A Survey on Web Personalisation and Recommendation Techniques

Dhaval Patel*, Amit Ganatra**, C. K. Bhensdadia***

Abstract

()

The quantity of accessible information on the web continues to grow rapidly and has exceeded human processing capabilities. The sheer amount of the information increases the complexity for users from discovering desired information. Recommendation systems have become a valuable resource for users seeking intelligent ways to search through enormous volume of information available to them. Web logs are important information repository which records users activates on search results. The mining of these logs can improve the performance of search engines, since user has a specific goal when searching for information. In this paper, a survey is provided on the different recommendation techniques with their advantage and drawbacks. A brief comparison of different personalisation techniques based on certain parameters is done.

Keywords: Log Mining, Personalisation, Recommendation Techniques, Web Usage Mining

Introduction

World Wide Web (WWW) has been proving to be source for tremendous amount of data; also data on WWW is growing exponentially in terms of both their size and its usage with respect to time. To obtain some interesting knowledge, access (browsing) patterns, user behaviours, the detailed and in-depth study of this data source are essential. Using this better design, structure and content can be created for website and represented to the user. In contrast to the standard data mining methods need to deal with the heterogeneous, semi structured or unstructured data (Sharma et al., 2011). Web mining aims to discover the useful information or knowledge from the web hyperlink structure, page content and usage data. Web mining can be categorised based on the interest and what kind of knowledge to be mined from the web data (Kosla & Blockeel, 2000).

1. Web Content Mining: It refers to mining or extracting the useful data from the web pages i.e. text, images and records etc. (Singh & Singh, 2010)

- 2. Web Structure Mining: It aims to discover the knowledge from the hyperlinks which improves the quality of website (Singh & Singh, 2010).
- 3. Web Usage Mining: It discovers users' access patterns from patterns from the web usage logs which records every click made by the users.

Web usage Mining can be divided into following stages in order to get the knowledge 1) Data Collection, 2) Data Preprocessing, 3) Pattern Discovery, and 4) Pattern Analysis. This extracted knowledge can be applied for efficient re-organisation of website, better personalisation and recommendation, improvement in links, and navigation (Singh & Singh, 2010).

In this paper the overview of what is web data mining, process of Web Usage Mining (WUM) and in-depth review of personalisation and recommendation methods for WUM is provided. The second section discusses about the categories of the recommended system. The third section discusses the related work done in recommended system. The fourth section gives issues and challenges

۲

^{*} Dharamsinh Desai University, Nadiad, District Kheda, Gujarat, India. E-mail: dhav.p07@gmail.com

^{**} CHARUSAT, Changa, District Anand, Gujarat, India. E-mail: amitganatra.ce@charusat.ac.in

^{***} Dharamsinh Desai University, Nadiad, District Kheda, Gujarat, India. E-mail: ckbhensdadia@ddu.ac.in

of the system. The fifth section gives the comparative analysis of various approaches. The sixth section gives conclusion and directs towards the future work.

Recommendation System

A recommendation system is a software program which attempts to narrow down the selections for the users based on their expressed preferences, past behaviour and other data which can be mined about the users or other users with the similar interests. The current recommendation system can be divided into four categories based on the knowledge sources they use to make recommendation (Anand & Mobasher, 2005).

- 1) Content based recommendation,
- 2) Collaborative recommendation,
- 3) Demographic recommendation,
- 4) Hybrid Recommendations.

Content based filtering systems have their roots in information retrieval. The approach to recommendation generation is based around the analysis of items previously rated by a user and generating a profile for a user based on the content descriptions of these items. The profile is then used to predict a rating for previously unseen items and those deemed as being potentially interesting are presented to the user (Anand & Mobasher, 2005). The items that have high degree of similarity to the users preference would be recommended.

Collaborative filtering recommend items to the active user, previous feedback is used to find other likeminded users which are known as user's neighborhood. These are users who have provided similar feedback to a large number of the items that have been consumed by the active user. Items that have been consumed by like minded users but not by the current user are candidates for recommendation. The assumption made by these systems is that users that have had common interests in the past, will have similar tastes in the future (Anand & Mobasher, 2005).

Demographic recommendations categorise the user based on personal attributes and makes recommendations based on demographic classes, e.g. college students, teenagers, women, men, etc. The advantages is that it does not depend upon any database of users rating and disadvantage of this system is that the knowledge base needs to remain

۲

updated to keep an eye on the changing preferences of customer.

The hybrid recommenders have been motivated by the observation that single recommendation techniques have some difficulties which is needed to be overcome. These can be done by the approach of combining two recommended system. The most common form of hybrid recommender combines the content based and collaborative based filtering (Anand & Mobasher, 2005).

Related Work

Fuzzy Logic based Product Filtering Algorithm for Web Personalisation (Hua et al., 2007)

In this approach, the authors proposed a method based on the fuzzy logic for the product filtering instead of the collaborative filtering and rule based filtering. Two information segments and two processing segments are considered in this approach. Information module collects the user's details and processing modules are used to measure client linkage.

The preference learning uses the fuzzy logic method which deals with vague data from users. The author suggests to add the fuzzy logic for measurement of users' likeness. The membership function of fuzzy logic describes how each fact of input mapped to the membership value in the interval [0, 1].

The proposed method deals with the unclear data of user's activities; the suggested system produced most appropriate and meaningful value based on user's behaviour and access time. Thus the fuzzy logic system can be used as an alternative web personalisation system but the difficulty is in selection of appropriate membership function for fuzzy logic.

A Web Personalisation System based on Users Interested Domains

(Lei & Fan, 2008) In this paper authors proposed a model which can recommend the web pages to the users interested in. This model finds out the users interest based on the past behaviour on the web. In this approach the training data needs to be collected like past click history, users' interested domain then these data and store it into the knowledge base for better recommendations. The authors use the k-means algorithm to cluster the data. The algorithm starts by portioning the input points into k initial sets randomly and then calculates the mean point or centroid of each set. The second task is to find out the cluster representative the new user similar to which can be achieved using the k-Nearest Neighbor (k-NN). When the user comes, his or her click history is collected by the system and using k-NN approach could get the k similar users' interested domain lists from knowledge base. If more than one domains are selected, the system gives a user list of union of all the interested domains from the list of domains.

Because this model is just based on the user's interested domain based on the past click history without considering time constraint the user spent on each page or the content, the system can't detect the actual interest of particular user so the recommended services may not be relevant for the particular user.

Recommendation of Optimised Web Pages to users Using Web Log Mining Techniques (Bhushan & Nath, 2012b)

In this paper the authors proposed a web recommendation system which is based on the learning from web logs and recommends user a list of pages which are relevant to him by comparing with users historic patterns. The proposed architecture in paper can be divided into two main phases back end and front end phase.

In back end phase there are two modules: Data Preprocessing and Sequential Pattern Mining. Preprocessing is the web usage mining aims to reformat the original web logs to identify user's access sessions and provide the data in the required form in which the mining techniques can be applied. Sequential Pattern Mining technique is to determine the sequential patterns in each cluster. A web access sequence is an ordered sequence of access events. For all the patterns P in the candidate set with length n, all URL sequences are processed once and the count is incremented for each detected pattern in the candidate set. Candidate sets whose support is less than the support threshold are eliminated by the module.

In the front end phase, URL request given by the user is processed by search engine and captures the recommendation list of web pages relevant to the user query and the rank updating algorithm is applied on them.

۲

Matching query algorithm submits the query on query interface and compares with the already existing index document from the knowledge base by using the similarity function and based on this the matching document is used for recommendation list. Rank updating algorithm takes in the set of matched documents returned from matching query algorithm. Then for each page from the sequential pattern it calculates its value based on the order it is accessed and the importance of the user. Based on this, the new improved rank of pages is calculated and thus the popular and relevant pages gain the upwards position in the recommendation list.

The result shows that the proposed approach improves the relevancy of the pages and thus reduces time spend in getting the required information. Future extension suggest to give perfect relevancy of the query in terms to the user.

Web Personalisation Based on Weighted Association Rules (Forsati, Meybodi & Neiat, 2009)

In this approach, authors extended the generalised association rule method by assigning a significant weight to the each element in a transaction to examine the importance of each item within the transaction and extend a new process based on the projected weighted association rule mining method. In association rule based recommendation, the researches do not consider the time spent by the user on each page as an implicit rating for the web page. They only assumed all the access web pages are equally important to the users' navigation pattern.

In weighted association based rule miner, the author considered the time spent by each user on each page and visiting frequency of each page are used to assign the quantitative weight to the pages instead of the traditional binary weights. The methodology is like this: first, the weighted association rules of each URL will be extracted from the web log data and similarity between active user sessions will be calculated upon the weighted rules instead of an exact match for finding the best rule. Duration is the time spent on each page, it reveals the significance of each page to the individual user, because a customer spend extra time on more interesting item, if he/she is not interested in a page then generally switch to interesting or new page rapidly. Frequency is the count that a web page is retrieved by different customers. Higher frequency of web pages indicates the higher interest to the users about

those pages. Author used time spent on the page and visiting frequency by user askey statistics in evaluating the customer's attention on the web page.

In this approach high value of weights are assigned to more interesting items according to the individual user, this weight can be used to calculate the importance of a web page to a user and give better recommendation which is highly useful to the customer. But the approach has the high complexity due to increase in the parameter to calculate weights of pages.

An Efficient Algorithm for Web Recommendation Systems (Forsati Meybodi & Rahbar, 2009)

In this paper the authors propose the effective and scalable techniques to solve the web page recommendation problem for that uses the distributed learning automata to learn the behaviour of previous users and cluster pages based on learned patterns. The authors deal with the unvisited or newly added pages as they would never be recommended. The opportunity should be included into the recommended set. The author stryto quantitative weight to each page taking into the account the degree of interest.

The methodology is like this: first, cluster the pages based on the user's usage pattern. Second, the weighted associations rules of each URL will be extracted from the web log data and similarity between the active user session will be calculated upon the weighted rules instead of the exact match for finding the best rule. The recommendation engine will then find the most similar rules to the active user sessions with the highest weighted confidence. Finally, HITS algorithm is applied to rank the candidate set based on the authority and hub values from the given cluster. A good authority is a page pointed by many good hub pages and a good hub is a page that points to many good authority pages. As a result hub values are used to rank the candidate's recommendation set. The highest recommendation score is chosen as the recommendation to the active user.

The algorithm solves the problem of recommending rarely visited or newly added pages and provides the opportunity for these rarely visited or newly added pages to be

۲

included in the recommendation list. The results show that method is better in precision and coverage rates than the conventional association rule based recommendation.

Automatic Recommendation of Web Pages for Online User Using Web Usage Mining (Bhushan & Nath, 2012a)

In this paper the authors proposed a web recommendation system which is based on the learning from web logs and recommends user a list of pages based on users historic patterns and list of web pages which have not been visited yet. The proposed architecture in paper can be divided into two main phases: back end and front end phase.

In back end phase there are two modules: Data Preprocessing and User Navigation Pattern Mining. Preprocessing is the web usage mining which aims to reformat the original web logs to identify user's access sessions and provide the data in the required form in which the mining techniques can be applied. User Navigation Pattern Mining is used to determine sessions with similar navigation patterns/ interest from user session files. A model based acceptation maximisation clustering techniques is used to determine the user session classes. Each cluster represents several user session of similar usage pattern. Hence clusters of patterns for user session are obtained and fed into knowledge base for future analysis.

In the front end phase URL request given by the user is processed by intelligent agent engine and captures the intuition pages of the user. Matching profile algorithm captures the active session of the user dynamically. The new session is compared with the already existing usage profile from the knowledge base and then the similarity to the active session with the discovered sessions is applied and selected based on the threshold. The matching clusters are used for the recommendations. Searching of unvisited pages consist of unvisited pages being compared to other users pattern and best possible profile is achieved by considering the maximum scalability. Subsequently pages are removed and unvisited page list is created.

The result shows that the proposed approach improves the accuracy to satisfactory level. Future extension suggests to give perfect accuracy to the user.

Application of Neural Network and Kano's Method to Content Recommendation in Web Personalisation (Chang *et al.*, 2009)

In this paper the authors have an objective to give users a website they want or need and thus knowing the needs of users an important task for the recommendation in web personalisation. The authors proposed an approach that trains the artificial neural network to group users into the different clusters and applies the well-established Kano's method to extract the implicit needs from users in different clusters.

The proposed approach for creating the personalised website consists of three phases, which is based on the integration of the ART (Adaptive Resonance Theory) network and Kano's method. The first phase involves the constructing and collecting data from the users regard to their browsing features. An ART network is applied in the second phase to build the group of user based on collected samples in first phase. The ART is an unsupervised learning theory. It has two layers - (i) comparison layer to receive the input vector and move the input to the finest match in the recognition layer, and (ii) recognition layer to enhance the major output and suppress others. The final phase includes the user's features of each cluster and categorises their requirements using Kano's method. The Kano-ANN method arises from the theory of collaborative filtering method by the synthesis of Kano's technique that is well organised psychology-based client satisfaction approach. This method covers the system to grab the linking and requirements of an individual customer.

The key role of this concept is to deal with the problem product and service recommendation in tailored manner from the viewpoint of psychology-based user satisfaction. Also, concerning artificial neural network to the customer clustering instead of classical approach like K-means algorithm, leads this method more adaptive to the new customers.

Techniques for Adaptive Website and Web Personalisation Without any Users Effort (Arora & Kant, 2012)

In this paper authors describe various techniques for web personalisation and then try to find out the problems associated with these techniques and further describe new techniques in order to overcome the problem.

All the techniques for recommendation require efforts from user in terms of ratings, feedback, registration, reactions etc. These types of methods may become time consuming and mostly users are unwilling to fill the information. These techniques are dependent on the users efforts and thus modifications are needed so that web personalisation can be done but without the efforts from the web users. In this approach, authors propose page gather algorithm to obtain the personalised results. In this algorithm applicant link sets are recognised without human intention and integrated with the catalog or index based on the user access log. The algorithm automatically records the user access patterns at the website. This algorithm takes a web server log as an input and maps it into a form ready clustering and then applies cluster mining to the data and produces candidate index page content as an output. Various new approaches to handle personalisation of website that do not require user efforts must be developed.

Issues and Challenges

Web personalisation aims to provide the users with what they need without requiring them to ask for it explicitly which means that a personalisation system must somehow know what the users need based on the previous or current interactions with the user. Studies have brought into light a number of issues and challenges that must be addressed (Ozgobek *et al.*, 2012; Soanpet & AdiLakshmi, 2014) ()

1. Cold Start Problem

The problem that the system cannot recommend new items if they do not have any clicks from other users or when there is no data from a completely new user then it is not possible to make recommendation.

2. Data Scarcity

۲

If the number of items is more than the number of users then it requires too many ratings to fill the item-user matrix.

3. Implicit User feedback

It is not practical for the system to interact with the user continuously so the system should be able to collect implicit feedbacks efficiently while protecting the user privacy.

18 International Journal of Knowledge Based Computer System

Parameters/Papers	Techniques used	Advantages	Drawbacks	Complexity
Fuzzy logic based product filtering algorithm for web personalisation (Hua et al., 2007)	Fuzzy logic for filtering	Deals with uncer- tainty and vagueness of users activities	Correctness and defining of the membership function	Medium
Web Personalisation system based on user's interested Domain (Lei & Fan, 2008)	Clustering Techniques	Recommendation is given based on user's interest only	Generates recommendation based on past history only and can't detect actual interest of user.	Medium
Recommendation of optimised web pages to users using web log min- ing techniques (Bhushan & Nath, 2012b)	Rank update and Matching query	Good relevancy and reduces the time spent of getting the required information	No recommendation of unvis- ited pages or new users	Medium
Application of Neural Network and Kano's method to content Recom- mendation in web personalisation (Chang et al., 2009)	Artificial Reso- nance Theory (ART) and Ar- tificial Neural Network	Recommendation based on the psychol- ogy of the user.	Complexity in implementation of neural network.	High
Web personalisation based on weighted rules (Forsati, Meybodi & Neiat, 2009)	Weighted Schema of as- sociation Rules	Gives recommenda- tion which has high weighted value	Complexity is more due to the increase in parameter	High
An efficient algorithm for web recommendation (Forsati Meybodi & Rahbar, 2009)	Distributed learning au- tomata	Able to recommend rarely visited or newly added web pages	Less accurate and not good relevancy	High
Automatic recommendation of web pages for online users using web usage mining (Bhushan & Nath, 2012a)	Matching pro- file algorithm applied on the user navigation pattern	Improves accuracy and gives unvisited pages	Classification of user naviga- tional pattern	High
Techniques for adaptive website and personalisation (Arora & Kant, 2012)	Path generator algorithm	Less user effort required	Issue in Overlapping of clusters	Low

Table 1: Comparative Analysis of Various Approaches

۲

4. Scalability

((()

Recommended systems are aimed to serve millions of users and number of items to be recommended can be very high. It is the most important problem of the recommended systems.

5. Accuracy

Recommendation link must be generated in such a way that it should provide the users with accurate recommendation.

6. Privacy Problem

To make proper recommendation system should know about users past preferences, interests and even the relations with other people. This data about the user often cause the privacy problem.

Analysis of Approaches

In Table 1 comparisons of the various approaches with their advantages and disadvantages are discussed. The complexity refers to the number of iterations required for the computation which is decided as low, medium and high.

Conclusion

In this paper, various approaches for web personalisation and recommendation techniques have been discussed. All the approaches have some advantages and disadvantages but the goal is to provide relevant information and services

۲

to the users. A comparative analysis on the basis of certain parameters has been discussed in Table 1.

References

- Anand, S. S., & Mobasher, B. (2005). Intelligent techniques for web personalization.Springer
- Arora, K., & Kant, K. (2012). Techniques for Adaptive Websites and Personalization without any User Effort. IEEE Students Conference on Electrical, Electronics and Computer Science.
- Bhushan, R., & Nath, R. (2012a). Automatic Recommendation of Web Pages for Online Users using Web Usage Mining. International Conference on Computing Sciences.
- Bhushan, R., & Nath, R. (2012b). Recommendation of optimized Web pages to users using web log mining Techniques. IEEE 3rd International Conference on Advance Computing.
- Chang, C. C., Chen, P. L., Chiu, F. R., & Chen, Y. K. (2009). Application of neural networks and Kano's method to content recommendation in web personalization. *Experts Systems with Applications*, 36(3), 5310-5316.
- Forsati, R., Meybodi, M. R., & Neiat, A. G. (2009). Web Page Personalization based on Weighted Association Rules. International Conference on Electronic Computer Technology.

- Forsati, R., Meybodi, M. R., & Rahbar, A. (2009). An Efficient Algorithm for Web Recommendation Systems. IEEE/ACS International Conference on Computer Systems and Applications, (pp. 579-586).
- Hua, B., Wong, K. W., & Fung, C. C. (2007). Fuzzy Logic Based Product Filtering for Web Personalization in E- Commerce. In IEEE Region 10 Conference.
- Kosla, R., & Blockeel, H. (2000). Web mining research-A survey. Sigkdd exploration, newsletter of sig on knowledge discovery and data mining. *SIGKDD Explorations*, 2(1), 1-15.
- Lei, M., & Fan, L. (2008). A Web Personalization System Based on the Users' Interested Domains. In Proceedings of 7th International Conference on Cognitive Informatics.
- Ozgobek, O., Gulla, J. A., & Erdur, R. C. (2012). A Survey on Challenges and Methods in News Recommendation.
- Sharma, K., Shrivastav, G., & Kumar, V. (2011). *Web Mining- Today and Tomorrow*.3rd International Conference on Electronics Computer Technology.
- Singh, B., & Singh, H. K. (2010). Web Data Mining Research: A Survey. IEEE Conference on Computational Intelligence and Computing Research, (pp. 1-10)
- Soanpet, S. L., & AdiLakshmi, T. (2014). A survey on challenges and methods in news recommendation. *International Journal of Computer Science and Information Technologies*, 5(4), 5771-5772.