Volume 5, Issue 4, April 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Appraising and Exploring Similarity Measurement Approaches in Recommendation Systems by means of Momentous

Prof. K. V. Bhosle*, Kautkar Rohit Anil
Computer Science and Technology, MIT, Aurangabad,
Maharashtra, India

Abstract— Nowadays, WWW is emerged very rapidly and it leads to creating internet as a significant place exchanging information over web. As a vast amount of data arising over web continuously, the information overloading problem arisen. It leads to difficulty in getting accurate information without ambiguity. So for the rescue, the Recommendation systems (RS) came. The RS carrying more significant in today's world. The basic principle of RS is based on similarity by which it suggests items that are closest to user interest. So various approaches are existing which are helpful for implementing RS. There are two major tasks like prediction of items and actual recommendation. Here, the task prediction is uses the concept of similarity for discovering most similar items. The major techniques like Content Based Filtering, Collaborative Filtering and Hybrid filtering employs various similarity measurement approaches. As recommendations generated based on similarity, the approaches are very vital. In this paper, the various similarity measurement techniques explored.

Keywords— Recommendation Systems (RS), Information Overloading, Content Based Filtering, Collaborative Filtering, Hybrid Filtering

I. INTRODUCTION

Recent years saw a rapid growth of internet which leads to creation of internet as a significant place for sharing out information [1]. As data on internet getable in structure of text, image, videos, etc. and each day it is rising by means of brisk speed in size, volume and velocity (3V's) referred as BigData. Also the dilemma of information overloading is key trouble that exists in many domains. There are plentiful choices available for users to choose from. So Recommendation Systems (RS) came for liberate [2]. RS are responsible for presenting catalog of elements that may similar to user's interest. RS basically do two major tasks like prediction of items according to user's interest and recommendation with appropriate ranking. So fundamental thing concerning RS is generating recommendations based on similarity principles [3]. While generating recommendations the similarity among exists items or users is calculated. Fig 1 shows various similarity measurement approaches.

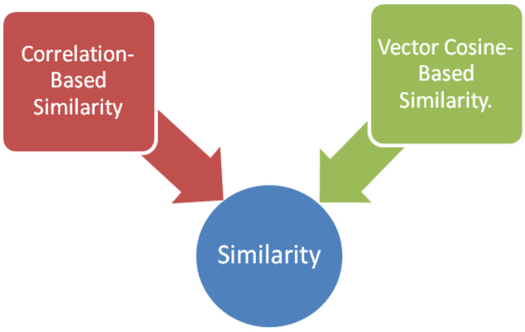

Fig 1: Similarity Measure Approaches

The neighboring items to current items are recommended and suggested. As there are a variety of key approaches for RS like Content Based Filtering (CBF), Collaborative Filtering (CF) or Hybrid Filtering (Combines one or more approaches) [4]. Content Based Filtering creates user, items profiles and make an effort to match the genuine contents to users and items profiles. This is one of the most extensively used techniques. The Collaborative Filtering remarked as most successful techniques in recommendation system field [2]. In this, the recommendations are produced by employing ratings reported by other similar taste users. It is also referred as Social Filtering [5]. The ratings of users having similar features or interest are considered so it is collaborative filtering. There are several shortcomings or advantages of both approaches [6]. For illustration, consider the Collaborative Filtering face crisis of cold start (item, user or community) where the data like rating regarding user, item or community is not available. So finding neighbors is great challenge. In contrast, the Content Based Filtering made the overspecialization. So user can't get other recommendations that one area. Now such shortcoming can be conquer by Hybrid Filtering where more than one technique can be united to produce recommendations. Also various challenges mentioned in Table I that confronts in Recommendation Systems.

Table I Various challenges of similarity measurements

Challenges of Similarity Measure	Description	Solution
▪ Data Sparsity	Users are lazy. They avoid providing ratings. So Data Sparsity	▪ Dimensionality Reduction ▪ Singular Value Decomposition (SVD) ▪ Principle Component Analysis (PCA)
▪ Scalability	Dealing with large number of items	▪ Clustering
▪ Improving Performance	As Recommendations are real time, so improving performance of techniques is crucial	▪ Threshold based Similarity Transitivity (TST)

The success of RS is very important as it decides future behavior of customer. As stated above, the user having lot of choices. A RS totally depend on concept of similarity measurement [7], [8], [9], [10] between existing items for recommendations. It is one of the most difficult tasks it has to achieve. The similarity is measured in terms of ratings provided by user. The various approaches uses various techniques for measuring similarity like Content based filtering[1],[2],[3] uses TF-IDF [2],[8],[11] for measuring similarity between users profile and actual contents. As shown in Table II, Various similarity computation approaches like Item To Item or User To User.

Table II Various similarity approaches

Similarity Approach	Underlying principle	Shortcomings
Item to Item	The similarity between two items are computed and explored	Can't Process offline
User to User	The similarity between tow users are computed and explored	Problem with scalability

Also on other side, Collaborative Filtering uses the Pearson Correlation Coefficient (PCC) [8], [3], [12] or Cosine Similarity [5], [11], [8], Measure. Also other approaches like Jaccard similarity measure [11], Spearman correlation Similarity [12], Log likelihood Similarity and Euclidean Distance [12], [13] Similarity that may employ to compute similarities etc. In Collaborative Filtering, there are two techniques like Memory Based [2], [6] and Model Based Filtering [2], [5], [6]. In this, the similarity can be measure based on the item or user.

In this paper, Section II describes related efforts carried out in this area. Section III presents various similarity measurement techniques in great extends. Section IV confronts the assorted experiments done in concern with similarity measurements specifically with Cosine Similarity. Section V presents Conclusion of presented work.

II. RELATED WORK

As Recommendation Systems carries significant in today's WWW, there are several authors who had proposed work in this field. Florent et al states, that there are two major categories of Recommendation Systems like Collaborative Filtering based which computes recommendations based on other similar users and Content Based Filtering which computed recommendations based on actual content of items [14]. Dhruv Gupta et al mentioned system Jester 2.0 which used the Euclidean distance for measuring similarity between users [15]. Resnick et al given the Pearson correlation coefficient in grouplense project [4], [12]. Pazzani et al. Proposed web page recommender system where the rating of webpage link is considered [6]. Das et al. proposed Google news for clustering news items and recommendations based on Collaborative Filtering [16].

In this, they employed MinHash and PLSI for clustering news items. NewsDude [11], [14] by Billsus and Pizzani and YourNews [11], [13], [2] by Ahn et al exercise the TF-IDF notion for evaluate similarity between documents. Herlocker et al modified the PCC for performance improvement [3], [9], [13]. PRES by van Meteren et al employed TF-IDF and the cosine similarity measure for news recommendation [11]. Flavius Frasincar et al used TF-IDF, Cosine and semantic approach for business news recommendations. Jiahui Liu et al used Bayesian framework with CTR approach for tracking user's interest and recommending news.

III. VARIOUS SIMILARITY MEASURMENT TECHNIQUES

As affirmed over, a variety of approaches for RS existing and each uses diverse techniques for calculation of similarity between items (i1, i2). This segment outline various similarity measurement techniques according to approach of RS like Content Based Filtering and Collaborative Filtering as shown in subsequent Table III. Figure 2 shows two major tasks like similarity computation [11],[13],[17],[18],[19] and prediction computation [3],[19] with their related approaches.

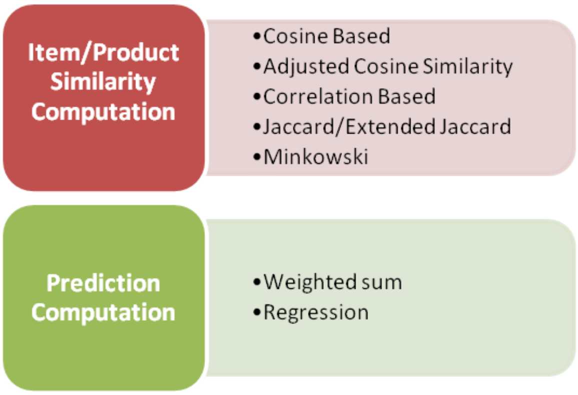
Item/Product Similarity Computation
•Cosine Based
•Adjusted Cosine Similarity
•Correlation Based
•Jaccard/Extended Jaccard
•Minkowski
Prediction Computation
•Weighted sum
•Regression

Fig 2: Two major tasks with approaches

Table III Overview of similarity techniques

Main Approach	Sub-Categories	Representative Techniques	Extensions
Collaborative Filtering	Memory-based/ Neighborhood-based	▪ Pearson Correlation Coefficient (PCC) ▪ Cosine Similarity Measure ▪ Neighbor-Based CF ✓ Item-Based ✓ User-Based ▪ Top-N Recommendations ✓ Item-Based ✓ User-Based	▪ Default Voting ▪ Inverse user Frequency ▪ Case amplification
	Model-based	▪ Latent Semantics Models ✓ PLA/PLSI ✓ DLA ✓ SVD ▪ Association Rule mining ▪ Clustering ▪ Bayesian Classifiers	▪ Not available
Content Based Filtering	--	▪ TF-IDF ▪ LSA ▪ SVD ▪ Jaccard Distance ▪ Cosine similarity	▪ Not available

Similarity grants a way for measuring the associations between user-item, user-user or item etc. Here similarity means the maximum time occurrences of terms similar to user's profile.

Table IV Approaches with similarity measure notions

Approach	Similarity Measure Concept
Content Based Filtering	▪ Similarity among two items (Vectors) calculated and most relevant items returned ▪ Based on Properties
Collaborative Filtering	▪ Similarity based on ratings given by other similar user
Hybrid Filtering	▪ Depends on incorporated techniques
Demographic Filtering	▪ Similarity based on users demographic characteristics

As revealed in above Table IV, various approaches and their way of measuring similarity.

1. Collaborative Filtering

In Recommendation System [2], [3], [6], [13] the Collaborative Filtering [2], [6], [9], [19] is one of the most broadly exercised approach for produce recommendations. Here, the trouble is viewed in form of user-item matrix where each row corresponds to a user and each column corresponds to an item [6]. Here, system predicts what rating a user may provide to missing cells by using ratings from other similar users. Means the prediction can be produced using collaboration of other similar user. Here, the interest of user is recognized by ratings given by them.

There are following two approaches in Collaborative Filtering techniques like Memory Based and Model Based [2], [5], and [6]. In Memory Based CF, all ratings from memory directly can be used for making predictions where in Model Based CF, the models are constructed offline using ratings (Web Usage Data) exists in database [9], [13], [19]. There are following Similarity measurement techniques available in Collaborative filtering-

A. Memory Based

In memory based approach, commonly the Pearson Correlation Coefficient (PCC) [8], [3], [9], [12], [19] and Cosine Similarity measurement is used. The Similarity Measurement is one of the complex and significant step in prediction of items.

- *Pearson Correlation Coefficient (PCC)*

The Pearson Correlation Coefficient (PCC)[8], [3], [12] is one of the most widely used similarity measurement technique with CF. Here the rating similarity between two items can be computed from -1 to +1 range. The value -1 means total dissimilarity where +1 means total similarity between item i and j. Grouplense utilized the PCC to measure similarity between users. It is based on linear relationships between two items which is termed as Correlation. Also it considers series of preferences of active user for making predictions. The formula of PCC uses actual rating values that are considered as preferences of user.

$$P_{sim\,(u,v)} = \frac{\sum_{i=0}^{n}(r_{u,i}-\bar{r}_u)\,(r_{v,i}-\bar{r}_v)}{\sqrt{\sum_{i=0}^{n}(r_{u,i}-\bar{r}_u)^2}\,\sqrt{\sum_{i=0}^{n}(r_{v,i}-\bar{r}_v)^2}}$$

The notion of PCC is based on correlation between users or items. The similarity $P_{sim\,(u,v)}$ between two items like u and v can be computed. PCC is not working efficient for item to item similarity as contrast to cosine similarity. PCC make use of the mean of ratings for making computation of similarity between two users.

The number of users with which the active user's ratings will be compared referred as neighborhood size. In point of fact, PCC represents the correlation between two variables. So here, the two variables represent the values of ratings. In this, the value is obtained as an average of ratings of each neighbor of a current item. Neighbors with similar taste or attribute will produce higher value which is signal of their closeness.

- *Shortcomings*
  - Not good for item-item similarity computation
  - Not considering overlapping preferences/ ratings

The prediction of user's ratings can be achieved with following formulae

$$Pred_{(u,i)} = \bar{r}_u + \frac{\sum n\; neighborsuserSim(u,n)\cdot(r_{n,i}-\bar{r}_n)}{\sum n\; neighborsuserSim(u,n)}$$

Where is $\bar{\bar{r}}_u$ the mean rating of user u, $userSim(u,n)$ is the similarity value between u and n. Also $r_{n,i}$ is the rating of user n for item i and $\bar{r}_n$ is the average rating of n user.

- *Cosine Similarity Measure*

In Collaborative Filtering, the similarity between two users u and v can be directly measure using the Cosine Similarity [5], [8], [9], [11], [19]. It is based on concept of Vector Space which uses Linear Algebra for computing similarity between two users or items. Here the two items can be thought as two vectors in n dimensional space. The Similarity between two item is depend on the Cosine angle between them.

$$Cosine_{sim(u,v)} = \frac{\sum_{i=0}^{n}(u_i\cdot v_i)}{\sqrt{\sum_{i=0}^{n}(u_i)^2}\,\sqrt{\sum_{i=0}^{n}(v_i)^2}}$$

Here users are represented as vector in which the similarity is measured by cosine distance between two vectors which represents ratings. It can be achieved by taking dot product of two vectors and divided by the Euclidean distance form. The ratings can be addressed as vector in m dimensional space and figure out similarity using the Cosine of angle between them. In cosine similarity, the negative ratings are not considered and blank space is considered as zero.

The items are represented as vector in coordinated space and the angle between two vectors computed. If two vectors are more similar the angle will be smaller otherwise larger. Another version of cosine similarity measure is adjusted cosine measure which is useful in web usage data where scale of rating is different.

$$Adjusted\_Cosine_{sim(u,v)} = \frac{\sum_{i=0}^{n}(r_{u,i} - \bar{r}_u)\,(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i=0}^{n}(r_{u,i} - \bar{r}_u)^2}\,\sqrt{\sum_{i=0}^{n}(r_{v,i} - \bar{r}_v)^2}}$$

- *Shortcomings*
    - Depend on users rating
    - The problem when data is sparse
    - Can't handle scalability

- *Neighbor-Based CF*

The Nearest neighbor approach [19], in which the whole user-item database to generate predictions directly so there is no models. There are two variations like

✓ *Item-Based*

The two items compared based on their patterns of ratings across users. Principally similarity between items is static. So it made possible the pre-computing of item-item similarity.

✓ *User-Based*

There are two key steps, like formation of neighborhood and generating recommendation. For formation of neighborhood, the Pearson Correlation Coefficient (PCC) can be used. To achieve this, real time calculation is requisite. So offline processing is not possible.

B. Model Based

- *Latent Semantics Models*

1. DLA

DLA [14], [19], [20] is one of the most primitive and popular topic modeling approaches. As in all places the information overloaded in form of text, audio, video, news etc. and we required an algorithmic approach to process, understand, and examine such bulky quantity of information. Topic modeling is one of the major methods for analyzing the large quantities of unlabeled data [14]. Here the topic is a probability distribution over collection of words. The topics are basically conveys the central theme of article [19], [20].

- *Association Rule mining*

Association rules [1], [13], [19] fundamentally provide the relation between diverse attributes and their values [1], [3], [19], [21]. Gobs of useful associations can be discovered by analyzing the data sets. Association Rules utilizes the concept of Support and Confidence to determine the usefulness of rule. Patterns are practicable for probing the buying behavior of customer for business analysis. Support for items X1 and Y1 for particular transaction can be given by percentage of transactions from data set which let in both items i.e. X1 U Y1. Confidence for association rule X1->Y1 is given by percentage of transaction from data set containing X1 also with Y1 [1], [19], [21]. The rule is genuinely helpful if it has Support and Confidence value above the defined support count for that dataset. The support count can fix by domain expert [1].

- *Clustering*

Clustering [1], [6] is also referred as unsupervised learning [19]. There may be circumstances in which there may be numerous items in which have to calculate similarity. This crisis may be overcome by dropping number of features. But then as well sometime there is probability of remaining many items. So cluster helpful to reduce number of computations [3]. Here, the users are clustered based on similar characteristics or features. The group of similar user's described as cluster [13]. The key goal is to assemble group in which users will be alike to apiece other but dissimilar to other users in dissimilar group. These are based on classification notion. Fig 3 shows various clustering approaches.

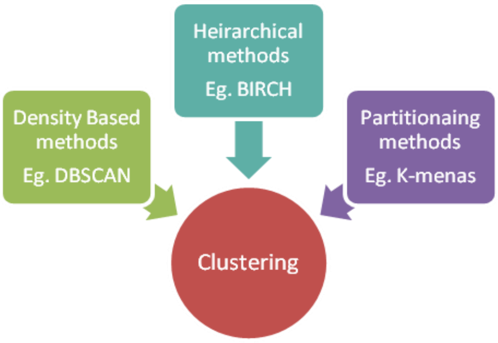

Fig3: Various Clustering Approaches

- *Bayesian Classifiers*

To solve classification problems, the Bayesian classifiers [3], [4], [5], [14], [19] are useful. Bayesian classifiers are depending on conditional probability and Bayes theorem [13]. There are various classifiers like naïve Bayes classifiers which assume probabilistic independence of variable. It means that, presence or absence of any variable is not depends on presence or absence of other variable. Bayesian classifiers are mostly used in model based collaborative filtering. The class labels and attributes are considered as random variable.

2. Collaborative Filtering

Another very widely used recommendation system implementation approach is Content Based Filtering [1], [2], [3], [13], [19]. In this, the profiles for user and items are generated which used to determine the similarity between the documents and user interest. The profile is nothing but set of keywords which depicts the interest of user. User's interest can be modeled in form of features [2], [3], [9]. In concern with Content based filtering there are following methods can be employed for the measuring similarity between the items

- *Term Frequency-Inverse Document Frequency*

TF-IDF (Term Frequency- Inverse Document Frequency) is one of major term weighting method in the IR field. It is basically used to measure the importance of words in specific document relative to frequency of word in collection of document [2], [8], [9], [11]. TF-IDF can be used with conjunction of Cosine Similarity as described above. Basically there are two concepts in TF-IDF like [10]:

1) Remove stop words like a, an, the etc. from the documents.

2) Removing stop words followed by stemming of remaining words.eg. Computerization became computer. Actually Stemming mean finding roots of words.

There are two steps like first compute the Term-Frequency (TF) and then compute Inverse Document Frequency (IDF). While using with Content Based Filtering, the user profile can be generated with help of TF-IDF by calculating for all important words related to user profile.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

In above equation, t_i is term frequency for term *i* document a_j where $n_{i,j}$ are number of occurrences of term t_i in document a_j.

- *Latent Semantic Analysis*

Latent Semantic Analysis [19], [20], [22] is about to finding hidden meanings of terms occurred in the documents. Here Latent means finding association explicitly in terms of documents and semantic means presenting those relationships. Latent Semantic Analysis is for dimensionality reduction in form of truncated Singular Value Decomposition (SVD) to extract Semantic relationships between documents. In IR field, it is referred as Probabilistic Latent Semantic Indexing (PLSI) [20], [22]. It is based on Probabilistic Theory rather than the Linear Algebra. In this, the task of document retrieval is considered in terms of what probability the document is relevant to given user query. These techniques can be used to measure the relatedness between the documents. It is solution for many challenges in field of IR.

Actually in SVD the problem of polysemy (more than one meaning) and synonymy (many words single meaning) occurs [11], [13], [19], [20]. It can be overcome with the LSA. Here the words and documents are mapped over the semantic space. PLSI process Term document structure to explores the relationships between them. The LSA takes advantage of Latent Semantic in the association of terms with documents [22].

- *Singular Value Decomposition*

Singular Value Decomposition [17], [3], [12], [19], [20] is one of the powerful techniques for dimensionality reduction. It is based on matrix factorization approach. Here the problem is to find the lower dimensional feature space where the concepts are shown with their strengths. The idea of SVD is to decompose rectangular metric into three various matrices like two orthogonal and one diagonal [19],[20].

$$A = u \sum V^T$$

Where u orthogonal, Σ is diagonal and *V* is orthogonal.

- *Jaccard Coefficients and Distance*

The functional and description similarity can be computed using the Jaccard distance. The Jaccard index also known as the Jaccard Similarity Coefficient [8], [9], [11], [19]. It is basically used for computing the similarity and also diversity between sample sets. It is basically defined as cardinality of intersection divided by the cardinality of union of the sample sets.

$$J(U,A) = \frac{|U \cap A|}{|U \cup A|}$$

In above formulae, $|U \cap A|$ denote the intersections between set U and A where $|U \cup A|$ denotes union of both sets. In Content Based Filtering, the user profiles and news items intersections. The Jaccard Coefficient is used to measure the Similarity and Jaccard distance is used measure the Dissimilarity [11].

The Extended Jaccard Similarity can be given by following formula-

$$Extended\_Jaccard_{sim\,(u,v)} = \frac{\sum_{i=0}^{n}(u_i \cdot v_i)}{\sqrt{\sum_{i=0}^{n}(u_i)^2}\ \sqrt{\sum_{i=0}^{n}(u_i)^2} - (u_i \cdot v_i)}$$

Various users may use various rating scales so in that case the extended Jaccard can be useful. In above formulae, u_i is the rating of user u for item i and v_i is the rating of user v for item i.

IV. EXPERIMENTS AND RESULTS

- *Dataset*

Here, in this work the Clicks and Views of news portal users considered having account on portal. The Clicks and Views of 19 users were recorded for period of 9/10/2014 to 09/04/2015. The dataset consist 313 instances of Clicks, Views and CTR of various categories. Also, 171 instances of Recommendation Panel views, clicks and CTR. As a Dataset cleaning process, all users having zero clicks and views are excluded in concern with more precise results.

- *Experiments and Results*

As in this work, we are investigating the various Similarity techniques like Pearson Correlation Coefficient (PCC), Cosine Similarity, Adjusted Cosine Similarity, Euclidean Distance, Minkowski, Jaccard Similarity Distance etc.

We performed experiments with various techniques. Here the presented results are generated with Cosine Similarity measurement between two users. Various performance metrics like Precision, Recall and F1 [24] with others [24] are used to compute the result which helps to distinguish between relevant and non relevant items. For clarity of figures, only few user's with their web usage data considered for obtaining following results. The following experiments performed on the extension (With User Similarity Computation) of Personalized News Recommendation Framework [2] proposed earlier.

Figure 4 shows various users id with their values of Precision and Recall. For some users, Precision is reached up to value one means recommended items liked and used by them. Also relative Recall value can be seen. Figure 5 and subsequently Figure 6 shows the Recall, Precision with F1 measure for various user ids. The F1 measure gives the overall tradeoff of precision and recall and carry more significant in recommendation systems to measure performance.

Figure 7 shows the fallout and Missrate which tells irrelevant items recommended to user and items not recommended but actually relevant to total number of relevant items respectively. Missrate varies from 0.5 to 0 which is indicates that it suggesting right content to user. Figure 8 depicts Precision, Recall and F1 with various users. Figure 9 shows Markedness and Informedness which produced unbiased results and does not depend on the recommender accuracy. The metrics Markedness and Informedness returns values between -1 to 1. For both, 0.60 is highest obtainable value which is good.

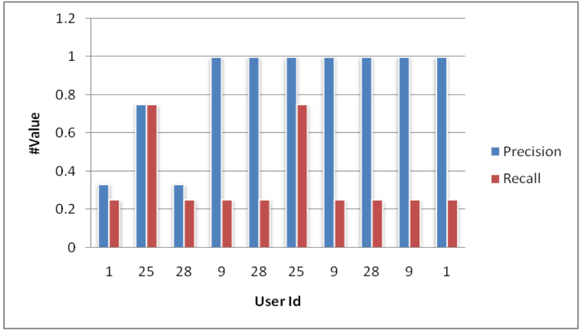

Fig 4: Precision and Recall

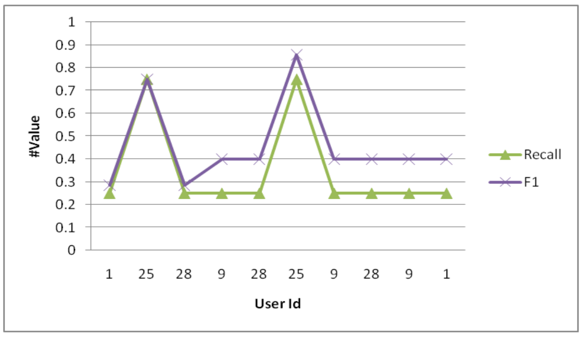

Fig 5: F1 Measure and Recall

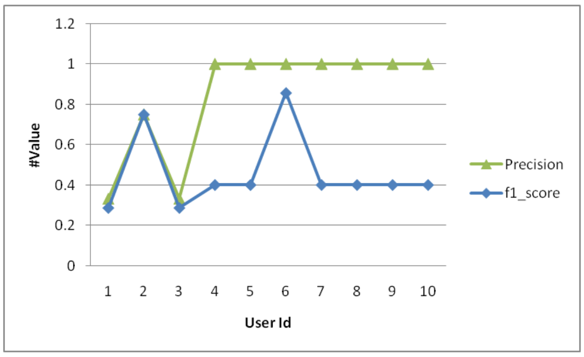

Fig 6: Precision Vs F1

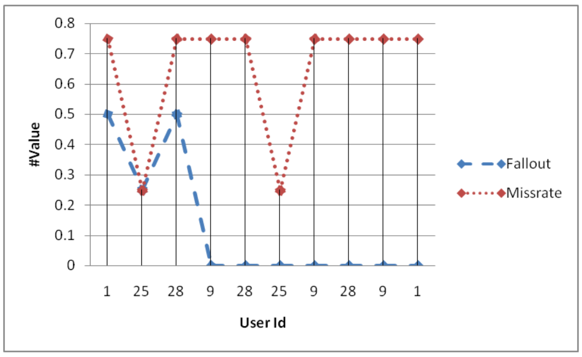

Fig 7: Fallout and Missrate

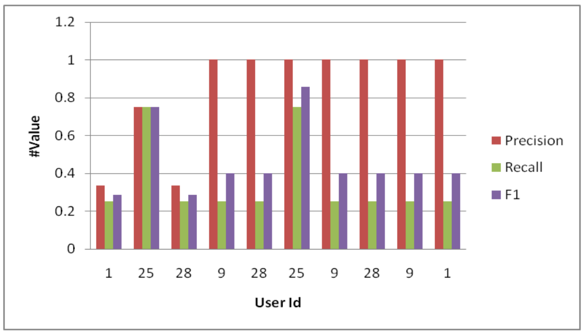

Fig 8: Precision Vs Recall Vs F1

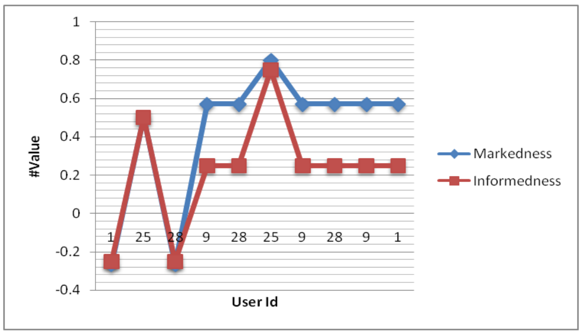

Fig 9: Markedness and Informedness

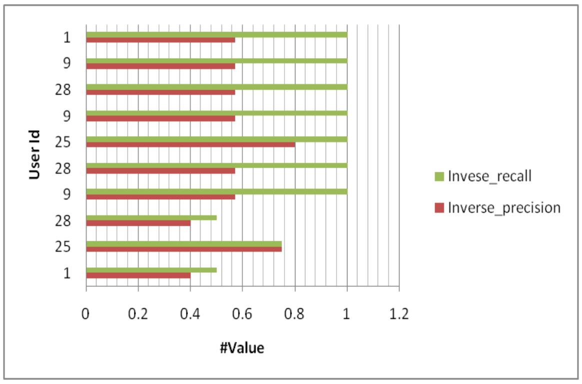

Fig 10: Inverse Recall and Inverse Precision across various users

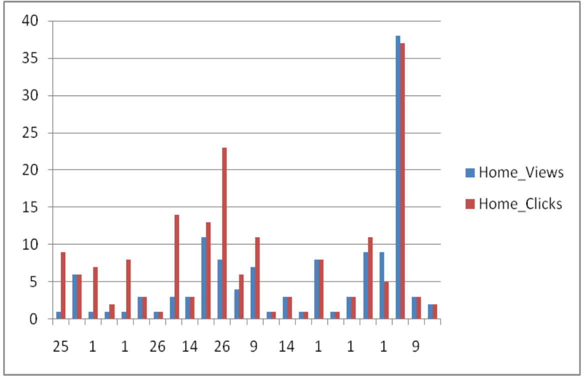

Fig 11: Home Page Views and Clicks across various users

Figure 10 shows the values of inverse Precision and Inverse Recall. Inverse Recall for many users touches value 1 which indicates that the accurate items were recommended and irrelevant items really skipped. Figure 11 shows clicks and views of home page which are important for measuring the CTR (Click-through-Rate) value which is important metric to measure user's interest as shown in Figure 12. In Figure, the variation of CTR values can be observed. In Figure 13, the accuracy of recommendation for various users can be observed. As in Figure, 0.61% average accuracy is obtained. In Figure 14, the various performance measure metrics of Recommendation System are shown. Their values are described in percent where Inverse Recall recorded with highest value 80% which is good indicator of systems good prediction. Also the next highest value 77% marked by important metric like precision. Fallout is 11% means the system really ignores irrelevant items to users interest.

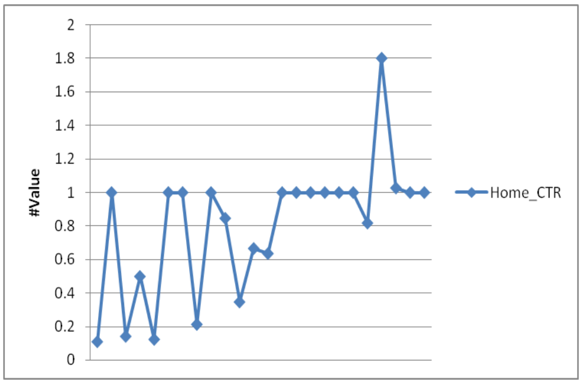

Fig 12: Variation of Home Page CTR

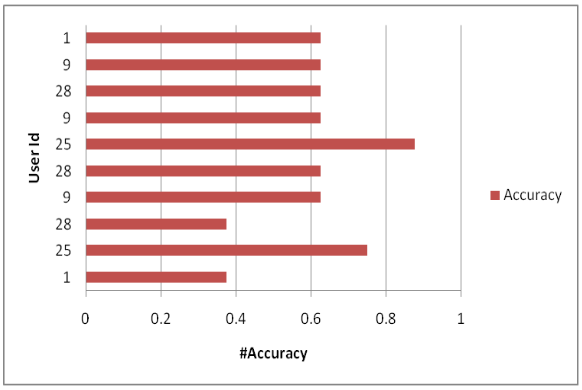

Fig 13: Accuracy of Recommendation

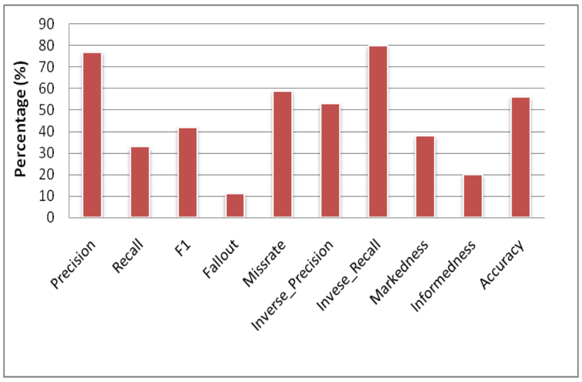

Fig 14: Various Performance Metrics (%)

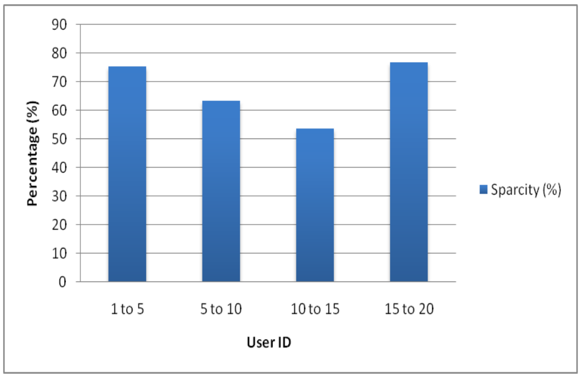

Fig 15: User Groups and Sparsity

Figure 15 depicts various user groups and percentage of sparsity. The sparsity percentage affects the accuracy of recommendations and is quite difficult to produce recommendations from data sparsity.

V. CONCLUSION AND FUTURE SCOPE

This work provides the important concepts of Recommendations i.e. similarity measurement concept with respective approaches. The Recommendation Systems that are using Collaborative filtering, their success is totally depends on results of similarity computation between user's and item's. So it's necessary to choose right similarity measurement techniques otherwise the results will to too far from actual expectations. So in this work, we addressed various similarity measures according to filtering techniques like Content Based Filtering and Collaborative Filtering. The various similarity techniques experimented on mentioned dataset. Some similarity techniques like Cosine Similarity, Adjusted Cosine Similarity, Jaccard Similarity (Also Extended Jaccard) can be used in both approaches. In experiments, for Precision and Recall the values are satisfactory and still can be improved with lot of efficiency. In future, this work can be extended with performance analysis of other similarity measurement techniques with large number of users.Also there is scope for demographic segmentation with analysis for performance improvement in concern with processing time, number of records, accuracy, scalability.

ACKNOWLEDGEMENT

First I would like to express deeper sense of appreciation for Prof. K.V. Bhosale for her support, encouragement and guidance. Also, I am enormously grateful Prof. V. Kala, Prof. B. Sonawane, Prof. R.B. Mapari, Prof. S. Kankal, Prof. J. Dhage, Prof. B. Chaudhari, Prof. S. Bangar and Prof. Gurav of Computer Science and Technology Department, MIT, Aurangabad for their advice, inspiration and remarks.

This work is incomplete without expressing words of thanks for Dr. S. J. Gorane, Dr. K. Nemade, Prof. N. Anwat significant guidance and comments. Also thankful to Prof. R.D.Kalambe, Prof. M. Arde, Prof. Borde, Prof. Nikale, Prof. Dudhe and all Staff Members of Computer and Information Technology department, Government Polytechnic, Nashik for their help and kind support.

Also expressing gratitude towards my dear parents Sangeeta and Anil for their love and care.

REFERENCES

[1] Kautkar Rohit, A. A COMPREHENSIVE SURVEY ON DATA MINING.

[2] Anil, K. R. Content Optimization for Personalized News Recommendation: An Experimental CTR Based Approach.

[3] Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In The adaptive web (pp. 291-324). Springer Berlin Heidelberg.

[4] Breese, J. S., Heckerman, D., & Kadie, C. (1998, July). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (pp. 43-52). Morgan Kaufmann Publishers Inc..

[5] Qian, F., Zhang, Y., & Duan, Z. (2013). Community-based user domain model collaborative recommendation algorithm. Tsinghua Science and Technology, 18(4).

[6] http://www.ijteee.org/final-print/mar2015/Exploring-Approaches-Of-Recommendation-System-In-Support-Of-Verdict-And-Comparison-A-Per-sonalized-Prospect.pdf

[7] Xie, F., Chen, Z., Xu, H., Feng, X., & Hou, Q. (2013). Tst: Threshold based similarity transitivity method in collaborative filtering with cloud computing. Tsinghua Science and Technology, 18(3), 318-327.

[8] Hu, L., Song, G., Xie, Z., & Zhao, K. (2014). Personalized recommendation algorithm based on preference features. Tsinghua Science and Technology, 19(3), 293-299.

[9] Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In Recommender systems handbook (pp. 257-297). Springer US.

[10] Li, L., Wang, D. D., Zhu, S. Z., & Li, T. (2011). Personalized news recommendation: a review and an experimental investigation. Journal of Computer Science and Technology, 26(5), 754-766.

[11] IJntema, W., Goossen, F., Frasincar, F., & Hogenboom, F. (2010, March). Ontology-based news recommendation. In Proceedings of the 2010 EDBT/ICDT Workshops (p. 16). ACM.

[12] Said, A., Jain, B. J., & Albayrak, S. (2012, March). Analyzing weighting schemes in collaborative filtering: cold start, post cold start and power users. In Proceedings of the 27th Annual ACM Symposium on Applied Computing (pp. 2035-2040). ACM.

[13] Rokach, L., Shapira, B., & Kantor, P. B. (2011). Recommender systems handbook (Vol. 1). New York: Springer.

[14] Garcin, F., Dimitrakakis, C., & Faltings, B. (2013, October). Personalized news recommendation with context trees. In Proceedings of the 7th ACM conference on Recommender systems (pp. 105-112). ACM.

[15] Gupta, D., Digiovanni, M., Narita, H., & Goldberg, K. (1999, August). Jester 2.0 (poster abstract): evaluation of an new linear time collaborative filtering algorithm. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 291-292). ACM.

[16] Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007, May). Google news personalization: scalable online collaborative filtering. In Proceedings of the 16th international conference on World Wide Web (pp. 271-280). ACM.

[17] Xie, F., Chen, Z., Xu, H., Feng, X., & Hou, Q. (2013). Tst: Threshold based similarity transitivity method in collaborative filtering with cloud computing. Tsinghua Science and Technology, 18(3), 318-327.

[18] Hu, R., Dou, W., & Liu, J. (2014). ClubCF: A Clustering-based Collaborative Filtering Approach for Big Data Application.

[19] Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. Advances in artificial intelligence, 2009, 4.

[20] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[21] Han, E. H. S., & Karypis, G. (2005, October). Feature-based recommendation system. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 446-452). ACM.

[22] Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. Foundations and Trends in Human-Computer Interaction, 4(2), 81-173.

[23] Schröder, G., Thiele, M., & Lehner, W. (2011, October). Setting Goals and Choosing Metrics for Recommender System Evaluations. In UCERSTI2 Workshop at the 5th ACM Conference on Recommender Systems, Chicago, USA (Vol. 23).

[24] Schröder, G., Thiele, M., & Lehner, W. (2011, October). Setting Goals and Choosing Metrics for Recommender System Evaluations. In *UCERSTI2 Workshop at the 5th ACM Conference on Recommender Systems, Chicago, USA* (Vol. 23).