The Recommendation System for User Interest on Web

Priyanka D. Khawale[1], V. S. Nandedkar[2]

[1]Pune university, PVPIT Pune/Pune, India

[2]Professor, PVPIT Pune/Pune, India

Abstract: ***A web personalization system provides most relevant pages to user according to the user interest domain. The system obtains the knowledge user interest domain by analyzing the user browsing history. Web services are interested in learning user interest, so they can better target the product according to user interest. The browsing history is obtained by accessing the web log data at server and cookies. The web personalization system is also known as recommender system. The motivation for this hybrid approach comes from the observation that the existing system provides the relevant data, but the system should be effective and fast. To achieve this, it combines the usage mining along with content mining with content caching. This approach improves the system performance with the help of content caching and content filtering. Thus user can obtain the relevant information on web as per user interest.***

Keywords: Web service recommendation, user interest, usage history

1. Introduction

To personalize means to make or change something according to the individual need. Personalization is the ability to provide the relevant information based on the user interest. The main goal of personalization is to help users find the information they are interested in. Most of the personalization systems try to filter available content found potentially interesting for that particular user. Extraction process information from the log files of web site is used to identify the usage patterns and profile of user.

Web personalization system incorporates the web mining concepts: Web content mining, web structure mining and web usage mining. The motivation for this hybrid approach comes from the observation that personalized content on the web is relevant. Nowadays, when information overload is one of the common problems of web uses, it is difficult for the users to find the relevant information. This issue is solved with personalization. Personalization is used to provide the relevant information on the web. It provides information based on the user browsing history.

Knowledge obtained by studying the preferences of web users can be used to improve the effectiveness of the website. More web services are interested in learning user interest, so they can better target the product according to user interest. This paper is aimed to identify the patterns in sequences from web log data and cookies in specific period using Apriori-all algorithm. Patterns will be analyzed for information from data. Then by applying the content search by crawler it makes the cluster according to the content. Each cluster is get assigned with the cluster grade and each page in that cluster is get assigned with page rank based on the session time of the user for specific page. The following section discusses Methodology which can be used for the system development.

2. System Architecture

The web personalization system architecture consists of the 5 sub-systems which are the parts of system, represented as modules. These sub-systems are: Sequential analysis, clustering, content filtering, Content Caching and page rank. An overview of the architecture of the proposed system is given in Fig. 1. First, all users' web access activities of a website are recorded by the WWW server of the website and stored into the Web Server Logs. Each user access record contains the client IP address, request time, requested URL, user ID, HTTP status code, etc.

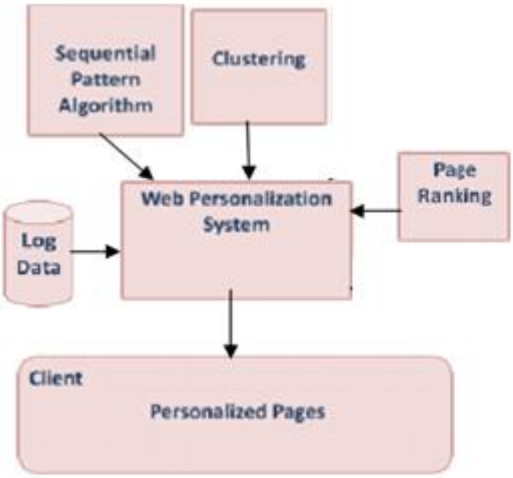

Figure 1: System Architecture of web personalization system

When a user visits the website, the user's HTTP requests in the current browsing session are recorded in order, and the current access sequence is constructed. Each user accessing the website can be identified using his/her IP address [1, 3]. The system accesses the data from the log data then it applies the sequential pattern analysis for finding the succeeded are the pages which are accesses sequentially. Clustering applies for making the classification of the web pages according to the content and page rank. The finally page rank and is used to number or weight page according to number of visitors and the content is stored to the cache memory.

3. Methodology

This paper incorporates the hybrid approach for web personalization. It takes input as the web log data and it analyses the data using sequential pattern analysis using Apriori –all algorithm. Sequential pattern analysis is: given a database of sequences where each sequence is an ordered list of user access based on access time and each access consist of a collection of useful information, and then searched the entire access pattern with minimum support by the user, where supports a number of database sequence contain the pattern. After sorting the pages it applies the content search by crawler. Then makes the cluster according to the contents. Each cluster is get assigned with the cluster grade and each page in that cluster is get assigned with page rank based on the session time of the user for specific page. The pages after the sequential analysis, it crawls for the content with the help of crawler or by Wrapper generation. Web crawler takes data form user, search it and get well selected pages using breadth first search algorithm. In Wrapper generation, web page data can be extracted using HTML wrapper. Here the data is DOM tree which is constructed by web browser [1, 2, 6].

The training data is get find out from the extraction of content by crawling. Based on this content or pattern, a system does cluster the data by applying the association rule mining and by construction of pattern tree and by using the K-means algorithm (EM). This cluster is based on content that represents the interest of the user. Then each cluster is assigned with cluster grade for getting distinguished from other cluster. So each cluster represents the different unique interest domains. The pages in each cluster are assigned with numerical weight based on the number time for which the page is accessed. Then at final, pages with rank are get displayed to user as the personalization result based on the user interest based on user browsing history. Finally we are caching the content for improvement of the personalization for future accesses.

The following section discusses a detailed about the techniques.

Data Extraction and User Identification:
Data on web is classified as structured data, semi structured data and Unstructured data.

Structured Data Extraction:
Some structured data are list tree and data in table. It can be extracted by using the wrapper generation.

Unstructured Data Extraction:
It is in the form of text document. It is related to text mining.

Semi structured Data Extraction:
It is hierarchical structure. It can be extracted by using NLP, TINTIN.
There are various techniques to extract the data using web content mining. These are: Web Crawler and Wrapper generation. For extracting the data Web crawler takes data form user, search it and get well selected pages. It uses breadth first search algorithm. In Wrapper generation, web page data can be extracted using HTML wrapper. Here the data is DOM tree which is constructed by web browser. Using the web log and cookies we can extract the data about the user. It identifies the user by accessing the cookies and by accessing the web log [2].

Sequential Pattern Analysis

It takes input as the web log data and it analyses the data using sequential pattern analysis using Apriori–all algorithm. Sequential pattern analysis is: given a database of sequences where each sequence is an ordered list of user access based on access time and each access consist of a collection of useful information, and then searched the entire access pattern with minimum support by the user, where supports a number of database sequence contain the pattern. Apriori-all algorithm as follows:

1) Sort the pattern on user id and time of reference on page
2) Calculate Support.
3) Find maximum reference sequence
4) Applying a priori algorithm.

Web usage mining has three activities: Preprocessing, Discovery of pattern, Analyzing patterns. The algorithm as follows.

```
INPUT:
D = (s1, s2, .., sk) / / Database of the session (session)
s / / Support

OUTPUT: Sequential Pattern
Sequential Pattern Algorithm:
D = D sorting on User ID and time of reference on the first page in each session.
Find L1in D;
L = Apriori All (D, s, L1)
Find a maximal reference sequence of the L;
```

After sorting the pages it applies the content search by crawler and by generating wrapper generation [1, 3].

4. Web Content Mining (Extract Patterns)

Web content mining confronts this problem gathering explicit information from different web sites for its access and knowledge discovery. Basically, web mining is concerned with the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services.

To extract information from deep web that is a large collection of dynamic query able databases, we need a system that can extract automatically. For this purpose we use web content mining techniques that uses XML version of HTML query interfaces. Web content mining is a form of text mining and can take advantage of the semi structured nature of web page text.

In Wrapper generation, web page data can be extracted using HTML wrapper. Here the data is DOM tree which is constructed by web browser. DOM is a standard language that gets a web page as an input and shows it in a structured tree from interfaces, objects and relations between them as an output.

Steps for wrapper generation:

Step 1. Creates filter from a visual interaction with a human wrapper designer.
Step 2.Then user will give his response which is equivalent to marking of nodes in DOM tree.
Step 3.Find out the filter that identifies the entire designer pattern.
Step 4. Select an input instance and mark out missing instance.
Step 5. Matches all intended instance of current input, user will decide to continue with input instance or HTML document.

This wrapper generation algorithm uses clustering and attributes classification. Cluster is similar to their tree structure. It will build the list of feature used for classification of filter. Now the list constructed from their attributes and values, construct the training database, for every customer build a decision tree based on attribute classifier. Then it will build a tree based attribute classifier. Each cluster is divided into blocks. Each cluster defines its extraction rule which is core Xpath expression and an attribute classifier. Instance of input DOM is found by Xpath expression, which matches particular tree shape of the cluster. Attribute classifier will sort out the instance [6].

5. Clustering Pages According To Extracted Patterns

The training data is get find out from the extraction of content by crawling. Based on this content or pattern, a system does cluster the data by construction of pattern tree and by using the K-means algorithm. In statistics and data mining, *k*-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. This results into a partitioning of the data space into Voronoi cells.

((f1 log(n/df1), (tf2, log(n/df2), ... , (tfn ,log(nldfn)).

Where tfi is the frequency of the ith term in the document and dfi is the number of documents that contain the ith term. To account for documents of different lengths, the length of each document vector is normalized so that it is of unit length (‖ dtfidf = 1‖). The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the *k*-means algorithm; it is also referred to as Lloyd's algorithm particularly in the computer science community.

Given an initial set of *k* means **m**1(1),…,**m***k*(1) (see below), the algorithm proceeds by alternating between two steps[3,4].

- Assignment step: Assign each observation to the cluster with the closest mean (i.e. partition the observations according to the Vornoi diagram generated by the means).
- Update step: Calculate the new means to be the centroid of the observations in the cluster.

This cluster is based on content that represents the interest of the user. In this system we are proposing the variation to the k-means as follows:

Step 1. Extracting the content using crawler
Step 2.Consider any content
Step 3. Finds the related web pages
Step 4. Associae the pages.
Step 5. Form the cluster.

To assign weight to the content it first consider a pattern then it looks for the content in the other pages in cluster, according to the number of content patterns in the web page it assigns the weight to the content. The weight will be more when numbers of patterns are more in the page. The steps will be:

Step1: Take a content pattern
Step2: Crawls the other web pages in the same cluster for the same content pattern.
Step3: Find the number of same content pattern in page. Step4: Assign the weight to the page in the cluster. After the cluster is formed, each cluster is assigned with the unique number [1, 3].

Ranking the Page in Cluster

The pages in each cluster are assigned with numerical weight based on the number time for which the page is accessed. This system will refer the weighted page rank algorithm. The weighted Page Rank algorithm (WPR), an extension to the standard Page Rank algorithm, is introduced. WPR takes into account the importance of both the in links and the out links of the pages and distributes rank scores based on the popularity of the pages.

Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some ranking algorithms depend only on the link structure of the documents i.e. their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining), while some use a combination of both i.e. they use content of the document as well as the link structure to assign a rank value for a given document [2, 3, 4].

Finally system displays the pages which are clustered and scored according to user interest and system caching the content for improvement of performance of web personalization systems.

Content Filtering

Web personalization techniques are classified in five classes:

content based filtering, traditional collaborative filtering, model based techniques, hybrid techniques and semantic techniques. Content based filtering uses an individual approach which relies on user's ratings and item descriptions. Items having similar properties as items positively rated by user are being recommended to the user. The most common problem of content based filtering is the new user problem. This problem occurs when a new user is added to the system, hence has an empty profile (without ratings) and cannot receive recommendations. Finally system displays the pages which are clustered and scored according to user interest and system caching the content for improvement of performance of web personalization systems [1, 2, 3].

6. Conclusion

Using hybrid approach for personalization, we can conclude that the pages produced after the clustering, are displayed to user. User will get more relevant information and pages according to the user interest domain. Using sequential pattern mining web logs can explore the patterns which explores the habit of user that is ordered patterns. After identification of the patterns in sequences from web log data and cookies in specific period using Apriori-all algorithm. Patterns will be analyzed for information from data. Then by applying the content search by crawler it makes the cluster according to the content. Each cluster is get assigned with the cluster grade and each page in that cluster is get assigned with page rank based on the session time of the user for specific page.

7. Acknowledgement

I have taken efforts in this paper. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to place on record my deep sense of gratitude to Prof. Vijay Patil, Dept. of Computer Science and Engineering, MIT, Aurangabad, India for her generous guidance, help, useful suggestions, continuous encouragement and supervision throughout the course of present work. I would like to extend my sincere thanks to Prof. S.A. Kinariwala, Prof. K.V.Bhosle, Prof. S. P. Abhang, Dr. C.G. Desai and other colleagues for their support and guidance. I would like to express my gratitude towards my parents & member of Marathwada Institute of Technology, Aurangabad for their kind co- operation and encouragement which help me in completion of this paper.

References

[1] Minxiao Lei, Lisa Fan Department of Computer Science, University of Regina, Regina, Saskatchewan, "A Web Personalization System Based on Users' Interested Domains", Proc. 7th IEEE Int. Conf. on Cognitive Informatics (ICCI'08) Y. Wang, D. Zhang, J.-C. Latombe, and W. Kinsner (Eds.) 978-1-4244-2538-9/08 ©2008 IEEE

[2] Ford Lumban Gaol Faculty of Computer Science Bina Nusantara University Indonesia , " Exploring the patterns of Habits of users using web log sequential pattern",© 2010 IEEE DOI10.1109/ACT.2010.37

[3] Dario Vuljani_, Lidia Rovan, Mirta Baranovi_ Faculty of Electrical Engineering and Computing, University of Zagreb Unska 3, HR-10000 Zagreb, Croatia, "Semantically Enhanced web personalization Approaches and techniques", ITI 2010 32nd Int. Conf. on Information Technology Interfaces, June 21-24, 2010, Cavtat, Croatia.

[4] Matthew Fredrikson University of Wisconsin, "Re-Imagining content personalization and in-browser Privacy", 1081-6011/11 © 2011 IEEE DOI 10.1109/SP.2011.37

[5] Kshitija Pol Datta Meghe College of Engineering, Nita Patil Datta Meghe College of Engineering,, "A Survey on Web Content Mining and extraction of Structured and Semi structured data", 978-0-7695-3267-7/08$25.00©2008 IEEE