International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-6, Issue-5, 2019 Regular Edition E-ISSN: 2349-7084

HYBRIDIZATION OF WEB PAGE RECOMMENDER SYSTEMS BASED ON ML TECHNIQUES

S. R. Patil[1*]

[1*]*Information Technology Department, D.K.T.E.'s Textile and Engineering Institute Ichalkaranji India*

e-mail: shrenikrpatil@gmail.com

Available online at: http://www.ijcert.org
Received:15/05/2019, Revised: 24/05/2017, Accepted: 25/05/2019, Published: 29/05/2019

Abstract: - World Wide Web is the most significant source of information. Though the World Wide Web contains a tremendous amount of data, most of the data is irrelevant and inaccurate from users' point of view. Consequently, it has become increasingly necessary for users to utilize automated tools such as recommender systems to discover, extract, filter, and evaluate the desired information and resources. Recommender systems (RS) are widely used in e-commerce, social networks, and several other domains. Web page recommender systems predict the information needs of users and provide them with recommendations to facilitate their navigation. Web content and Web usage mining techniques are employed as conventional methods for the recommendation. Machine Learning techniques used for recommender system are Clustering, Association rules, and Markov models. These techniques have strengths and weaknesses. Combining different methods to overcome the disadvantages and limitations of a single system may improve the performance of recommenders. Hybrid recommender systems can be used to avoid the drawbacks or limitations of previous recommendation method. They combine two or more methods to improve recommender performance. In this paper, the four recommender systems are combined by using different hybridization methods. The effects of the hybrid recommenders are examined by comparing the results of a hybrid system against the results of a single recommendation method. The result shows that the hybrid recommender provides strong recommendation when all the systems of the hybrid generate the recommended page.

Keywords: Recommender Systems, Clustering, Association Rule Discovery, Machine Learning techniques, Hybridization Methods.

---

1. Introduction

The introduction should lead the reader to the importance of the study; tie-up published literature with the aims of the study and clearly, Machine learning is a research field that includes algorithms whose objective is to predict the outcome of data processing.

ML uses computers to pretend human learning and allows computers to recognize and obtain knowledge from the real world, and improve performance on some tasks based on this new knowledge. Today, there are a massive number of ML algorithms proposed in the literature. They can be classified based on the approach used for the learning process. There are four main classifications: supervised,

 https://doi.org/10.22362/ijcert/2019/v6/i05/v6i0501

unsupervised, semi-supervised, and reinforcement learning. Different Machine learning techniques have been used to develop efficient and effective recommendation systems. User satisfaction is an essential part of the recommender system. Today the quality of recommendations and the user satisfaction with such systems are still not most favourable. Recommender systems are not favourable for the quality of recommendations and user satisfaction. Methods used for the recommender system focuses on the different characteristics of the user. As a result, for the same data set, two recommender systems show the two different results. The most common Web usage mining techniques used for recommender system are Markov models, Association rules, and Clustering. These techniques have strengths and weaknesses. For example, lower order Markov models lack accuracy because of the limitation in covering enough browsing history; whereas higher order Markov models usually result in higher state space complexity. Association rule mining is a major pattern discovery technique. The main limitation of association rule mining is that many rules are generated, which result in contradictory predictions for a user session. The second limitation is that association rule mining is a non-sequential mining technique that does not preserve the ordering information among page views in user sessions. Recommendation system based clustering can capture a broader range of recommendations, though this is sometimes at the cost of lower prediction accuracy. Another drawback is Clustering methods are unsupervised methods, and usually, are not used for classification directly. Consequently, combining different systems to overcome the disadvantages and limitations of a single system may improve the performance of recommenders. Hybrid recommender systems can be used to avoid the drawbacks or limitations of previous recommendation method. They combine two or more systems with improving recommender performance. In this paper, hybrid recommender methods combining the results of different recommender systems are constructed in the following way: Initially, recommender system is implemented separately then the resulting predictions are combined by using hybrid recommender methods. Three hybridization methods are used, namely, Hit-Ratio based Ranking method, frequency-based ranking, and switching. In this paper, the effects of the hybrid recommenders are examined. This is achieved by comparing the results of the hybrid system against the consequences of a single recommendation method, and its performance is evaluated based on the correct prediction of the next request of a user, namely Hit-Ratio. Our detailed experimental results show that when choosing appropriate combination methods and modules, hybrid approaches achieve better prediction accuracy.

2. Related Work

There is a large body of work on Web usage mining, recommender system, and hybridization of a recommender system. The comprehensive review of some main work is done as follows.

Data mining associated with the Web, called Web mining, is divided into three domains: Web usage mining, Web content mining, and Web structure mining. The new classification of Web mining is provided in [8]. Various Web usage mining techniques have been used to develop efficient and effective recommendation systems. Resnick and Varian proposed the term recommender system to represent a system that takes user recommendations of items as inputs and uses these recommendations as a basis for making recommendations to other users.

Lü L.et.al. [12] reviewed recent developments in recommender systems and discussed the major challenges. In this paper, they have compared and evaluated available algorithms and examined their roles in future developments.

Burke, R. [7] proposed the landscape of actual and possible hybrid recommenders and introduced a novel hybrid, "EntreeC", a system that combines knowledge-based recommendation and collaborative filtering to recommend restaurants. The six hybridization techniques are surveyed and implemented in this work: weighted, mixed, switching, feature combination, feature augmentation and meta-level.

Uyar, A. S. et.al. [13] analyzed the effects of the different recommendation models which take into account different characteristics of user sessions. For this purpose, they used three different recommender models. The first one considers only the existence of the visited pages in a session and the view time of each page. In the second recommender model, only the order of the visited pages in each session is considered. The third model is based on the co-occurrence of the visited pages among user sessions. Their experimental results show that using the ordering information improves the prediction accuracy of the next request.

3. Proposed system

Offline phase and Online phase are present in the proposed system.

3.1 Offline phase:

The offline phase contains two components as follows

3.1.1 Data pre-processing:

The aim of data pre-processing is to collect and clean the data, identify the users, and create the sessions.

3.1.2 Pattern extraction component:

The pattern extraction component consists of four different modules, each of which is a recommender system using a different technique. These modules are clustering, association rule discovery, Markov model, and click-stream tree.

3.2 Online Phase:

The Online Phase also consists of two components:

3.2.1 Recommendation Engine:

This work consists of the implementation of a Recommendation engine, which consists of four recommender techniques, namely Recommender model based on clustering user sessions, Association Rule discovery, click-stream tree, and Markov model.

3.2.2 Hybridization Component:

The system combines results of multiple recommender models to produce a single output. Hit-Ratio based raking method; switching and frequency based ranking hybridization methods are used to combine recommendation.

4 Implementation

4.1 Data Pre-processing:

The data pre-processing technique contains four processes. These processes are:

4.1.1 Data Collection

The primary data source in Web Usage Mining is information residing on the Web sites logs. Server logs are the key input to the pre-processing phase. There are three most common sources of data, namely server side, proxy side, client side.

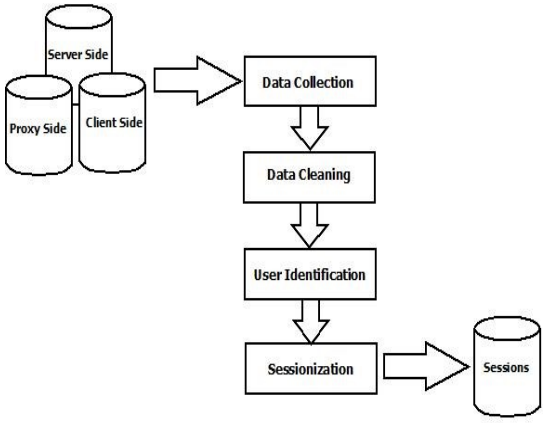

Fig. 1Phases of data pre-processing technique

4.1.2Data cleaning:

The data in the original Web user log files are raw; hence, not all the log entries are valid for Web Usage Mining. Thus the primary purpose of this process is to clean all the log files. All log entries with file name suffixes such as gif, JPEG, jpeg, GIF, jpg, JPG removed. As well as the entire request from the Web spiders is also removed from Web log files.

4.1.3User Identification

The same IP but different Web browsers, or different operating systems, in terms of type and version, means a new user.

The user identification process is used, to identify the user based on methods as mentioned above.

4.1.4 Sessionization

This process is used as one of the time-oriented heuristic methods for session identification. In this system, session-duration based heuristic method is used for the sessionization. The session-duration-based method aims to set a session duration threshold. If the duration of a session exceeds a certain limit, it could be considered that there is another access session of the user. Discovered from empirical findings, a 30-min threshold for total session duration has been recommended [9]. Result of session identification is sessions as shown in figure 1.

4.2 Pattern extraction component

The pattern extraction component consists of four modules.

4.2.1 Recommender system based on Clustering

A cluster is a collection of objects that are similar to each other and are dissimilar to the objects belonging to other clusters. Clustering is the technique used to group together items that have similar characteristics.

The main task in the session clustering is to assign a weight to Web pages visited in a session. The weight needs to be well determined to analyze a user's interest in a Web page.

Let P be the set of Web pages accessed by the user in Web server logs, P = {p1, p2...pm} each of which is uniquely represented by its URL. Let S be a set of user access sessions. S={s1,s2,…..sn}, Representation of each session is as vector model s_j= {w(p_1,s_j), w(p_2,s_j),….. w(p_m, s_j)}, where w(p_i,s_j) is weight assigned to the i^{th} Web page in j^{th} session. The w (p_i, s_j) needs to be determined to capture user interest in a Web page in the user session. The interest of a Web page is calculated by using frequency and duration. Frequency is the number of visits of a Web page and is given by Equation [12],

$$\text{Frequency} = \frac{\text{NumberOfVisit(Page)}}{\sum_{\text{Pages}\in\text{VisitedPages}}\left(\text{NumberOfVisit(Page)}\right)}$$

Duration is defined as the time spent on a page, i.e. the difference between the requested times of two adjacent entries in session. Duration is calculated as [12],

$$\text{Duration(Page)} = \frac{\frac{\text{TotalDuration(Page)}}{\text{Length(Page)}}}{\max_{\text{Page}\in\text{VisitedPages}}\left(\frac{\text{TotalDuration(Page)}}{\text{Length(Page)}}\right)}$$

where Duration of a Web page is further normalized by the max ''Duration" of pages in the session. The system uses the average duration of the relevant session as "Duration" of the last accessed Web page.

User's interest is always calculated with two strong indicators i.e. "Frequency" and "Duration".Interest degree of a Web page in the users is given by [12],

$$\text{Interest(Page)} = \frac{2 \times \text{Frequency(Page)} \times \text{Duration(Page)}}{\text{Frequency(Page)} + \text{Duration(Page)}}$$

Every user access session is transformed into an m-dimensional vector of weights of Web pages, i.e. s= {w1, w2, wm}, where m is the number of Web pages visited in all user access sessions [6][7].

To generate cluster vectorized sessions, K-means clustering algorithm is used. K-means is a prototype-based, simple partitioned clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids (a cluster centroid is typically the mean of the points in the cluster).

The clustering process of K-means is as follows:

1. The algorithm is composed of the following steps:
2. Partition object into k non-empty subsets randomly.
3. Compute the centroids of the clusters
4. The set membership of each object is decided by assigning that object to the nearest cluster centroid.
5. When all objects have been assigned, the value of the k centroids recalculated.
6. If none of the objects changed membership in the iteration, then generate final sets of the clusters otherwise repeat steps 3 and 4.

The center of that cluster represents the usage pattern for each cluster. The center of a cluster ct can be computed by calculating the mean vectors of the sessions assigned to the cluster:

$$\bar{\mu}_t = \langle w(p_1), w(p_2),,, w(p_n)\rangle$$

where w (pj) for cluster ct is given by

$$w(p_j)=\frac{1}{|c_t|}\sum_{s_i\epsilon c_t} w\left(p_j,s_i\right)$$

In the recommendation step, the cosine similarity metric is used to find a similarity value sim ($\bar{s_a}$,$\bar{\mu}$) between each cluster center$\bar{\mu}$and the active user session $\bar{s_a}$ given by,

$$\bar{s_a}=\langle w\left(p_1,s_a\right),w\left(p_2,s_a\right),,,w\left(p_n,s_a\right)\rangle$$

The best matching cluster is selected if that cluster has the highest similarity value, sim ($\bar{s_a}$,$\bar{\mu}$). A recommendation score is calculated by multiplying each weight in the cluster center vector by the similarity value of that cluster. The recommendation score of a page $p_i = p$ is calculated as follows

$$\text{rec}\ (\bar{s_a},p_i)=\sqrt{w(p_i)\times\text{sim}(\bar{s_a},\bar{\mu}\ )}$$

In this way, the recommendation score is generated for each page and the first k pages with the highest recommendation score are added to the recommendation set.

4.2.2 Recommender system based on Association rule discovery

Association rules capture the relationships among items based on their patterns of co-occurrence across transactions. Association rules that reveal similarities between the Web pages derived from user behavior can be simply utilized in recommender systems. To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on

support and confidence. In Web Usage Mining the support is defined as follows. The Support for a page is the number of sessions that contain the page whereas confidence of the association rule (X→Y) is the conditional probability that a session having X also contains Y.
The system makes use of the Apriori algorithm to find the groups of pages frequently occurring together in many user sessions. The basic intuition is that; any subset of a large itemset must be large. Therefore, the candidate itemsets having k items can be generated by joining large item sets having k-1 items, and deleting those that contain any subset that is not large. This procedure results in the generation of a much smaller number of candidate itemsets.

Candidate item sets generated from the previous step are used as input for recommendation engine to make a recommendation. The system uses a fixed-size sliding window over the current active session to capture the current user's history depth. For example, if the current session (with a window size of 3) is< *A, B, C* >, and the user references the page-view D, then the new active session becomes <*B, C, D*>. The recommendation engine matches the current user session window with item sets to find candidate page-views for giving recommendations [15]. The recommendation value of each candidate page-view is based on the confidence of the corresponding association rule whose consequent is the singleton containing the page-view to be recommended. If the rule satisfies a specified confidence threshold requirement, then the candidate page-view is added to the recommendation set.

4.2.3 Recommender system based on the Markov model:

Markov models are well-suited for modeling and predicting a user's browsing behavior on a Website. The Markov model mainly targets the User's navigation behavior. This denotes the input for the Markov model is the User's navigation behavior i.e., the user's sequentially accessed Web pages and the goal is to recommend the Web pages to the user. Three parameters are used to represent Markov model. i.e. < *A; S; T* >, where A denotes the set of all probable actions that can be performed by the user; S denotes set of all probable states used to build Markov, model,; and T is a |S| × |A| Transition Probability Matrix (TPM), where each entry t_{ij} corresponds to the probability of performing the action j when the process is in state i. Once the states of the Markov model have been identified, the transition probability matrix can be generated. Markov model uses the training set to generate the transition probability matrix. The transition probability matrix used to make a prediction for Web sessions by only considering the user's previous activity. The first k pages with the highest transition probability are added to the recommendation set.

5 Hybridization Techniques

The purpose of a hybridization block is to combine multiple recommender sets together to produce a single output. Hybridization process contains multiple techniques. These techniques are as follows

5.1 Ranking based on Occurrences(RBO)

The Ranking based on Occurrences hybrid first create one recommendation set by combining the individual recommendation sets of its modules and then applies a ranking method to sort the pages in this set. First, each of the modules is the hybrid generated a recommendation set. The recommendation set (RS) is obtained by the merging of the individual recommendation sets.
The system computes the scores for the pages by using the ranking method and based on these scores the pages are ranked. The final recommendation set is generated from the first k pages and recommended to the active user.
The score of each page depends on the total number of visits on that page. The Score of the page is defined as the ratio of the total number of visits on the page and number of pages. Once the scores are computed rank is assigned to each page and the final recommendation set is generated from top k rankers. The ranking method is also called as a Web page scoring method since this method assigns the score to every page on the Web site that reflects its popularity.

5.2 Ranking based on Hit-Ratio (RBHR)

A Hit-Ratio based Rankinghybrid presents recommendations of its different modules side-by-side in a combined list. However, the challenge of these types of hybrids is the integration of ranked pages in each recommendation set into the final recommendation set. Three phases same as previous technique are used to generate final recommender set.
Initially, in the training phase, training data are applied to each recommender system. In candidate set generation phase each recommendation module generates a candidate set consisting of k pages, based on an active session. The system assumes that each module generates uniformly

accurate recommendations so that it assigns equal weight to every module. The system finds the best and worst modules according to their Hit-Ratio for the last page of the user session. Hit-Ratio is defined as follows: A hit is declared if any one of the four recommended pages is the next requests of the user. The Hit-Ratio is the number of hits divided by the total number of recommendations made by the system. The system selects the two best modules and combines the individual candidate sets to get a final recommendation set, which consists of k pages.

6 Experiments details

For experiments, Synthetic dataset for dktes.com (SDD) [1] and hyperreal.org (SDH) [2] are used. Log data of dktes.com and hyperreal.org site is present in extended log format, which is supported by Microsoft Internet Information Server (IIS).

Total 16693 log entries from SDH dataset and 284187 log entries from SDD dataset are processed for the system. In the data cleaning step, first the irrelevant log entries with filename suffixes such as gif, jpeg, GIF, JPEG, jpg, JPG are eliminated, and all the log files are cleaned.
Table I presents some statistics of the pre-processed experimental dataset, including both training and testing sets.

Table 1: Statistics of the experimental dataset

Attributes	SDD	SDH
Total access entries	284187	16693
Clean access entries	55883	8968
Different access users	10000	1979
Accessed web pages	895	876
Identified sessions	1491	996
Sessions for the	1151	754
Sessions for the	340	242

Clusters are created by using the K-Means algorithm. WEKA machine learning tool is used to implement this clustering method. For the SDD, a total of nine clusters whereas for SDH total six clusters are created.
Association rules are generated by using the Apriori Algorithm. Apriori Algorithm available in the WEKA machine learning tool is used. Total 60,000 rules for SDH and 25000 rules for SDD are generated. Following table shows a sample of generated rules by Apriori algorithm.
A set of experiments are conducted with all of the recommender systems. Table 2 shows the results of these experiments as the Hit-Ratio of each recommender system. As shown in the graph, Clustering and Association Rule (AR) are having less Hit-Ratio compared with Markov Model (MM) The reason for this could be those recommender systems that consider the order of visiting pages have a better performance compared with the other models that represent user sessions differently (e.g., time spent on page or co-occurred pages).

Table 2: Hit-Ratio for the recommender systems

Methods	SDH	SDD
Clustering	35	30
AR	39	35
MM	82	84

Table 3: Hit-Ratio for the hybrid recommenders

Methods	SDH	SDD
RBHR	92	94
RBO	80	84

As can be seen from the Table3, the Hit-Ratio based ranking hybridization method has the highest Hit-Ratio, whereas frequency based Ranking method has the lowest hit ratio among the Hybridization methods. These results also show that recommendation accuracy is directly proportional to the switching criteria. This paper aims to examine the effects of hybrid recommenders. For this reason, the system is used to take the results of the hybrid recommender against the results of its modules.

7. Conclusion

By analysing the results of the hybridization methods, the following conclusions are drawn:
1. Modules of hybrid give better result than the result of the individual recommender system.
2. Comparison between Hit-Ratio of Recommender systems and Hit-Ratio of Hybridization methods shows that there is a correlation between the performance of the modules and the performance of the hybrid recommender methods. Any improvement of the Hit-Ratio of the modules will also have a positive impact on the performance of the hybrid recommender that uses these modules.

[1] http://www.dktes.com/
[2] http://www.hyperreal.com/

3. Choice of hybridization method is critical, but it increases the performance of the hybrid recommender system.

7.References

[1] Agrawal R., & Srikant R., "Fast algorithms for mining association rules," in J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), Proceedings of the 20th international conference on extensive databases, VLDB, 1994, pp. 487–499.

[2] Agrawal R., Swami A, Imieliński T., "Mining association rules between sets of items in large databases", Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93, 1993, p. 207.

[3] Barragáns-Martínez A. B., Costa-Montenegro E., Burguillo J. C., Rey-López, M., Mikic-Fonte, F. A., & Peleteiro, A. "A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition." Information Sciences, 180(22), 2010 p. 4290–4311.

[4] Burke R. "Hybrid recommender systems: Survey and experiments". User Modeling and User-Adapted Interaction, 12(4), 2007, 331–370.

[5] Deshpande M., & Karypis G., "Selective Markov models for predicting Web page accesses," ACM Transactions on Internet Technology (TOIT), 4(2), 2004, 163–184.

[6] Ericson, K., & Pallickara, S. (2011, December). On the performance of distributed clustering algorithms in file and streaming processing systems. In Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on (pp. 33-40). IEEE.

[7] Ericson, K., & Pallickara, S. (2013). On the performance of high dimensional data clustering and classification algorithms. Future Generation Computer Systems, 29(4), 1024-1034.

[8] Gündüz S. & Özsu M. T., "A Web page prediction model based on Click-Stream Tree representation of user behavior", in Proceedings of 9th ACM international conference on knowledge discovery and data mining (KDD), Washington, DC, USA, August 2003.

[9] Magdalini Eirinaki and Michalis Vazirgiannis, "Web Mining for Web Personalization" Communications of the ACM, vol. 3, No. 1, Feb. 2003 pp.2-21.

[10] Mobasher B., Dai H., Luo T., & Nakagawa M., "Effective personalization based on association rule discovery from Web usage data" in Web information and data management, 2001, pp. 9–15.

[11] Tao Luo, Bamshad Mobasher, Honghua Dai, Miki Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization,"Data Mining and Knowledge Discovery6(1), 2002,p.61–82.

[12] Vlado Kes˘elj, Haibin Liu "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests". Data & Knowledge Engineering 61, 2007, 304–330.

[13] Wang Q., Makarov D. J., and Edwards H. K., "Characterizing customer groups for an e-commerce website," EC'04, USA, 2004, p. 218-227.

[14] Lu, J., Shambour Q., Xu Y., Lin Q., & Zhang, G., "A Web-Based Personalized Business Partner Recommendation System Using Fuzzy Semantic Techniques. Computational Intelligence," in Press, 2012.

[15] Lucas, J. P., Segrera, S., & Moreno, M. N. (2012). Making use of associative classifiers to alleviate typical drawbacks in recommender systems. Expert Systems with Applications, 39(1), 1273-1283.

[16] Uyar A. S., Demir G. N., Goksedef M., "Effects of session representation models on the performance of web recommender systems," In Proceedings of the workshop on data mining and business intelligence, 2007, pp. 931–936.

Authors Profile

Prof. Mr. S. R. Patil received B.E. degree in Information Technology from DKTE'S Textile and Engineering Institute Ichalkaranji and M.Tech. in Computer science and Technology from Department of Technology Shivaji University, Kolhapur. He is currently working as Assistant Professor in DKTE Society's Textile and Engineering Institute, Ichalkaranji, Maharashtra. Research interests are in the field of Data mining, Computer Networks, and Web mining.

