ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

Precise Recommendation System for the Long Tail Problem Using Adaptive Clustering Technique

S.Jeyshirii [1], Dr.T.K.Thivakaran[2]

P.G. Student, Department of Computer Science & Engineering,Sri Venkateswara College of Engineering , Chennai, Tamilnadu, India[1]

Professor, Department of Computer Science & Engineering,Sri Venkateswara College of Engineering , Chennai, Tamilnadu, India[2]

ABSTRACT*:* The Total Sale of large number of non-hit items is called “the long tail”. Long Tail Problem is one major issue in providing effective recommendation system; since long tail problem have only a few ratings. To solve this issue the proposed system identifies the clustering of items according to their popularity, It is identified that tail items are clustered based on the ratings of clustered groups while the head items are based on the ratings of individual item or groups. This method is applied to movie lens data sets and the results are compared with those of the non grouping and fully grouped methods in terms of recommendation accuracy and scalability. The results show that by implementing proposed adaptive clustering technique it reduces the recommendation error rates for the tail items while maintaining reasonable computational performance

KEYWORDS*:* Adaptive Clustering, Recommender Systems, Long Tail Problem.

I. INTRODUCTION

A. Long Tail Problem:

The term long tail has gained popularity in recent times as describing the retailing strategy of selling a large number of unique items with relatively small quantities sold of each-usually in addition to selling fewer popular items in large quantities. Anderson elaborated the concept of long tail. The distribution and inventory costs of businesses successfully applying this strategy allow them to realize significant profit out of selling small volumes of hard –to-find items to many customers instead of only selling large volumes of a reduced number of popular items. The total sale of this large number of “non-hit items” is called “the long tail”. The long tail concept has found some ground for application, research, and experimentation. It is a term used in online business, mass media, micro-finance, user-driven innovation, and social network mechanisms economic models, and marketing a frequency distribution with a long tail has been studied by statisticians since at least 1946.The term has also been used in the finance and insurance business for many years.

1) Objective:

- To address the long Tail Recommendation Problem (LTRP) in the recommendation system in order to reduce the error rate.

- To overcome the problem of insufficiency of tail items in Each Item (EI) method and to overcome the problem of excessive items of head items inTotal Clustering (TC) method by implementing adaptive proposed technique.

2) *Clustering*: Clustering can be considered the most important unsupervised learning problem so as every other problem of this kind it deals with finding a structure in a collection of unlabeled data.A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way.A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

3) *Recommended Systems:* Seek to predict the 'rating' or 'preference' that user would give to an item such as music, books, or movies or social element they had not yet considered, using a model built from the characteristics of an item content based approaches or the user's social environment collaborative filtering approaches. Recommender systems have become extremely common in recent years. Fig 1 refers the workflow diagram of the recommender system.

B. *Adaptive Clustering*

Adaptive Clustering method clusters the item with other similar items when it has only small amount of data, but groups to lesser extent or does not group when it has considerable amount of data.
The main features of adaptive clustering are

- It provides only the necessary amount of data to both the head and tail items and does not exceed this amount.
- It builds predictive rating models for EI(unlike TC method) by providing more data to the tail items(unlike the EI method)
- One item can be clustered into several different groups redundantly as similar items.

II. Related Work

Research on recommender systems has focused on the problem of estimating ratings for items that have not been seen by the user. Since recommender systems have become an important research topic in this e-commerce era, there are many studies proposing various recommending methods, such as the collaborative filtering (CF) approach content-based recommendations and hybrid methods. Data mining techniques are also used with recommender systems in many cases, since prediction abilities are easily applicable to ratings estimation for recommendation with or without using the profile information of customers and items. There are some other previous studies proposing hybrid data mining techniques and other methods.

The long tail problem in the context of recommender systems (LTRP) has been addressed previously in other studies. The impact of recommender systems on sales concentration and develop an analytical model of consumer purchases that follow product recommendations provided by a recommender system. The recommender system follows a popularity rule, recommending the bestselling products to all consumers. In their study, the process tends to increase the concentration of sales. As a result, the treatment is somewhat akin to providing product popularity information. Another related problem is the cold start [25], since our approach can be viewed as a solution to the cold start problem for the items in the long tails that have very few ratings. A popular solution to the cold start problem utilizes content-based methods when two items with no or only a few ratings are inferred to be similar based on their content .To improve the prediction accuracy of CF in which items are divided into smaller groups, and existing CF algorithms are applied to each group category separately. In a previous work related to this study, Park and Tuzhilin [22] use clustering ideas to solve the LTRP as well by cutting the item set into the head and the tail parts and apply the conventional clustering technique called expectation- maximization [32] in the tail part. This method is referred to as the clustered tail (CT) recommendation method. However, the CT method does not adaptively cluster the items according to their number of ratings (unlike the proposed AC method in this study) and needs to determine the head/tail splitting points manually. . Cluster models can perform much of the computation offline, but recommendation quality is relatively poor. To improve it, it's possible to increase the number of segments, but this makes the online user–segment classification expensive. Search-based models build keyword, category, and author indexes offline, but fail to provide recommendations with interesting, targeted titles. They also scale poorly for customers with numerous purchases and ratings.

III.PROBLEM STATEMENT

Many recommender systems ignore unpopular or newly introduced items, having only a few ratings and focusing on those items with enough ratings to be of real use in the recommendation algorithms. Alternatively, such unpopular or newly introduced items can remain in the system but require special handling.

C. Existing System Architecture

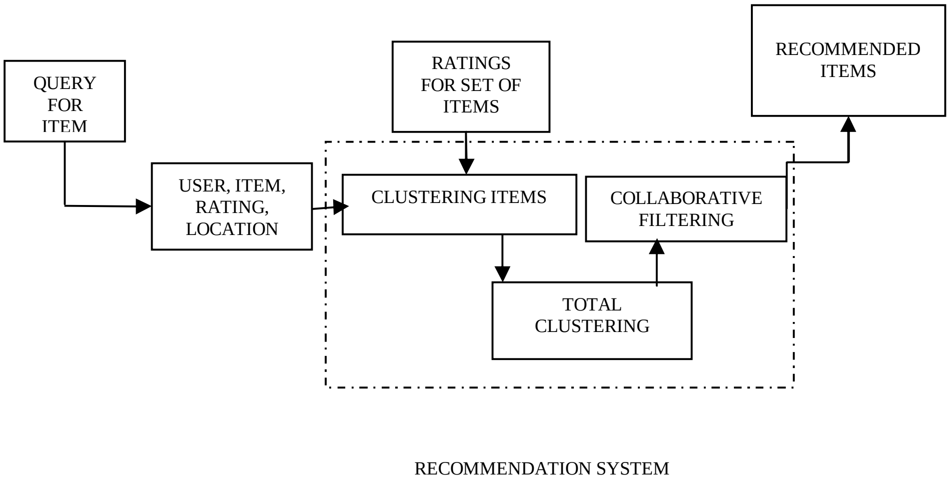


Figure1 Workflow of Existing Recommender System

D. Methods in Existing System

Two conventional recommender systems are the each item (EI) method and the total clustering (TC) method.

1) *EI Method*: The EI method builds rating-predictive models for each individual item, thus resulting models are highly customized for EI.
2) *Total Clustering (TC) Method*: To solve the long tail recommendation problem (LTRP) that arises in the EI method, we next cluster the whole item set into different groups and build rating-predictive models for the resulting group using more data than with the EI. We call this method the TC recommendation method. The error rates of the TC method are significantly lower than the basic EI method in many cases, especially for the

items in the long tail. In other words, the TC method does not have the LTRP. However, the TC method has the limitation that overly increases the data size for the head items, which already have an adequate amount of data. This excessive data impede the scalability of the TC method without significant performance improvement in the head.

E. Drawbacks

- EI method often has the long tail problem because of lack of data to build good predictive models in tail items. TC method has the limitation that overly increases the data size for the head items, which already have an adequate amount of data.

IV. PROPOSED SYSTEM

A. *Adaptive Clustering*

A new recommender method called the adaptive clustering (AC) recommendation method that adaptively groups items according to their popularities. Popularity is defined as the number of ratings provided by customers for that item. In other words, if the item has only a small amount of rating data, then the AC method clusters it with other similar items more intensively; on the other hand, if it has a large amount of data, then the AC method clusters it to a much lesser extent. The suggested AC method solves the LTRP in the sense that the error rates of the items, especially in the tail, are significantly lower than those for the basic EI method without overly increasing the amount of data for the head items. Moreover, the Adaptive Clustering (AC) method builds a more customized predictive rating model of EI than the TC method.Fig 2 represents the proposed architecture of recommender system.

1) *Advantages:* AC method adaptively clusters the items according to their popularities overcoming the problem of EI and TC method.
2) *Architecture*:

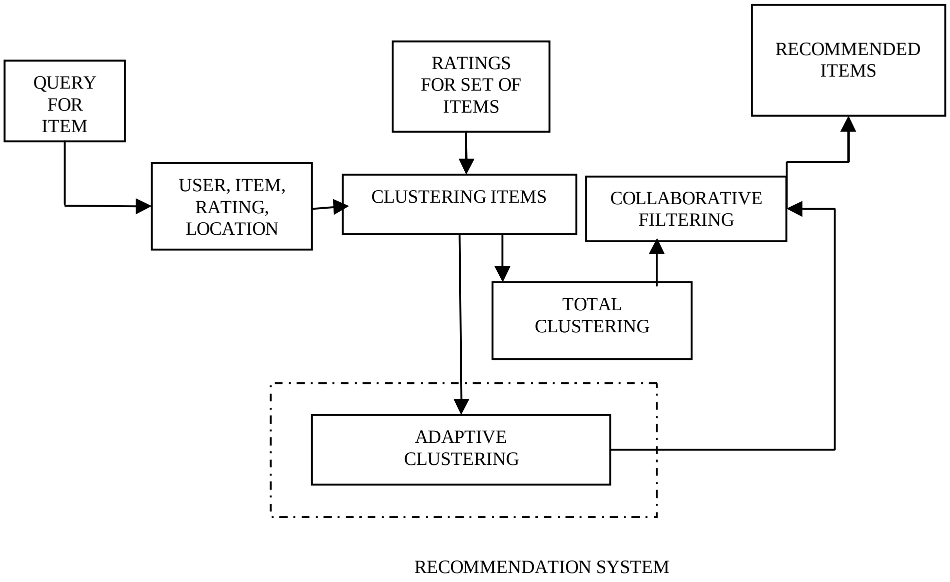


Figure2 Proposed Architecture of recommender system

IV. Each Item Method

A. *Description*

The EI recommendation method builds data mining models for each individual item i in I to estimate unspecified ratings for i. In other words, the EI method does not group an item with the other similar items at all and builds predictive models only by using the data in EI. For example, in the case of the Movie Lens data set, the EI method builds a predictive model for each of the 841 movies using the ratings of each particular movie. The main problem with the EI recommendation method is that only a few ratings are available in the long tail, so the predictive models for the tail items are learned from only a few training examples using the EI method.Fig 3 shows the rating prediction model for the targeted movie and user which represents the user defined collaborative count to predict the similarity and neighbor items to provide original rating and fig 4 shows the required graph for each item method.

B. *Algorithm*

1) *Input*:
   Target Movie
   Target User
   CF Count – user defined
2) *Output*:
   Predicted Rating

Original Rating
Long tail

3) *Process:*

- Select co-rated users
- Select the rating (Rit) for the target movie (t) provided by the co-rated users
- Select the rating (Rir) for the remaining movies (r) provided by the co-rated users (m)
- Calculate similarity as follows

$$sim(t,r) = \frac{\sum_{i=1}^{m} R_{it} R_{ir}}{\sqrt{\sum_{i=1}^{m} R_{it}^{2} \sum_{i=1}^{m} R_{ir}^{2}}}$$

Where

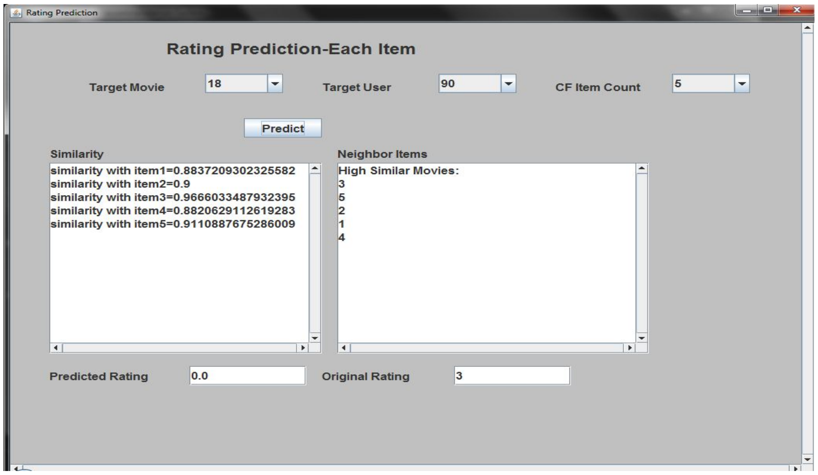
- Rit is the rating of the target item t by user i,
- Rir is the rating of the remaining item r by user i, and
- m is the number of all rating users to the item t and item r.
- Sim (t, r) – similarity between target item and remaining item
  - The rating of the target user u to the target item t is as following:

$$P_{ut} = \frac{\sum_{i=1}^{c} R_{ui} \times sim(t,i)}{\sum_{i=1}^{c} sim(t,i)}$$

Where

- Rui is the rating of the target user u to the neighbour item i,
- sim(t, i) is the similarity of the target item t and the neighbour it user i for all the co-rated items, and
- m is the number of all rating users to the item t and item r.
- Put – Predicted rating for the target item for the target user
- neighbor item – high similar item to target item

C. *Screen Shorts*



Rating Prediction

Rating Prediction-Each Item

Target Movie 18 Target User 90 CF Item Count 5

Similarity

similarity with item1=0.8837209302325582
similarity with item2=0.9
similarity with item3=0.9666033487932395
similarity with item4=0.8820629112619283
similarity with item5=0.9110887675286009

Neighbor Items

High Similar Movies:
3
5
2
1
4

Predicted Rating 0.0 Original Rating 3

Figure3 Each item Method

Fig 3 shows the rating for the targeted movie and user by setting the collaborative count which user can prefer from the predictive model for similar and neighbor items is calculated to predict the original rating and similarly fig 4 shows the reduction of long tail.

D.Graph

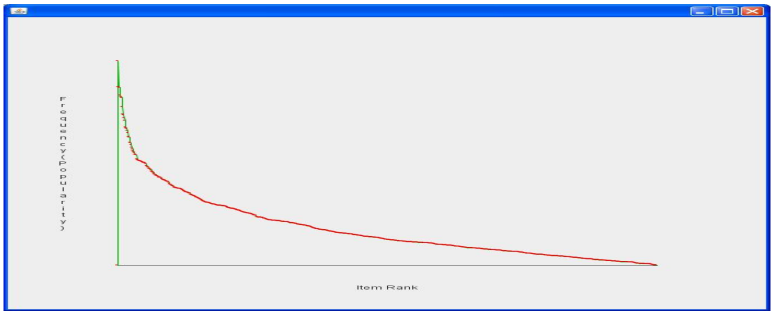


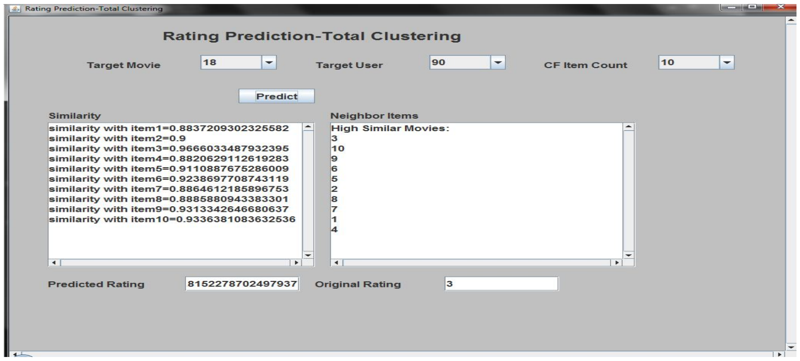
Figure4 Graph for Each Item Method

V. TOTAL CLUSTERING METHOD

The LTRP problem is caused by a lack of data to build good predictive models in the tail, and therefore, clustering items can be a reasonable solution. The TC recommendation method clusters the whole item set I into different groups by applying conventional clustering methods such as k-means clustering and building rating-predictive models for each resulting group. For example, in the case of the Movie Lens data set, the TC method clusters 841 movies into groups using the k-means clustering method and builds a predictive model.

It is concluded that the TC method can significantly outperform the EI method by providing more data to tail items. However, it attains this achievement inefficiently by providing an excessive amount of training data to the head items without performance improvements. Fig 5 shows the predicted rating for the targeted movie and user by setting the collaborative filtering count constant to provide original rating from the rating prediction model and fig 6 shows the required graph for the total clustering method.

A Screen Shorts


Rating Prediction-Total Clustering

Rating Prediction-Total Clustering

Target Movie 18 Target User 90 CF Item Count 10

Similarity

```
similarity with item1=0.8837209302325582
similarity with item2=0.9
similarity with item3=0.9666033487932395
similarity with item4=0.8820629112619283
similarity with item5=0.9110887675286009
similarity with item6=0.9238697708743119
similarity with item7=0.8864612185896753
similarity with item8=0.8885880943383301
similarity with item9=0.9313342646680637
similarity with item10=0.9336381083632536
```

Neighbor Items

```
High Similar Movies:
3
10
9
6
5
2
8
7
1
4
```

Predicted Rating 8152278702497937 Original Rating 3

Figure 5 Total Clustering Method

A. *Algorithm:*

1) *Input*:
Target Movie
Target User
CF Count - 100

2) *Output:*
Predicted Rating
Original Rating
Reduced Long tail compared to Each Item

3) *Process*:
- Select co-rated users
- Select the rating (Rit) for the target movie (t) provided by the co-rated users
- Select the rating (Rir) for the remaining movies (r) provided by the co-rated users (m)
- Calculate similarity as followed by the steps in the Each Item method.

B) *Graph*

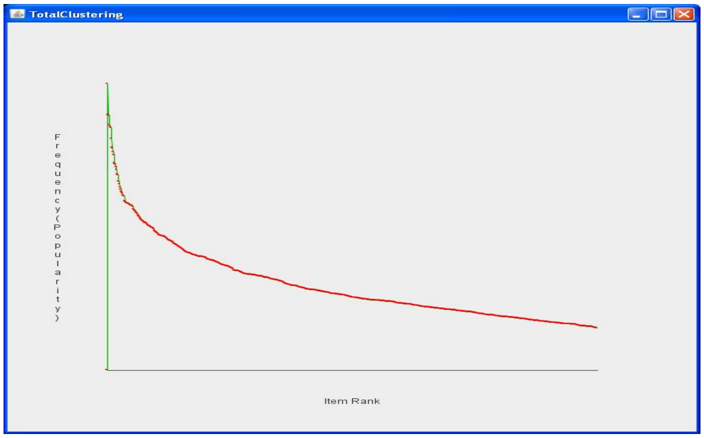


Figure 6 Graph for Total Clustering Method

VI. ADAPTIVE CLUSTERING

A new recommender system called the AC method that clusters items according to their popularities. The suggested method clusters an item with other similar items one by one until the resulting group size reaches the criterion number of ratingα.

The AC method clusters the item with other similar items when it has only a small amount of data, but groups to a lesser extent or does not group at all when it has a considerable amount of data. In the case of the Movie Lens data set, all movies from that data set are ordered based on the popularity for each movie. Then, the popularity of each movie is compared with the criterion number of ratingsα.

If it is larger thanα, then the AC method does not apply any clustering method; instead, it keeps the basic EI approach. However, if it is smaller thanα, then the AC method clusters the movie with other similar movies one by one until the resulting group size reaches α. After that, the AC method builds rating predictive models using the resulting group for EI.

Fig 7 represents the rating for the targeted movie and user by setting a criterion number to predict the original rating from the predictive model and fig 8 shows the required graph for the adaptive clustering technique then table1 shows the comparison about the three different methods and represents the count of the popular and unpopular items and table 2 represents how the popularity and rank is increased when compared it to the existing methods.

A) Algorithm

1) *Input*:

   Target Movie
   Target User
   CF Count – 100 (popularity < 200)
   CF Count – 75 (popularity > 200 &&< 400)
   CF Count – 50 (popularity > 400 &&< 600)
   CF Count – 25 (popularity > 600)

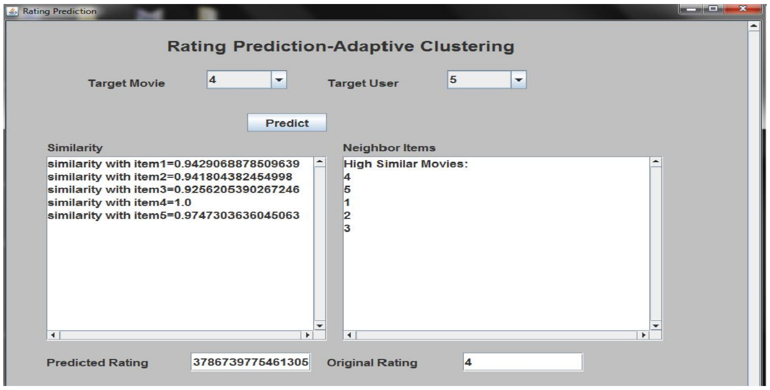
2) *Output*:

   Predicted Rating
   Original Rating
   Reduced Long tail compared to Each Item and Total Clustering

3) *Process*:

   - Select co-rated users
   - Select the rating (Rit) for the target movie (t) provided by the co-rated users
   - Select the rating (Rir) for the remaining movies (r) provided by the co-rated users (m)
   - Calculate similarity as followed by the steps in the Total Clustering Method.

B) Screen Shorts



Target Movie: 4
Target User: 5

Similarity

```
similarity with item1=0.9429068878509639
similarity with item2=0.941804382454998
similarity with item3=0.9256205390267246
similarity with item4=1.0
similarity with item5=0.9747303636045063
```

Neighbor Items

```
High Similar Movies:
4
5
1
2
3
```

Predicted Rating: 3786739775461305
Original Rating: 4

Figure7 Adaptive Clustering Method

Figure7 shows the original rating of the targeted movie and user by setting the criterion number to find the original rating from the predictive model .Fig 8 shows the perfect reduction of tail to solve the problem.

C) Graph

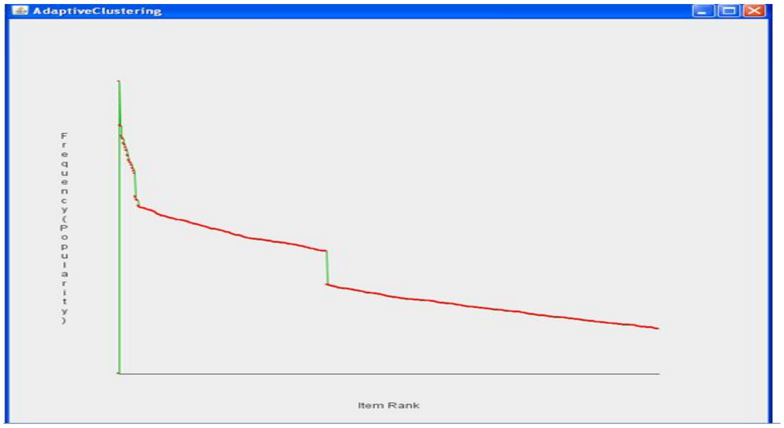


Figure 8 Graph for Adaptive Clustering Method

TABLE 1
COMPARISON TABLE

Algorithm	No of Movies	Category	Popular	Unpopular
Each Item Method(EI)	1048	50	32	809
Total Clustering(TC)	1048	50	117	724
Adaptive Clustering(AC)	1048	50	221	620

From the above table it is inferred that Adaptive Clustering reduces the tail portion of long tail by bringing the unpopular to popular side. Sample of data of ten values is shown from the entire table.

TABLE 2

COMPARISON TABLE FOR THREE METHODS

POPULARITY TABLE											
EACH ITEM METHOD				TOTAL CLUSTERING METHOD				ADAPTIVE CLUSTERING METHOD			
MOVIEID	POPULARITY	RANK	CATEGORY	MOVIE ID	POPULARITY	RANK	CATEGORY	MOVIE ID	POPULARITY	RANK	CATEGORY
210	331	24	POPULAR	210	431	24	POPULAR	210	456	24	POPULAR
211	206	114	UNPOPULAR	211	306	114	POPULAR	211	283	300	UNPOPULAR
212	92	352	UNPOPULAR	212	192	352	UNPOPULAR	212	192	352	POPULAR

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

213	134	231	UNPOPULAR	213	234	231	UNPOPULAR	213	329	145	POPULAR
214	114	295	UNPOPULAR	214	214	295	UNPOPULAR	214	294	254	UNPOPULAR
215	212	105	UNPOPULAR	215	312	105	POPULAR	215	287	284	UNPOPULAR
216	290	45	UNPOPULAR	216	390	45	POPULAR	216	342	117	POPULAR
217	120	278	UNPOPULAR	217	220	278	UNPOPULAR	217	295	245	UNPOPULAR
218	171	162	UNPOPULAR	218	271	162	UNPOPULAR	218	346	108	UNPOPULAR
219	111	302	UNPOPULAR	219	211	302	UNPOPULAR	219	286	287	POPULAR

VII. CONCLUSION AND FUTURE WORK

Proposed work deals with the LTRP (Long Tail Recommendation Problem) responsible for improving the error rates in the tail of the item distribution of the recommendation system. The error rates of the non clustering EI recommender method increase for the low-ranked items in the tail part. This problem occurs because rating-prediction models do not have enough data for the less popular items. Therefore, the whole items are clustered into different groups to provide more data and this recommender method is referred as TC (Total Clustering) method. This fully grouping TC method solves the LTRP in many cases; however, it is not an efficient solution because it generates an excess amount of training data for the head items. Therefore, a new recommender method called the AC (Adaptive clustering) method is proposed that clusters items according to their popularity. This strategy solves the LTRP problem in the sense that the error rates of the items, especially in the tail, are significantly lower than those for the basic EI method while using a reasonable amount of training data. The contributions of this proposed work lie in showing that 1) the item-based long tail of the ratings distribution does matter; 2) the items in the long tail can be used productively by AC considering the item's rating numbers, and 3) there are optimal criterions α for AC which are simultaneously performance effective and cost efficient.

In Future long tail can still be reduced by using the latent dirichlet allocation and matrix factorization through latent dirichlet allocation and further also to add location of the user as feature in order to provide location based recommendation.

REFERENCES

1. G. Linden, B. Smith, and J. York, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan. /Feb. 2003.
2. C. Basu, H. Hirsh, and W. Cohen, "Recommendation as Classification: Using Social and Content Based Information in Recommendation," Proc. Nat'l Conf. Artificial Intelligence (AAAI '98), pp. 714- 720, 1998.
3. R.M. Bell and K. Yehuda, "Improved Neighborhood-based Collaborative Filtering," Proc. KDD Cup Workshop, pp. 7-14, 2007.
4. A. Hervas-Drane, "Word of Mouth and Recommender Systems: A Theory of the Long Tail," working paper, Harvard Business School, 2007.
5. Y.J. Park and A. Tuzhilin, "The Long Tail of Recommender Systems and How to Leverage It," Proc. ACM Conf. Recommender Systems, pp. 11-18, 2008.
6. K.Q. Truong, F. Ishikawa, and S. Honiden, "Improving Accuracy of Recommender Systems by Item Clustering," IEICE Trans. Information and Systems, vol. E90-D, no. 9, pp. 1363-1373, 2007.
7. SongJie Gong,"A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering", pp. 697-712, 2009.

8. W. Iba and P. Langley, “Induction of One-Level Decision Trees,” Proc. Ninth Int’l Conf. Machine Learning, pp. 233-240, 1992.
9. D.Agarwal and B.-C. Chen.flda: matrix factorization through latent dirichlet allocation. In WSDM, pages 91-100,2010.
10. Y.Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In KDD, pages 426-434, 2008.
11. D.Agarwal and B.-C. Chen. Regression-based latent factor models. In KDD, pages 19-28, 2009.