www.ijemr.net ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Volume-7, Issue-2, March-April 2017

International Journal of Engineering and Management Research

Page Number: 309-314

K-Means Clustering based Solution of Sparsity Problem in Rating based Movie Recommendation System

Rahul Shrivastava[1], Himanshu Singh[2]

[1,2]Assistant Professor, School of Engineering & Technology, Jagran Lakecity University, Bhopal, INDIA

ABSTRACT

Movie Recommendation is more useful in our community life due to its strength in giving enhanced entertainment. Recommendation system can advise a collection of movies to users depend on their choice, or the popularities of the movies. while, a set of motion picture recommendation systems have been planned, mainly of these either cannot advise a movie to the presented users powerfully. In this paper we propose to solve the sparsity problem in movie recommendation system that has the ability to recommend movies to a new user as well as the others. It mines movie databases to collect all the important information, such as, popularity and attractiveness, required for recommendation. But in Recommendation system has many problems like sparsity , cold start , first Rater problem , Unusual user problem. K- mean clustering is the most successful method of Recommender System. K-means clustering also K-Means Clustering. The Algorithm K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

Keywords-- k-mean clustering, euclidean distance, k-mediod clustering**.**

I. INTRODUCTION

Recommendation system are used for many purposes. It is the type of filtering the information it means that it is used for predict the Rating of item given by the user. In Movie Recommendation system, we try to recommend movies to users based on their interests.

But in the field of movie Recommendation system we have many problems like cold start problem, sparsity problem etc. There are three points to focus for movie Recommendation system: -

Why: Movie Recommendation system is required because of movie information are overload.

Where: it is used in social site, box offices and all types of area like Bollywood, Hollywood etc.

What: it suggests movies to users for watching, Rating or purchasing the movie. if the users are interested.

II. BASIC PRELIMINARIES

Some methods given below are used to predict the items for users. Although there is large list of methods but we are discussing some methods which are of prime importance to movie recommendation.

A. Fuzzy c- mean clustering

Fuzzy clustering is a one of the type of clustering in which every data point can belong to more than one cluster.[4][10]

Clustering or cluster analysis involves transmission data points to clusters (also called buckets, bins, or classes), or homogeneous classes, such that items in the same class or cluster are as similar as possible, while items belonging to different classes are as dissimilar as possible. Clusters are identified via similarity measures. These similarity measures include distance, connectivity, and intensity. Different similarity measures may be chosen based on the data or the application.

Advantages

1) gives most excellent effect for overlap data set and moderately improved after that k-means algorithm.

2) nothing like k-means anywhere data point must completely be in the right place to one cluster hub here data point is assigned association to each cluster middle as a consequence of which facts end may be in the right place to supplementary subsequently single come together center.

Disadvantages

1) Apriori requirement of the number of clusters.

2) With lesser value of β we obtain the improved end result but at the expenditure of more numeral of iteration.

3) Euclidean detachment procedures can inequitably weight fundamental factors.

B. Gath – Geva Clustering

The Gath-Geva algorithm is an addition of Gustafson-Kessel algorithm that take the volume and density of the cluster into report.[5]

- The distance function is preferred ultimately comparative to the (un normalized) a posteriori opportunity, because a small distance way a

high probability and a big distance way a low probability of association.

- In contrast to FCM algorithm and the Gustafson-Kessel algorithm, the Gath-Geva algorithm is not based on an objective function, but is a fuzzification of statistical estimators. [6]

Gath –Geva Algorithm

The Gath-Geva Algorithm become new rational to local minima with growing complication. [7]

- Designed for dissimilar initializations of the prototype, the partition of the Gath-Geva algorithm can be very dissimilar.
- Present can be without difficulty occur floating-point overflow since of the exponential function. Therefore, it is suitable to use a customized exponential meaning that provides normal or linearly growing principles when the point of view faces an run over.[8,9]

III. PROBLEM IDENTIFICATION

Online shops today contain lot of items and users. In order to relate users to items they need association of user interest on particular items. But due to time or other constraints it is generally not possible to have enough ratings on particular items by users. By means of collaborative and other approach recommender systems normally produce locality of users by means of their profile.

If a client has to evaluate just scarcely any items, then it's pretty not easy to establish his experience and he/she might be connected to the incorrect area. Sparsity is the difficulty of need of information. In real world sparsity is very common. We often do not have enough ratings to make our highly efficient Recommendation Algorithm to work.

It is common in e-business supplies that even the majority active customers have purchase or rate a extremely incomplete proportion of products, when compare to the obtainable total. That lead to sparse user-item matrices, incapability to establish victorious neighbors and to finish, the age group of weak recommendations. As a result, techniques to reduce the sparsity in user-item matrices should be proposed.

Data sparsity is single of the biggest difficulty and excellence of recommendation is extremely dependent relative on sparsity of obtainable data. The main motive for occur sparsity problems are:

- The amount of items that contain be rate by the users would be too small. This can make our Recommendation algorithms fail.
- Similarly, the number of users who have rate one exacting item might be too small compare to the total no. of users connecting in the system. These situations provide rise to sparse ranking matrix.

A widely-used topic modeling algorithm is Latent Dirichlet Allocation (LDA). It is used in profiling documents. [15] This algorithm can also be used to solve some level of sparsity. But we have not included it in our methodology.

IV. METHODOLOGY

Figure 1 shows the process of solving sparsity problem in rating based Recommendation system.

1. In the first step the dataset will collect information from IMDb (Internet Movie Database), All the required information about the movie will be available.

All information of movie and user are presents; in the first step we can gather the all data set that requires for solving sparsity problem in movie recommendation system.

2. After that the process of Rating and Review are started we can generate the review and rating matrix and apply k- mean (Object clustering) clustering in both matrix . The k-mean clustering is simply solving clustering problem. It will make cluster of similar object but it has no any predefined classes. And classification of Reviews is based on good, bad and average comments of movies we can take the 29*100 matrix for Rating and 29*3 are matrix for Review.

3. Both data of matrix is convert into relational data using Euclidean distance. Euclidean distance used to find the distance between two points in Euclidean distance. then apply the harmonic mean for calculate the average set of number. after that we can apply the k-mean clustering for n*n matrix, if we have the same number of column and same number of row than we can apply k mediod clustering.

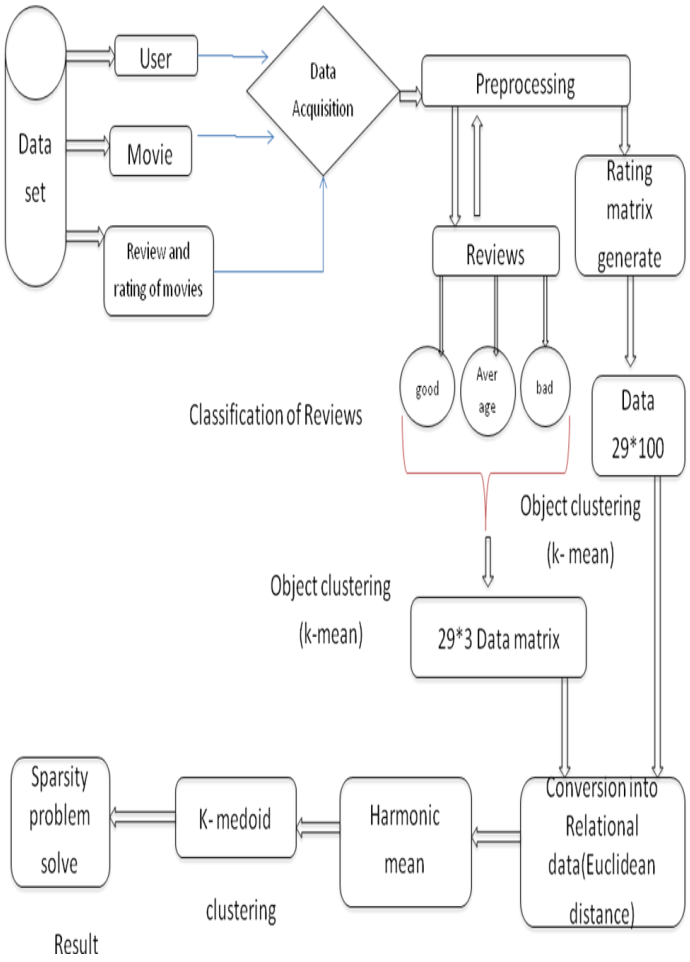

Fig1: The Process for Solving Sparsity Problem in Rating Based Movie Recommendation System

A. K- Mean Clustering

k-means is used for solving clustering problem. [11][13] It is the unsupervised leaning. No classes are defined previously. The process follows a straightforward and simple method to categorize a certain data set from side to side a definite amount of clusters (assume k clusters) predetermined Apriori. The main idea is to classify k centers, one for each cluster. These center should be located in a craftiness method since of different position cause different result. So, the enhanced choice is to put them as a group as potential far absent beginning each other. The next step is to get each point belong to a known data set and transmit it to the adjacent core. while no point is awaiting, the primary step is finished and an untimely cluster period is complete. on this spot we require to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be completed between the similar data set point and the adjacent novel center.[14] A disk has been generated. As a result of this disk we may observe that the k centers modify their position step by step until no extra change are completed or in other words centers do not go any other.

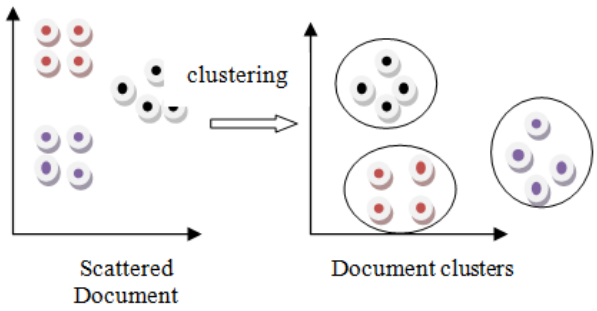

Fig2: K – Mean Clustering

Algorithm of K- Mean Clustering

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

B. Euclidean Distance

In terms of mathematics, the Euclidean distance is the distance between two points in Euclidean space. With this distance, Euclidean space makes a metric space. [12] The associated norm is called the Euclidean norm**.** In our project first loaded the Rate matrix then Rating matrix and applying Euclidean distance in both matrixes.

C. Harmonic Mean

In mathematics, the harmonic mean contains several types of average and Pythagorean mean. naturally, it is suitable for conditions when the normal of rates is required.

Two Number

The special case of just two numbers, x_1 and x_2, the harmonic mean can be written

$$H = \frac{2x_1 x_2}{X_1+X_2}$$

k-medoid is based on centroids (or medoids) calculating by minimizing the absolute distance between the points and the selected centroid, rather than minimizing the square distance. As a result, it's more robust to noise and outliers than k-means.

D. K - Mediod clustering

The k-medoids algorithm is a similar as a k-mean clustering. *K-means* and k-medoids algorithms are splits in to some parts (breaking the dataset up into groups) and both challenge to decrease the distance between points labeled to the center of the cluster.

k-medoid is a classical partitioning technique of clustering that divided the n object in to k cluster.

It is more robust to sound and outliers as well as k-means because it minimizes a sum of pair wise dissimilarities instead of a sum of squared Euclidean distances.

A k-mediod can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. i.e. it is a most centrally located point in the cluster.

Algorithm of K-mediod clustering

The most common realization of *k*-medoid clustering is the **Partitioning Around Medoids (PAM)** algorithm. PAM uses a greedy search which may not find the optimum solution, but it is faster than exhaustive search. It works as follow.

1. Initialize: *k* randomly select (without replacement) from the *n* data points as the medoids
2. Each data point associate with the closest medoid.
3. While the configuration cost decreases:
    1. For each medoid *m*, for each non-medoid data point *o*:
        1. Swap *m* and *o*, recompute the cost (sum of distances of points to their medoid)
        2. If the total cost of the configuration increased in the previous step, undo the swap

Other algorithms than PAM have been suggested in the literature, including the following Voronoi iteration method.

1. Select initial medoids
2. Iterate while the cost decreases:
    1. In each cluster, make the point that minimizes the sum of distances within the cluster the medoid
    2. Reassign each point to the cluster defined by the closest medoid determined in the previous step.

V. CHALLENGES AND ISSUES

A. Cold start

Its very complicated to give the recommendation to new customer as his profile is empty and he is not rated any item over the available item . this is called cold start problem. And this problem is solved by combination of k – mean clustering, k-mediod clustering and Euclidean distance and harmonic mean.

B. Scalability

When increase the number of customer and items ,the system require more number of processing the information of the users and items for recommendation . many number of resource are used for determining the user with similar teste ,goods and similar description. This types of problem is also solved by various types of method used in this paper.

C. Sparsity

In online shopping, there are the more number of users rated the few number of items over the total number of available items. Using another approaches like collaborative filtering and association retrieval. In this approach generally created neighbourhood of the user according to their profile .if the user evaluate the few number of item , it's the difficult to evaluate similar taste with users. sparsity is a problem they occur for lack of information.

VI. EXPERIMENTS AND ANALYSIS

Experiment Data:

We have implemented the solution in MATLAB. Some important findings were very positive. The method chosen and discussed above seem to solve the sparsity problem in this domain.

The dataset was collected by the IMDb. in this dataset consist of 29 movies and 100 users .and each users rating at least 7 movie. The sparsity degree is 99.23%.

The experiments are performed as follows:

- Initially, submit the data of movie from IMDb, and obtain the original Rating of movie results.
- Now, submit the original Rating of movie result for obtain the accurate rating of movie.
- Re-rank the Rating of movie results according to our algorithm.
- Compare the Rating of movie results with our algorithm.

Data Set

Original Movie Data Set

Actual data are given by the no. of users given the rating of movie, in this project we can divide in to three clustering of rating the range of 0 %to 33%, 34% to 65% and 66 to 100%. In this group we can easily classify the how no. of users rating the same data . in the actual data after observation the data give [3,13,13] clustering.

Object Data Set

Object data is classify by the k-mean clustering, k-means is used for solving clustering problem. It is the unsupervised leaning. No any classes are define previously. The process follow a straightforward and simple method to categorize a certain data set from side to side a definite amount of clusters (assume k clusters) predetermined Apriori. The main idea is to classify k centers, one for each cluster. These center should be located in a craftiness method since of different position cause different result.

Relational Data Set

Relational data are given by k-mediod clustering, The *k*-medoids algorithm is a similar as a k-mean clustering. *K-means* and *k*-medoids algorithms are splits in to some parts (breaking the dataset up into groups) and both challenge to decrease the distance between points labeled to the center of the cluster.

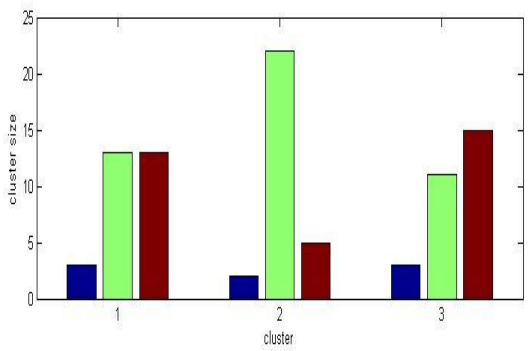

Fig 3: Graph of result of sparsity problem solved (Generated in MATLAB)

In the above figure graph shows the result of sparsity problem solved, in this graph x label shows the three no. of clustering and y label shows the cluster size . In the first three bar shows the actual data in second cluster shows the object data and third cluster shows the Relational data, the cluster of actual data are [3,13,13] and relational data are[3,11,15]. The comparison between both clustering first cluster are same, second cluster has 2 differant and 3rd cluster has 2 different. total different are 4. so the dataset contain 29 movies. So the actual data and relational data are adjacent value .so the k-mediod clustering is the better than the k- mean clustering. and 70.22% sparsity problem is solving.

VII. CONCLUSION

In this paper, we aimed to solving sparsity problem in rating based movie recommendation system and improve the performance of movie Recommendation system. We use the k- mean clustering, k- mediod clustering and combination of harmonic mean and Euclidean distance method to solving sparsity problem. The effectiveness of the approach was evaluated experimentally using data from IMDb Dataset. the experiment indicated that our approach solves the sparsity problem and achieved significantly better Recommendation quality then the other sparsity problem solving method. Our approach is used for solving sparsity problem and achieved significantly better movie recommendation. The volume of data will continue to increase over the time.

REFERENCES

[1] Yibo chen et al Progress of solving sparsity problem in recommendation system using association retrieval process in journal of computers vol 6 published on 9th September year 2011.

[2] Lalita sharma, anju gera et al. Progress of hybrid approaches to reduce the sparsity problem . Progress on Mtech. Scholar BSAITM faridabad. vol 6 on 8 july 2013.

[3] yu rong ,xiao wen, hong cheng, et al an monte carlo algorithm for cold start problem at International world wide web conference committee on April 7-11-2014.

[4] Mohammed mahmuda rahumen rahumen lecture, et al contextual recommendation system using multidimensional approch. International journal of intelligent information system august20,2013.

[5] Zuping liu sichuon et al. Recommendation algorithm based on user interest, advanced science and technology letters vol. 53, 2014.

[6] Manos papagelis, dimitris plexousakis Alleviating the sparsity problems of collaborative filtering using trust inferences Institutes of computer science , foundation for research and technology- hellas Years:2004.

[7] Andy yuanxue, jianzhong Qi, Solving the data sparsity problem in destination prediction University of Melbourne , Australia Year: 2013.

[8] Beau piccart, jan struf Alleviating the sparsity problem in collaborative filtering by using an adapted distance and a graph based method. IEEE computer technology Year:2007.
[9] Badrul sarwar, george karypis , joseph konstan Item based collaborative filtering recommendation algorithm . Department of computer science and engineering, University of Minnesota Year:2006 .
[10] Badrul sarwar, ,joseph konstan john riedl Using filtering agent to improve prediction quality in the grouoplen research collaborative filtering Department of computer science and engineering , University of minnesota in year 2008 .
[11] Chrsistian Desrosiers, George Karypis. Solving the Sparsity Problem: Collaborative Filtering via Indirect Similarities. Technical Report. Department of Computer Science and Engineering University of Minnesota 4-192 EECS Building 200 Union Street SE Minneapolis, MN 55455-0159 USA. 2008
[12] Zan Huang, Hsinchun Chen, et al. Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering. ACM Transactions on Information Systems, Vol. 22, No. 1, January 2004, 116–142. http://dx.doi.org/10.1145/963770.963775.
[13] Sanghack Lee and Jihoon Yang and Sung-Yong Park, Discovery of Hidden Similarity on Collaborative Filtering to Overcome Sparsity Problem, Discovery Science, 2007.
[14] Rong Jin, Luo Si, et al. Collaborative Filtering with Decoupled Models for Preferences and Ratings. CIKM '03, New Orleans, Louisiana, USA, November 3-8, 2003.
[15] Cheng-kang-Hsieh, et al. Immersive Recommendation: News and Event Recommendations Using Personal Digital Traces (2016) published in Proceedings of the 25th International Conference on World Wide Web on April 11–15, 2016 at Montréal, Québec, Canada.