Clustering and Recommending Services based on ClubCF approach for Big Data Application

G.Uma Mahesh
Department of Computer Science and Engineering
Vemu Institute of Technology, P.Kothakota, Andhra Pradesh,India

G.Lokesh
Department of Computer Science and Engineering
Vemu Institute of Technology, P.Kothakota, Andhra Pradesh,India

Abstract: The number of services emerging on the Internet are generating huge amount of data leading to big data. Storing such data using traditional storage approaches is impractical which can be solved using Big Table capable of storing number of services in the form of multi dimensional sorted map again searching for a services and to recommend it to the new users requires large computations. In the present work these problems are solved by using the Clustering based Collaborative Filtering (ClubCF) approach and Mash Up data set with 6888 services along with their description and their functionality is considered for clustering with the help of agglomerative hierarchical clustering algorithm.

Key Words: Clustering, Collaborative Filtering, Recommendations, Big Data, Mash Up data set.

I. INTRODUCTION

The Collaborative filtering technique is an important approach to develop recommender systems that helps in predicting the interests of active users based on the preferences or interests of previous users provided in the form of ratings and reviews.

Traditional collaborative filtering techniques include Item based Collaborative Filtering and User based Collaborative Filtering techniques [1], both the CF techniques suffer from the data sparsity problem because the users rate only few set of services if further the services increase it will be a problem due to the limited users. In Item based Collaborative Filtering [2] the similarity between the services will be calculated and recommended to the new user with the most rated services. In user based Collaborative Filtering technique the similarity between different users is calculated and recommend to the new user with the service that most preferred by the previous users. The Collaborative Filtering techniques will have some challenges for the Big Data applications such as: performance decreases when the data was sparse; cannot provide recommendations to new users and new items because of limited users who will rate the limited number of items; limited scalability for large data sets.

The problems occurred during traditional Collaborative Filtering techniques can be solved by using ClubCF approach which helps in decreasing the number of services that needs to be processed and reduces the online computation time. The ClubCF approach consists of two steps; Clustering and the Collaborative Filtering, during the clustering step different services can be clustered or grouped in to different clusters by calculating their similarities between them and during the Collaborative Filtering step helps to recommend the services to the users by calculating the rating similarities between different users. The benefits of using Club CF approach are; better addresses the sparsity, scalability and other problems; to handle large amount of services in effective manner; to decrease the number of services needed to be processed; improve prediction performance and improve efficiency of recommendations.

II. SERVICE BIG TABLE

Big table was designed as a distributed storage system aims at storing structured data having ability to vary Peta bytes of information [3]. Some examples of services that are stored by Google people in big table are like Google Maps, Google Book History, Google Earth, Google Code History, You Tube, Gmail etc. The benefits of Big Table includes: distributed, Sparsity, Multidimensional sorted map, Scalability, easy to add new services in an effective manner.

The Google Big Table was indexed by the Row Key, Column Key and a time stamp where each service id or service number can be considered as row key and the column keys which deals with the service grouped together represents as column families, time stamp will focus on The time when the service actually recorded in the big table . The Table 1 will describe an example of how services are stored in big table which represents the 4 wheels route map service indexed by service id or row key as s1 and column families includes description, functionality and ratings provided by appropriate users after using the service along with the time stamps.

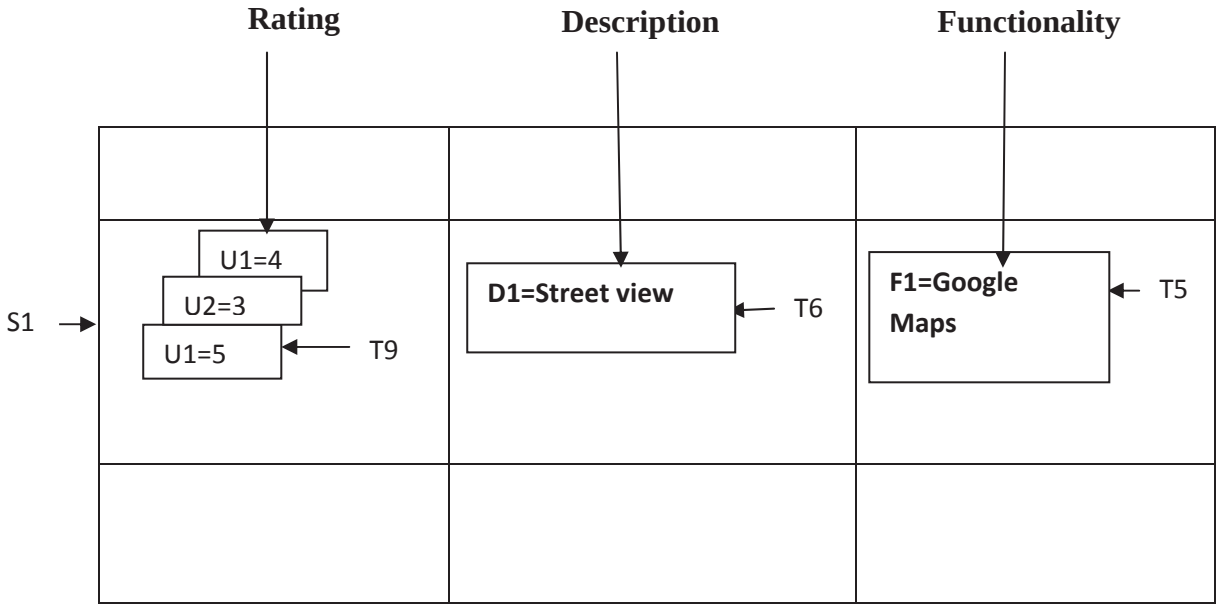


Table 1: Slice of Service Big Table

III. PROPOSED SYSTEM

The proposed method is implemented using Clustering based Collaborative Filtering (ClubCF) approach, as it reduces the online computation time by clustering the services based on their similarities and then recommends the services to the active users. ClubCF approach focuses on two stages Clustering and the Collaborative Filtering.

Algorithm for ClubCF approach:

1. Calculate description similarity, functionality similarity and characteristic similarity between possible pairs of services.

1.1 Compute Description and Functionality similarities using 'Jaccard similarity coefficient' [4] by using Equations 1 and 2.

$$D\_sim\ (st, sj) = \frac{|Dt' \cap Dj'|}{|Dt' \cup Dj'|} \qquad (1)$$

$$F\_sim\ (st, sj) = \frac{|Ft \cap Fj|}{|Ft \cup Fj|} \qquad (2)$$

It can be inferred from the above Equations 1 and 2 s*t*, sj referred to as services, consider an example of description similarity and functionality similarity between two mashup services 4 wheels route map (s1) and 100 destinations (s3) was

D_sim (s1, s3) $\frac{|D1' \cap D3|}{|D1' \cup D3'|}$ = 1/8 = 0.125

F_sim (s1, s3) = $\frac{|F1 \cap F3|}{|F1 \cup F3|}$ = ½ = 0.5

1.2 Calculate characteristic similarity between s*t*, sj using Equation 3.

$(st,sj)=\alpha \times D\_(st,sj)+\beta \times F\_sim(st,sj)$ (3)

In the above Equation, α belongs to the weight of description similarity, β belongs to the weight of functionality similarity and $\alpha+ \beta = 1$. The weights express the relative importance between two measures. Example of Characteristic similarity between two services s1 and s3 is

$C\_(s1, s3) =0.5 \times D\_si(s1, s3) +0.5 \times F\_sim\ (s1,s3)$

= 0.5* 0.125+ 0.5* 0.5= 0.3125

1.3 Repeat steps 1.1 and 1.2 for all the pairs of services and form the distance matrix (D_m).

2 Apply Agglomerative hierarchical clustering algorithm [5], [6] on Distance matrix (D_m) and group services with maximum distance as shown the procedure in below figure 1.

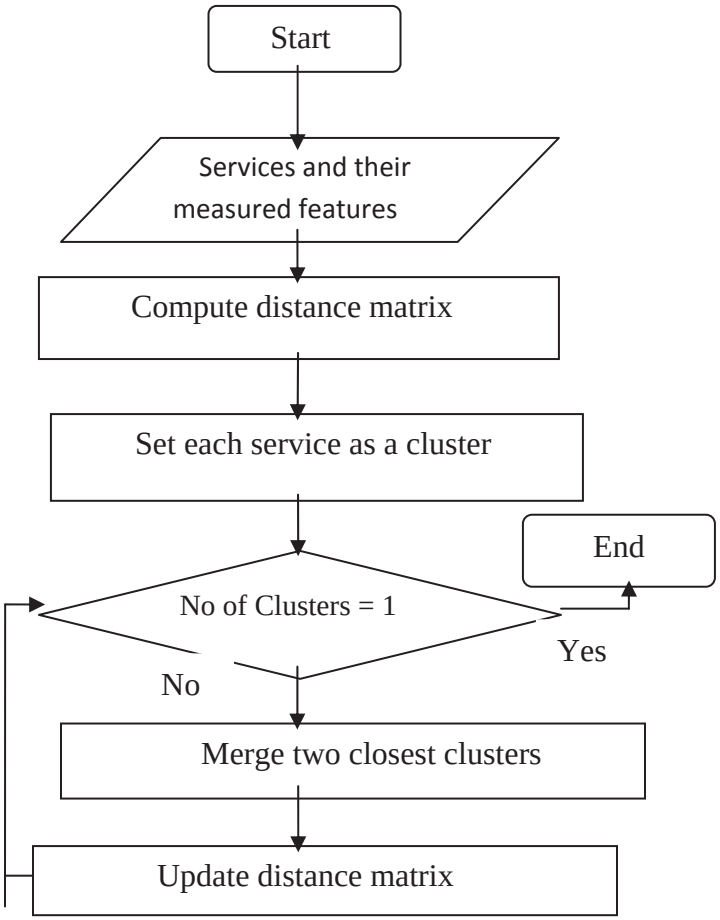


Fig 1: Flow Chart for AHC algorithm

3 Collect the Ratings given by the users to the service clusters formed in step 2.

3.1 Calculate rating similarities between the users using pearson correlation[7] coefficientgiven in Equation 4.

$$R\_sim(s_t,s_j)=\frac{\sum u_i s u_i s u_i (r_{u_i s_t}-\overline{r_{s_t}})(r_{u_i s_j}-\overline{r_{s_j}})}{\sqrt{\sum u_i s u_i s u_i (r_{u_i s_t}-\overline{r_{s_t}})^2}\sqrt{\sum u_i s u_i s u_i (r_{u_i s_j}-\overline{r_{s_j}})^2}} \quad (4)$$

The Equation 4 depicts u_t is a set of users who rated s_t while u_j is set of users who rateds_j, u_i is a user who rated both $s_t$ands_j. $r_{ui,s_t}$is the rating of s_t given by u_i.

3.2 If the number of users is limited then use enhanced rating similarities [8] for similarity calculations using Equation 5

$$R\_sim'(s_t,s_j)=\frac{2n|u_t\cap u_j|}{|u_t|+|u_j|}*R\_sim(s_t,s_j) \quad (5)$$

Where$u_t \cap u_j$ is the number of users who rated both services s_t,s_j, $u_t$and $u_j$are number of users who rated services s_t,s_j, respectively.

3.3 Calculate Neighbor's similarity of ratings for active users using the Equation 6.

$$N(s_t)=\{s_j | R'_{sim}(s_t,s_j)>\gamma, s_t\neq s_j\} \quad (6)$$

The above Equation depicts R _sim **(s_t,s_j)** as the enhanced rating similarity between service $s_t$and s_j,γ is the rating similarity threshold.

3.4 Compute predicted ratings for new users and also for the un rated users by using Equation 7.

$$p_{u_a,s_t}=\overline{r_{s_t}}+\frac{\sum_{s_j\in N(s_t)}(r_{u_a,s_j}-\overline{r}_{s_j})*R\_sim'(s_t,s_j)}{\sum_{s_j\in N(s_t)}R\_sim'(s_t,s_j)} \quad (7)$$

Where $\overline{r}$is the average rating of s_t ,N(s_t) is the neighbor set of s_t, $s_j \in N(s_t)$ denotes Sj is the neighbor of terget service s_t, r_{u_a,s_j} is the average rating of active user $u_a$gave to s_j,$\overline{r}_{s_j}$is the average rating of Sj, and R_sim'(St,Sj) is the enhanced rating similarity between services s_t,s_j computed using above Equation 7.

4 Calculate the MAE of ClubCF and compare the results with the IBCF approach.

IV. RESULTS

The proposed algorithm is implemented using Intel Dual Core processor with 2 GB RAM and Java (JDK 1.6) for coding. The experimental results of ClubCF approach are compared with the tradtional Item based Collaboraive Filtering which shows the computation time decreases as the number of services increases.

All the services are stored in the big table were labelled with row key column key, time stamp of their records along with their description and functionality. The major difference between the services storing in big table and the traditional Data Base Management systems was online computation time. The computation time decreases with in Big Table when compared to the DBMS storage mechanism. The following Figure shows the graphical representations of their comparision.

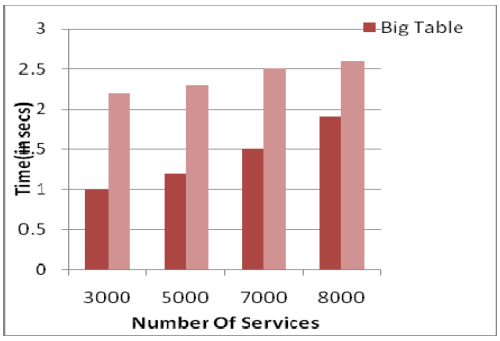


Fig 2: Big table vs traditional Data Base

The Big Table stores the mash up services along with their functionality and description whose similarities needs to be calculated in order to obtain the distance matrix which is the input for he Agglomerative hierarchical clustering algorithm.

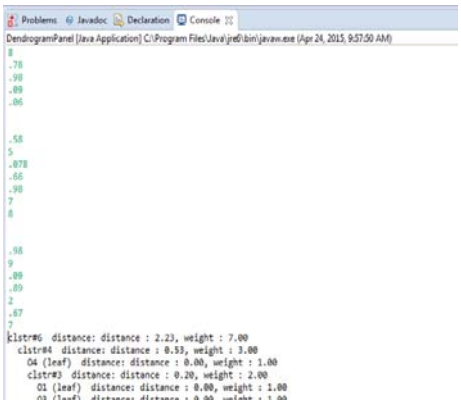


Fig 3: Input matrix

After providing the distance matrix as an input we will get the dendogram panel representation of different clusters based on minimum or maximum distance. The following Figure represents the dendogram panel of heirarchical clusters.

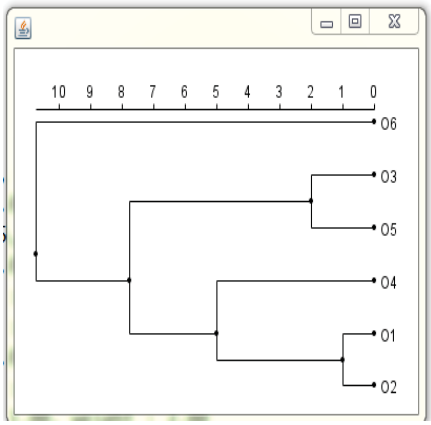


Fig 4: Dendogram panel

After that the users need to rate the services with in the different clusters between 0 to 5 rating system. Then the recommendation mecahnism came into picture in order to help the active users who don't know any thing about the service by calculating the rating similarities between different usrers and compute predicted ratings for new users.

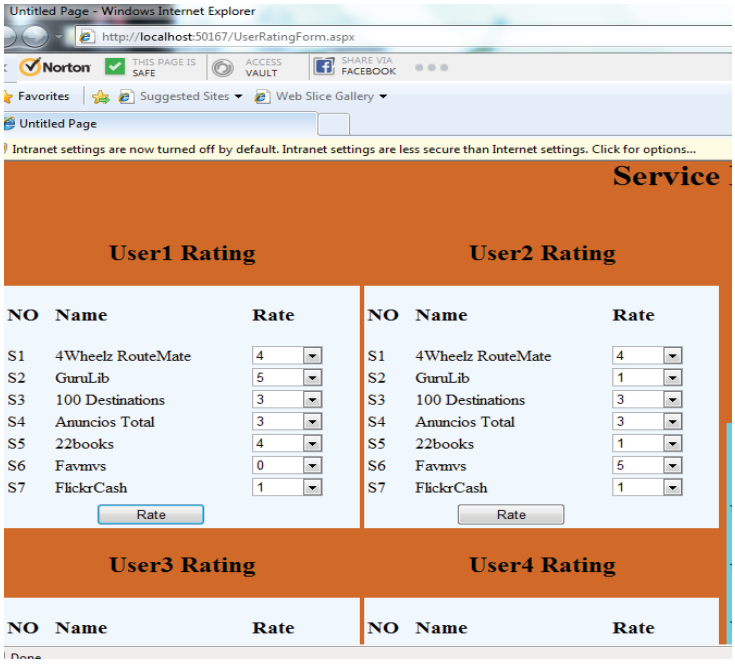
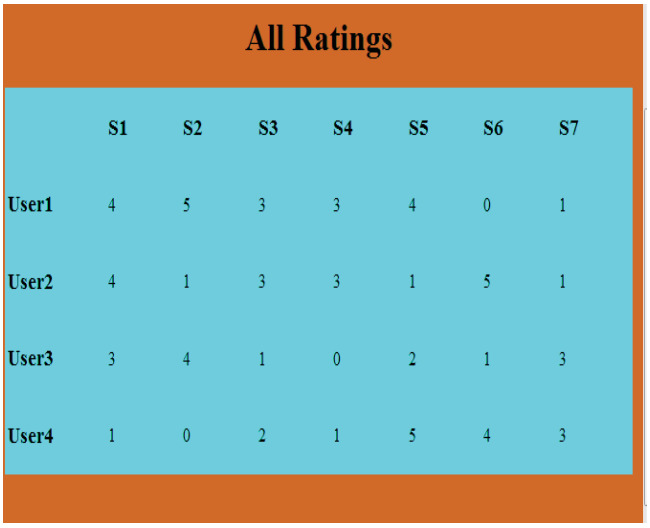


Fig 5: Rating for the services



All Ratings

	S1	S2	S3	S4	S5	S6	S7
User1	4	5	3	3	4	0	1
User2	4	1	3	3	1	5	1
User3	3	4	1	0	2	1	3
User4	1	0	2	1	5	4	3

Fig 6: Ratings of users to services

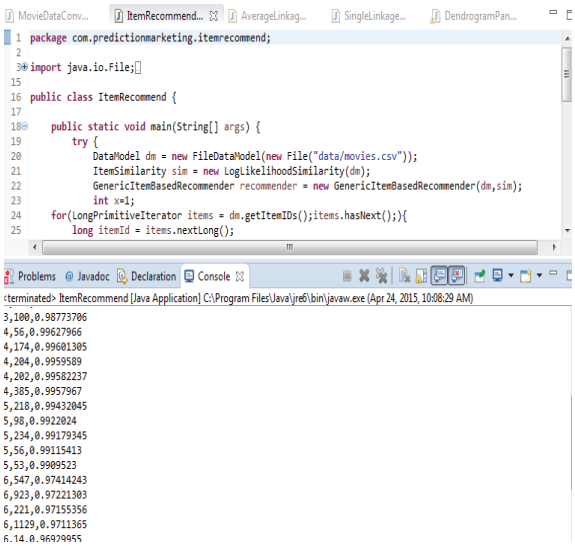


Fig 7: Recommendations

After calculating the rating similarities and generating recommendations for the un rated users we need to check for the accuracy of the ClubCF approach by comparing the mean absolute errors of ClubCF approach and the IBCF approach which leads the better accuracy of using the ClubCF approach. The below Figure represents the comparision between ClubCF and IBCF approaches.

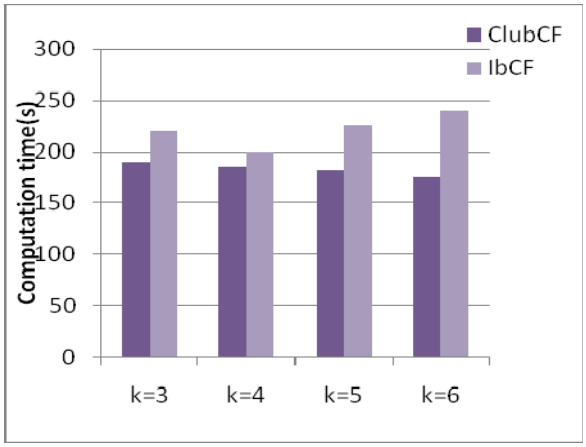


Fig 8: ClubCF vs IBCF approach

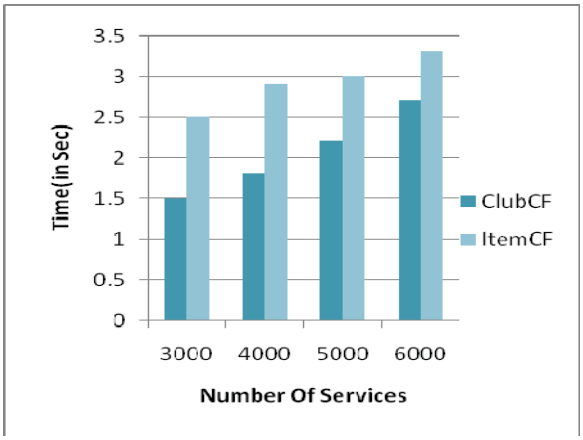


Fig 9: ClubCF vs IBCF Services

V. CONCLUSIONS AND FUTURE WORK

The clustering of services is performed by using AHC algorithm over 6888 services. The actual rating similarities and predicted ratings are are computed using PCC and enhanced rating similarities. The obtained results will be compared with the IBCF approach. The experimental results shows that computation time is approximately reduces by 30% in proposed method over IBCF even with increae in number of services. The proposed method also overcomes the problem of data Sparsity.

The future work can be extended to user based records and mining users interests and automatically provides recommendations to the users based on their explicit interests also, to develop methods to provide recommendations to new users even when few ratings are available.

REFERENCES

[1] Xiaoyuan Su and Taghi M. Khoshgoftaar .,"A survey of Collaborative Filtering Technique" A Review article, Hindawi Publishing Corporation Advances in Artificial Intelligence, Received 9 February 2009; Accepted 3 August 2009.

[2] A. Yamashita, H. Kawamura, and K. Suzuki, "Adaptive Fusion Method for User-based and Item-based Collaborative Filtering," *Advances in Complex Systems*, vol. 14, no. 2, pp. 133-149, May 2011.

[3] F. Chang, J. Dean, S. mawat, et al.,"Big table: A distributed storage system for Structured data," *ACM Trans. on Computer Systems*, vol. 26, no. 2, pp. 1-39, June2008.

[4] A. Rodriguez, W. A. Chaovalitwongse, L. Zhe L, et al., "Master defect record retrieval using network-based feature association," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 3, pp. 319-329, October 2010.

[5] C. Platzer, F. Rosenberg, and S. Dustdar, "Web service clustering using multidimensional angles as proximity measures," *ACM Trans. on Internet Technology*, vol. 9, no. 3, pp. 11:1-11:26, July, 2009.

[6] G. Thilagavathi, D. Srivaishnavi, N. Aparna, et al., "A Survey on Efficient Hierarchical Algorithm used in Clustering," *International Journal of Engineering*, vol. 2, no. 9, September 2013.

[7] G. Adomavicius, and J. Zhang, "Stability of Recommendation Algorithms," *ACM Trans. On Information Systems*, vol. 30, no. 4, pp. 23:1-23:31, August 2012.

[8] D. Julie, and K. A. Kumar, "Optimal Web Service Selection Scheme With Dynamic QoS Property Assignment," *International Journal of Advanced research in technology, vol.2,no.2, pp. 69-75, may 2012.*

.