WEB PERSONALIZATION

Mrs. M. Kiruthika¹, Nitin Cyriac², Aditya Mandhare³, Soniya Nemade⁴

DEPARTMENT OF COMPUTER ENGINEERING
Fr. CONCEICAO RODRIGUES INSTITUTE OF TECHNOLOGY, VASHI
Email- ¹venkatr20032002@gmail.com, ²nitincyriac@yahoo.in, ³adityamandhare@gmail.com, ⁴sonianemade@gmail.com

Abstract:

Web personalization is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user’s navigational behaviour in correlation with other information collected in the Web context, namely structure, content and user profile data. Due to the explosive growth of the Web, the domain of Web personalization has gained great momentum both in the research and the commercial area. In this paper we would like to discuss different categories of web personalization, its phases and the techniques used in it. It also highlights the applications and tools of web personalization.

1. Introduction

Interest in the analysis of user behaviour on the Web has been increasing rapidly. This increase stems from the realization that added value for Web site visitors is not gained merely through larger quantities of data on a site, but through easier access to the required information at the right time and in the most suitable form. To improve the Internet service quality and increase the user click rate on a specific website, thus, it is necessary for a Web developer or designer to know what the user really wants to do, predict which pages the user is potentially interested in, and present the customized web pages to the user by learning user navigational pattern knowledge.

Web personalization can be seen as an interdisciplinary field that includes several research domains from user modelling, social networks, web data mining, human-machine interactions to Web usage mining. Personalization process is about offering ‘tailor-made’ environment to a visitor with dynamic content, which varies as per the needs, selection and user behaviour. The process of website personalization starts with the collection of user data, which is then integrated with user needs, demands and the result is a specifically designed website. Some basic elements of website personalization include content, imagery and even the theme of the website. Benefits offered by Website Personalization include a customer-focused and a much comfortable website, Information catered and presented in a much better manner resulting in saving of time, productive results, and higher call-to-action.

The tremendous growth in the number and the complexity of information resources and services on the Web has made Web personalization an indispensable tool for both Web-based organizations and for the end users. The ability of a site to engage visitors at a deeper level, and to successfully guide them to useful and pertinent information, is now viewed as one of the key factors in the site’s ultimate success.

2. WEB DATA

Web data are those that can be collected and used in the context of web personalization.

2.1 Categories of web data

2.1.1 Content data:
Presented to the end-user appropriately structured. It can be simple text, images, or structured data, such as information retrieved from databases.

2.1.2 Structure data:
Represent the way content is organized. They can be either data entities used within a Web page, such as HTML or XML tags, or data entities used to put a Web site together, such as hyperlinks connecting one page to another.
2.1.3 Usage data:
Represent a Web site’s usage, such as a visitor’s IP address, time and date of access, complete path (files or directories) accessed, referrers’ address, and other attributes that can be included in a Web access log.

2.1.4 User profile data:
Provide information about the users of a Web site. A user profile contains demographic information for each user of a Web site, as well as information about users’ interests and preferences. Such information is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs.

3. WEB MINING

Web mining is the application of data mining techniques to extract knowledge from web data.

3.1 Types of web mining:

3.1.1 Web content mining:
It is the process of discovering useful information from the content of the web page. The web content may include, text, image, audio, video. Technologies used in web content mining are, Natural Language Processing (NLP) and Information Retrieval (IR).

3.1.2 Web structure mining:
It is a process of discovering structure information from the web. This type of mining is done either at the document (intra-page) or at the hyperlink level (inter-page). The research at hyperlink level is also called as hyperlink analysis.

3.1.3 Web usage mining:
It is a type of web mining activity that involves the automatic discovery of user access patterns from one or more web servers. It uses data mining to analyze and discover interesting patterns of user’s usage data on the web.

The steps in web usage mining are:

3.1.3.1 Data collection:
Data collection is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service’s quality. Therefore it must use scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data (agent server data and package detecting).

3.1.3.2 Data preprocessing:
Some databases are insufficient, inconsistent and including noise. The data pretreatment is to carry on a unification transformation to those databases. The data pretreatment work, mainly includes data cleaning, user identification, session identification and path completion.

a) Data Cleaning:
The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning.

b) User and Session Identification:
The task of user and session identification is find out the different user sessions from the original web access log. User’s identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access.

c) Path completion:
There are some reasons that result in path’s incompleteness, for instance, local cache, agent cache, “post” technique and browser’s “back” button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource locators (URL) recorded in log may be less than the real one. As a result, the user access paths are incompletely preserved in the web access log. To discover user’s travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task.

3.1.3.3 Knowledge discovery:
Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community’s interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

4.1.4 Pattern analysis:
Challenges of Pattern Analysis is to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered...
data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

4. PERSONALIZATION STRATEGIES

The web personalization strategies are as follows in the increasing order of importance and complexity.

4.1 Memorization:
In this simplest and most widespread form of personalization, user information such as name and browsing history is stored (e.g. using cookies), to be later used to recognize and greet the returning user. It is usually implemented on the Web server. This mode depends more on Web technology than on any kind of adaptive or intelligent learning. It can also pose a threat to user privacy.

4.2 Customization:
This form of personalization takes as input a user’s preferences from registration forms in order to customize the content and structure of a web page. This process tends to be static and manual or at best semi-automatic. It is usually implemented on the Web server. Typical examples include personalized web portals such as My Yahoo and Google.

4.3 Guidance or recommendation system:
A guidance based system tries to automatically recommend hyperlinks that are deemed to be relevant to the user’s interests, in order to facilitate access to the needed information on a large website. It is usually implemented on the Web server, and relies on data that reflects the user’s interest implicitly or explicitly. This approach forms the focus of Web personalization.

4.4 Task performance support:
In these clientside personalization systems, a personal assistant executes actions on behalf of the user, in order to facilitate access to relevant information. This approach requires heavy involvement on the part of the user, including access, installation, and maintenance of the personal assistant software. It also has very limited scope in the sense that it cannot use information about other users with similar interests.

5. PHASES OF WEB PERSONALIZATION

5.1 Collection of web data:
Data can be collected either in an implicit or an explicit manner.

5.1.1 Implicit data:
It includes past activities/clickstreams as recorded in Web server logs and/or via cookies or session tracking modules. This helps to study user’s behaviour at the website.

5.1.2 Explicit data:
It comes from registration forms which the user fills while signing up with the website. The user rating questionnaires. Additional data such as demographic (i.e. study of the characteristics of visitors) and application data (example: e-commerce transactions) can also be used. In some cases, Web content, structure, and application data can be added as additional sources of data, to shed more light on the next stages.

5.2 Pre-processing of web data:

Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step. Pre-processing may include the following steps:

1. Cleaning data of inconsistencies
2. Filtering out irrelevant information,
3. Completing the missing links in incomplete click through paths.

Most importantly, unique sessions need to be identified from the different requests, based on a heuristic, such as requests originating from an identical IP address within a given time period.

5.3 Analysis of web data:

Also known as Web Usage Mining, this step applies machine learning or Data Mining techniques to discover interesting usage patterns and statistical correlations between web pages and user groups. This step frequently results in automatic user profiling, and is typically applied offline, so that it does not add a burden on the web server.

5.4 Decision making and recommendation stage:

The last phase in personalization makes use of the results of the previous analysis step to deliver recommendations to the user. The recommendation process typically involves generating dynamic Web content on the fly, such as adding hyperlinks to the last web page requested by the user. This can be accomplished using a variety of Web technology options such as CGI programming. “CGI” stands for “Common Gateway Interface.” CGI is one method by which a web server can obtain data from (or send data to) databases, documents, and other
Information about the visitor’s identification is stored, along with password information. Additional information such as credit card details, if one is used during a transaction, as well as details concerning the visitor’s activities at the Web site, for example which pages where

6. COMMONLY USED IMPORTANT TECHNIQUES

6.1 USER PROFILING

In order to personalize a Web site, the system should be able to distinguish between different users or groups of users. This process is called user profiling and its objective is the creation of an information base that contains the preferences, characteristics and activities of the users. A user profile can be either static, when the information it contains is never or rarely altered (e.g. demographic information), or dynamic when the user profile’s data change frequently. Such information is obtained either explicitly, using online registration forms and questionnaires resulting in static user profiles, or implicitly, by recording the navigational behaviour and/or the preferences of each user, resulting in dynamic user profiles.

6.1.1 Data Collection

A way of uniquely identifying a visitor through a session is by using cookies. W3C [WCA] defines cookie as “the data sent by a Web server to a Web client, stored locally by the client and sent back to the server on subsequent requests”. In other words, a cookie is simply an HTTP header that consists of a text-only string, which is inserted into the memory of a browser. It is used to uniquely identify a user during Web interactions within a site and contains data parameters that allow the remote HTML server to keep a record of the user identity, and what actions she/he takes at the remote Web site. The contents of a cookie file depend on the Web site that is being visited. In general, visited, which purchases where made or which advertisements were selected, can also be included. Often, cookies point back to more detailed customer information stored at the Web server.

A user can be identified making the assumption that each IP corresponds to one user. In some cases, IP addresses are resolved into domain names that are registered to a person or a company, thus more specific information is gathered.

6.1.2 Issues

6.1.2.1 Overdependence on browser cookies

First of all, in case a system depends on cookies for gathering user information, there exists the possibility of the user having turned off cookie support on their browser. Other problems that may occur when using cookies’ technology are the fact that since a cookie file is stored locally in the user’s computer, the user might delete it and when she/he revisits a Web site will be regarded as a new visitor. Furthermore, if no additional information is provided (for example some logon id), there occurs an identification problem if more than one user browses the Web using the same computer.

6.2 LOG ANALYSIS

By applying statistical and data mining methods to the Web log data, interesting patterns concerning the users’ navigational behaviour can be identified, such as user and page clusters, as well as possible correlations between Web pages and user groups.

6.2.1 Web Log

Each access to a Web page is recorded in the access log of the Web server that hosts it. The entries of a Web log file consist of fields that follow a predefined format. The fields of the common log format are:

```
remotehost rfc931 authuser date "request" status bytes
```

where remotehost is the remote hostname or IP number if DNS hostname is not available, rfc931 is the remote log name of the user, authuser the username as which the user has authenticated himself, available when using password protected WWW pages, date the date and time of the request, “request” the request line exactly as it came from the client (the file, the name and the method used to retrieve it), status the HTTP status code returned to the client, indicating whether or not the file was successfully retrieved and if not, what error message was returned, and bytes the content-length of the document transferred. If any of the fields cannot be determined a minus sign (-) is placed in this field.

6.2.2 Data Pre-processing

There are some important technical issues that must be taken into consideration during this phase in the context of the Web personalization process, since it is necessary for Web log data to be prepared and pre-processed in order to use it in the consequent phases of the process. The first issue in the pre-processing phase is data preparation.
Depending on the application, Web log data may need to be cleaned from entries involving pages that returned an error or graphics’ file accesses. In some cases such information might be useful, but in others such data should be eliminated from a log file.

Most important of all is the user identification issue. There are several ways to identify individual visitors. The most obvious solution is to assume that each IP address (or each IP address/client agent pair) identifies a single visitor. Nonetheless, this is not very accurate since for example, a visitor may access the Web from different computers, or many users may use the same IP address (if a proxy is used). A further assumption can then be made, that consecutive accesses from the same host during a certain time interval come from the same user.

More accurate approaches for a priori identification of unique visitors are the use of cookies or similar mechanisms or the requirement for user registration.

6.2.3 Log analysis

Log analysis tools (also called traffic analysis tools), take as input, raw Web data and process them in order to extract statistical information. Such information includes statistics for the site activity (such as total number of visits, average number of hits, successful / failed /redirected/cached hits, average view time, average length of a path through a site), diagnostic statistics (such as server errors, page not found errors), server statistics (such as top pages visited, entry / exit pages, single access pages), referrers statistics (such as top referring sites, search engines, keywords), user demographics, client statistics etc. Some tools also perform click-stream analysis, which refers to identifying paths through the site followed by individual visitors by grouping together consecutive hits from the same IP, or include limited low-level error analysis, such as detect in unauthorized entry points or finding the most common invalid URL.

These statistics are usually outputted to reports and can also be displayed as diagrams. This information is used by administrators for improving the system performance, facilitating the site modification task and providing support for marketing decisions. However, most advanced Web mining systems further process this information to extract more complex observations that convey knowledge, utilizing data mining techniques such as association rule and sequential pattern discovery, clustering and classification.

6.2.4 Web usage mining in log analysis

Log analysis is regarded as the simplest method used in the Web usage mining process. The purpose of Web usage mining is to apply statistical and data mining techniques to the pre-processed Web log data, in order to discover useful patterns. Association rule mining is a technique for finding frequent patterns, associations and correlations among sets of items. Association rules are used in order to reveal correlations between pages accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. Aside from being exploited for business applications, such observations also can be used as a guide for Web site restructuring, for example by adding links that interconnect pages often viewed together, or as a way to improve the system’s performance through prefetching Web data.

Sequential pattern discovery is an extension of association rules mining in that it reveals patterns of co-occurrence incorporating the notion of time sequence. In the Web domain such a pattern might be a Web page or a set of pages accessed immediately after another set of pages. Using this approach, useful users’ trends can be discovered, and predictions concerning visit patterns can be made.

Clustering is used to group together items that have similar characteristics. After discovering patterns from usage data, a further analysis has to be conducted. Additionally, visualization techniques are used for an easier interpretation of the results. Using these results in association with content and structure information concerning the Web site there can be extracted useful knowledge for modifying the site according to the correlation between user and content groups.

7. CONCLUSION

Web personalization is the process of customizing the content and the structure of a Web site to the specific and individual needs of each user, without requiring from them to ask for it explicitly.

Enterprises expect that by exploiting the information hidden in their Web server logs they could discover the interactions between their Web site visitors and the products offered through their Web site. Using such information, they can optimise their site in order to increase sales and ensure customer retention. Apart from Web usage mining, user profiling techniques are also employed in order to form a complete customer profile. Lately, there is an effort to incorporate web content in the recommendation process, in order to enhance the effectiveness of personalization. Thus web personalization has made web activities user-
centric and has made the users an integrated part of the web environment.

References

[1] Information Customization, Helping online customers decide through web personalization, Sung Ho Ha, Kyungpook National University, 1094-7167/02/$17.00 © 2002 IEEE.


