Scalable Hierarchical Recommendations Using Spatial Autocorrelation

Ayushi Dalmia * Joydeep Das +
Prosenjit Gupta ++ Subhasish Majumder ++ Debarshi Dutta **

*International Institute of Information Technology, Hyderabad, India

+The Heritage Academy, Kolkata, India ++Heritage Institute of Technology, Kolkata, India

**University of Southampton, Highfield Campus, Southampton, UK

ABSTRACT

Collaborative Filtering (CF) is one of the most successful and widely used approaches behind Recommendation Algorithms. In this paper we deal with the *scalability* issue which is one of the main challenges to the CF process. In CF, finding similarity amongst N users is an $O(N^2)$ process. In this work, we propose a *Quadtree* based user partitioning technique that partitions the entire users' space into regions based on the location. We develop a Spatially Aware Recommendation Algorithm, where the Recommendation Algorithm is applied separately to each region and therefore allows us to reduce the quadratic complexity associated with the CF process. While recommending, our approach uses the location and ratings of the target user as well as the rating history of the other users in that region. The proposed work tries to measure the Spatial Autocorrelation indices, such as Geary's index in the regions or cells formed by the Quadtree decomposition. One of the main objectives of our work is to reduce the running time as well as maintain a good quality of recommendation. This approach of recommending using the decomposition method makes our algorithm feasible to work with large datasets.

PRELIMINARIES

Quadtree: A Quadtree is a tree data structure which is used to partition a two-dimensional space and has the following features:

- Each cell decomposes space into four adaptable cells (regions).
- Each cell (or bucket) has a maximum capacity. When maximum capacity is reached, the bucket splits.
- The tree directory follows the spatial decomposition of the Quadtree.

Spatial Autocorrelation: Spatial Autocorrelation [2] measures the co-variance of properties within geographic space and it deals with both attributes and locations of spatial features. A commonly used measure of Spatial Autocorrelation is Geary's index (c)

Collaborative Filtering: Collaborative Filtering (CF) is based on the principle of finding a subset of users who have similar taste and preferences to that of the active user. The idea is that given an active user u, compute her n similar users $\{u_1, u_2, \cdots, u_n\}$ and predict u's preference based on the preferences of $\{u_1, u_2, \cdots, u_n\}$.

OUR CONTRIBUTION

The motivation of our work comes from the property of *Preference Neighborhood* which suggests that users from a spatial region (e.g., locality) prefer items (e.g., movies, destinations) that are noticeably different from items preferred by users from other, even adjacent regions. Our goal is to partition the space into smaller manageable regions and in turn reduce the overall running time without sacrificing recommendation quality.

The Decomposition Algorithm: The space partitioning algorithm finds the spatial correlation value for a region by using Geary's Index, and then applies some splitting criteria to split the region into four regions. This process is continued as long as the splitting criterion is satisfied. The scheme is detailed as follows:

Algorithm Quadtree_Decomposition

1. Represent user location (city) as coordinates (longitude-latitude).
2. Find the spatial autocorrelation value of the entire region (level-0 of the tree).
3. Build the tree using splitting criteria.
   - 3.1 If correlation is good and number of users in the region is low (below the threshold limit), we do not split the region.
   - 3.2 If the number of users in a region is high (above the threshold limit), then irrespective of the correlation value we split the region.
   - 3.3 If both the number of users and correlation value of a region is low (below threshold limit), we apply look ahead criteria, and consequently split(or do not split).
4. Repeat steps 3 and 4 for each of the regions (tree nodes) as long as the splitting criterion is met.

The Recommendation Algorithm The Recommendation Algorithm emphasizes on applying the CF algorithm separately to each region, and in turn reduce the running time. The algorithm is briefly described below:

Algorithm Recommend_Item

1. Select a user for recommendation.
2. Identify the location (longitude and latitude) of the user.
3. Map the user in the exact region (node) of the Quadtree according to his/her location.
4. Find a subset of users in the region who share similar preferences for items with the active user. Select *top-10* similar users.
5. Select top 5 highly rated item from each of these *top-10* users to from a *top_set* of 50 items.
6. Recommend *top-10* items from the *top_set* by averaging the rating of the items in the region.

RESULTS

We have tested our decomposition algorithm on the Book-Crossing dataset and the MovieLens Dataset to validate our scheme with the following different threshold parameters: User Threshold u_1 and u_2, Correlation Threshold C_T and Item Threshold f. $n_1$and n_2 represent the minimum and maximum number of users in a region respectively, C_T is the minimum correlation value a region must have for not being decomposed further and f is the fraction of the total number of items used for correlation calculation. We report the summary of the results of the experiments performed on the Book-Crossing Dataset and MovieLens Dataset in Tables 1 and 2 respectively. Here we have shown the average correlation values across all the regions using different threshold values. We can observe that the decomposition algorithm produces better results (in terms of correlation values) when the fraction of items(f) is less. For testing our recommendation algorithm, we randomly split the user ratings into two sets - observed items (80%) and held-out items (20%). Ratings for the held-out items were to be predicted. We use two commonly used metrics for evaluating the prediction accuracy of traditional collaborative filtering algorithms are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) while recommendation quality is measured using the Recall metric [1] . We report the results of the Recommendation Algorithm performed on the Book-Crossing Dataset and Movie Lens Dataset in Tables 3 and 4 respectively. Here we have reported the Recall, MAE and RMSE values averaged over all the regions.

Table 1: Summary of Spatial Decomposition with various threshold values [*BookCrossing Data*]

n_1	n_2	CT	f	No. of Regions	% < 0.75 (Average)	% < 1.0 (Average)
1000	3000	0.5	0.250	12	98.65	98.87
1000	3000	0.5	0.125	12	99.07	99.33
1000	5000	0.5	0.250	8	99.03	99.22
1000	5000	0.5	0.125	8	99.29	99.55

Table 2: Summary of Spatial Decomposition with various threshold values [*MovieLens Data*]

n_1	n_2	CT	f	No. of Regions	% < 0.75 (Average)	% < 1.0 (Average)
500	1000	0.5	0.250	19	29.38	54.59
500	1000	0.5	0.125	19	31.64	50.87
1000	3000	0.5	0.250	4	22.38	49.99
1000	3000	0.5	0.125	4	27.46	53.37

Table 3: Summary of Recommendation Results with various Threshold Values [*BookCrossing Data*]

n_1	n_2	CT	f	No. of Regions	*Recall* (Average)	*MAE* (Average)	*RMSE* (Average)
1000	3000	0.5	0.250	12	0.9530	0.7632	0.8544
1000	3000	0.5	0.125	12	0.9175	0.7620	0.8980
1000	5000	0.5	0.250	8	0.9540	0.7775	0.8797
1000	5000	0.5	0.125	8	0.9125	0.7779	0.9363

Table 4: Summary of Recommendation Results with various Threshold Values [*MovieLens Data*]

n_1	n_2	CT	f	No. of Regions	*Recall* (Average)	*MAE* (Average)	*RMSE* (Average)
500	1000	0.5	0.250	19	0.8824	0.4435	0.6147
500	1000	0.5	0.125	19	0.9041	0.4447	0.6274
1000	3000	0.5	0.250	4	0.8731	0.4941	0.6856
1000	3000	0.5	0.125	4	0.9008	0.4982	0.6956

CONCLUSION

In our proposed Spatially Aware Recommender System we have employed a splitting technique to first divide the locations based on the correlation value and then we have mapped each user to a particular location according to the split criteria. Experimental analysis using real datasets shows that our model is efficient and scalable. Further, it provides quality recommendations and also minimizes the computations over irrelevant or less significant data to a large extent without degrading the efficiency of the Recommender System. Online experimentation of the split algorithm and the Recommendation Algorithm will be the focus of our future work.

REFERENCES

[1] B. Bhasker and K. Srikumar. Recommender Systems in e-Commerce. McGraw-Hill Education, 2010.

[2] C. P. Lo and A. K. W. Yeung. Concepts and Techniques of Geographic Information Systems. Prentice Hall, 2007.