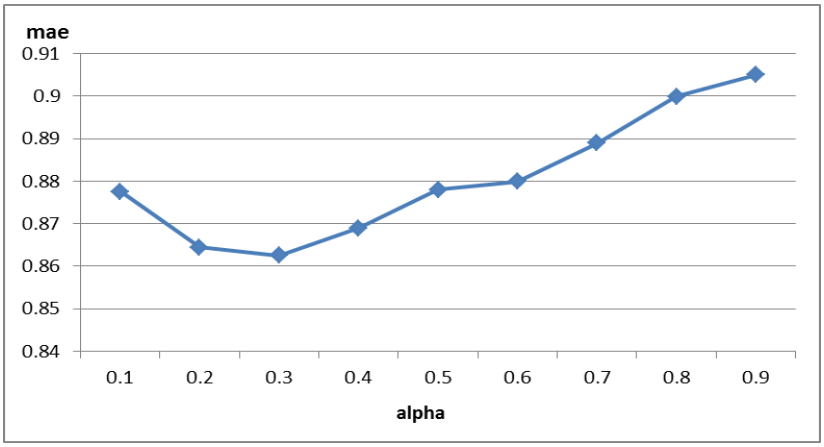

Fig. 3: Mean Absolute Error (MAE) of prediction ratings of the combined CF model with parameter α spanned from 0.1 to 0.9.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 2, pp: (191-198), Month: April - June 2016, Available at: www.researchpublish.com

At that point, we test the impact of neighborhood parameter k to the combined CF model. We alter our combined model parameter α to 0.3, look at the execution of three distinctive CF strategies: item focused CF calculation, user focused CF calculation and our combined CF display with various neighborhood parameter k, as delineated in Fig.4. Point by point MAE estimations of three diverse CF models are spoken to in Table.1. Clearly, our proposed CF display has the best result.

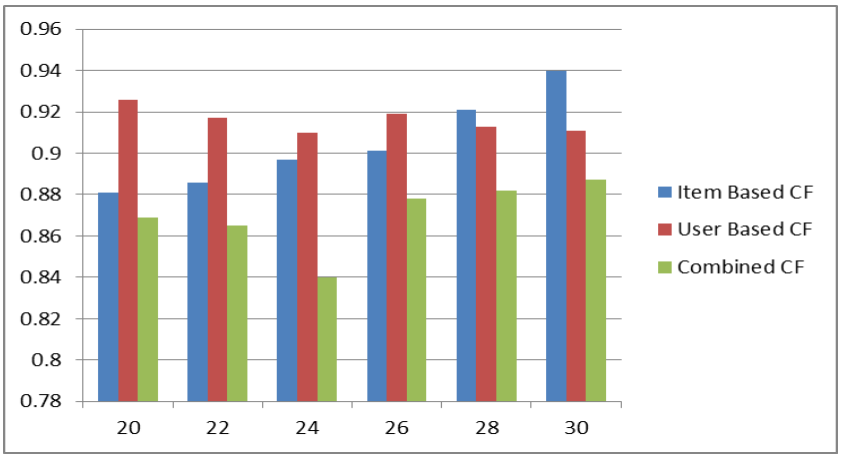

Fig. 4: Mean Absolute Error(MAE) of prediction ratings of three different CF models with neighborhood parameter *k* spanned from 20 to 30.

The MAE values of the three CF models are mentioned in the table 1 below.

TABLE 1: Detailed MAE values of the three CF models: item oriented CF algorithm, user-oriented CF algorithm and our proposed CF model

k	user-based CF	item-based CF	Proposed Combined CF
20	0.881	0.926	**0.869**
22	0.886	0.917	**0.865**
24	0.897	0.91	**0.84**
26	0.901	0.919	**0.878**
28	0.921	0.913	**0.882**
30	0.94	0.911	**0.887**

The figure 5 shows the performance analysis of proposed system. It is examined that the proposed system yields better performance than existing systems.

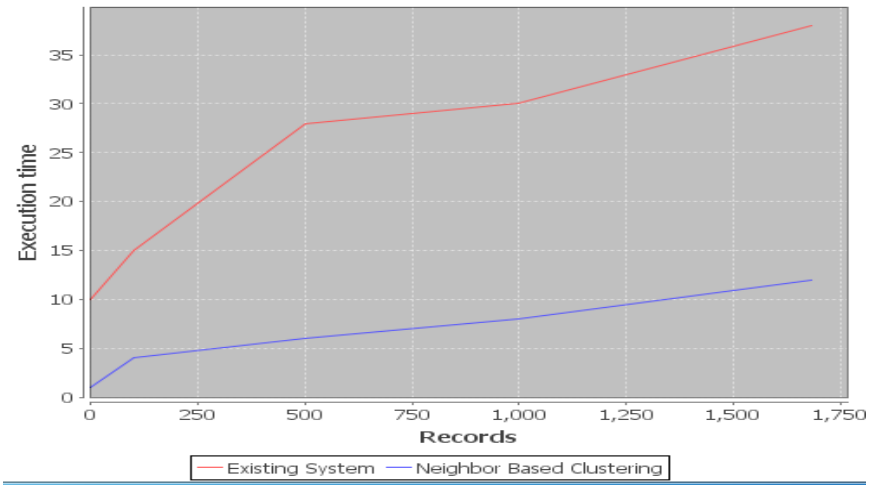

Fig 5. Performance analysis of proposed system

The X and Y axis denotes the number of movies in the data set used and the time required to perform the required processing of data, similarity computation, matrix formation, average weight calculation, nearest neighbor calculation and determining the recommendations of each data set and finally merging the results of two data sets and ranking them with an highest ranking of 5 and lowest ranking of 1 based on collaborative filtering and content analysis methods.

V. CONCLUSION

The proposed system is a complete collaborative and content analysis Recommender System that is specifically fit for those cases where information is large. The recommendation system is based on combined collaborative filtering and content analysis with k-nearest neighbour approach and cosine similarity to create a product bundling strategy. Since the recommendation system applies collaborative filtering, it achieves a pretty good accuracy since the pre-processing of data, similarity computation, matrix formation, average weight calculation, nearest neighbor calculation and determining the recommendations is performed on each data set (item and user data set) and then finally the results of both data sets are merged and then ranked to provide the final recommendations to the user. The user gets the most accurate movie recommendation based on his preference and the past user ratings and similar users rating available in the data set.

REFERENCES

[1] Gediminas Adomavicius, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", IEEE transactions on knowledge and data engineering, vol 17, No .6, June 2005

[2] Felix (Guohan) Gao, Calvin (Kim-pang) Lei, "Movie Recommendation System", University of California, Los Angeles.

[3] Prof .Ujwal U.J, Dr.Antony P.J, Abhilash K.R, "Implementation Of Recommendation System in a Web Browser with Help of Cloud Computing", International Conference on Emerging Research in Computing Information, Communication and Applications.

[4] Dan R. Greening "Building Consumer Trust with Accurate Product Recommendations", A White Paper on LikeMinds Personalization Server.

[5] Songjie Gong, "A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering ", Journal of Software, Vol 5, No 7 (2010), 745-752, Jul 2010.

[6] Ms. Smita Krishna Patil, Mrs. Yogita Deepak Mane, Mrs. Kanchan Rufus Dabre, "An Efficient Recommender System using Collaborative Filtering Methods with K-separability Approach" , International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622

[7] Taher Niknam, Elahe Taherian Fard, Shervin Ehrampoosh And Alireza Rousta , "A new hybrid imperialist competitive algorithm on data clustering", Vol 36, Part 3, pp.293-315, June 2011.

[8] Http://grouplens.org/datasets/movielens/