Survey on Data Mining Using Recommendation System Methods

Paritosh Nagarnaik
Department of Computer Science and Engineering
G. H. Raisoni College of Engineering
Nagpur (M.S), India
paritoshnagarnaik91@gmail.com

AkashPatil
Department of Computer Science and Engineering
G. H. RaisoniPolytechnic
Nagpur (M.S), India
kash12itengg@gmail.com

***Abstract*–In recent years recommendation systems have changed the way of communication between both websites and users. Recommendation system sorts through massive amounts of data to identify interest of users and makes the information search easier. For that purpose many methods have been used. Collaborative Filtering (CF) is a method of making automatic predictions about the interests of customers by collecting information from number of other customers, for that purpose many collaborative base algorithms are used. CHARM algorithm is one of the frequent patterns finding algorithm which is capable to handle huge dataset, unlike all previous association mining algorithms which do not support huge dataset. This paper covers different techniques which are used in recommendation system and also proposes a new system for efficient web page recommendation based on hybrid collaborative filtering i.e. using collaborative technique and CHARM algorithm which are coupled with the pattern discovery algorithms such as clustering and association rule mining.**

Keywords: Web Recommendation, Collaborative base filtering, preprocessing technique, Association rule.

I. INTRODUCTION

Web mining plays very important role for finding the frequent data pattern from Internet, data set, data mart etc. World Wide Web has become a powerful platform to store and retrieve information as well as mine useful knowledge and use that knowledge to predict the interest of people. Due to the huge, dynamic and unstructured nature of Web data, Web data researchers have used recommendation system to prediction of user interest.

Web recommendation systems help the website visitors for easy navigation of web pages, quickly reaching their destination and to obtain relevant information. There are two types of approaches to developrecommendation system [1].

i. Content based filtering method [2],
ii. Collaborative filtering method [2].

In some cases combination of both the approaches [2] is preferred by the researchers.

In Content-based filtering technique [2] filtering is done based on customer's interested items. In content-based filtering technique, the web pages are recommended for a user very quickly from ancient database. In that database different content of items are added that the user has used in the ancient times and/or user's personal information and preferences. The user's data files can be constructed by using responses to questions, item ratings, or the user's navigation information to infer the user's preferences and interests. Using this method, recommendation can be done mainly from the available data-base and past experience of the website visitor. The disadvantage of this method is, not all the times users give their ratings properly for a website or web pages [2].

In collaborative filtering approach, [3] web pages are recommended to a particular user when other similar kind of users also prefers those web pages. For example, it may be defined as users having similar ratings of web pages or websites or users having related navigation behavior.

A collaborative filtering system collects all information about user's interest on the web site from the web servers/database and calculates the similarity among the users interest. Users have similar characteristics will be categorized to the same group. This method had two disadvantages: i. Sparsity[4] ii. Scalability[5]. Recommendation systems using collaborative filtering approach to find the neighborhood generally necessitate very long computation instance that grows linearly with both the number of customers and the number of products or web pages.

II. RELATED WORK

Bamshadet al [6] have proposed a new algorithm based on web usage mining called Profile Aggregations. In that algorithm Clustering is done on database with respect to similar kind of transactions and also page view clustering is applied to predict the similar pages in each transaction.

Yoon Ho Cho et al [7] has used decision tree induction method, association rule mining algorithms and data warehousing technologies to solve the problem of sparsity and scalability in collaborative filtering technique. Because of that new hybrid technique has improved the efficiency of the collaborative filtering approach by using web usage mining. Authors have used web logs as a database to find the frequent patterns using Apriori algorithm. And to classify the customers author use Decision tree induction method.

OlfaNasraouiet al [8] proposed Fuzzy approximation reasoning method on intelligent web recommendation system. They have extracted the user profile using used web usage mining and also they apply clustering method for grouping the user information on user database. For clustering they have used hierarchical unsupervised clustering method. And for recommendation they have used Fuzzy approximation reasoning techniques.

MagdaliniEirinaki et al [9] have proposeda clustering method to produce better and quick recommendations to the end user. For clustering they used semantically coherent clusters. Also for recommendation they use Domain ontology which is based on the keywords extracted from the web contents.

Feng Hsu Wanga et al [10] have used clustering and association rule mining using web usage mining for better recommendation. For clustering they used Hierarchical Bisecting Mediods.

Baoyao Zhou et al [11] have used sequential pattern mining technique for predicting the next web pages. In second step they used model base filtering technique, which stores the sequential web access patterns, and also useful for user pattern matching and recommendation rules generation.

Sumathi et al [12] have used Web usage mining techniques for determining the interest of “similar” Users. Authors divide it into two main parts:

i. Offline: for that part authors use Data Preprocessing, Pattern Discovery and Pattern Analysis methods.
ii. Online: In online part match the current user’s profile to the aggregate usage profiles is done. For that purpose use different recommendations techniques. In this paper for improving recommendation system quality, use active user search pattern which is comparison of different active users search pattern.

Mohammad et al [2] have used collaborative technique and content mining filtering techniques for web page Recommendation. They also use Fuzzy C-Mean and Ant colony clustering techniques on offline process (i.e. page matching with the previous similar users and recommendation similar products information).

Haibo Liu et al [13] have used the data structures such as Web-Interest Matrix, User-Interest Matrix, Class-Interest Matrix and Frequent-Path Matrix for recommendation and personalization of websites based on the user interest. Because of that product recommendation is developed on the basis of user interest.

FlorentGarcin et al [14] have used context tree technique for better recommendation of news and stories. This system provides better recommendation on the basis of user’s interest. Because of that improved prediction accuracy and recommendation quality.

III. COLLABORATIVE FILTERING TECHNIQUES

In recommendation system Collaborative filtering technique plays very important role. It uses only the rating data across huge dataset. In CF different customers rates ‘N’ items or have similar behaviors so CF will rate or act on other items similarly. CF technique use already available information from log servers related to items/user interest to predict items/user interest to different active (new) users which might like active (new) user. [1] To handle the increasing number of users and items, to make effective commendation in a short time period and also to deal with other problems like cold star problem, synonymy, data noise, CF algorithms must deal with highly sparse data. Basically CF techniques are divided into three parts:

- Memory-based collaborative filtering technique
- Model-based collaborative filtering technique
- Hybrid recommendationtechnique

a. Memory-based collaborative filtering technique:

The memory base collaborative filtering technique is use complete dataset related to user-item dataset. As described by Breese et. al Memory-based CF algorithms generally use rating matrix to store user-item database to generate recommendation. Mostly in memory base collaborative filtering technique use neighbors item datasets to find the interest of user, which use in future for all the ratings by referring to users or items whose ratings are similar to the other user or items.

b. Model based collaborative filtering technique:

The main drawback of memory base collaborative filtering technique is it use complete dataset related to user-item datasets and because of that this system is not work as fast as other collaborative system and also occurs scalability problem when generate real-time entries in recommendations system database. To overcome those problems, model-based recommendation systems are introduced by researchers. In Model-based recommendation systems use some small datasets called model. This model is design using extracting some information from the huge database related to particular parameter/attribute and uses this model every time without using huge database, because of that models increases both speed and scalability of recommendation system.

The design and progress of models permit the system to recognize somewhat complex patterns based on the training data, and then issue recommendations for the collaborative filtering tasks for testing data or real-world data, based on the fitted models. A model-based CF algorithm includes Bayesian models, cluster-based CF and regression-based methods to solve the shortcomings of memory-based CF algorithms.

c. Hybrid recommendation technique:

Now a day's hybrid collaborative filtering is more popular because it improves quality of web page recommendation or user interest recommendation. Hybrid Collaborative Filtering systems combine Collaborative Filtering with other recommendation techniques, to make better predictions or recommendations of web pages to new users [7]. The hybrid recommendation techniques are basically divided into two parts first that include all preprocessing methods and second that includes all rule finding. Because of that hybrid recommendation system improves the predication scalability and quality

.

Sr. No.	**Author Name & Title**	**Approach used**	**Techniques and algorithms used**	**Advantages**	**Disadvantages**
1.	BamshadMobasher, Honghua Dai, Tao Luo, Miki Nakagawa. " Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data".	Collaborative Filtering approach	Profile aggregation clustering	Fast results show because of clustering, Scalability problem solved.	Gray Sheep problem and cold star problems
2.	Yoon Ho Cho, Jae Kyeong Kim, SoungHie Kim. "A personalized recommender system based on web usage mining and decision tree induction. Expert Systems with Applications".	Collaborative Filtering approach	Decision tree induction and Association rule mining	overcome the problem of sparsity and scalability.	Gray Sheep problem and cold star problems occurs.
3.	Feng Hsu Wanga, Hsiu-Mei Shao. "Effective personalized recommendation based on time-framed navigation clustering and association mining. Expert Systems with Applications".	Collaborative Filtering approach	Hierarchical bisecting clustering and Association rule mining	Improve prediction quality.	Gray Sheep problem and cold star problems occurs.
4.	Harita Mehta, ShvetaKundraBhatia,	Collaborative Filtering	Entropy based similarity	Improve prediction quality	only for the trustworthy

	PunamBedi and Dixit V. S.“Collaborative Personalized Web Recommender System using Entropy based Similarity Measure”.	approach	measure		customers
5.	Shiva Nadi, Mohammad Hossein Saraee, AyoubBagheri. “A Hybrid Recommender System for Dynamic Web Users, International Journal Multimedia and Image Processing”.	Content base filtering approach	Rating techniques is used	Improve prediction quality	Gray Sheep problem
6.	OlfaNasraoui and Chris Petenes. “An Intelligent Web Recommendation Engine Based on Fuzzy Approximate Reasoning”.	Web usage mining	Fuzzy approximation reasoning techniques	Improve recommendation	Scalability problem
7.	Baoyao Zhou, Siu Cheung Hui and Kuiyu Chang. “An Intelligent Recommender System using Sequential Web Access Patterns. Cybernetics and Intelligent Systems”.	Web usage mining	sequential pattern mining technique and user pattern matching techniques	Improve prediction quality using recommendation rules generation method	Scalability problem
8.	Mohamed KoutheaïrKhribi. “Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval”.	Content and Collaborative filtering approach	Browsing History for e-learning	Improve efficiency, Scalability	Gray Sheep problem, cold start problem.

Table 1: Authors and their Techniques used for Web Recommendation

IV. PROPOSED METHODOLOGY

In this paper new system is proposed for more effective web page recommendation with the help hybrid collaborative filtering technique which is combination of collaborative filtering technique and pattern finding algorithms. For better patter discovery used CHARM algorithm has been used. To improve the result of CHARM algorithm, it has been applied on clustered datasets, because of that it captured all closed item sets very quick and perfectly from cluster dataset.The new recommendation system is developed with the following four processes:

(i) Data Preparation (also termed as data preprocessing in web usage mining field).

(ii) Clustering the web log files user wise.
(iii) Determining the associative patterns.
(iv) Web page recommendation.

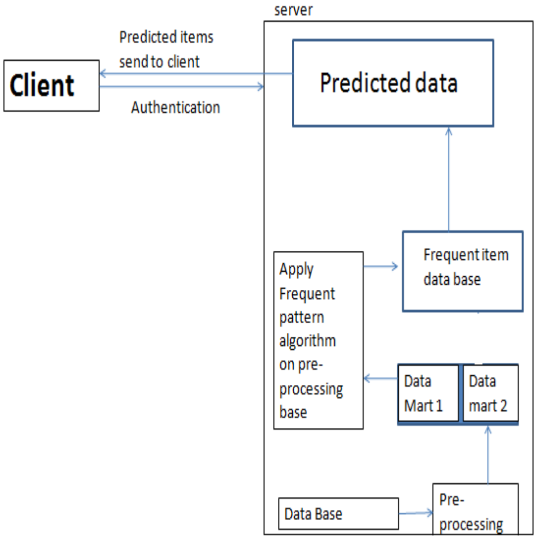

Fig 1: general architecture of proposed recommendation system.

4.1Data Preparation

Whenever the user interacts with the website, the interaction details are recorded in the web server in the form of web log files. Web log files [15] are maintained in the web servers in the form of plain text files. It is too difficult to use the web log files directly.

Preprocessing techniques are necessary for the web logs to discover the knowledge from them. Commonly the web log files are maintained in Web Servers (Web Logs for all the users), Proxy Servers (Maintained Somewhere) and Browser (Web Logs for the particular client).

The data preprocessing is considered as the important activity in web usage mining technique and treated as a key to success. It consists of the following steps:

Data Collection

Data collection is the first step in web log preprocessing. The web logs may be collected from web server, proxy server or client machine.

Data Cleaning

Data cleaning is the important step in preprocessing. In this step, irrelevant and noisy data are removed from the log files. The result of data cleaning has the fields like date, time, client ip, URL access, Referrer and Access log files.

4.2 Clustering the web log files based on user wise

For cloistering a datasets use K-mean algorithm.

K-means Algorithm

In K-mean algorithm data is clustered into different groups based on attributes/features into K number of groups. K is positive integer number. K-mean prototype (center-based) clustering technique which is one of the algorithm that solve the well-known clustering problem. It creates a one-level partitioning of the data objects.

K-means (KM) define a prototype in terms of a centroid, which is the mean of a group of points and is applied to dimensional continuous space. Another technique as prominent as K-means is K-medoid, which defines a prototype that is the most representative point for a group and can be applied to a wide range of data since it needs a proximity measure for a pair of objects. The difference with centroid is the medoid correspond to an actual data point [15].

4.3 Determining the associative patterns

For finding association rules use CHARM algorithm which is efficient algorithm for mining all closed frequent itemsets.

Association rule mining basically divided into two steps.

i. Finding the set of all frequent itemsets.
j. Testing and generating all basic rules among itemsets.

The CHARM algorithm is basically divided into two steps: in first step find some mine all frequent itemsets, in second step it mine the set of *closed* frequent itemsets, which is mostly used than the set of all frequent itemsets. The main advantage of CHARM algorithm is redundant rules can be eliminated, because of that prediction time and quality are improved.

3.4 Web page recommendation

For web page recommendation used any model base system to recommend the web pages. This model system is used to identify the next pages based on the sequence of previously visited pages by the users. When a new user enters to obtain the suggestion of web page, the sequence path of that user is compared with the CHARM algorithm pattern and it would recommend the web page using the probability definition.

V. CONCLUSION

Recently several recommendation systems have been proposed, that are based on collaborative filtering, content based filtering and hybrid recommendation technique. Collaborative filtering technique (CF) is one of the most successful recommendation techniques to solve the scalability problem related to recommendation system and also providing better recommendation. In this paper cover all collaborative based recommendation techniques which are used for better recommendation. Also proposed new improve collaborative filtering technique using Hybrid recommendation which is combination of both K-mean algorithm and CHARM algorithm. This Hybrid recommendation method improves the prediction quality of recommendation system.

REFERENCES

[1]Mobasher, B., Cooley, R. and Srivastava, J. 2000a. Automatic personalization based on web usage mining. Communications of the ACM, 43(8), 142–151.

[2]Shiva Nadi, Mohammad Hossein Saraee, AyoubBagheri. 2011. A Hybrid Recommender System for Dynamic Web Users, International Journal Multimedia and Image Processing (IJMIP), 1(1).

[3]Riecken, D. 2000. Personalized Views of Personalization. Communications of the ACM 43, (8) 27–28.

[4]Claypool M., Gokhale. A., Miranda. T., Murnikov, P., Netes, D., and Sartin, M. 1999. Combining content-based and collaborative filters in an online newspaper. ACM SIGIR '99 Workshop on Recommender Systems, Berkely.

[5]Sarwar. B., Karypis. G., Konstan. J., and Riedl. J. 2000. Analysis of recommendation algorithms for e-commerce. Proceedings of ACM Ecommerce Conference. 158–167.

[6]BamshadMobasher, Honghua Dai, Tao Luo, Miki Nakagawa. 2002. Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data.

[7]Yoon Ho Cho, Jae Kyeong Kim, SoungHie Kim. 2002. A personalized recommender system based on web usage mining and decision tree induction. Expert Systems with Applications 23, 329–342.

[8]OlfaNasraoui and Chris Petenes. 2003. An Intelligent Web Recommendation Engine Based on Fuzzy Approximate Reasoning. Proceedings of the IEEE International Conference on Fuzzy Systems - Special Track on Fuzzy Logic and the Internet.

[9]MagdaliniEirinaki, CharalamposLampos, StratosPaulakis, Michalis Vazirgiannis. 2004. Web Personalization Integrating Content Semantics and Navigational Patterns. WIDM'04, November 12-13. Washington, DC, USA. Copyright 2004 ACM 1-58113-978-0/04/0011.

[10]Feng Hsu Wanga, Hsiu-Mei Shao. 2004. Effective personalized recommendation based on time-framed navigation clustering and association mining. Expert Systems with Applications. 27, 365–377.

[11]Baoyao Zhou, Siu Cheung Hui and Kuiyu Chang. 2004. An Intelligent Recommender System using Sequential Web Access Patterns. Cybernetics and Intelligent Systems. IEEE Conference on Cybernetics and Intelligent Systems.

[12]Sumathi, C., P., PadmajaValli, R., and Santhanam, T. Automatic Recommendation of Web Pages in Web Usage Mining. (IJCSE) International Journal on Computer Science and Engineering 02(09),3046-3052.

[13]Haibo Liu, Hongjie Xing, Fang Zhang. 2012. Web Personalized Recommendation Algorithm Incorporated with User Interest Change. Journal of Computational Information Systems 8(4), 1383-1390.

[14]FlorentGarcin, Christos Dimitrakakis, Boi Faltings,2013. Personalized NewRecommendation with Context Trees. ACM Journal.

[15]Cooley, R., Mobasher, B., and Srivastava, J. 1999. Data preparation for mining world wide web browsing patterns. Journal of Knowledge and Information Systems.

[16]Mobasher, B., Dai, H., Luo, T., Sun, Y., and Zhu, J. 2000b. Integrating web usage and content mining for more effective personalization. Proceedings of the EC-Web 2000, 165–176.

[17]Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. 2000.Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, 1(2), 1–12.

[18]N. Chawla, S. Eschrich, L.O. Hall, "Creating Ensembles of Classifiers", IEEE International Conference on Data Mining, pp. 580-581, 2001.

[19]R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and DataMining, pages 307–328. MIT Press, 1996.

[20]M. DjamilaMokeddem, HafidaBelbachir ,"Distributed Classification using Class-Association Rules Mining Algorithm",IEEE 2010.

[21]Mohamed KoutheaïrKhribi, Mohamed Jemni1 and OlfaNasraoui. 2009. Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. Educational Technology and Society, 12 (4), 30–42.

[22]Harita Mehta, ShvetaKundra Bhatia, PunamBedi and Dixit, V., S. 2011. Collaborative Personalized Web Recommender System using Entropy based Similarity Measure. IJCSI International Journal of Computer Science. 8(6),3, 1694-0814.

[23]C.Borgelt. Efficient Implementations of Apriori and Eclat.Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations (FIMI 2003, Melbourne, FL).CEUR Workshop Proceedings 90, Aachen, Germany 2003.

[24]Mishra Shesh Narayan et al (2012), "An Effective algorithm for web mining based on Topic sensitive page rank algorithm", International journal of computer science and software engineering.

[25]Dimitris Antoniou, Mersini Paschou, EfrosiniSourla, Athanasios Tsakalidis, "A Semantic Web Personalizing Technique The case of bursts in web visits", IEEE 2010.

[26]Dr.Kanwal Garg," Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms",(IJRITCC)Volume 3, Issue 6, June 2013.

[27]Zhuang Chen ,Chongqing, "An improved Apriori algorithm based on pruning optimization and transaction reduction", China,(ETCSIT2012) .

[28]Dr. Deepak Garg,"FP-Tree Based Algorithms Analysis: FP-Growth, COFI-Tree and CT-PRO", (IJCSE)Vol. 3 No. 7 July 2011.

[29] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.

[30]K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.-P. Kriegel, "Probabilistic memory-based collaborative filtering," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 56–69, 2004.

[31]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, 2003.

[32]M. Claypool, A. Gokhale, T. Miranda, et al., "Combining content-based and collaborative filters in an online newspaper," in Proceedings of the SIGIR Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, Calif, USA, 1999.

[33]Jaideep Srivastava, Robert Cooleyz, Mukund Deshpande, Pang-Ning Tan proposed , "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data",IEEE 2000.

[34]Yogita S. Pagar, Vishakha. R. Mote, Rahul S. Bramhane, "Web Personalization using Web Mining Techniques", Emerging Trends in Computer Science and Information Technol2012 (ETCSIT2012) .