International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

A Novel Approach for Smart Shopping Using Clustering-Based Collaborative Filtering

Samiksha M. Pande[1], Ashwini Gaikwad[2]

[1]*Department of Computer Science and Engineering, DIEMS, Aurangabad, MH, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, DIEMS, Aurangabad, MH, India*

Abstract - ***In today's fast developing Internet technology era, e-commerce is taking more and more market share from physical stores. With this development the user is overwhelmed by the huge amount of choices that are provided for searching item and the information overloading problem has been at its peak. This is when Recommendation System comes in handy, which is the system that produces individualized recommendations by searching through the large volume of dynamically generated information to provide users with personalized possible options based on certain reference characteristics. But data become too big to handle and process effectively by traditional approaches. So, to attenuate the impact of dense data, a clustering based collaborative filtering algorithm using user interest information is proposed in this paper. This method aims at recruiting similar data in the same cluster to recommend an item & personalized services for users collaboratively.***

Key Words: E-commerce, Recommendation system, user interest, clustering, collaborative filtering

1. INTRODUCTION

Big data has come into view as a widely recognized trend, attracting courtesy of the government, multinational companies, banking sectors, academia and etc. Generally speaking, big data is the act of collecting and storing large volume, complex, enormously growing information for eventual analysis with multiple, autonomous sources whose global level is almost inconceivable.

So, where this Big Data comes from? These Big Data comes from its various applications like social networking sites, E-commerce sites, Telecom companies, health and life sciences, weather stations, share market, and etc. Due to all the applications of Big data where data collection has grown enormously and it is exceeding the ability of traditional data processing application software tools to capture, store, analyze, manage and process the data, that too in tolerable elapsed time [1].

With the development of web technology and advancement in the area of E-commerce, most people go for Online shopping rather than the window (retail) shopping. As the competition between businesses has become increasingly strong, customers are facing information overloading problem. It occurs as users are provided with a huge amount of product options available to choose from, most of which may not be relevant to what they are searching for and that makes users overwhelmed and indecisive.

Hence, Recommendation Systems (RSs) are techniques and an intelligent application that tries to predict items out of large pool a user may be interested in and recommends the best option to the target user. So, to explore the large volume of data and mine useful information or knowledge for further actions, according to user interest, preferences and behavior that are being captured from his/her previous purchase history, which then are stored and use the same for the personalized recommendation in the future.

Recommendation techniques are applied to almost all large E-commerce platforms like, Amazon, eBay, etc., which are supposed to personalize services for customers by reducing transaction costs of searching and choosing items of interest in an Online shopping environment and are also beneficial to service providers. To, the service providers, they have increased revenue for the fact that advertisements are the effective means of selling more products by directly reaching customer of interest [2].

Various methods are applied in RSs which can take either of two basic approaches: Collaborative Filtering (CF) or Content-Based Filtering. Collaborative Filtering is the most dominant technique used in RSs that do not need any external information about either the user or the item. CF technique basically assumes that if two users have similar behavior, i.e. watching, buying, or have similarity in rating n items, and hence will act or rate on other items similarly. So, we can say CF when appeared at recommendation is based on a model of user's prior behavior. The model can be created simply from a single user behavior or also from the behavior of other users having similar traits. Here, when other user's behavior is taken into account, CF uses group knowledge to form a recommendation based on similar users.

Collaborative Filtering are required to have the ability to deal with highly sparse data, to scale with increasing number of items and users, to deal with the problems like similar items having different names, noisy data, privacy protection, shilling attacks and to make adequate recommendations in a short time period, which

may reduce recommendation accuracy by keeping it far behind the expectation of customers as well as businesses. To solve those problems by reducing their impact, improving the scalability and accuracy to some extent a number of CF algorithms such as user-based, item-based, content-based, model-based and so on, are proposed [3].

The basic assumption behind the item based algorithm is that it recommends the user the item that is similar to what she/he has preferred before. Alternate to item-based User-based algorithm assumes that people who tend to agree in the past may agree again in the future.

2. LITERATURE SURVEY

Nowadays Service relevant data have grown tremendously, become complex, and it is beyond the ability of traditional approaches to effectively capture, manage, and process that Big data, so one solution to this challenge is Clustering Based Collaborative Filtering (CLUBCF) [1]. CLUBCF recruits similar services in the same cluster to recommend services collaboratively and contains two modules. The first step is Clustering, in this, services are clustered depending on the similarity in the description, functionality which is Computed by using the Jaccard similarity coefficient (JSC), & Characteristic similarity between the two services is computed by using the weighted sum of Description Similarity and Functionality Similarity then Agglomerative Hierarchical Clustering (AHC) Algorithm is used for clustering. The second step is of Collaborative Filtering, in this initially, rating similarity is computed by using Pearson correlation coefficient (PCC) & then predicted rating is given to the clustered services. Lastly, according to the predicted ratings, all recommended services are ranked in descending order.

Collaborative filtering (CF) such as user based and item based methods are popular techniques to retrieve the services from overwhelming services but it consumes a lot of time. Clustering techniques are the solution to decrease the data size of service. Bottom-up hierarchical Clustering based collaborative filtering approach is used in [4]. In this approach, similar services are grouped into clusters by their functionalities and classifications, and recommendations are made based on the similar services under the same cluster where K-means clustering algorithm issued for recommending. It is one of the partition-based clustering algorithms. It was applied to the partition of services based on the user's preference. Even though K-means is one of the partition-based clustering algorithms, but for clustering process, it requires additional information from users.

Collaborative filtering (CF) is widely used in many domains with numerous algorithms for personalized recommendation. Despite of many advantages CF suffers from some issues like data sparsity, scalability, cold start problem, shilling attacks and so on. To improve the accuracy many of the researchers have proposed some new measures of similarity. For example, H. J. Ahn (2008) proposed a new similarity of CF that is called Proximity-Impact-Popularity (PIP) that combined item content, information and popularity with user-behavior data to get good results [5].

In CF to reduce the impact of data sparsity number of methods are put forward by researchers. Existing research on data sparsity problem can be largely divided as i) Model-based CF and hybrid recommender algorithm, ii) improved user similarity calculated methods in memory-based CF, which includes item-based and user-based CF algorithm. To improve the insufficient of traditional similarity measure method an item-based CF algorithm was proposed by Sarwar B. & et al. (2001) for when user rating is extremely sparse [6]. Ma. & et al (2007) proposed a CF algorithm based on the nearest neighbor set of target users and items with adjustable parameters to control the weight of the two parts to generate recommendation results [7]. iii) The other stream of research is by analyzing the characteristics of users and items thoroughly to improve the performance of CF algorithm.

To increase the performance of Collaborative filtering algorithm extra information is used by some researchers such as user activity, user location and user interest [8]. Liu. Q & et. Al (2012) calculates user similarity [9] and to do so, user interest based on the relationship between items is expanded. Based on dynamic recording within the nearest neighbor set Yehuda Koren (2009) proposed a CF algorithm where based on the user activity, weight of neighboring user's set is adjusted dynamically according to different target items [10]. These studies improve scalability and accuracy by reducing the impact of data sparseness of CF algorithm in some extent.

An improved user-based clustering CF algorithm combining with user interest information is proposed in [11] which improved basically by two ways: improving user similarity calculating method and extending user-item rating matrix. Also to solve cold-start problem in the CF algorithm [12-13] for calculating user similarity some researchers extend the user-item rating matrix by using user' attributes and item content information. Because of these methods, the calculation methods of similarity in the algorithm are improved.

An improved similarity model is proposed in [14] to calculate similarity between items and users, and improve recommendation performance in memory-based CF algorithm. The disadvantage of the Pearson correlation coefficient [15] and cosine similarity [16] is that, they are not capable enough for users who only rate a small number of items to capture similar users effectively, so the improved similarity model combines global preference of user behavior with local information of user rating. This disadvantage is also analyzed by Ahn [5] which considers 3 aspects as impact, proximity, and popularity of the user ratings. But, this similarity does not consider the global preference of the user ratings and considers only the local information about the ratings. The Weighted Pearson

correlation coefficient has been proposed [17] to solve the problem of traditional Pearson correlation coefficient that it does not consider the size of the set of common users.

a) Existing System

Amazon uses Item Clustering Collaborative Filtering technique for recommending books and all other products. LinkedIn, Facebook, MySpace uses collaborative Filtering technique to make friend suggestions, groups and other social connections by observing the network of connections between a user and people present in their connections. Recommender System in Twitter is used for suggesting whom to follow which makes use of several signals and in-memory calculations.

YouTube also uses the Item Clustering CF technique for the recommendation. Netflix and YouTube are hybrid systems as they recommends by comparing the habits of searching and watching of similar user, i.e. Collaborative Filtering combined with Content based Filtering i.e. by offering movies that a user has rated highly and that share similar characteristics.

Last.fm and Reddit uses User-based Collaborative Filtering technique as suggestions are made by considering user choice. In Last.fm, a station is created in which songs are suggested based on what other people did with that song, what the people you follow, who is similar to you, and what music they listen to/ rate/like. But for making reliable suggestions Last.fm needs large amount of data related to the user, this makes it suffer the cold start problem.

Pandora uses Content based approach. Pandora uses the qualities and the features of a song or artist for tuning into a station playing similar featured music. It uses the feedback given by users for considering the feature like liking a particular song Feedback given by the users is used to tune the station instead of considering some features as like and dislikes. Pandora can get started with little information and has limited scope.

3. PROPOSED SYSTEM

i) User/ Item Based Collaborative filtering

Collaborative filtering (CF) also referred to as social filtering is a technique used by recommender systems which filters information by using the recommendation of people. CF deals with the problem of effective extraction of useful information from vast available data, leads to the concept of CF. It is a technique used in Recommender System (RS) which filters information by using the recommendation of people having same taste.

Collaborative filtering such as the User-based and Item-based methods are mostly implemented techniques in RSs. The basic idea behind User-Based is that people who agreed in the past are likely to agree in future in their evaluation of certain items. Suppose if we want to predict how user U will rate item I, we can check how other users who are similar to user U (i.e. Like minded users) have rated that item. It is possible that the user will rate items similar to users with similar taste than that of randomly chosen user from the crowd for ex. A CF Recommender System for music taste could make predictions about which type of music a user should like, given a partial list of that user's likes or dislikes. Item-based CF algorithm recommends a user the items similar to that of she/he has preferred before. Suppose if we want to predict how user U will buy or rate item I, we can check how he has bought or rated other items, which are similar to an item I. It is possible that the user will buy/rate similar items similarly for ex. A user who bought an item X also tends to buy item Y.

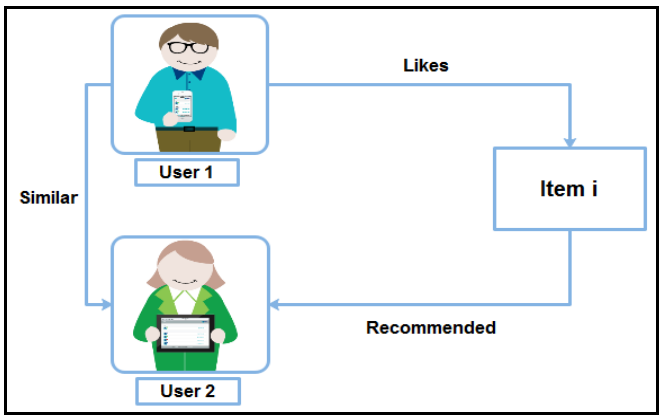

Fig -1: User-Based Recommendation

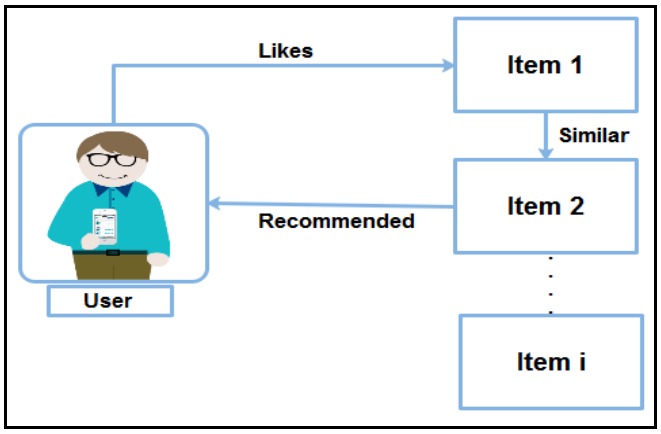

Fig -2: Item-Based Recommendation

Although many e-commerce RSs has successfully implemented traditional CF techniques, they have to face some challenges for Big data like i) to give an ideal recommendation from a large number of items and ii) to give an ideal decision in acceptable time.

In the traditional CF algorithm, to calculate similarity between every pair of users or items may take too much time, many times it may go beyond the processing capability of current recommendation systems. Therefore,

items recommendation based on user similarity or item similarity may not be done completely or may exceed the time. Also, when considered ratings in traditional CF algorithm all items are considered when computing rating similarity while most of them are different to the target item. The accuracy of predicted rating is affected due to these dissimilar ones. So as to decrease the number of items that need to be processed in real time Clustering can be used.

Clustering is a technique which reduces the data size by making a group of like-minded users or similar items. This gives the result as less number of items/users in a cluster is much less than the total number of items/users. Thus the computation time of the CF algorithm can be reduced and accuracy may be enhanced.

Agglomerative Hierarchical clustering algorithm (AHC) is applied for clustering and for optimization cuckoo search optimization algorithm is implemented. The cuckoo search when compared with other algorithms such as particle swarm optimization and genetic algorithms has shown better performance. A novel recommender system is proposed for recommending best suitable item with AHC algorithm & cuckoo search optimization.

For stemming that is for getting a common root format of word transformed from variant word forms, various kinds of stemming algorithms, such as Lovins stemmer, Dawson Stemmer, Paice/Husk Stemmer, and Porter Stemmer, have been proposed [18]. One of the most widely used stemming algorithms among them is Porter Stemmer. It does not require the use of a lexicon [19] and applies cascaded rewrite rules that can be run very quickly.

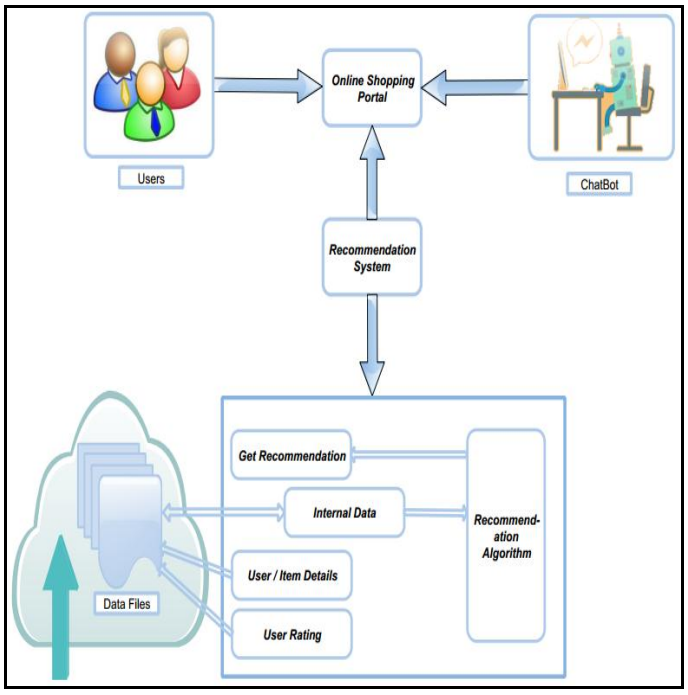

Fig -3: System Flow

ii) MS Chat BOT for User response

A Chatbot is a computer program that carries out a conversation by textual or auditory methods. Simply it can be a basic pattern matching with a response or highly developed Artificial Intelligence techniques with complex conversational state tracking and integration of existing business services. Today Chatbots are accessed by many websites, applications and different platforms.

In order to study we are going to use Microsoft Chatbot framework which enables us to build bots that supports different types of interaction with users. We can design conversation in bot to be freeform. The conversation can use simple text string or more complex rich cards that contain images, text and action buttons.

We can let users interact with our bot by adding natural language interactions.

iii) Natural Language Understanding Response

While interacting with computers, humans face a problem regarding the ability of the computer to understand what a person wants. So, developers are building smart applications with the help of Natural Language Understanding Response (LUIS) which enables applications to understand human language and react accordingly.

Any client application having a conversation with user input like Chatbot or any other dialog system can pass user input to a LUIS app and receives a result that provides natural language understanding. LUIS uses the power of Machine Learning to solve difficult problem of extracting meaning from natural language input.

3. CONCLUSION

In today's fast life online shopping is considered as a more preferable option of shopping rather than window shopping. For this online shopping, Recommendation Systems are the techniques used for filtering and retrieving relevant items that reduces the problem of overloaded product options provided to users. We have studied different methods used in RSs, out of which Collaborative Filtering approach is most successful and popular of all other methods which is used in our smart shopping portal. Some problems arise with Collaborative Filtering and to solve them number of CF algorithms are proposed like User-based, Item-based, Content-based, Model-based and so on are. In the proposed system we are focusing on User-based and Item-based CF techniques.

Initially, the Agglomerative hierarchical algorithm is used for clustering of items, where one group items are more similar than that of other groups and this reduces the amount of items to be worked on, then the rating similarity of users and items is calculated within the same cluster. This helps to generate more accurate recommendations from a

large amount of items than based on all similar or dissimilar items in all clusters, which finally reduces our result time. Along with this, to make the smart shopping system more interactive we are going to provide a Chatbot support system that helps the users to solve the query by answering their questions.

REFERENCES

[1] RONG HU, (Member, IEEE), WANCHUN DOU, (Member, IEEE), and JIANXUN LIU, (Member, IEEE): ClubCF: "A Clustering-Based collaborative Filtering Approach for Big Data Application," Digital Object Identi_er 10.1109/ TET.2014. 2310485.

[2] Parul Aggarwal, Vishal Tomar, Aditya Kathuria: "Comparing Content based and collaborative filtering in Recommender System," International Journal of New Technology and Research (IJNTR) ISSN: 2454-4116, Volume-3, Issue-4, April 2017 Pages 65-67.

[3] FIDEL C.,V´ICTOR C.,DIEGO F.& VREIXO O. Comparison of Collaborative Filtering Algorithms: Limitations of Current Techniques and Proposals for Scalable, High-Performance Recommender Systems [J]. ACM Transactions on the Web, Vol. 5(1), pp.2-33, 2011.

[4] V. Devi, R. Kanaga Selvi: "An Innovative Approach to Endorse the Best Web Services to Craft Mashup Web Applications Using Bigtable," DOI 10.4010/2014.268 ISSN-2321-3361 © 2014 IJESC.

[5] H. J. Ahn. "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," Information Sciences 178, pp.37-51, 2008.

[6] Sarwar B, Karypis G, Konstan J, et al. "Item-based collaborative filtering recommendation Algorithms" [C].Proceedings of the 10th international conference on World Wide Web. ACM, pp.285-295, 2001.

[7] Ma Hao, King Irwin, Lyu Michael. "Effective missing data prediction for collaborative filtering"[C].Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, Netherlands, pp.39-46, 2007.

[8] Akihiro Y., Hidenori K.&Keiji Suzuki. "Adaptive Fusion Method for User-Based and Item-Based Collaborative Filtering"[J]. Advances in Complex Systems, 14(2), pp.133- 149, 2011.

[9] Liu, Q., Chen, E., Xiong, H., Ding, C. H., & Chen, J. Enhancing collaborative filtering by user interest expansion via personalized ranking. Systems, Man, and Cybernetics [J]. Part B: Cybernetics, IEEE Transactions on, 42(1), pp.218-233,2012.

[10] Yehuda Koren. Collaborative filtering with temporal dynamics. 15th ACM SIGKDD'2009, pp. 447-456, 2009.

[11] Li Zhang, Tao Qin PiQiang Teng: An Improved Collaborative Filtering Algorithm Based on User Interest, JOURNAL OF SOFTWARE, VOL. 9, NO. 4, APRIL 2014, doi:10.4304/jsw.9.4.999-1006.

[12] Iaquinta L, Semeraro G. Lightweight approach to the cold start problem in the video lecture recommendation[C]. Proceedings of the ECML/PKDD Discovery Challenge Workshop, CEUR Workshop Proceedings. CEUR Workshop Proceedings.770, pp.83-94, 2011.

[13] Qiu T, Chen G, Zhang Z K, et al. An item-oriented recommendation algorithm on cold-start problem [J]. EPL (Europhysics Letters), 95(5): 58003, 2011.

[14] Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, Xuzhen Zhu: "A new user similarity model to improve the accuracy of collaborative filtering," Knowledge-Based Systems 56 (2014) 156–166.

[15] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in: Proceeding of the ACM Conference on Computer Supported Cooperative Work, 1994, pp. 175–186.

[16] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 734–749.

[17] J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing collaborative filtering, in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 230–237.

[18] R. S. Sandeep, C. Vinay, and S. M. Hemant, "Strength and accuracy analysis of af_x removal stemming algorithms," Int. J. Comput. Sci. Inf. Technol., vol. 4, no. 2, pp. 265_269, Apr. 2013.

[19] V. Gupta and G. S. Lehal, "A survey of common stemming techniques and existing stemmers for Indian languages," J. Emerging Technol.Web Intell., vol. 5, no. 2, pp. 157_161, May 2013.

BIOGRAPHIES

Ms. Samiksha Mahavir Pande is currently studying in Masters of Computer Science and Engineering at DIEMS, Aurangabad. She has completed her B.E. from SPWEC, Aurangabad in 2013. Her research topic is Novel approach for recommendation system using Clustering-based collaborative Filtering approach.

Ms. Ashwini Sanjay Gaikwad has received the B.E. degree from JNEC, Aurangabad in 2002 and M.E. degree in Computer Science and Engineering at Government Engineering College, Aurangabad in 2008. Currently she is pursuing her Ph.D in Computer Science and Engineering with the topic Gender and Age Recognition System based on Fingerprints' and is working as Asst. professor in DIEMS, Aurangabad. Her specialization includes Digital Image processing, Artificial Intelligence, Computer Vision.