ISSN: 2321-7782 (Online)
e-ISJN: A4372-3114
Impact Factor: 7.327

Volume 7, Issue 3, March 2019

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Collaborative web recommendation systems based on Data Mining

Priya Shah[1]
Assistant Professor
Department of Computer Application
Patel Group of Institutions, GTU
Mehsana, Gujarat, – India

Vikash Katariya[2]
Assistant Professor
Department of Computer Application
Patel Group of Institutions, GTU
Mehsana, Gujarat, – India

Abstract: ***The World Wide Web has evolved in less than two decades as the major source of data and information for all domains. Web has become today not only an accessible and searchable information source but also one of the most important communication channels, almost a virtual society. Web mining is a challenging activity that aims to discover new, relevant, and reliable information and knowledge by investigating the web structure, its content, and its usage. Though the web mining process is similar to data mining, the techniques, algorithms, and methodologies used to mine the web contains those specific to data mining, mainly because the web has a great amount of unstructured data and the changes are frequent and rapid. Personalization tools rely on click stream data captured in Web Server logs. The lack of user rating, sparse nature and large volume of data poses serious challenges to standard collaborative filtering techniques in terms of efficiency and performance. Web personalization can be effective if it is based on Association rule discovery from usage data.***

Key Words: web usage mining, Apriori algorithm, Collabrative filtering, Recommendation system.

I. INTRODUCTION

Collaborative filtering (CF) is a technique used by some recommender systems. Collaborative filtering is the process of filtering information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc. It is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users.

Collaborative filtering systems have many forms, but many common systems can be reduced to two steps:

1. Look for users who share the same rating patterns with the active user (the user whom the prediction is for).
2. Use the ratings from those like-minded users found in step 1 to calculate a prediction for the active user.

The collaborative filtering suffers a lot of limitations like-

- **Data sparsity**

Many commercial recommender systems are based on large datasets. As a result, the user-item matrix used for collaborative filtering could be extremely large and sparse, which brings about the challenges in the performances of the recommendation. One typical problem caused by the data sparsity is the cold start problem. As collaborative filtering methods recommend items based on users' past preferences, new users will need to rate sufficient number of items to enable the system to capture their preferences accurately and thus provides reliable recommendations.

Similarly, new items also have the same problem. When new items are added to system, they need to be rated by substantial number of users before they could be recommended to users who have similar tastes with the ones rated them. The new item problem does not limit the Content based recommendation, because the recommendation of an item is based on its discrete set of descriptive qualities rather than its ratings.

- **Scalability**

As the numbers of users and items grow, traditional CF algorithms will suffer serious scalability problems. For example, with tens of millions of customers $O(M)$and millions of items$O(N)$, a CF algorithm with the complexity of n is already too large. As well, many systems need to react immediately to online requirements and make recommendations for all users regardless of their purchases and ratings history, which demands a higher scalability of a CF system.

- **Synonyms**

Synonyms refer to the tendency of a number of the same or very similar items to have different names or entries. Most recommender systems are unable to discover this latent association and thus treat these products differently.

- **Grey Sheep**

Gray Sheep refers to the users whose opinions do not consistently agree or disagree with any group of people and thus do not benefit from collaborative filtering. Black Sheep are the opposite group whose idiosyncratic tastes make recommendations nearly impossible.

- **Shilling attacks**

In a recommendation system where everyone can give the ratings, people may give lots of positive ratings for their own items and negative ratings for their competitors. It is often necessary for the collaborative filtering systems to introduce precautions to discourage such kind of manipulations.

The challenge in designing of web personalization system is to improve the scalability of the recommendation system. One of the most successful and widely used technologies for building personalization and recommendation system is K –nearest neighbor approach. The K-nearest –Neighbor (kNN) approach compares those records of other users in order to find the top k users who have similar interest.

The work presents in this paper can improve the recommender system with the help of association rule mining from click stream data. First of all the frequent itemset is generated with the help of Apriori algorithm and it is stored in the frequent itemset graph. Then the recommendation engine matches the current user session window with itemsets to find candidate page views for giving recommendations.

II. WEB USAGE MINING

Web data mining is the process of extracting structured information from unstructured or semi-structured web data sources. Web Extraction also referred as Web Data Mining or Web Scrapping. It is done by creating programmed or script written in any programming language that processes the unstructured or semi-structured html web pages of a target web site to extract information or data for converting unstructured data into structured format. Web data mining scripts and applications will simulate a person viewing a web site with a browser. With help of web data Mining, We can connect to a website's web pages and request information or a pages, exactly as a browser would do. The web server will send back the html web page which we can then extract specific information from that web page.

Therefore, Web Data Mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web .Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site.

Web mining can be divided into three different types,

1. Web Content Mining
2. Web usage Mining and
3. Web structure mining

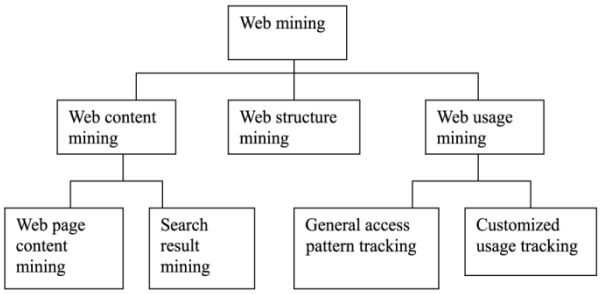

Fig 1. Different types of Web Mining

Web content mining examines the content of Web pages as well as results of Web searching. The content includes text as well as graphics data. It is further divided into Web page content mining and search results mining. It is traditional searching of Web pages via content, while Search results mining is a further search of pages found from a previous search.

The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining. Hyperlink hierarchy is also determined to path the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links.

- **Architecture Of Web Usage Mining**

Web usage mining is one of the prominent research areas due to these following reasons-

a) One can keep track of previously accessed pages of a user. These pages can be used to identify the typical behavior of the user and to make prediction about desired pages. Thus personalization for a user can be achieved through web usage mining.

b) Frequent access behavior for the users can be used to identify needed links to improve the overall performance of future accesses. Perfecting and caching policies can be made on the basis of frequently accessed pages to improve latency time.

c) Common access behaviors of the users can be used to improve the actual design of web pages and for making other modifications to a Web site.

d) Usage patterns can be used for business intelligence in order to improve sales and advertisement by providing product recommendations.

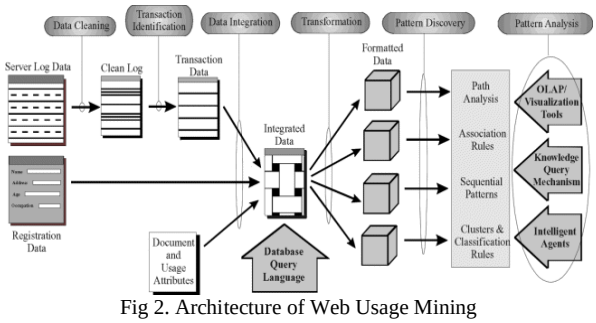

Fig 2. Architecture of Web Usage Mining

➢ **Different steps in Web Usage Mining**

There are five major steps followed in web usage mining are-

1. Data collection – Web log files, which keeps track of visits of all the visitors
2. Data Integration – Integrate multiple log files into a single file
3. Data preprocessing – Cleaning and structuring data to prepare for pattern extraction
4. Pattern extraction – Extracting interesting patterns
5. Pattern analysis and visualization – Analyze the extracted pattern
6. Pattern applications – Apply the pattern in real world problems

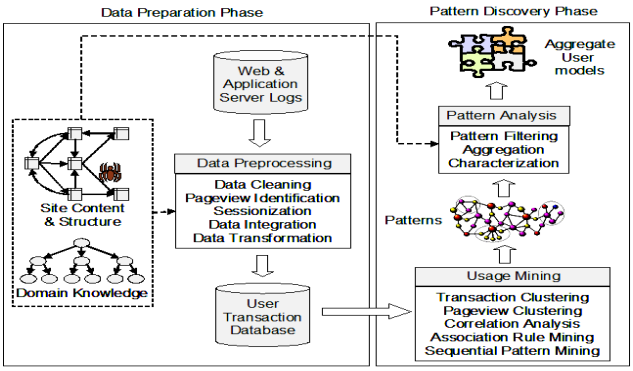

Fig 3 Different steps in Web Usage Mining

III. WEB RECOMMENDATION

Personalization is a process of gathering and storing information about visitors of a web site, analyzing the stored information, and, based on this analysis, delivering the right information to each visitor at the right time. A personalization component should be capable to recommend documents and/or other web sites, promote products, make appropriate advice, target e-mail, etc. A personalization component builds and exploits models or profiles of the users interacting with the system. A user profile is a (possibly structured) representation of characteristics of that user, in order to take into accounts his or her needs, goals, and interests.

Classification Of Recommendation System:-

1. Recommendation System Using Apriori Algorithm
2. Rule-Based Techniques
3. Item-Based Collaborative Filtering
4. Content-Boosted Collaborative Filtering
5. The Weighted Combination of Content-based and Collaborative Filtering

IV. RELATED WORK

According to the past research in the field of web personalization the most widely used technologies for building personalization and recommendation system is collaborative filtering.

One of the approaches used in collaborative filtering is the K-nearest neighbor approach. The K-nearest neighbor approach compare that record with the historical data of other users in order to find the top k users who have similar taste of interests. The mapping of a visitor record to its neighborhood could be based on similarity in ratings of items access to similar pages or

purchase of similar items. The identified neighborhood is then used to recommend items not already accessed or purchased by the active user.

The CF-based techniques has a lots of limitations that the it requires the neighborhood formation phase be performed as an online process and for very large data set this may not be acceptable. A number of optimization strategies has also used as the remedy of this shortcoming like similarity indexing and dimensionality reduction to reduce real time search costs.

In recent years there has been a lot of amount of work done in the area of web usage mining and finding the interesting patterns of user behavior. A lot of work has been proposed in the area of web personalization.

The web usage mining technique like clustering depends on the offline pattern discovery from user transactions and can be use to improve the scalability of collaborative filtering. Algorithms of some previous research work have also proposed the automatic personalization based on clustering of user transaction and page views but this approach reduce the recommendation accuracy.

Table 1: Sample Web transaction with pageviews A, B, C, D, and E

T1:	{A B D E}
T2:	{A B E C D }
T3:	{A B E C}
T4:	{B E B A C }
T5:	{D A B E C }

V. Data Preparation and Recommendation Generation

The proposed work is divided into three parts

1. Data Preparation ,
2. Pattern Discovery, and
3. Recommendation generation

- **Data Preparation**:

Data preprocessing is considered to be the very important step in data preparation. The data preprocessing has to be done carefully. The data has to preprocess. The first step in data preprocessing is data cleaning. In data cleaning step all the unnecessary fields of log files like the images, fields with extension of .html has been removed.

The next step in data preprocessing is the user identification then session identification and the pageview identification.

The above preprocessing tasks finally results into the set of n pageviews,

P={(p1,p2,p3…….pn)} and a set of m user transaction s, T= {t1,t2,t3……,tm}

Where each ti € T is a subset of P. Therefore each transaction t can be viewed as an l-length sequence of ordered pairs.

$T= ((p_1^1, w(p_1^1)), ((p_2^2, w(p_2^2), \ldots\ldots\ldots\ldots ((p_t^t, w(p_t^t))$

Where $p_1^1 = p_j$ for some j € {1,2,3,……n} and $w(pi^i)$ is the weight associated with the pageview.In this paper only binary weights on pageviews within the user transaction has considered.(Binary represents the existence or non-existence of the product-purchase or document access in the transaction.

After data preprocessing we have finally get a set of transactions with different pageviews as shown in below table 1.

➢ **Pattern Discovery:**

Using Apriori algorithm the frequent itemset is generated. The Apriori algorithm initially finds groups of items i.e. pageviews occurring frequently together in many transactions.

Given a transaction T and a set I={I_1,i_2…….I_k} of frequent itemset the support is define as

$A(I_k) = \{t \in T: I_k\ t\}/ T$

For example: - The frequent itemset for the above shown transactions is shown in the below table 2-

Size 1	Size2	Size 3	Size 4
{A}(5)	{A,B}(5)	{A,B,C}(4)	{A,B,C,E}(4)
{B}(6)	{A,C}(4)	{A,B,E}(5)	
{C}(4)	{A,E}(5)	{A,C,E}(4)	
{D}(5)	{B,C}(4)	{B,C,E}(4)	
	{B,E}(5)		
	{C,E}(4)		

Table 2 : Frequent Itemset generated by Apriori algorithm

The support of an itemset is defined as the minimum support of all items contained in the itemset.

➢ **The Recommendation Engine:**

Here we are using a fixed size sliding window over the current active session to capture the current user's history depth. For the above example it he current session with window size 3 is suppose <A,B,C> and now the user reference the pageview D then the new active session becomes <B,C,D> .

The recommendation engine matches the current user session window with the itemsets to find candidate pageview for giving recommendation. The recommendation value of each candidate pageview is based on the confidence of the corresponding association rule whose consequent is the singleton containing the pageview to be recommended. If the rule satisfied the specified confidence threshold requirement then the candidate pageview is added to the recommendation set. Given the frequent itemsets is stored in frequent item set graph of level k. Both the user active session window and the frequent itemset graph is stored in lexographical order. For a given active user session window *w* , a depth first search is performed to level *w,* If match is found then the children of the matching node *n* containing *w* are use to generate candidate set. For a given example ,if the user active session window <B,E> the recommendation generation algorithm finds items A and C as candidate recommendation .The recommendation scores of item A and C are 1 and 4/5 corresponding to the confidences of the rule {B,E}-{A} and{B,E}-{C}.

VI. CONCLUSION

This paper has presented the work that can increase the efficiency of web recommendation. Since web recommendation can help the web site developer to improve the web site and provide good and efficient material to its user. The algorithm can be enhanced by using more effective techniques for generation of frequent itemsets.

References

1. Mobasher, R. Colley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining", ACM, volume 48,
2. Agrawal R, Srikant R., "Fast Algorithms for Mining Association Rules", VLDB. Sep 12-15 1994, Chile, 487-99, pdf, ISBN 1-55860-153-8.
3. Mannila H, Toivonen H, Verkamo A I., "Efficient algorithms for discovering association rules." AAAI Workshop on Knowledge Discovery in Databases (SIGKDD). July 1994, Seattle, 181-92.
4. P. Becuzzi, M. Coppola, and M. Vanneschi, ì Mining of Association Rules in Very Large Databases: A Structured Parallel Approach," Proc. Europar-99, vol. 1685, pp. 1441-1450, Aug. 1999.

5. Junjie Chen and Wei Liu, "Research for Web Usage Mining Model", International Conference on Computational Intelligence for Modeling Control and Automation, and International Conference onIntelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06) 0-7695-2731-0/06 ©2006 IEEE
6. Olfa Nasraoui, Maha SolimanEsin Saka, Antonio Badia, Member and Richard Germain "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 2, February 2008
7. Sutheera Puntheeranurak, Hidekazu Tsuji, "Mining Web logs for a Personalized Recommender System", 0-7803-8932-S/05/ 2005 IEEE
8. B.Santhosh Kumar, K.V.Rukmani," Implementation of Web Usage Mining Using Apriori and FP Growth Algorithms",Int.J.of Advanced Networking and Applications, Volume:01, Issue:06, (2010)
9. S.Veeramalai, N.Jaisankar and A.Kannan," Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy", International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010
10. A.kumar, Dr. P. Thambidurai," Collaborative web recommendation systems based on an effective fuzzy association rule mining algorithm", Indian Journal of Computer Science and Engineering Vol1 No 3 184-191
11. C-H Lee, Y.H.-Kim, P.-K. Rhee," Web Personalization expert with combining Collaborative filtering and Association rule mining technique", Expert systems with Applications 21(2001)

