ISBN 978-93-84422-71-4
Proceedings of 2016 International Conference on Image Processing, Production and Computer Science (ICIPCS'2016)
London (UK), March 26-27, 2016 pp.57-61

Neighbourhood-Based Collaborative Filtering Approach Using K-means Clustering

Ozge Makul[1], M. Cemil Aydogdu[1], Murat Aykut[1] and Murat Ekinci[1]

[1]Departmant of Computer Engineering Karadeniz Technical University Trabzon, Turkey

Abstract**:** *Nowadays the widespread use of technology caused trade to move into the internet environment. Therefore several methods have been developed for commercial to provide convenience to their users and to be superior against their rivals. One of the most common practices is recommendation systems. Recommendation systems are developed to provide convenience to users and suggests alternatives to their users. Collaborative Filtering (CF) is popular approach of recommendation system. These approach gives recommendations based on users' behaviors. This paper proposes to use clustering with the Neighbourhood based CF methods.In this approach, firstly similarities between users is computed by using user-item rating matrix is. Then K-means clustering algorithm is used to determine the neighbourhood by using these similarities. Weighted Mean is calculated by using neighbour similarities for prediction. The performance of the proposed approach is evaluated by comparing with the other neighbourhood based CF approaches in literature.*

Keywords: *recommendation systems, neighbourhood based collaborative filtering, pearson similarity, k-means clustering*

1. Introduction

Nowadays the use of internet is increasing and storage is not a critical issue no more. This situation increases amount of data which is stored on electronic environment and makes difficult to access and interpret them. To overcome this, recommendation systems are used to process user data, extract information and interpret. Recommendation systems try to find users' interest on item which users haven't seen it before by using user's behaviours to other items [1].

Recommendation systems are separated into different classes based on input data. The more common and successful classes are Content Based Filtering (CBF) and Collaborative Filtering (CF) [2, 3]. CBF techniques recommend users according to relations between items which are generated using items content. If there is a high relation between items, the users who are interested in one of the items are recommended the other one. And as for CF techniques, they recommend the most suitable item to active user by using user's interest similarities or gather remarks on specific item. The main idea of CF is user who have same remarks with his neighbour at the past having same remark in the future with high possibility.

The CF algorithm is composed of the following steps:

1) Active user notifies his interest by rating items.

2) System finds users which rate items similarly with active user.

3) System recommends items which active user doesn't rate but similar users rate to active user by using similar user's ratings.

The CF algorithm needs user-item rating matrix. The example of user-item matrix is as follows:

http://dx.doi.org/10.17758/UR.U0316019

TABLE I: Example of rating table

	Item1	Item2	Item3	Item4	Item5
User 1	4	2	1	5	**?**
User 2	2	5	**?**	3	4
User 3	**?**	4	2	5	3
User 4	5	?	**?**	4	2

The aim of CF algorithm is to find value of unknown cells of matrix which are shown as '?'. After finding the value of unknown cells, matrix is filled and recommendation is done.

CF techniques are separated into two types. These are Model Based CF and Memory Based CF. Model based technique is used for big database and generates model of all users' interest by using machine learning algorithm. Model is used for recommendation in this technique [4, 5, 6, 7, 8]. Memory Based technique uses users' ratings to compute similarity between users and items. This technique uses all users in database to compute similarity. The most common one of Memory Based CF is Neighbourhood Based CF [9]. Neighbourhood Based CF uses similarities between active users and active users' neighbours to make a recommendation by using user-item rating matrixes [10].

In this paper, we present hybrid CF approach by using Neighbourhood based CF and clustering. The main idea of this paper depends on Neighbourhood based CF. As first step similarities between active user and users are calculated. Pearson Correlation Coefficient approach (PCC) is used as a similarity measurement due to having an important success in literature [9]. Then by using similarities active user's neighbours are determined. Clustering is used to find active neighbours as a part of Neighbourhood based CF as distinct from other studies in literature. Because of success and easy to implement K-means are choosed for clustering (K=2). After applying K-means, neighbour cluster is determined in two clusters by using most similar user's similarity value. This point is also different from other studies in literature [8, 10, 11]. After all, predicted value is calculated for specific item by using ratings of neighbour users who rated this item from neighbour cluster. In this step Weighted Sum Method is used for prediction. Mean Absolute Error is computed for evaluation of system performance.

The rest of the paper is organized as follows. Section 2 gives proposed work with an introduction to similarities measurement, K-means and Weighted Sum shortly. The experimental results and conclusions are given in Section 3 and Section 4 respectively.

2. Proposed Method

In this study,there are three basic steps listed below

1) Similarities are calculated between active user and other all users with PCC.
2) Active user's neighbours are determined according to user's similarities by using K-means clustering method.
3) Predicted value of item is calculated using Weighted Average method.

2.1. Similarity Computation

First step of proposed method is similarity computation. Similarity value is used as a weight which is computed between users using ratings of specific item. While similarities are computing, ratings of users who rated specific item commonly are used [12]. There are several approaches for similarity computation. The most common ones are Pearson Correlation Coefficient and Cosine Similarity. In this paper, two approaches are used and PCC is choosed after comparison of results.

http://dx.doi.org/10.17758/UR.U0316019

2.1.1. Pearson Correlation Coefficient (Pearson Similarity)

Users can rate the items using different scale. So it causes not to rate items in common range. PCC computes similarity measure to normalize user's ratings as zero mean and unit variance. Because of using user's ratings, mean of user's ratings are projected common range. PCC computation is as follows:

$$sim(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1}$$

$sim(x,y)$ is Pearson similarity between active user x and user y, x_i and y_i are ratings of item i voting by user x and y, $\bar{x}$ and $\bar{y}$ are mean ratings of user x and user y, n is number of items which user x and user y rates commonly. PCC produces values between -1 and 1. As $sim(x,y)$ is close to 1, so users are similar.

2.1.2. Cosine Similarity

Cosine Similarity (CS) evaluates user's rating as a user vector. The cosine value of angle between two vector is a similarity of two vector in this approach. CS computation is as follows:

$$sim(x,y) = \frac{\boldsymbol{x}.\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \tag{2}$$

$sim(x,y)$ is CS between active user x and user y, x and y are vector which consist of user’s ratings. CS produes similarity values range of 0 and 1. As $sim(x,y)$ is close to 1, so users are similar as well as PCC similarity.

2.2. Neighbours Determination

Neighbourhoods determination is an important step of Neighbourhoods based CF. Neighbourhoods are determined by using user's similarities which are calculated with similarity measurement. The best determination allows to obtain the best prediction value at the end of the algorihtm. In this work, two different approaches which are threshold method and clustering are used for determination of the neighbours. The clustering method is proposed in this paper.

2.2.1. Threshold Method

In this method active user's neighbours whose similarity measure are above threshold value are determined. The threshold value varies according to similarity function. In this work, after many testing threshold value are defined divergently for both similarity measure.

2.2.2. Clustering Method

This method is proposed in this paper because of robustness of clustering. The main idea of this method is to find similar cluster as a neighbour cluster by using similarities of users. Because of this reason K-means clustering is used for determining neighbours. The algorithm of K-means is shown as follows [13]:

1) Initialize K cluster centroids by randomly choosing from data.

2) Compute Euclidean Distance between centroids and data and cluster data to the closest cluster centers by using distances.

3) Compute new cluster centroids after clustering.

4) Repeat steps 2 and 3 until centers remain stable.

The number of k was selected as 2 because of determination of users which are active user’s neighbour or not is enough for this study. After clustering of user, it is necessary to find best cluster as a neighbour in two clusters. For this purpose, the neighbours are defined as cluster whose label was same as user having maximum similarity value with active user. So using most similar user to find neighbour cluster incerase robustness of predicted value.

2.3. Prediction Computation

In this step, determining neighbour cluster's rates which was given to specific item are combined to make a prediction on item for active user. The higher similarities between active user and other user in neighbour cluster, the more contrubution occurs for prediction [12].

The common method in literature for prediction is Weighted Average Method (WAM). This method uses user's ratings (user's behaviours) and rating range to produce result [14,15]. So method obtains prediction close to reality. WAM computation is as follows:

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{m}\left(r_{u,i} - \bar{r}_u\right) * sim_{a,u}}{\sum_{u=1}^{m}|sim(a,u)|} \tag{3}$$

$\bar{r}_a$ is active user's ratings mean, $r_{u,i}$ is the rating of user u to item i, $\bar{r}_u$ is mean ratings of user u, m is number of active user's neighbour, $sim_{a,u}$ is similarity coefficient of between active user and user u.

3. Experimental Results

In this paper, Movilens database which belongs to GroupLens consisting of 943 users and 1682 films is used. There are one million ratings and users rate at least 20 films 1 to 5 in this database.

Mean Absolute Error (MAE) [8] which is statistical measure of accuracy is used as an evaluation metric. This metric computes system performance by measuring predicted value how close to actual value. MAE is calculated by obtain predicted results for users whose ratings are known. MAE computation is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|r_{u,i} - r'_u\right| \tag{4}$$

N is number of rated items, $r_{u,i}$ is actual rating value of user u for item i, r'_u is predicted value. The mean of MAE is computed for every user to obtain error rate of system. Results are obtained from four different approaches as Pearson similarity with threshold method (PST), Cosine similarity with threshold method (CMT), Pearson similarity with K-means (PSK), Cosine similarity with K-means (CSK) and are compared. These results are shown below.

TABLE II: MAE results

Method	MSE
PST	0.5449
CST	0.5765
PSK	0.5291
CSK	0.5314

When examining results in TABLE II, it is seen that PSK approach producing the least error rate of system. Results shows that clustering approach produces least MAE than threshold method. The reason is that all neighbours which have similar behaviours are determined by using K-means. However in threshold method, it is not guaranteed to determine all neighbours which have similar behaviours because of using only users whose similarites exceeds threshold value.

CF approaches are offered as a solution to situations which users don't rate items to predict using similar neighbours' ratings and problems of missing values. So this work can computed prediction values in these situations, too.

4. Conclusions

In this paper Neighbourhood based CF approaches are analyzed and compared. New clustering approach in neighbourhood determination step is used and obtained results compared other methods in literature. These results show proposed method’s robustness and success comparing to others.

5. References

[1] Firan C. S., Nejdl W., Paiu R., "The benefit of using tag-based profiles", Fifth Latin American Web Congress, Santiago de Chile, Chile, 31 October-2 November 2007.

http://dx.doi.org/10.1109/la-web.2007.13

[2] Yu X., Sun S., "Research on personalized recommendation system based on web mining", International Conference on E-Business and E-Goverment, DC, USA, 7-9 May 2010.

http://dx.doi.org/10.1109/icee.2010.95

[3] Pazzani M. S., "A framework for collaborative, contend-based and demographic filtering", Artifical Intelligence Review, 1999, 13, 393-408.

http://dx.doi.org/10.1023/A:1006544522159

[4] Breese J. ,Heckerman D., Kadie C., "Empirical analysis of predictive algorithms for collaborative filtering", 14th Conference On Uncertainty in Artificial Intelligence, Madison, USA, 26-30 July 1998.

[5] Basu C., Hirsh H., Cohen W., "Recommendation as classification: using social and content-based information in recommendation", Fifteenth National Conference on Artifical Intelligence, Monona Terrace, USA, 26-30 July 1998.

[6] Ungar L. H., Foster D.P., "Clustering methods for collaborative filtering", Workshop on Recommender Systems, 15th National Conference on Artificial Intelligence, Monona Terrace, USA, 26-30 July 1998.

[7] Sarwar B. M., Karypis G., Konstan J. A., Riedi J., "Application of dimensionality reduction in recommender system-A case study", ACM WebKDD Workshop at the ACM-SIGKDD Conference on Knowledge Discovery in Databases, Boston, Massachusetts, USA, 20-23 August 2000.

[8] Chee, S. H. S., Han, J. and Wang, K., "RecTree: an efficient collaborative filtering method, " in Proceedings of the 3rd International Conference on DataWarehousing and Knowledge Discovery, pp. 141–151, 2001.

http://dx.doi.org/10.1007/3-540-44801-2_15

[9] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J., "Item-based collaborative filtering recommendation algorithms", In Proceedings of the 10th international conference on World Wide Web, 285-295, 2001.

http://dx.doi.org/10.1145/371920.372071

[10] Su, X. and Khoshgoftar, M., "A survey of collaborative filtering techniques", Advances in artificial intelligence, 2009, 4.

http://dx.doi.org/10.1155/2009/421425

[11] Dakhel, G. M., and Mahdavi, M., "A new collaborative filtering algorithm using K-means clustering and neighbours' voting". In Hybrid Intelligent Systems (HIS), 2011 11th International Conference on IEEE, 179-184, 2011.

[12] Herlocker L. J., Konstan J. A., Borchers A., Riedl J., "An algorithmic framework for performing collaborative filtering", 22nd Annual International ACM SIGIR Conference, Berkeley, CA, USA, 15-18 August 1999.

http://dx.doi.org/10.1145/312624.312682

[13] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms", the textbook, IEEE Press & John Wiley, (First edition, November 2002; Second Edition, August 2011).

[14] Adomavicius G, Tuzhilin A. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions". IEEE Transactions on Knowledge and Data Engineering, 17(6), 734-749, 2005.

http://dx.doi.org/10.1109/TKDE.2005.99

[15] Ding Y, Xue Li, Orlowska ME. "Recency-Based Collaborative Filtering". Proceedings of the 17th Australian Database Conference, ADC ‘06, 49, 99-107, 2006.