MyWeb: A Novel Approach for the Personalisation Of World Wide Web by Learning User Preferences

Vimal Jose
Amal Jyothi College of engineering
Mahatma Gandhi University
Kerala, India, +91 9446784708
vimaljosehere@gmail.com

Maria Joseph K
Amal Jyothi College of engineering
Mahatma Gandhi University
Kerala, India, +91 9446784674
mariajkarippan@gmail.com

Shaan Geo, Mentor
Amal Jyothi College of engineering
Mahatma Gandhi University
Kerala, India, +91 9400418132
shaangeo@amaljyothi.ac.in

ABSTRACT

The World Wide Web has been evolving ever since its invention. The next step in this evolution being personalisation, the need of the hour is practical and efficient methods to implement the same. In this paper, we present a novel way of personalising websites by supplying the user preferences, both visual and content based, to web servers automatically. These preferences are learned by assessing user behaviour and identifying the preferences intelligently and also by explicit questionnaires. The learned user preferences are made available to the web servers when the user visits a website and the website modifies itself accordingly. By this mechanism personalisation of the World Wide Web is made possible.

General Terms

User preference learning, user adaptive websites, personalization of web

Keywords

Personalized websites, learning user preferences, intelligent web

1. INTRODUCTION

The next step in the evolution of the computing being personalisation, it is important to personalise the World Wide Web for enhanced usability and ease of finding the correct data. In this paper, we present a novel way of personalising websites by supplying the user preferences, both visual and content based, to web servers automatically. Personalised websites should be able to modify itself according to the preferences of the user. For this, a detailed knowledge of the interests and preferences of the user is essential. This should include content preferences like areas of interest, type of content preferred and presentation preferences ranging from favourite font to the background colour preferred.

These details about a user should be made available to all the websites that the client accesses. The web servers use this data along with the client request to prepare the target page.

From the user's point of view, the website shows relevant information of the preferred content type, in a style which is visually appealing to him/her. Thus MyWeb helps the user to easily find the required information on the web as well as viewing the web in his/her preferred visual themes.

From the webmasters point of view, MyWeb helps them to serve customers better by knowing the services he actually needs. It also helps webmasters by giving them a chance to generate better user interest by providing the exact data that he is looking for.

Using MyWeb, online advertising can be modified in such a way as to show only those advertisements in which the user is interested in, rather than showing the same advertisement for all users. Thus online advertising can be transformed from general advertising to targeted advertising, which will also benefit the advertisers and also the webmasters to achieve higher click-through rates.

2 COLLECTION OF RAW DATA

Collection of user preference data is the longest and toughest tasks of the proposed MyWeb system. MyWeb has to detect and record the actions of the user ranging from page views to mouse movements. It should also extract and record keywords from web pages and also the type of contents preferred.

The data collection about the user can be broadly classified into

(1) Implicit data collection
(2) Explicit data collection

2.1 Implicit Data Collection

Implicit data collection means the automatic detection and recording of relevant data about the user without the knowledge and intervention of the user. The implicit data collection process works in background while the user browses the internet. The methods for implicit data collection include page content analysis, context recording and action detection for the estimation of degree of user interest on a page [1]. When a page is opened in browser by the user, the contents of the page analysed to extract the keywords which define the topic described in the page. For this purpose, any one of the existing algorithms is employed.

The context under which a page is opened is an important criterion for a better understanding of the content and relevance of the page. We assume that the pages simultaneously opened by a user have better chances of being related to the same topic. For this a timeline of the page views can be made.

Table 1: Example timeline of page views by a user

Address	Page load time	Page-close time
http://www.website1.com	0900	0904
http://www.website2.com	0903	0909
http://www.website3.com	0910	0918
http://www.website4.com	0911	0918

Action detection means the identification and recording of specific actions done by the user. The common actions that show user interest/ disinterest according to the study given in [1] is

1. Text selection: Selecting text by dragging a mouse
2. Text tracing: Moving mouse pointer along a sentence while reading.
3. Link clicking: Clicking on a link to move to another page.
4. Link positioning: Positioning a mouse over a link, but not clicking the link.
5. Scrolling: Scrolling a window at a certain speed.
6. Circling: Moving a mouse pointer in circles.
7. Bookmark registration: Registering the page as bookmark.
8. Saving of image data: Saving images from the page to system.
9. Page printing: Printing the page
10. Opening a page in a new window.
11. Total time spent viewing a page

These parameters can be used to reveal a lot of information about a user.

2.2 Explicit Data Collection

In the explicit data collection method, there are three ways of getting user data from user.

2.2.1 By using preliminary questionnaire [2]

The user is requested to answer a set of questions which describe the personal interests of a user. These may range from topics of interest to colours of preference.

2.2.2 By reviews of page

When a page is visited in the browser, the user is allowed to review the page by giving points to each attribute of the page. These may be content relevance, presentation style, colours used etc.

2.2.3 By refining results of implicit method [2.1]

In this method, the user is presented with the results of the implicit method and is allowed to modify any part of it. This method helps to reduce the errors in the results of implicit data collection and thereby enhancing the total performance of the system.

Explicit data collection has the advantage of being more accurate and direct from the user.

3 PROCESSING OF DATA AND GENERATION OF RESULTS

From the processing of collected data, accurate inferences about user interests and preferences are to be generated. This process consists of two steps.

3.1 Generating inferences about topics of interest

The main function of MyWeb is to give an idea of the user interests to the web servers with which the user is interacting. The most important information about a user is his topics of interest. The steps by which we generate inferences about topics of interest of the user are

3.1.1 Sorting the web pages based on relevance

In this step, we use the data collected in the data collection phase to estimate the relevance of the web pages, by two methods. A probabilistic method as described in [3] and a non-probabilistic method as described in [1]. The probabilistic method uses the time spent viewing a page as input and generates a relevance score R_{w1}, $0< R_{w1}<1$. The non-probabilistic method described in [1] estimates a user's degree of interest in page using both the explicit as well as the implicit parameters collected during the data collection phase. This method also returns a relevance score between 0 and 1 for each web page. Let this be R_{w2}. We employ both methods for calculating the relevance of the web page to overcome the short comings of each method. The final relevance score is calculated as.

$$R_w = \frac{N1Rw1+N2Rw2}{2} \qquad (1)$$

Where N_1 and N_2 are the normalisation factors for methods 1 and 2 and $0< N_1, N_2 <=1$. Normalisation factors are introduced to nullify the effect of the differences in finding the relevance in the two methods.

3.1.2 Sorting keywords on the basis of relevance

After the relevance of each page is found out, the relevance of each keyword in the page is to be calculated. For this, the first step is to fetch all the meta-keywords of each page. The relevance of the keyword is initialised to the relevance of the page at which it was found. If multiple pages contain the same keyword, the higher relevance is to be selected.

After initialising the keyword relevance with the highest relevance score of the web pages in which it appears, we modify this value according to

$$Rk_{i+1} = Rk_i + \frac{Rwi+1}{Rki.C} \quad \forall i, 0 < i < W \qquad (2)$$

Where C is a constant which can be modified according to the total number of websites containing the same keyword; higher the value of C, smaller will be the step at each iteration. W is the total number of web pages that contain the keyword under consideration.

The equation (2) is inclined to give a higher relevance score to keywords since it starts from the highest relevance score available. But this is justified with the requirement that the relevance score of a keyword must not be substantially reduced by a single irrelevant page, because the irrelevance of that page is possibly related to the unappealing appearance of the page. Though the vice versa is also possible, that is found to be more tolerable.

Repeating the same process for each keyword, we can obtain the list of keywords sorted in the decreasing order of relevance. The content preferences can be correctly identified from this list.

3.2 Generating inferences about visual preferences

This step is relatively easy when compared to 3.1, since no iterative computations and only comparisons are required. In this step, we extract and compare the visual properties of the web pages starting with that of highest relevance. We select the values with highest number of occurrences in highly relevant web pages for each property. This process will result in a number of possible values for each property along with a preference score for each value.

Storage details of this information are not discussed in this paper since a lot of methods are already available for the same.

4 IMPLEMENTATION

After the data analysis and computation phase, the results are to be made available to the websites that the client accesses. We make two types of information accessible to the websites.

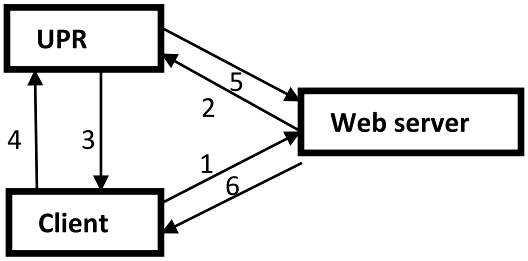
1. The sorted list of keywords and their relevance scores
2. The visual attributes and their preferred values

This data can be stored at the client or at a remote server. While the advantages of keeping the data at the client being enhanced privacy and security, storing at a remote server has the advantages that less client bandwidth usage and with such a system, the client can login from any computer around the world to continue using MyWeb. A viable option for this is to let the user choose which storage to use.

MyWeb can be implemented as a browser support application, for e.g. a plug-in, which is to be installed at the client system to start using the system. This application collects raw data with both implicit and explicit methods and process the data to generate inferences. The two pieces of information in list 1 will be store in a User Preference Repository (UPR) which is accessible to the websites that the client accesses. The location of the UPR can be the client computer itself or a remote server. The websites use the data from UPR to generate highly personalized web pages.

4.1 Working

The overall working of the system is



1. Client request to web server
2. Server request to UPR to access information about user
3. Request for permission to grant rights for the web server to access information
4. Permission granted/ not granted
5. If permission is obtained, give information to web server

5 MYWEB: FUTURE WORKS

The next phase of the internet being the personalisation of the web, MyWeb has enormous potential in the present scenario. The disadvantages of MyWeb system at present are problems of security, loss of privacy and high bandwidth usage. Another problem is that when using the client computer as the storage location, the user cannot switch computers while using MyWeb. This can be solved by storing the data remote server and restricting its usage based on a secure login. The security and privacy issues can be resolved to some extent by using encryption algorithms and selective conditional data access to servers and the use of certificates.

6 CONCLUSION

MyWeb is a tool for personalising the World Wide Web according to the preferences of the user. The preferences are automatically detected using the browsing behaviour of the user in the past. The preferences are dynamically updated according to the web browsing behaviour of the user. It is beneficial to both the clients and the web masters. The next trend of World Wide Web being personalisation, MyWeb is a step towards the future.

REFERENCES

[1] Yoshinori Hijikata. Estimating a Users Degree of Interest in a Page during Web Browsing, IBM Research, Tokyo Research Laboratory

[2] Norikatsu Nagino and Seiji Yamada. Future View Web Navigation based on Learning Users Browsing Patterns, CISS, IGSSE, Tokyo Institute of Technology and National Institute of Informatics

[3] K. Voigt. SKIPPER A Tool that Lets Browsers Adapt to Changes in Document Relevance to Its User. Department of Computer Science, California State University San Bernardino

[4] Jorge Bergasa-Suso, Member, IEEE, David A. Sanders, Member, IEEE, and Giles E. Tewkesbury. Intelligent Browser-Based Systems to Assist Internet Users

[5] A. Eckhardt, T. Horváth and P. Vojtáš. PHASES A User Profile Learning Approach for Web Search - Charles University, Prague, Czech Republic.

[6] Janez Brank, Natasa Milic Frayling, Anthony Frayling and Gavin Smyth. Predictive Algorithms for Browser Support of Habitual User Activities on the Web