AIVP ADVANCES IN IMAGE AND VIDEO PROCESSING
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 4, ISSUE 3
ISSN 2054-7412

Solving Sparsity Problem in Movie Based Recommendation System

Pratibha Bargah and Nitin Mishra
Rungta collage engineering & Technology Bhilai (c.g), India
pratibha23bargah@gmail.com, drnitinmishra10@gmail.com

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 a 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 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, Harmonic Mean.

1 Introduction

Recommendation system are use for many purpose, it is the type of filtering the information it means that it is used for predict the Rating of item given by the user [1].

In Movie Recommendation system, Recommend the movie for watching or Rating the movie by the user. But in the field of movie Recommendation system it has many problems like cold start problem, sparsity problem etc[2] .there are three points for movie Recommendation system: -

Why:- Movie Recommendation system are Required because of movie information are overload.

Where :- used in social site, box offices and all types of area like bollybood, hollybood etc.

What :- suggest item to users for watching, Rating or purchasing the movie.if the users are interested.

2 The Existing Ranking Methods

2.1 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[1]. The main idea is to classify k centers, one

DOI: 10.14738/aivp.43.1997
Publication Date: 23rd July, 2016
URL: http://dx.doi.org/10.14738/aivp.43.1997

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[2]. 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[8].

2.1.1 Steps of K- Mean Clustering

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

2.2 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[3]. 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 matrix .

In general, for an n-dimensional space, the distance is

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

2.3 Harmonic Mean

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

2.4 Two Number

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

$$H = \frac{2x_1x_2}{x_1 + x_2}.$$

In this special case, the harmonic mean is related to the arithmetic mean $A = \frac{x_1 + x_2}{2}$ and the geometric mean $G = \sqrt{x_1x_2}$ by

$$H = \frac{G^2}{A} = G \cdot \left(\frac{G}{A}\right).$$

In the above figure shows the process of solving sparsity problem in rating based Recommendation system. In the first step the dataset collect from IMDb(Internet Movie Database) ,the all require information of movie 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[4]. Afterthat 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 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. 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[5] .

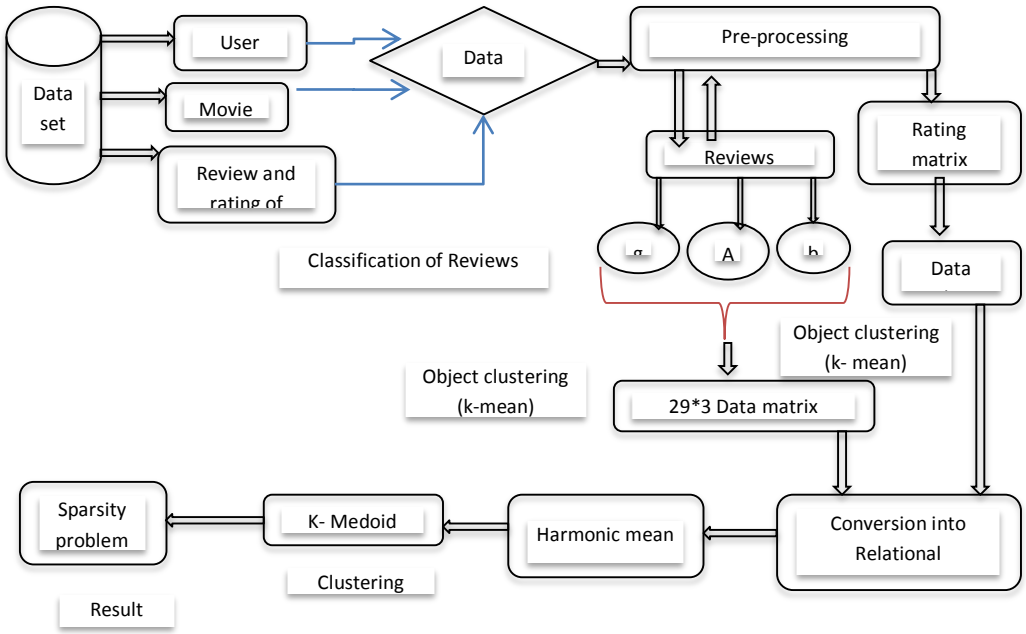

Figure 1: The Process for Solving Sparsity Problem in Rating Based Movie Recommendation System

In fig. shows steps of Solving Sparsity Problem In Recommendation System for movie recommendation system . in this steps show the dataset from IMDb. and using k- mean clustering for better movie Recommendation. In this steps it contain Review and Rating matrix. The Review matrix has 3 *29 matrix and Rating matrix contain 29*100 matrix. The Review are based on good comments, ba comments and average comments. and for generate the sparsity by applying the k- mean clustering. Afterthat we can apply Euclidean distance for converting Relational data . then for solving sparsity problem we apply Harmonic mean and for n*n Matrix you can apply the k- medoid clustering. Afterthat the sparsity problem are solved in Rating based movie Recommendation system. .

2.5 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[10].

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.

2.5.1 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.

3 User Query Intent and Storage of Tags

In the simplest technique used to reduce the sparsity of the user-item matrix, we simply insert a default rating, *d*, for appropriate items for which there exist no explicit ratings. ”Appropriate” is the key word here, meaning that it is wise to choose the matrix entries where the default ratings would be inserted. Nevertheless in where this techniqueis proposed. In recommendation system Yu Rong Xiao,Wen Hong Chenj the Chinese university of hongkong Year 2014b had done best performance.This authors already worked on Monte Carlo Method which is get working on future enhancement also

3.1 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.

3.2 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.

3.3 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 , its the difficult to evaluate similar taste with users. sparsity is a problem they occur for lack of information.

4 K- Mediod Clustering Algorithm

4.1 K- Mediod Clustering Algorithm

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.

4. Initialize: k randomly select(without replacement) from the n data points as the medoids
5. Each data point associate with the closest medoid.
6. 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.

3. Select initial medoids
4. 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.

5 Experiments and Analysis

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.

5.1 Data Set

5.1.1 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.

5.1.2 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.

5.1.3 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.

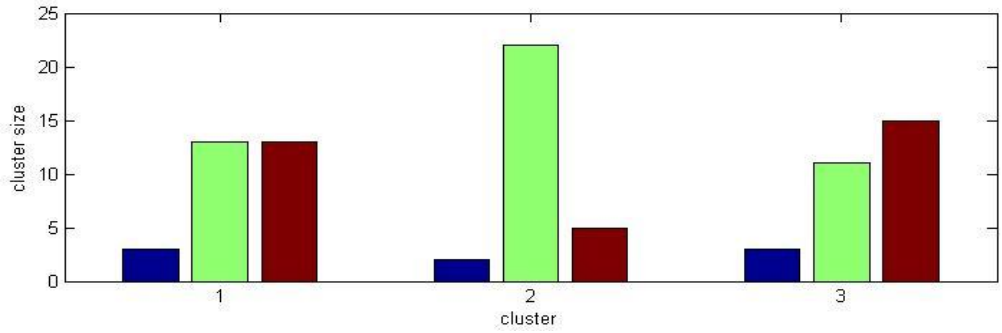

Figure 2:-Graph of result of sparsity problem solved

In this 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]. So the actual data and relational data are adjacent value. So the k-mediod clusteing is the better than the k- mean clustering.

6 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 probem. The effectiveness of the approach was evaluated experimently using data from IMDb Dataset . the experiment indicated that our approach solve the sparsity problem and achieved significantly better Recommendation quality then the other sparsity problem solving method.

REFERENCES

[1] Yibo chen et al progress of solving sparsity problem in recommendation system using association retrival .progress journal of computers vol 6 9september year:2011.

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

[3] yu rong ,xiao wen, hong cheng, et al an monte carlo algorithm for cold start problem International world wide web conference committee 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 intrest ,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 recommedation 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 gruoplen research collaborative filtering Department of computer science and engineering , University of minnesota 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] Liu Jianguo, Zhou Tao, et al. Overview of the Evaluated Algorithms for the Personal Recommendation Systems. Complex System and Complexity Science. 2009, Vol.6, No.3, 1-10

