A Process Oriented Perception of Personalization Techniques in Web Mining

Gopal Pandey, Swati Patel, Vidhu Singhal, Akshay Kansara

Abstract*-Web personalization is an approach, a marketing tool and a fine art. With the rapid development of Deep Web, a large number of web information often lead to "information overload" and "information disorientated ", yet, personalized techniques can solve this problem. Personalized techniques are one such software tool used to help users obtain recommendations for unseen items based on their preferences. The commonly used personalized techniques are content based filtering, collaborative filtering and rule based filtering. In this paper, we present a survey on a personalized collaborative filtering method combining the association rule mining focusing on the problems that have been identifying and the solution that have been proposed.

Index Term* — *Association rule mining, collaborative filtering, personalization, web mining, web usage mining

I. INTRODUCTION

Web mining uses many data mining techniques; it is not purely an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of the Web data. Web mining process is similar to the data mining process. The difference is usually in the data collection. In traditional data mining, the data is often already collected and stored in a data warehouse. For Web mining, data collection can be a substantial task, especially for Web structure and content mining, which involves crawling a large number of target Web pages. Once the data is collected, we go through the same three-step process: data pre-processing, Web data mining and post-processing. However, the techniques used for each step can be quite different from those used in traditional data mining.

Web mining is the application of data mining techniques to extract knowledge from web data, [1] i.e. web content, web structure, and web usage data mining.

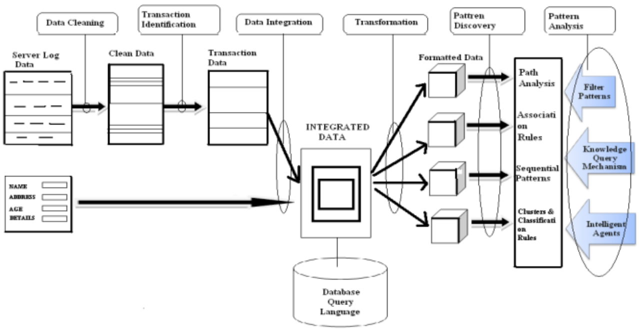

Fig 1: Web Mining Architecture

Table: 1Comparison of Web Mining Approaches

Mining Type	View Of Data	Source	Object	Collection
USAGE	Interactivity	Access	Behavior	Server/Browser Logs
CONTENT	Semi Structured, Unstructured	Pages	Index	Text/Hypertext Documents
STRUCTURE	Link Structure	Map	Map	Link Structure

II. WEB USAGE MINING

Web-usage mining (WUM), an emergent domain in web mining that has greatly concerned both academia and industry in recent years. One of many possible applications of Web Usage mining, which is the process of applying data mining techniques to the discovery of usage patterns from Web data, targeted towards various applications.[3] WUM is the process of discovering and interpreting patterns of user access to web information systems by mining the data collected from user interactions with the system. A typical WUM system consists of two tiers: 1) tracking, in which user interactions are captured and acquired, and 2) analysis, in which user access patterns are discovered and interpreted by applying typical data-mining techniques to the acquired data. There are three main tasks for performing WUM— pre-processing, pattern discovery and pattern analysis. As below:-

Pre-processing: It is generally used as groundwork of data mining practice, data pre-processing cleaned/filtered the raw data to eliminate outliers or irrelevant items, grouping individual page accesses into semantic units for the purpose of the user. The different types of pre-processing in WUM are— usage, content, and structure pre-processing.

Manuscript Received on January 2013.

Prof. Gopal Pandey, Department of Information Technology, Sir Bhavsinhji Polytechnic Institute, Bhavnagar, Gujarat, India

Prof. Swati Patel , Department of Computer Science, L.D.College of Engineering, Ahmedabad, Gujarat, India

Vidhu Singhal, Department of Information Technology, Shantilal Shah Engineering College, Bhavnagar, Gujarat, India

Akshay Kansara, Department of Information Technology, L.D.College of Engineering, Ahmedabad, Gujarat, India

Pattern Discovery: In this, WUM can be able to unearth patterns in server logs and carried out only on samples of data.

Interpretation and evaluation of results be done on samples of data. The various pattern discovery methods are— Statistical Analysis, Association Rules, Clustering, Classification, Sequential Patterns, and Dependency Modeling.

Pattern Analysis: The need behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. Most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

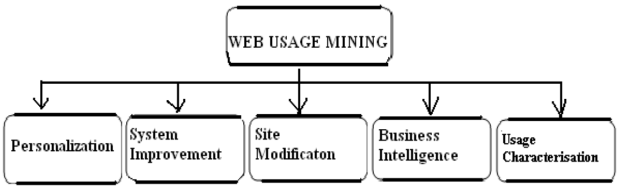

Fig 2: Domains of Web Usage Mining

III.PERSONALIZATION

Personalization is all about edifice customer loyalty by building a meaningful one-to-one relationship by understanding the needs of each individual and satisfying a objective that efficiently and knowledgeably addresses each individual’s need in a given perspective. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. [4] Personalization for its own sake has the potential to increase the complexity of site interface and drive inefficiency into the architecture. It is the capability to customize customer communication based on knowledge preferences and behaviours at the time of interaction. Web personalization process includes:

A. Personalization Techniques:

Personalization techniques can be divided in three parts:

Content-Based Filtering

In this, the user model includes information about the content of items of interest- whether these are web pages, movies, music, or anything else. Using these items as a basis, the technique identifies similar items that are returned as recommendations. One of the limitations when using content-based techniques is that it leads to over - specialization.

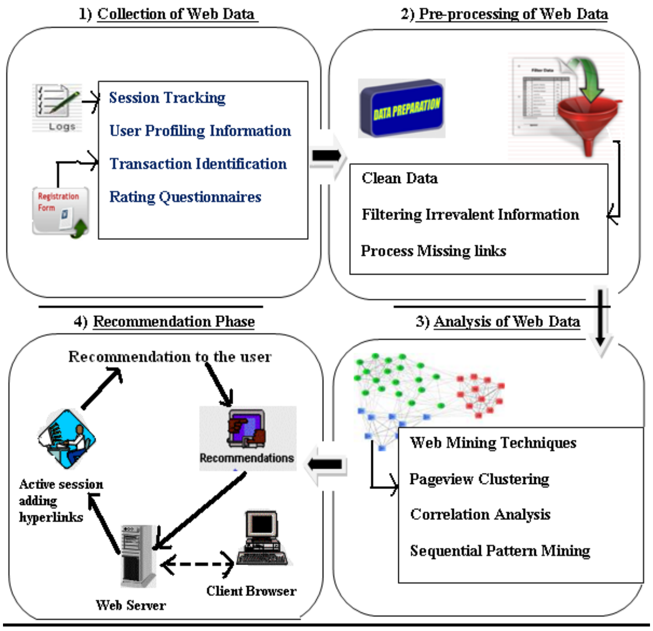

Fig 3: Web Personalization Process

Collaborative -Based filtering

In social or collaborative filtering, the system constructs rating profiles of its users, locates other users with similar rating profiles and returns items that the similar users rated highly. Scalability is a problem because computation grows linearly with the number of users and items. The advantage of social filtering, compared to content-based techniques, is that the pool from which recommendations originate is not restricted to items for which the active user has demonstrated interest. The pool will also include items that other users, users that are in some respect similar, have rated highly. This can prove to be instrumental in enhancing the user’s model: social filtering systems give the user the opportunity to explore new topics and items.

Rule-Based filtering

It selects only the appropriate service in formations by comparing the query result produced from the Search Manager (Search Module) which is the user interface of the search engine and the rule of the user fetched from the user profile registry

IV. COLLABORATIVE FILTERING TECHNIQUES

Collaborative filtering system is a challenging task in itself. It uses only the rating matrix across diverse domains. The elementary assumption of CF is that if different users rate n items similarly, or have analogous behaviours and so will rate or act on other items similarly. CF techniques use a database of preferences for items by users to predict additional topics or products a new user might like. [2]

CF algorithms requisite to have the skill of dealing with highly sparse data, to balance with the increasing numbers of users and items, to make acceptable commendation in a short time period, and to deal with other problems like synonymy, shilling attacks, data noise, and privacy protection problems.

Basically CF techniques divided into three parts:

- Memory-based collaborative filtering
- Method-based collaborative filtering
- Hybrid recommenders

A. *Memory-based collaborative filtering:*

Memory-based CF algorithm memorizes the rating matrix and uses the more or less of the user-item database to generate/issue recommendation. The most popular memory-based CF methods are neighbourhood-based methods, which foresee all the ratings by referring to users whose ratings are similar to the queried user, or to items that are similar to the queried item. The neighbourhood-based CF algorithm uses the following steps: [8]

# Compute the similarity or weight which reflects distance, correlation or weight between two users or two items.

# Produce a prediction for the active user by taking the weighted average of all the ratings of the user or item on a certain item or user, or using a simple weighted average.

# When the task is to generate a top-N recommendation, we need to find k most similar users or items (nearest neighbours) after computing the similarities, then aggregate the neighbours to get the top-N most frequent items as the recommendation.

B. *Model based collaborative filtering:*

The design and progress of models permit the system to recognize somewhat complex patterns based on the training data, and then issue recommendations for the collaborative filtering tasks for testing data or real-world data, based on the fitted models. A model-based CF algorithm includes Bayesian models, cluster-based CF and regression-based methods to solve the shortcomings of memory-based CF algorithms. The A recent class of successful CF models are based on low-rank matrix factorization. A Bayesian Belief net (BN) is a directed, acyclic graph (DAG) with a triplet N,A, Θ, where each node n Є N represents a random variable, each directed arc a Є A between nodes is a probabilistic association between variables, and Θ is a conditional probability table quantifying how much a node depends on its parents.

C. *Hybrid recommenders:*

Hybrid CF systems combine CF with other recommendation techniques (typically with content-based systems) to make predictions or recommendations. [8] The two chief modules of CF approaches, memory-based and model-based CF, are capable to be combined to form hybrid CF approaches. The commendable performance of the algorithms are generally enhanced than some pure memory-based CF algorithms and model-based CF algorithms.

V. ASSOCIATION RULE MINING

Association rule mining are one of the major techniques of data mining and it is the most common form of local-pattern discovery in unsupervised learning systems. It serves as a useful tool for finding correlations between items in large databases. The terms used in these rule are : [10,11]

Support: The support supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.
supp(X) = no. of transactions which contain the itemset X / total no. of transactions

Confidence: The measurement of certainty coupled with each and every discovered pattern. The confidence for an association rule X implies Y is the ratio of the number of transaction that contains X U Y to the number of transaction that contains X.
con f (X->Y) = supp (XUY) / supp(X)

Large Item Set: A large item set is an item set whose number of occurrences is above a threshold or support.

Association rule mining is the most used technique in Web Usage Mining generally applied to databases of transactions where each transaction consists of a set of items. When applied to Web Usage Mining, association rules are used to find associations among web pages that frequently appear together in users' sessions. When applied to Web Usage Mining, association rules are used to find associations among web pages that frequently appear together in users' sessions. The most common approach to finding association rules is to break up the problem into two parts:

1. Finding every occurred frequent itemsets.
2. Generating strong association rules commencing the frequent itemsets satisfying minimum support and minimum confidence.

Many algorithms have been proposed to solve the problem of detecting frequent itemsets in transaction database. Most of them can be classified into two categories, candidate generation and pattern growth.

Apriori represents the candidate generation approach. Apriori is a Breadth First Search Algorithm (BFS) which generates candidate k+1-itemsets based on frequent k-itemsets. The key idea of Apriori algorithm is to make multiple passes over the database.

The advantages of using apriori algorithm are

- ➢ Uses large item set property.
- ➢ Easily parallelized.
- ➢ Easy to implement.

FP-growth is a representative pattern growth approach. It is a Depth First Approach (DFS) and uses a special data structure, FP-Tree, for compact representation of the original database. FP-growth detects the frequent itemsets by recursively finding all frequent 1-itemsets in the conditional pattern base that is efficiently constructed based on the node link structure associated with FP-Tree. FP-growth doesn't explicitly generate candidates; its detection of the item supports is equivalent to generating 1-itemset candidates implicitly.

The advantages of FP-Growth algorithm are

- ➢ Uses compact data structure.
- ➢ Eliminates repeated database scan.

A. Apriori Algorithm

Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, “if an itemset is not frequent, any of its superset is never frequent”. By principle, Apriori assume that items contained by transaction or itemset are sorted in lexicographic strategy.

Let the set of frequent itemsets of size k be L_k and their candidates be N_k. Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for respected item and gathering that one which assure the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets:

1. Generate N_{k+1}, candidates of frequent itemsets of size k +1, from the frequent itemsets of size k.

2. Scan the database and calculate the support of each candidate of frequent itemsets.

3. Add those itemsets that satisfies the minimum support requirement to L_{k+1}.

The Apriori algorithm is shown below. Function apriori-gen in line 3 generates N_{k+1} from L_k in the following two step process:

1. *Join step*: Generate R_{k+1}, the initial candidates of frequent itemsets of size k + 1 by taking the union of the two frequent itemsets of size k, P_k and Q_k that have the first k−1 elements in common.

$$R_{k+1} = P_k \cup Q_k = \{iteml, itemk-1, itemk, itemk'\}$$
$$P_k = \{iteml, item2, \ldots, itemk-1, itemk\}$$
$$Q_k = \{iteml, item2, \ldots, itemk-1, itemk'\}$$

where, $iteml < item2 < \cdots < itemk < itemk'$.

2. *Prune step*: Check if all the itemsets of size k in R_{k+1} are frequent and generate N_{k+1} by removing those that do not pass this requirement from R_{k+1}. This is because if any subset of size k of N_{k+1} is not recurrent then it cannot be a subset of frequent itemset of size k + 1.

It is evident that Apriori scans the database at most k_{max+1} times when the maximum size of frequent itemsets is set at k_{max}.

Algorithm:

```
    $L_1$= (Frequent itemsets of cardinality 1);
     for(k=1;$L_k$ !=0;k++) do begin
        $N_{k+1}$ = apriori-gen($L_k$);//New candidates for all transactions t Є Database do begin
      N'$_k$ =subset($N_{k+1}$, t);//$C_{k+1}$ that are contained in t for all
                              candidate n Є N'$_t$ do
                          n.count++;
end
$L_{k+1}$ = candidates in $N_{k+1}$ with min_support
end
end
      return $U_k$ $L_{k;}$
```

Limitations of Apriori Algorithm:

Apriori algorithm, in spite of being simple, has some limitation. They are:

- ➢ It is costly to handle a huge number of candidate sets.[12]
- ➢ It is tiresome to scan the database repetitively and checking large set of candidates by pattern matching, which is chiefly true for mining long patterns.

Thus to prevail over the drawback inherited in Apriori, FP-growth is used which is an efficient FP-tree based mining method, It is faster than Apriori, and is also from the perspective of tree-projection which contains two phases, where the first phase constructs an FP tree, and the second phase recursively researches the FP tree and outputs all frequent patterns.

B. FP Growth Algorithm

FP-growth algorithm is an efficient method of mining all frequent itemsets without candidate‘s generation. FP-growth utilizes a combination of the vertical and horizontal database layout to store the database in main memory. Instead of storing the cover for every item in the database, it stores the actual transactions from the database in a tree structure and every item has a linked list going through all transactions that contain that item.

The algorithm mines the frequent itemsets by using a divide and conquer strategy as follows: In the first phase FP-growth constrict the database that represent frequent itemset into frequent-pattern tree i.e. FP-tree. In the next phase, it divide a constrict database into set of conditional databases, each and every linked with one frequent item. Finally, mine each such database separately. Particularly, the construction of FP-tree and the mining of FP-tree are the main steps in FP-growth algorithm.

A frequent pattern tree is a tree structure defined below.

1. It consists of one root labeled as “root”, a set of item prefix sub-trees as the children of the root, and a frequent-item header table.

2. Each node in the item prefix sub-tree consists of three fields: item-name, count, and node-link, where item-name registers which item this node represents, count registers the number of transactions represented by the portion of the path reaching this node, and node-link links to the next node in the FP-tree carrying the same item-name, or null if there is none.

3. Each entry in the frequent-item header table consists of two fields, (1) item-name and (2) head of node-link, which points to the first node in the FP-tree carrying the item-name.

Structure of FP-Tree:

•First, creating the root of the tree, labeled with “null”.

• Then scan the database D to create 1-itemset and scan again for n times till all itemsets get all together paired.

• The items in each transaction are processed in L order or we can say in sorted order.

• A branch is formed for every transaction with items having their support count separated by colon.

• Whenever the same node come across in another transaction, then just increment support count of the common node or Prefix.

• To assist tree traversal, an item header table is constructed so that each and every item points to its occurrences in the tree via a chain of node-links.

• Now, the dilemma of mining frequent patterns in database is altered to that of mining the FP-Tree.

VI. CONCLUSION

We have taken the view that Web personalization is an application of data mining and therefore must be supported during the various phases of a typical data mining cycle. We defined a framework for web personalization expert based on web mining and collaborative filtering techniques.

We adopted collaborative filtering technique combined with association rule mining technique, especially apriori algorithm respectively to associate the usage pattern of the clients in particular website.

The main drawback of Apriori algorithm is that the candidate set generation is costly, particularly when immense number of patterns or long patterns subsist. The main drawback of FP-growth algorithm is the explosive quantity of lacks a good candidate generation method. Future research can combine FP-Tree with Apriori candidate generation method to solve the disadvantages of both apriori and FP-growth. The new algorithm will reduce the storage space, improves the efficiency and accuracy of the algorithm.

ACKNOWLEDGMENT

Authors desire to show gratitude to our guide Prof.Gopal Pandey and Prof. Swati Patel for giving their valuable time and resources that helped us to implement research work. This research was carried out as part of the activities of, dissertation work in Post-Graduation.

REFERENCES

1. Jiawei Han, Micheline Kamber, “Data mining concepts and techniques”, Elsevier Inc., Second Edition, San Francisco, 2006
2. Charalampos Vassiliou, Dimitrios Stamoulis, Anastasios, “*Creating Adaptive Web Sites Using Personalization Techniques: A Unified, Integrated Approach and the Role of Evaluation”,* Greece, Idea Group Publishing, 2003, pp. 261-285,ch 12
3. Jaideep Srivastava, Robert Cooleyz, Mukund Deshpande, Pang-Ning Tan proposed “*Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*”, 2000.
4. Yogita S. Pagar, Vishakha. R. Mote, Rahul S. Bramhane, “*Web Personalization using Web Mining Techniques”,* Emerging Trends in Computer Science and Information Technol2012 (ETCSIT2012)
5. Liana Razmerita, Thierry Nabeth, Kathrin Kirchner, ”*User Modeling and Attention Support: Towards a Framework of Personalization Techniques”,* The Fifth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, 2012
6. Elnaz Davoodi, Keivan Kianmehr, Mohsen Afsharchi, “*A semantic social network-based expert recommender system”,* Springer Science Business Media, LLC 2012
7. Ms.Kavita D.Satokar, Mr.S.Z.Gawali, “*Web Personalization Using Web Mining”,* International Journal of Engineering Science and Technology Vol. 2(3), 2010, 307-311.
8. Xiaoyuan Su and Taghi M. Khoshgoftaar, “*A Survey of Collaborative Filtering Techniques”,* Hindawi Publishing Corporation Advances in Artificial Intelligence Volume 2009, Article ID 421425, 19 pages
9. Hongwu Ye, “*A Personalized Collaborative Filtering Recommendation Using Association Rule Mining and Self-Organizing Map”,* JOURNAL OF SOFTWARE, VOL. 6, NO. 4, APRIL 2011
10. Rahul Mishra, Abha Choubey, “*Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data”,* International Journal of Computer Science and Information Technologies, Vol. 3 (4) , 2012,4662 – 4665
11. Sanjeev Rao, Priyanka Gupta, “*Implementing Improved Algorithm over APRIORI Data Mining Association Rule Algorithm”,* IJCST Vol. 3, Issue 1, Jan. - March 2012
12. B.Santhosh Kumar, K.V.Rukmani, “*Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms”,* Int. J. of Advanced Networking and Applications 400 Volume:01, Issue:06, Pages: 400-404 (2010)
13. [Online]Available: http://www.wikipedia.com/datamining
14. [Online]Available:http://www.en.wikipedia.org/wiki/Association_rule_learning

AUTHORS PROFILE

Prof. Gopal Pandey BE in Information Technology in 2002 from SVIT Vasad, ME, IT from GTU 2012. Area of interest Software Engineering, Semantic Web, Java Programming. Indian Society for Technical Education (ISTE) Life membership

Prof. Swati Patel M.E. in Computer Science and Engineering, A survey on Digital Video Watermarking IJCTA- Online Journal ISSN:2229-6093 Pg no: 3015 to 3018, A detailed study of digital image watermarking techniques research work IAIR - ICITEC-2011 HYDERABAD, INDIA ISBN 978-81-921178-1-2 Pg no: 81 to 84, Comparison between DCT and DWT algorithm of Digital Watermarking ATAST 2011, SURAT.

Ms. Vidhu Singhal (ME Scholar), M.E. in Information Technology . Area of interest is Web Mining.

Mr. Akshay Kansara,(ME Scholar**),** M.E. in Information Technology, working on various personalization techniques and also improving the algorithm based on the techniques