A Survey on Web Page Recommender Systems

Bhoomika AP [*], Dr. Selvarani R [**]

*Department of Computer Science, ACED, Alliance University
**Department of Computer Science, ACED, Alliance University

Abstract- The huge information in the web has created a big challenge for users to find relevant and useful information. Web page recommender systems solve this problem by suggesting the products or web pages of user interest. There are different types recommender systems generate useful recommendations based on navigational behavior of user. FAST algorithm is one of the sequential frequent pattern mining algorithms that is capable of mining complete set of patterns by greatly reducing the effort for support counting and candidate sequences generation. This paper discusses different types of recommender systems and proposes a hybrid context aware recommender system that is based on integrating semantic knowledge at different stages of web usage mining and FAST algorithm, coupled with clustering and sequential association rule mining.

Index Terms- web page recommendation, web usage mining, sequential patterns, association rule mining.

I. INTRODUCTION

World Wide Web (WWW) is the biggest source of information. The size and complexity of web is getting increased as it tends to grow at an exponential rate [1]. Although web contains huge amount of information that is useful to the users, there is a problem in finding desired information easily. Often users find it difficult to extract most relevant information in the right time from large information space. This is very crucial in e-commerce web sites where they can lose the customers very easily. So, it is necessary for users to make use of automated tools such as recommender systems in order to obtain desired information.

There are different approaches in recommending web pages.

1. Content based filtering
2. Collaborative filtering
3. Web usage based recommender system

In content based filtering the key words in the web pages are used to describe the items. In this approach, a user profile is built based on his preferences to indicate the type of items he likes [2]. Most similar items are recommended by comparing various items in the web pages with the items that are previously rated by the user. Collaborative filtering methods generate recommendations to a given user based on their similarity to other users. This method works by collecting and analyzing information on users' behaviors, activities or preferences in order to generate recommendations.

Web usage mining is a technique that analyses user information to find access pattern of web pages. The user related information can be usually obtained from web server log. The web server log is further mined to obtain usage patterns. Recommendations are generated to a given user by comparing the active user session with previously mined usage patterns.

There is hybrid recommender systems, which are the combination of the different approaches discussed above. Sparsity and scalability problems are the disadvantages of collaborative filtering. Content based filtering fail to predict the future interests of user since solely the content of the webpage is considered for recommendation. Recommendations generated from web usage patterns suffer from new item problem, where the recommender system fails to recommend newly added items to user. In recent days researches have been carried out to combine semantic information at various stages of web page recommender system based on web usage mining to generate meaningful and accurate recommendations.

Prefix span algorithm is used to generate frequent sequential navigational patterns. These patterns are in turn used to mine association rules and to generate web page recommendations. The proposed approach uses FAST algorithm to generate sequential patterns. Researches have showed that FAST algorithm outperforms prefix span algorithm and other existing sequential pattern mining algorithms in generating sequential patterns in terms of speed and memory. Lastly sequential rules can be mined from generated patterns. Meaningful and accurate recommendations are generated for given user by considering active user session, context information and previously mined sequential rules.

II. RELATED WORK

Robin van Meteren and Maarten van Someren presented [3] a recommender system that makes use of content-based filtering techniques to suggest items to users. A content-based filtering system selects items based on the correlation between the content of the items and the user's preferences. Explicit feedback and implicit feedback from user and tf-idf scheme are considered to find the document they are interested in. The recommender system provides dynamic hyperlinks to web pages that contain the items of user interest. The presented system failed to predict the future interests of user.

Badrul Sarwar et al presented a recommender system [4] based on item based collaborative filtering. The conventional collaborative algorithms suffered with scalability and sparsity problem. To alleviate these problems, item-based collaborative filtering considers the relationships between items over relationships between users. Cosine based similarity and adjusted cosine based similarity are used to find the similarity of items. The presented system produced much faster recommendations when compared to conventional user based collaborative filtering systems.

Sule Gunduz et al presented an approach [5] that considers order of pages in a session, the distance between identical pages, and the time spent on these pages for providing recommendations. The authors introduced a similarity metric to find pairwise similarities between user sessions. Clustering of user sessions is carried out based on pair wise similarity and the clusters are represented by using a click – stream tree. The click-stream tree is used to generate recommendations to the users.

Jia Li et al [6] present a framework for a combined web recommender system, in which users' navigational patterns are automatically learned from web usage data and content data. These navigational patterns are then used to generate recommendations based on a user's current status. The items in a recommendation list are ranked according to their importance, which is in turn computed based on web structure information. The presented system overcomes limited information problem, incorrect information problem and persistence problem observed in existing recommendation systems. There was an increase in recommendation accuracy and up to date recommendations were obtained.

Kim et al presented an approach that combines content-based filtering and collaborative filtering [7] to utilize the strengths of both approaches for achieving good performance. The recommender system used group rating matrix obtained from user profiles for generating web page recommendations. The presented system could overcome the disadvantages of individual content and collaborative filtering approaches and makes better use of strengths offered by each approach.

Taowei Wang et al presented an approach [8] for generating recommendations based on collaborative filtering and web usage mining. To enhance recommending quality, the recommender system made use of Uniform Resource Locator (URL) related analysis and K-means algorithm. The presented system could overcome the sparsity problem of collaborative systems.

Reza Samizadeh et al presented an approach [9] to overcome the drawbacks of user-based collaborative filtering such as sparsity, scalability, new item problem etc., by combining of semantic knowledge with web usage mining. Scalability problem was alleviated by using web usage mining in collaborative component. New item and sparsity problems were overcome by extracting and utilizing semantic information in the web pages. The semantic patterns were created by integrating domain knowledge in the form of ontology with navigation patterns. The presented system provided good recommendation accuracy.

C. Ramesh, et al presented a recommender system [10] that integrates semantic information at different stages of web usage mining process. The main advantage of presented approach is incorporating semantic information into web usage mining process which could provide more interesting patterns that consequently makes the recommendation system more functional, smarter and comprehensive. While the limitation is, web pages with semantic information cannot be validated as no standard parsers are available for validation of such web pages.

Soheila Abrishami et al presented a web page recommender system [11] based on semantic web usage mining. The presented system integrated sematic information at different stages of web usage mining. Frequent sequential patterns were generated using prefix span algorithm. Sequential association rules were generated by using Rule gen algorithm. The presented system generated accurate, more meaningful recommendations and showed high precision and coverage compared to other systems.

Mehdi Hosseinzadeh Aghdam [12] presented a context aware recommender system based on hierarchical hidden Markov model. Context aware recommender systems consider contextual information that affects user information and states. Many model-based recommender systems, such as matrix factorization or neighborhood-based methods, do not consider changes in user's interests to recommend items. Context-aware recommender systems take into account changes in user preferences by modeling them in time. Hierarchical hidden markov model identifies the changes in user’s preferences over time by modeling the latent context of the users. Using the user-selected items, the proposed method models the user as a hidden Markov process and considers the current context of the user as a hidden variable. The latent contexts are automatically learned for each user utilizing hidden Markov model on the data collected from the user’s feedback sequences. The incentive to use the latent context is privacy, data access, and usability considerations. The proposed model has a better performance ad showed diversity in recommendations compared to existing methods.

III. MOTIVATION

As discussed above, there are different types of recommender systems which generate relevant webpage recommendations to the user. In content based recommender systems, recommendations are provided solely based on content of the web pages, but these systems predict future interests of users. In collaborative filtering, products are recommended based on user- user or item- item similarity, but these systems do suffer from scalability and sparsity problems.

Table 1- Authors and their approaches for web page recommendation

Sr. No	Author Name and Title	Approach	Algorithms and Techniques	Advantages	Disadvantages
1	Robin van Meteren and Maarten van Someren Using content Based Filtering for Recommendation.	Content Based Filtering	tf-idf	User independence, Transparency, No cold start	Unable to predict future interests of user
2	Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl Item – Based Collaborative Filtering Recommendation.	Collaborative Filtering	K nearest neighbor approach,	No cold start and sparsity problem	Expensive model building
3	Sule Gunduz and M. Tamer Ozsu Web page prediction model based on click- stream tree Representation of User behavior.	Web Usage Mining	Click stream tree, clustering	No cold start and sparsity problem.	Unable to generate meaningful recommendations to user
4	Jia Li and Osmar R. Zaian Combining Usage, Content, and Structure Data to Improve Web Page Recommendation.	Content Based Filtering, Web usage Mining	Missions, HITS algorithm	Overcomes persistence problem, incorrect information problem	Unable to generate meaningful recommendations to user
5	Kim, Byeong Man, Qing Li, Chang Seok Park, Si Gwan Kim, and Ju Yeon Kim Combining content-based and collaborative filters in web page Recommendation.	Content based filtering, Collaborative filtering	Group rating matrix, Adjacent cosine algorithm	Overcomes the disadvantages of content based filtering and collaborative filtering	Scalability problem
6	Taowei Wang and Yibo Ren Web Page Recommendation Based on Web Usage Mining Using Collaborative Filtering Technique.	Collaborative Filtering, Web Usage Mining	K means clustering algorithm	Good recommendation accuracy	Scalability and sparsity problems
7	Reza Samizadeh and Babak Ghelichkhani Improving Web Page Recommendations using web usage Mining and Semantic information.	Semantic Web usage Mining	BOW	Good recommendation accuracy	No efficient standard parameters are defined for evaluating semantic similarities.
8	C.Ramesh, Dr. K. V. Chalapati Rao, and Dr. A. Goverdhan A Semantically Enriched Web Usage Based Recommendation Model	Semantic Web usage Mining	Clustering and association rule mining	Good recommendation accuracy	No standard parsers available for validation of web pages
9	Soheila Abrishami, Mahmoud Naghibzadeh, and Mehrdad Jalali Web Page Recommender System Based on Semantic Web Usage Mining	Semantic Web usage Mining	Prefix span, Rule gen	Good recommendation accuracy	No standard parsers available for validation of web pages
10	Mehdi Hosseinzadeh Aghdam Context-aware recommender systems using hierarchical hidden Markov model	Collaborative filtering, Context aware recommendations	Hierarchical Hidden Markov Model	Accurate and Diverse Recommendations	Scalability problem

In recent days, web usage mining has gained notable consideration for finding user behavioral patterns. Web usage mining is a technique for mining and analyzing the web server logs for finding

interesting usage patterns. Despite of this, one of the most important disadvantages of this approach is result is produced in the form of web page addresses so that common navigation profile does not have any semantic meaning. The patterns often fail to reason about user's underlying interests and preferences. Thus recommendations generated may not be accurate and the system would not be able to interpret and explain the recommendations. New item problem is the another disadvantage of web usage mining; it is the failure to recommend newly added pages or products to the visitors since these products or pages are not in the current common navigation profiles. This arises the need of a strategy that alleviates problems of existing recommender systems.

The combination of web usage mining and semantic web has created a new and fast emerging research area called semantic web usage mining, where semantic web is incorporated at different stages of web usage mining process. Recommender systems that are based on semantic web usage mining have capability to overcome all the drawbacks of existing recommender systems thereby providing accurate and meaningful recommendations.

Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between user and system. The context can be defined by a vector of features such as time and location.

In this paper a new system is proposed to generate comprehensive, diverse and accurate recommendations to user. The semantic information is incorporated at different stages of web usage mining to generate more meaningful recommendations to user. FAST algorithm is used to generate frequent sequential patterns. Context information along with sequential association rules mined from sequential patterns is used in generating recommendations.

IV. PROPOSED METHOD

The proposed system has four steps in recommending web pages to users.

A. Data preprocessing
B. Sequential Association Rule mining
C. Clustering webpages based on their semantic similarity.
D. Web page Recommendation

A. Data preprocessing

Data preprocessing is the first step in web page recommendation. Input for Pre-processing is raw web server logs [13]. Web log files are files that contain information about website visitor activity. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, etc.) from the site, information on his request is appended to a current log file.

As a first step in data pre-processing, data cleaning is carried out. In data cleaning invalid log entries are removed.[14] After data cleaning, the navigation history of each session from log files extracted. For extraction of navigation history user and session identification is carried out. In the next stage ontology mapping is done, where each page access in every transaction/session is mapped to the ontology instances defined in the ontology for the website. The Web server does not register semantic information about the request in the log file; instead it registers only the address of the request. Therefore, before starting the frequent sequence finding, mapping between ontology and the Web site address should be carried out. At the end of data pre-processing transactions consisting of ontology individuals are obtained.

B. Sequential pattern generation using FAST algorithm

After the preprocessing phase, the next step is the extraction of frequent navigation patterns. In this step sequential navigation patterns are created by using FAST algorithm. FAST algorithm quickly mines the complete set of patterns in a sequence database, reducing the effort for support counting and candidate sequence generation phases[15]. It employs new data representation of the dataset based on sparse id-lists and indexed vertical id-lists, which allows to quickly access an element and count its support without database scans. Researches have showed that FAST algorithm overcomes the limits of existing pattern mining methods like 1)the need of multiple scans of database 2)the generation of a potentially huge set of candidate sequences 3)the inefficiency of handling very long sequential patterns. The algorithm is mainly divided into two steps:

a. Item set extension
b. Sequence extension

FAST algorithm is applied on transaction containing ontology individuals to obtain semantically rich sequential patterns. In the next step association rules are mined from generated patterns. The main advantage of this algorithm is recommendation time and accuracy can be improved.

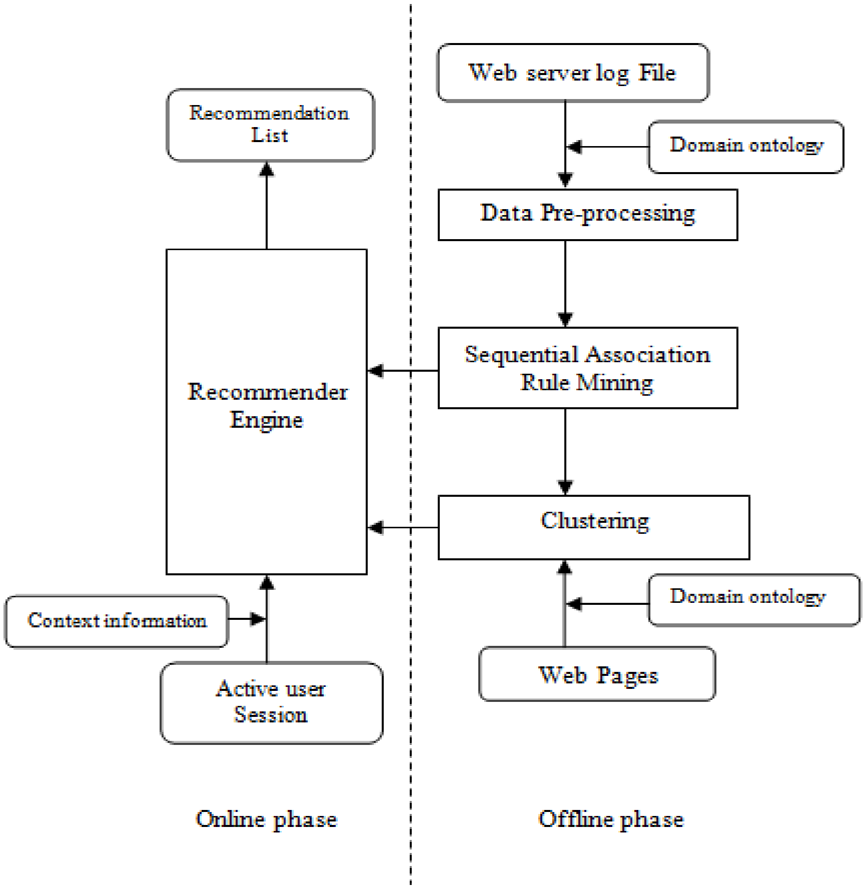

Fig. 1 General Architecture of proposed Hybrid Recommender System

C. Clustering webpages based on their semantic similarity.

In this step webpages are clustered based on semantic similarity. The distance between the two ontology concepts in calculated by concept match, which is based on upward cotopy – a semantic similarity measure [16]. By using these distance metric web pages are clustered using k- means clustering algorithm based on their semantic similarity.

D. Web page Recommendation

This is the final step in generating web page recommendations. This is an online phase in which active user session is matched with sequential association rules to generate web page recommendations to user by considering user context information as well. Each of the recommended webpage is checked to determine to which cluster they belong to. Finally the cluster with maximum number of webpages is added to the final recommendation set.

V. Conclusion

Web page recommender systems are the tools which recommend the web pages based on user interests. There are several recommender systems based on content based filtering, collaborative filtering and web usage mining. Researches have showed that incorporating semantic information at different stages of recommender system can generate meaningful and accurate recommendations. In this paper a new approach is proposed to recommend webpages, which used FAST algorithm and k-means clustering to generate recommendations. This hybrid recommender system improves the recommendation time and accuracy.

References

[1] Dietmar Jannach and Gerhard Friedrich, “Tutorial: Recommender Systems”, In *Proceedings of International Joint Conference on Artificial Intelligence Beijing*, pp. 1-144, 2013.

[2] http://en.wikipedia.org/wiki/Recommender_system.

[3] Robin van Meteren and Maarten van Someren, “Using Content-Based Filtering for Recommendation”, *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*, pp. 26 – 28, 2000.

[4] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, “Item-Based Collaborative Filtering Recommendation Algorithms”, *Proceedings of the 10th international conference on World Wide Web. ACM*, pp. 285– 295, 2001.

[5] Sule Gunduz and M. Tamer Ozsu, “A Web Page Prediction Model Based on Click Stream Tree Representation of User Behavior”, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535 – 540, 2003.

[6] Jia Li and Osmar R. Zaiane, “Combining Usage, Content, and Structure Data to improve Web Site Recommendation”, *5th International Conference, EC-Web, Zaragoza, Spain*, pp. 305 – 315, 2004.

[7] Kim, Byeong Man, Qing Li, Chang Seok Park, Si Gwan Kim, and Ju Yeon Kim., “A new approach for combining content-based and collaborative filters”, *Journal of Intelligent Information Systems*, Vol. 27, no. 1, pp. 79 – 91, 2006.

[8] Taowei Wang and Yibo Ren, "Research On Personalized Recommendation Based On Web Usage Mining Using Collaborative Filtering Technique", Wseas Transactions on Information Science and Applications, Vol. 6, no. 1, pp. 62 - 72, 2009.

[9] Reza Samizadeh and Babak Ghelichkhani, " Use of Semantic Similarity and Web Usage Mining to Alleviate the Drawbacks of User-Based Collaborative Filtering Recommender Systems", *International Journal of Industrial Engineering & Production Research*, Vol.21, no.3, pp. 137 – 136, 2010.

[10] C. Ramesh and Dr. K. V. Chalapati Rao and Dr. A. Goverdhan, "A Semantically Enriched Web Usage Based Recommendation Model", *International Journal of Computer Science & Information Technology*, Vol. 3, no. 5, pp. 193 – 202, 2011.

[11] Soheila Abrishami, Mahmoud Naghibzadeh, and Mehrdad Jalali, "Web Page Recommendation Based on Semantic Web Usage Mining", *Social Informatics; Springer Berlin Heidelberg*, Vol. 7710, no. 1, pp. 393 – 405, 2012.

[12] Mehdi Hosseinzadeh Aghdam, "Context-aware recommender systems using hierarchical hidden Markov model", Physica A (2018), https://doi.org/10.1016/j.physa.2018.11.037

[13] Vijayashri Losarwar and Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", *In Proceedings of International Conference on Artificial Intelligence and Embedded Systems*, pp. 1 - 5, 2012.

[14] Kim Shaily Langhnoja, Mehul Barot and Darshak Mehta, "Pre-Processing: Procedure on Web Log File for Web Usage Mining", *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 12, pp. 419 – 423, 2012.

[15] Alexander Maedche and Valentin Zacharias,"Clustering Ontology-based Metadata in the Semantic Web", *Springer, Heidelberg*, vol. 2431, no. 2, pp. 383–408, 2002.

[16] Eliana Salvemini, Fabio Fumarola, Donato Malerba, and Jiawei Han, "FAST Sequence Mining Based on Sparse Id-Lists".