Comparison of Similarity Measure Using Density Peak Clustering For Collaborative Filtering Techniques

Dr. S. Shajun Nisha[1], B.Rajeswari[2],
[1]Research Supervisor, [1]Assistant Professor & Head,
[2]Reg.No: 18211192162020, [2]Research Scholar,
[1,2]PG & Research Dept. of Computer Science,
[1,2]Sadakathullah Appa College, Tirunelveli,
[1,2]Affiliation of Manonmaniam Sundaranar University,
Abishekapatti, Tirunelveli, 627012, Tamilnadu, India.

Abstract

Collaborative Filtering (CF) filters the flow of data that can be recommended, by a Recommendation System (RS), to a target user according to his taste and his preferences. The target user's profile is built based on his similarity with other users. For this reason, CF technique is very sensitive to the similarity measure used to quantify the dependency strength between two users (or two items). In this paper compared two different types of similarity measures and find the best similarity techniques used for CF-based recommendation system. For each measure, we outline its fundamental background and we test its performance through an experimental study. Experiments are carried out on standard datasets (MovieLens100k) and reveal many important conclusions. Find the best similarity techniques for clustering algorithm in CF method.

Keywords: *Collaborative Filtering (CF), Recommendation System (RS), Movie Lens data set, Similarity Measure, Accuracy*

I. INTRODUCTION

CF-based Web service recommendation refers to recommending services according to the past composition history, the similarity of users, or the similarity of services. Collaborative filtering is a method of making assumption of user's interest by gathering the information about likes and dislikes from large number of users [3].The fundamental fact is that if a User A has the same opinion as like User B on the

same item, then User A has more chances to have similar opinion with User B for different item also. In this case, the recommendation system will suggests the item liked by User A to User B. At first, the user conveys their likes and dislikes by rating the items such as books, videos, movies or music etc. These ratings can be taken as the representation of user's interest for the particular item. The recommendation system matches the ratings of this user with the ratings of other users to find the users with most similar taste. Then, the system list out all the items which are rated higher by similar users. But those items are not yet rated by the current user. Even though, the recommendation system recommends those items to the current user as these are all rated by the user who have the similar taste. So, the collaborative based filtering system considers the past activities or behaviours of the user and it also use the similar decisions made by different users. Collaborative filtering was categorized into two types. They are as follows: First one is User based recommendation system, the User based recommendation system used to predict the items which the user might like based on the ratings given to the particular item by other similar users with similar taste with that of the current user. Second one is Item based recommendation system. Item based collaborative filtering is a type of collaborative filtering for recommendation systems. The ratings of the items which was given by the users was collected and using those ratings, the similarity between the items was calculated using the similarity measures such as Euclidean distance and Jaccard similarity etc. Similarity measurement is done prior to clustering using similarity measures. The closeness level of the objective items is measured with respect to the qualities that are accepted to recognize the cluster implanted in the information. The attributes may be based on the information and the text issued, hence there is no measure that is all around best for a wide range of clustering issues. Additionally, picking the proper closeness measurement is vital for examining the clusters, particularly for a specific kind of cluster algorithm. Reviewing the closeness as the separation parameter, a huge number of similarity measurements are needed to find the thick area and determine clustering task for new information. As a result, knowing the viability of various measurements is vital in picking the best option. When all is said in short, the similarity measurement acts as a separation among two items that is mapped into a solitary numeric esteem using two factors namely the properties of the two items and the measurement. As this research is concerned with Item based recommendation system, the item to item similarity has to be found for clustering and recommendation.

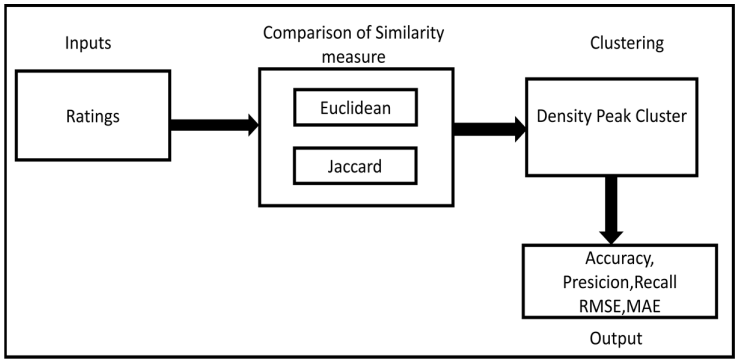
Let us discuss the role of similarity measures in recommendation system with an example. Let us take a sample dataset which consists of five movies which can otherwise be called as five items. The rating ranges from 5 (Strongly like) to 1 (Strongly dislike).Clustering Algorithms for Collaborative based recommendation**,** Clustering algorithm organizes the pattern collections most probably as a vector measures, or as a point in an n-dimension space in a cluster depending on the similarity measures. So, the input data has to be clustered based on the similarity measurements. The results of similarity are used for clustering algorithms to check the efficiency of the similarity measures along with the clustering algorithm. The density peak Clustering based approaches can be implemented and the performance comparison has been carried out to find the best similarity measure as well as the clustering algorithm.

II. LITERATURE REVIEW

In this work, the titles of the research papers were utilized to calculate similarity between papers [1].The measurement of Cosine quantifies the similarity between the two vectors as the cosine of the angle between two vectors[2]. There are commonly used Cosine similarity [5], Euclidean Similarity[6]. According to similarity between documents can be partitioned into three categories: first, string-based (character based and term-based) secondly, corpus-based and finally knowledge based (similarity, relatedness). This work utilises term-based similarity measures i.e. Cosine similarity, Euclidean distance, Jaccard similarity and Pearson coefficient similarity measure [4]. Performed a comparative systematic study on similarity measure for online documents. They compared four similarity measures (Euclidean, cosine, Pearson correlation and extended Jaccard) in conjunction a variety of clustering techniques (k-means, weighted graph partitioning, hyper-graph partitioning, self-organising feature map and random). Our research however follows a different direction in that it concentrates on the four similarity measures mentioned above, but they are used in conjunction with classification techniques (rpart, boosted and the random forest algorithms). In their work, the cosine similarity metric performed better than the rest, and the weighted-graph outperformed the other clustering techniques. Our work similarly compares the four similarity measures on how they perform with classification algorithms[7]. A. Huang Investigated partitioning clustering algorithms

with hierarchical clustering schemes and it was established that partitional clustering algorithms performed better. Further, similarity measures were utilised to compare and analyse the effectiveness of these similarity measures on document clustering. Their experiments established that three components ultimately affect the final result in a text clustering scenario: the objects, distance or similarity measures used and the clustering algorithm employed in the experiment. It has also been reported that with the given diversity set of distance and similarity measures available in data mining, their effectiveness in text classification is still not very clear[8]. It is widely used in data mining [10], recommendation [9].Clustering is the classification of data into separate classes or clusters based on a similarity measure and dissimilar data classification into separate clusters [11].

III. OUTLINE OF THE PAPER



IV. METHODOLOGY

4.1. Euclidean Distance

The Euclidean distance (a special case of the Minkowski distance with m = 2) is the notable distance measurement utilized in many numerical information. Its performances will be good if conveyed to dataset that incorporates minimum or isolate clustering. Euclidean distance has a disadvantage that when two vectors share no attributes, they would have a little separation than other vector pairs which contain a similar attributes. Next issue of Euclidean distance (being the Minkowski family) is its biggest scaled features that overrule the other measurement techniques. To mitigate this, the continuous features are normalized. It is a standard measurement in geometric analysis utilized with k-means clustering. The customary distance between

two vectors is effectively estimated in 2D/3D space and is utilized in most of the clustering (including text clustering).

The Euclidean distance between two movies characterized as vectors $\vec{v}_a$ and $\vec{v}_b$ is

$$D_E(\vec{v}_a, \vec{v}_b) = \sqrt{(\sum_{v=1}^{m} | W_{v,a} - W_{v,b}|)}$$

4.2. Jaccard Similarity measure

A measurement which is utilized to compare the similarity between model set can be a string or an entire report. The Jaccard coefficient estimates the comparability for a limited set and is characterized as the proportion of intersection size and the union size of the set.

The different types of similarity measurements are available to decide the level of similarity among movies. Among these, some measurements depend on the attributes present in every movies or nearness of common properties between the considered movies while some measurements consider both the existence and non appearance of characteristics in every movie. Such kinds of measurements are done by the general similarity measurement proposed by Tversky as

$$S_{v_1,v_2}(m_1, m_2) = \frac{f(m_1 \cap m_2)}{f(m_1 \cap m_2) + v_1.f(m_1 - m_2) + v_2.f(m_2 - m_1)}$$

Where v_1, v_2 positive real numbers are m_1, m_2 are the data items. When $v_1 = v_2 = 1$, the Jaccard similarity measurement is expressed as

$$S_{v_1,v_2} = \frac{|m_1 \cap m_2|}{|m_1 \cup m_2|}$$

4.3. Density Peak Clustering

The Density Peak Clustering (DPC) method is a novel clustering technique based on density peaks and distance. It primarily employs two factors, one is local density and another one is distance of the sample to the nearest neighbour with higher density to segregate and to identify the cluster centre. After finding the cluster centre, the data items are assigned to their respective nearest neighbour with higher density (Yewang, C et al., 2020). The steps involved in Density Peak Clustering algorithm is as follows

Algorithm:

Step 1: Calculating the distance d_{ij} between points and constructing similarity matrix

Step 2: Calculating local density β_i and high density distance δ_i for each data point based on the matrix which construct by step1 and parameters d_c of user input

$$\beta_i = \sum_j (d_{ij} - d_C)$$

Where d_c is defined as a cutoff distances. It is an adjustable parameter. Generally most of the cases it defined as: $d_c = d_{Nd} \times 2\%$

The δ_i is the minimum value of any point than its high density point distance. The δ_i is defined as the following formula

$$\delta_i = \min_{j:y_j > y_i} (d_{ij})$$

Step 3: Find $\mu(x_i)$ which the nearest higher density point of x_i

$$\mu(y_i) = \operatorname{argmin}_{j:y_j > y_i} (d_{ij})$$

Step 4: Taking the data points as the clustering center, whose two attribute values are all high.

Step 4: Remaining points can be classified according to the nearest neighbour classification algorithm

Step 5: Finally, filter the noise outlier data

V. EXPERIMENTAL RESULTS AND DISCUSSION

Similarity Measure	Accuracy	Precision	Recall	RMSE	MAE
Euclidean Method	93.18	90.78	89.777	0.294	0.082
Jaccard Method	**97.18**	**95.81**	**95.77**	**0.156**	**0.020**

Table 3.1 Performance comparison of similarity measure using Density Peak clustering algorithms for collaborative filtering

In this research work, experiments for different similarity measures with clustering algorithm. The performance of similarity measures was evaluated using metrics such as Accuracy, precision, Recall, MAE and RMSE. The performance is higher if the value MAE and RMSE is low evident that jaccard similarity performs better than other similarity measures for clusters. So, jaccard similarity measure is considered to be the better performing similarity measure and is used for the purpose of clustering.

Similarity Measure	**Clusters**	**Performance Metrics**	
		MAE	**RMSE**
Euclidean Distance	**Cluster=0**	**0.082**	**0.294**
	Cluster=1	0.106	0.333
	Cluster=2	0.126	0.362
Jaccard	**Cluster=0**	**0.020**	**0.156**
	Cluster=1	0.027	0.163
	Cluster=2	0.032	0.205

Table 3.2 Performance comparison of similarity measures

The experimental results for clustering algorithms which use jaccard similarity measure for clustering is shown in Table3.1. From the analysis of Table 3.2, it is obvious that Density Peak Clustering algorithm performs better than other clustering algorithms especially in terms of accuracy. It provides an average of 97.18% whereas other clustering algorithms provide lower average accuracy than Density Peak clustering algorithm. From the experimental results, it is concluded that Jaccard similarity with Density peak clustering provides the best clustering which in turn acts as a best recommendation system

VI. PERFORMANCE METRICS

Performance metrics are used to measure the performance in terms of metrics such as Accuracy, Precision, Recall, RMSE and MAE. The performance metrics used to evaluate the clustering algorithms are explained below.

6.1 Accuracy

Accuracy is the simplest intuitive performance metric, because it is just a ratio of accurately predicted observations to total observations. One would believe that if we have high accuracy, our model is the best. Accuracy is a fantastic measure, but only when you have symmetric datasets with almost equal values for false positives and false negatives.

$$Accuracy = \frac{True\ Negatives\ + True\ Positive}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

6.2 Precision

Precision is defined as the proportion of accurately predicted positive observations to the total number of expected positive observations.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

6.3 Recall

Recall is defined as the ratio of accurately predicted positive observations to all observations in the actual class.

$$Recall = \frac{TruePositive}{True\ Positive + False\ Negative}$$

6.4 Mean Absolute Error

Mean Absolute Error (MAE) is widely used metric to evaluate Recommender system. For each pair which the user rated, the absolute error is calculated. After

adding these pairs and dividing them by the total number of rating-prediction pairs, we can get Mean Absolute Error.

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n}$$

6.5 Root Mean Squared Error

Root Mean Squared Error (RMSE)is the widely used method for evaluating Recommender system. It calculates the deviation between predicted and real ratings. It gives more emphasis on deviation.

$$RMSE = \sqrt{\frac{\sum_{a,b}(x_i - y_i)}{N}}$$

Here, the overall number of rating is mentioned by *N*, *xi* and y_i represents the actual and predicted recommendation value in the recommender system.

VII CONCLUSION

To identify the efficient similarity algorithm for collaborative based clustering algorithms. A comparative analysis of Density peak clustering algorithms the experimental results, it is obvious that the combination of Jaccard algorithm with Density peak clustering algorithm overcomes other clustering algorithms. The average accuracy of Density Peak clustering is 97.18 %. Hence, it is decided to develop a modified version of Density Peak clustering algorithm which gives better accuracy than traditional Density Peak Clustering algorithms for movie recommendation.

REFERENCE:

[1] P. Jomsri, S. Sanguansintukul, and W. Choochaiwattana, "A comparison of search engine using “tag title and abstract” with CiteULike —An initial evaluation," in *Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for*, 2009, pp. 1-5.

[2].*Amine EL HADIa*, Youness MADANIb, Rachid EL AYACHIa ,Mohamed ERRITALIa”A new* semantic similarity approach for improving theresults of an Arabic search engine”,Procedia Computer Science 151 (2019) 1170–1175.

[3] .Su, X., & Khoshgoftaar, T. M.," A Survey of Collaborative Filtering Techniques",2009.

[4]. W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," International Journal of Computer Applications, vol. 68, 2013.

[5] M. Kuhn and K. Johnson. (2013). *Applied predictive modeling*. Available: http://dx.doi.org/10.1007/978- 1-

4614-6849-3

[6] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches,"*International Journal of Computer Applications,* vol. 68, 2013.

[7] A. Strehl and J. Ghosh, "Impact of similarity measures on web-page clustering," 2000.

[8] A. Huang, "Similarity measures for text document cluster," in *Proceedings of the sixth New Zealand* computer *science research student conference* (*NZCSRSC2008), Christchurch, New Zealand*, 2008, pp. 49-56.

[9]. Guo, G., Zhang, J., & Yorke-Smith, N. (2015). Leveraging multiviews of trust and similarity to enhance clustering-based recommendersystems. *Knowledge-Based Systems*, *74*, 1427. https://doi.org/10.1016/j.knosys.2014.10.016. [Crossref], [Web of Science ®], [Google Scholar]

[10].Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEET ransactions on EmergingTopicsinComputing*, *2*(3), 267279. https://doi.org/10.1109/TETC.2014.2330519. [Crossref], [Web of Science ®], [Google Scholar]

[11].Havens, T.C., Bezdek, J.C., Leckie, C., Hall, L.O.,& Palaniswami, M. (2012). Fuzzy c-means algorithms for very large data. *IEEE Transactions on Fuzzy Systems*, *20*(6), 11301146. https://doi.org/10.1109/TFUZZ.2012.2201485. [Crossref], [Web of Science ®], [Google Scholar]