CSE 5400 Interdisciplinary CS

Assignment 3:Product Recommendation

Instructor:Dr.Chan

Student name:Lingfeng Zhang

ID:9017490960

In this topic,firstly we studied some relative ideas about how a Search Engine ranks the advertisement while the user inputs the key words.I understand that the advertisement which has a high ranking is not contributing by the bid which is offered by the advertiser,but depends on the advertisement's relevance,Keywords relevance and the landing page experience.So that,using these information we can give a score to the advertisement,and a higher score advertisement will be given a advanced position on the web page.

Secondly,we studied about the Product Recommendation problem.We discussed several algorithms to predict the score of the movie for particular customer.And also learned the concept of the Root Mean Square Error(RMSE).While back to the implementation,in order to reduce the time complexity,we can converse the ascii file to the binary file.For reducing the memory complexity,we can use offset and array to store the main data.In the assignment,we will use two kinds of algorithms to calculate the rating for each customer.One is called intersection,which is using the superset,subset and disjoint set to predict the score.Another algorithm is called K-Nearest-Neighbor,which will depend on the nearest k similarity users to calculate the score.

A. Discuss the two key differences of the two algorithms.

i The first difference

1. For intersection algorithm,if a customer has superset,we must use his superset;if the customer dose not have a superset but subset,we use his subset.However,for K-Nearest-Neighbor algorithm,we don’t need to check whether a set is the superset or subset for the customer.What we do is to use all his neighbors,which means the one has at least one common watched movie with him,to calculate the similarity.

2. In this difference,I think the K-Nearest-Neighbor algorithm will have a better predictive performance than the intersection algorithm.Because in the intersection algorithm,for a customer,once he has a superset,he will use the superset to predict.This will cause the problem that he may have many subsets and each subset maybe have a plenty of common movies with him.But he can not use these data.This will lose the accuracy of the prediction.However,back to the K-Nearest-Neighbor algorithm,the customer will use all his potential neighbors.In this case,he will have lots of data and increase the

predictive performance by using these data.

ii The second difference

1. For intersection algorithm,a customer will use all his supersets or subsets to calculate the similarity,and use the similarity and linear equation to predict the rating.However,the K-Nearest-Neighbor just choose the top k highest similarity users to calculate the rating.

2. In this difference,I think the intersection algorithm has a better predictive performance than the K-Nearest-Neighbor algorithm.Because the intersection algorithm use all of supersets or subsets,but the KNN algorithm just use k sets.So the intersection algorithm is more accurate than the other.

B Compare the two algorithm

i RMSE performance

For toy-rating:

intersection algorithm gets the RMSE=1.0488

KNN gets the RMSE=1.0488

For nf-rating:

intersection algorithm gets the RMSE=1.1255

KNN gets the RMSE=1.1201

ii Time/speed

For intersection algorithm,we assume that there are N customers in the rating file,and the average movies for each customer is M.

1. we will visit all the rating file to figure out which one is a superset or subset or disjoint set.There will be (N-1)*M times.

2. For each chosen user,we will compare the common movies and figure out the distance and the similarity.So there will be M times for each one.And in the worst situation,this customer maybe have N-1 supersets or subsets.Which means we need to calculate M*(N-1) times.

3. After we get all the users' similarity,we will use the Linear Sum equation to calculate the rating.If in the worst situation,we may have (N-1) times to rate the score.So the total time will be M*(N-1)+M*(N-1)+(N-1). So that the time complexity will be O(M*N).

For the KNN algorithm,we use the same structure as the intersection algorithm.So that,we the step 1 and step 2 are the same as the intersection algorithm.So we have (M*N+M*N) to calculate the similarity.But in KNN algorithm,we just choose the top K highest similarities.So the total time will be M^2*N+M*N+K,and the time complexity will be O(M*N).

ii Space/memory

For the intersection algorithm,firstly we assume that there are N customers and the average movies for each customer is M.So that we have M*N space to store the main data for the rating file.In order to store the offset array,we need another N space.So the total space is M*N+N,and the memory complexity is O(M*N).

For the KNN algorithm,we use the same structure to store all the data.So the memory complexity is also O(M*N).