

Item based Collaborative filtering approach for Big Data application

K Sudha¹ M.Lavanya² A.Kanimozhi³

¹PG Student ^{2&3}Asst. Prof Dept. of CSE,

^{1&2&3}Affiliated to Anna University Chennai, Dept. of Computer Science and Engineering
Kongunadu College of Engineering and Technology

Abstract

Spurred by Service computing and cloud computing an increasing number of services are emerged in the Internet. As a result, resource -permissible data arise too big to be handled by already established techniques. Clustering along with Item based collaborative filtering has been proposed to reduce online execution time taken for processing services. Technically, these approaches sanction around two stages. In the first stage, services are divided among cluster and each cluster has similar services. At the second stage, item based collaborative filtering algorithm is enforced on one of the cluster. Real Dataset collected from the programming web has been used to conduct several experiments.

Index Terms: Collaborative Filtering, Clustering, Agglomerative Hierarchical Clustering Algorithm

1. INTRODUCTION

As internet users are increasing every year all the work gets digitalized today, as a result the data in the internet gets increased. Big Data is a term refers a large dataset. It consists of both structured and unstructured data. Unstructured data refers information that is not able to fit in row-column database. Social network like Facebook, Twitter, LinkedIn, Wikipedia and YouTube are some of the source for production of large amount of unstructured data. Big Data cannot be treated by existing database application. Some of the NoSQL database like MongoDB, MarkLogic, Apache Hadoop, Apache Cassandra, and IBMpureXML are used to analyze an unstructured data. The right Big Data analytics helps the enterprise to improve operation, customer service, to increase efficiency and risk management. Recommender Systems are method to serve users in a decision making process. Recommenders System is classified in to two groups. They are Content based system and Collaborative filtering system. The problem in the existing system is a large number of services are taken for analysis, thus it results in high computation time. A simple solution is to decrease the number of services that need to be processed in real time. Clustering are such techniques that can reduce the data size by grouping similar service.

2. PROPOSED SYSTEM

Clustering

Cluster models are widely used in recommender systems. Normally, a recommendation engine would cluster users based on their browsing or purchasing information, and recommend items to a user based on other users in the same cluster. In our case, the user-related information is too sparse for that to be truly

useful as mentioned above. Therefore we use properties of the item to cluster. AHC algorithm has been implemented to cluster the similar services. This algorithm helps to reduce idle service in recommendation.

AHC algorithm

Input: A set of items $S = \{I_1, \dots, I_n\}$, a characteristic similarity matrix $D = [d_{i,j}]_{n \times n}$ the number of required clusters k .

Output: Dendrogram for $k=1$ to $|I|$.

1. $c_i = \{I_i\}, \forall i$;
2. $d_{c_i, c_j} = d_{i,j}, \forall i, j$;
3. for $k = |I|$ down to K
4. Dendrogram $_k = \{c_1, \dots, c_k\}$;
5. $l, m = \text{argmax}_{i,j} d_{c_i, c_j}$;
6. $c_l = \text{Join}(c_l, c_m)$;
7. for each $c_h \in I$
8. if $c_h \neq c_l$ and $c_h \neq c_m$
9. $d_{c_l, c_h} = \text{Average}(d_{c_l, c_h}, d_{c_m, c_h})$;
10. end if
11. end for
12. $I = I - \{c_m\}$;
13. end for

Item based Collaborative filtering System

Item-based collaborative filtering is a type of model based algorithm for making recommendations. There are two stages in item based collaborative filtering. In the first stage similarities between items are calculated by using one of the number similarity measures. In the second stage the estimated similarity values are used to predict ratings for unknown item. In

item similarities measures, two items are considered as similar when two users rate both of them.

3. COMPUTE RATING SIMILARITY

Pearson (correlation)-based similarity

It measures the correlation between two items. It result 1if it exist any correlation between the items otherwise 1 to exhibit that there is no correlation between the items.

$$sim(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_i)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_j)^2}} \dots(1)$$

Enhanced rating similarity

Enhanced rating similarity between s_t and s_j is computed by formula

$$R_sim'_{s_t, s_j} = 2 \times U_t \cap U_j / (U_t + U_j \times R_sim(s_t, s_j)) \dots(2)$$

In this formula, $U_t \cap U_j$ is the number of users who rated services s_t and s_j , U_t and U_j are the number of users who rated service s_t and s_j , respectively.

Select Neighbors

Based on the enhanced rating similarities between services, the neighbors of a target service s_t are determined according to constraint formula

$$Ns_t = \{s_j | R_sim'_{s_t, s_j} > \gamma, s_t \neq s_j\} \dots(3)$$

γ is a rating similarity threshold. The bigger value of γ is, the chosen numbers of neighbors will relatively less but they may be more similar to the target service, thus the coverage of collaborative filtering will decrease but the accuracy may increase.

Compute Predicted Rating

Depending on predicted rating of target service s_t it is chosen as recommender for other active user u_a to whom the recommendation are being made.

$$Pu_{a, s_t} = \bar{r}_{s_t} + \frac{\sum_{s_j \in N(s_t)} (ru_{a, s_j} - \bar{r}_{s_j}) \times R_{sim}(s_t, s_j)}{\sum_{s_j \in N(s_t)} R'_{sim}(s_t, s_j)} \dots(4)$$

Architecture diagram of proposed system

The user has to upload the dataset on had oop software. Had oop split the dataset among clusters and then similarity algorithm has to be applied on data.



Fig

4. RELATED WORK

ManhCuongPham[12] used a clustering technique for collaborative filtering on venue recommendation system. The accuracy of collaborative filtering algorithm is increased by reducing scarcity problem. Gediminas Adomavicius[13] used a model based recommendation algorithms to consistently deliver high stability predictions. In stable technique addition of new ratings should not change the estimation of unknown ratings. The accuracy of prediction in recommendation system can be measured by Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) among many other technique. Akihiro Yamashita[14] proposed a unified method for user based and item based collaborative filtering to improve the accuracy of recommended system. The unifying method uses a constant value as weight parameter to unify two algorithms. BADRUL SARWAR used an item based collaborative filtering for high quality recommendation. As relationship between items is static, item based collaborative filtering algorithm may be able to provide same quality similar to user based algorithm with less computation time.

5. CONCLUSION

The computation time taken for processing services gets reduced through clustering. Most similar services can be recommended to user with clustering and collaborative filtering approach.

6. FUTURE WORK

Semantic analysis can be imposed on feedback of the item by user.

REFERENCES

[1] J. Herlocker, J. A. Konstan, and J. Riedl, "An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms," *Information retrieval*, vol. 5, no. 4, pp. 287-310, October 2002.

- [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] 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.
- [4] J. Wu, L. Chen, Y. Feng, et al., "Predicting quality of service for selection by neighborhood-based collaborative filtering," *IEEE Trans. on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 2, pp. 428-439, March 2013.
- [5] Y. Zhao, G. Karypis, and U. Fayyad, "Hierarchical clustering algorithms for document datasets," *Data Mining and Knowledge Discovery*, vol. 10, no. 2, pp. 141-168, November 2005.
- [6] Z. Zheng, H. Ma, M. R. Lyu, et al., "QoS-aware Web service recommendation by collaborative filtering," *IEEE Trans. on Services Computing*, vol. 4, no. 2, pp. 140-152, February 2011.
- [7] M. R. Catherine, and E. B. Edwin, "A Survey on Recent Trends in Cloud Computing and its Application for Multimedia," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 2, no. 1, pp. 304-309, January-February 2013.
- [8] X. Liu, Y. Hui, W. Sun, et al., "Towards service composition based on mashup," in *Proc. of IEEE Congress on Services*, pp. 332-339, July 2007.
- [9] X. Z. Liu, G. Huang, Q. Zhao, et al., "iMashup: a mashup-based framework for service composition," *Science China Information Sciences*, vol. 57, no. 1, pp. 1-20, January 2013.
- [10] H. Elmeleegy, A. Ivan, R. Akkiraju, et al., "Mashup advisor: A recommendation tool for mashup development," in *Proc. of IEEE Int'l Conf. on Web Services*, pp. 337-344, October 2008.
- [11] O. Greenshpan, T. Milo, and N. Polyzotis, "Autocompletion for mashups," in *Proc. of the VLDB Endowment*, vol. 2, no. 1, pp. 538-549, August 2009.
- [12] M. C. Pham, Y. Cao, R. Klamka, et al., "A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis," *Journal of Universal Computer Science*, vol. 17, no. 4, pp. 583-604, April 2011.
- [13] 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.
- [14] 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.