A Conceptual Framework for the Integration of Clustering and GA for a Web Personalization System

Mylini Munusamy
School of Information Technology
Monash University Malaysia
No. 2, Jln Kolej, Bandar Sunway,
46150 Petaling Jaya, Malaysia
mylini.munusamy@infotech.monash.edu.my

Leow Soo Kar
School of Information Technology
Monash University Malaysia
No. 2, Jln Kolej, Bandar Sunway,
46150 Petaling Jaya, Malaysia
leow.soo.kar@infotech.monash.edu.my

Abstract – **With the rapidly growing Web of information, web personalization has become one of the key research areas with the aim of studying methods to enhance the user's web browsing experience. One of the key benefits of web personalization systems is for making recommendations. These recommendations are usually made based on a single characteristic, either the web pages or the user browsing pattern. In our proposed framework, we suggest the use of a combination of characteristics for use in the personalization component, as a method for improving the accuracy of the recommendations made to the user. In this paper, the main objective is to propose a method of integrating both clustering and genetic algorithms (GA) for web personalization. In the research, we explore the use of GA as a possible solution for refining the clustering process. We also propose a conceptual framework for the use of GA in a web personalization system. Additionally, we discuss the user characteristics used for clustering and in the GA, an important consideration for the system.**

I. INTRODUCTION

As more and more information becomes available on the Web, there is a growing need to provide personalized web services that adapt to varying user requirements and preferences. The customization of services and information specific to different users is becoming an essential aid that enables users to browse more efficiently and effectively. Therefore, web personalization systems are getting more and more popular. Web personalization may be described as the process of enhancing a user's web browsing experience through customization according to the user's own preferences. Essentially, a web personalization system may be divided into three components: data pre-processing; usage pattern discovery and personalization.

In the first phase, the data is subjected to cleaning and filtering where unwanted data is removed. In the next phase, the necessary attributes or characteristics required for the pattern discovery are extracted and mapped to the characteristic representations. These characteristics may be normalized if necessary, and are then subjected to various pattern discovery techniques such as clustering, machine learning or predictive methods. Finally, the personalization will involve some form of customization based on the user's preferences derived, such as adapting web pages in the form of recommending pages or suggesting shortcuts to other pages within the site or even changing the layout and content of the page itself to accommodate the user's interests.

In this research, our focus is directed towards proposing the use of genetic algorithms (GA) [5, 6, 8] as a method for improving web personalization. Using the web log sessions as data, we obtain the necessary characteristics to perform the pattern discovery by clustering the web sessions. We then propose the use of GA for matching a user to a cluster for personalization. In this context, the clustering will be done offline using the existing web server logs and the matching of the user to a particular cluster will be accomplished during a user's browsing session by the GA component. Both processes will be done on the server side, thus providing a passive method for personalization that does not require explicit user input or feedback.

The remainder of this paper is organized as follows. In the next section, we discuss related work and provide a background on the work that we propose. In section III, we describe the characteristics that we will use to for clustering as well as some related research works that have been done. In section IV, we provide a brief description of the clustering component and a discussion of the GA component that we propose for our framework. Finally, we discuss our future work and plans for this research.

II. RELATED WORK AND BACKGROUND

Various Web personalization and recommender systems have been developed that employ varying methods for mining user web sessions. The multi-modal clustering algorithm [11] utilizes data features from content and structure, in addition to URL tokens and the sequence ordering as well as viewing time spent on each page. Fu et al. [9] proposed the generalization-based clustering of Web users based on their access patterns, where the sessions are generalized using an induction method. The dynamic profiler generates dynamic profiles for users based on the Web content using a supervised clustering scheme [17]. User classification is another method for categorizing users, as used in the SUGGEST system, where the clustering is done offline and the recommendations are accomplished on-the-fly [4]. Perkowitz et al. [15] proposes a web usage mining system that builds index pages that contain links to groups of similar pages, using the clustering of pages rather than sessions. In a related approach, Labroche et al. [12] proposed a clustering method based on the chemical recognition system of ants, called AntClust. In this paper, the clustering algorithm associates a data set to the odour of an ant to simulate the meeting of ants.

Whilst there have been various research works into this area of web usage mining and web personalization, however, most of which are directed towards the traditional clustering, machine learning or predictive

methods. In this research, our focus is directed towards exploring an alternative approach, the use of GA, in a web personalization framework. Our main purpose of introducing the GA component is to further refine the clustering process, where it's primary function is to match a user to the particular cluster for personalization using the user characteristics identified whilst the user is browsing.

With the proposed framework, we are able to combine the use of different user characteristics as a method to improve the accuracy of the personalization, therefore improving the recommendation to the user. In most of the research works that have been discussed, the personalization is usually based on the user's characteristics, such as the browsing pattern. However, in our proposed work, we combine both the user characteristics as well as the content interests derived from the web pages as a measure for use in both the clustering and GA algorithm. The basic idea behind this is to use the content interests – the contents of the pages browsed – derived from the pages visited by users of the site as a form of implicit weighting for the clustering algorithm, which is combined with the user characteristics derived from the browsing pattern for personalization. The identification of the user characteristics that would have a significant effect on the personalization process is discussed in the next section.

III. USER CHARACTERISTICS

There are two main categories of the web personalization systems – cognitive filtering and collaborative filtering. Cognitive filtering, also referred to as content-based filtering, is characterized as filtering that is based on the content of the data items and the user's interests. Cognitive filtering systems may vary depending on the attributes or characteristics of the user that are acquired and collaborative filtering may be described as filtering that functions through the interrelationships of a group of users.

In the proposed framework, we will employ the use of both content-based and collaborative filtering for personalization, where we combine both the benefits of collaborative and content-based filtering. To provide effective personalization, a system must be able to understand and derive the characteristics and behavior of a particular user. Usually, this is accomplished using some form of feedback mechanism, where users are generally required to rate a particular topic or page as an indication of the interestingness of that particular topic or page. There are also instances where users' are required to specify their profile at the beginning of the browsing session. However, this is often cumbersome to the user. Therefore, the data to be used for both purposes will be obtained through implicit methods.

Implicit methods are usually more discrete and do not require a user's active involvement. There are various characteristics or behaviors about a user that may be acquired for modeling through implicit methods. Specifically, these may include the number of hyperlinks clicked on a page; amount of scrolling performed; mouse activity; browsing history or navigational browsing behavior; time spent on a particular document; number of revisits made to the document; whether the user has mailed back the document to herself or not; and content interests.

Various research works on the attributes to be used for personalization have been conducted. Morita [14] proposed a profile acquisition technique to accumulate a user's preference based on the hypotheses that the amount of time spent on an article indicates the interestingness of an article. As a more accurate measure, CASPER [16] constructs user profiles by passively monitoring the click-stream and read time behavior of regular users. Specifically, the relevance of a document is determined by the amount of time spent reading the document, the number of revisits made to the document, and whether the user has mailed the document back to herself or clicked a button on the document.

Goecks [10] presents an approach that learns from the user's behavior, specifically the number of hyperlinks clicked on a page, the mouse activity and the amount of scrolling performed, based on the assumption that the user's interest corresponds to these behaviors. The web site personalizer, Proteus, by Anderson [2], mines the past interactions with the web site for the navigational patterns and the content viewed, from the access logs.

Considering the fact that web users tend to have routine web usage patterns, the Montage [1] system was developed to help users with routine web browsing, which is defined as the overall pattern of content access that a user performs whenever in the same or similar contexts. The primary knowledge acquired to build the user model is the sequence of pages that a user requests and the content interests.

Based on these works, our research will focus on the use of the following characteristics for each user in a session – the sequence of pages, the time spent on each page, the number of revisits to the page as well as the content interests. Here, we advocate that the characteristics that we will be using be using for clustering and in the GA, are all important characteristics for obtaining a better recommendation. Also, that the weights of the characteristics are related to one another. For example, a user will visit pages that contain materials that are deemed to be interesting. However, in the process of reaching a particular, the user may have to visit two other pages. But the time spent on those two pages may be very short as they function only as a bridge to the page of interest. Hence, upon reaching the page of interest, the user will spend a longer time on the particular page and may also revisit the page frequently for up-dates. Therefore, the number of revisits together with the time spent on the page as well as the content interests of the page will be an indication of the interestingness of a page to the particular user.

Using the web server logs, we can identify the transactions for each user within one session. The transactions are basically the links or pages that are visited by the user within a session, where each session will contain a minimum of one transaction. Each transaction record will show the page visited and the time spent on the page. Additionally, we can also derive the number of revisits made to the page in that particular session by counting the number of times the user has visited the page in that session. A session may be defined as the time from when a user starts browsing a page on a particular site to

the time the user leaves the site. In each session, there will be at least one transaction, where a transaction refers to a page that the user has visited, regardless of the time spent on the page.

In the first stage of pre-processing the data, the data is subjected to data cleaning, user identification, session identification and path completion [1]. Using the data preparation methods as discussed by Cooley [7] and Mobasher [13], the user sessions may be identified from the web log sessions.

There are some assumptions concerning the user characteristics that are made in this context. First of all, we assume that the time spent on a page is an indication of the user's interest on a particular page as well as the contents. Also, the number of revisits made to the page reflects the user's interest in the page – the more often the user revisits, the more interested the user. Finally, we assume a threshold period to derive each user session, where a long period of inactivity between the pages browsed may result in two or more separate sessions for the user. Therefore, each user identified would have at least one user session within the web log.

IV. USING GA FOR WEB PERSONALIZATION

A. Clustering Component

We will cluster users based on the browsing patterns and content interests where users with similar browsing activities will be clustered or grouped into classes. Specially, the goal of the clustering algorithm that we employ is to find small clusters according to the content interest as well as the browsing patterns. For this purpose, we identify three significant vectors: session vector, sequence vector and page vector.

The web server logs may be easily obtained from the server whilst the web pages from the web site may be obtained using a web crawler. From the web server logs, we perform the data cleaning and extract the user sessions as described in the previous section. Using the user sessions identified, we can construct both the sequence vector as well as the session vector. The reason for separating the session vector and the sequence vector is related to the GA component, where the purpose is to simplify the GA and reduce the time as well as the computational complexity. The sequence vector is essentially constructed from the link path as the user browses. For each user, there will be at least one session associated with the user. The weightings for the similarity metrics for the sequence vector will be derived from the time spent as well as the revisits made to the particular page.

From the web pages obtained by crawling, we can construct a page vector for each page within the web site. This page vector will consists of the keywords or terms extracted from the contents of the page. A simple Term Frequency – Inverse Document Frequency (TF-IDF) weighting scheme is used to obtain the frequency of the occurring words within a page [3]. This method provides a numerical weighting to a particular term on the pages indicating the relative importance of the term within the context of the page.

Using a combination of both the page vector and sequence vector, we can then define the similarity metrics and apply traditional clustering methods such as K-means to obtain the clusters. For each cluster identified, we can generate a list of index terms and the pages associated with these index terms as discussed in Baeza-Yates [3].

B. GA Component

A session vector that consists of the number of revisits made to a particular page and the time spent on the page is then used in this component. The session vector $S(j)$ may be defined as follows;

$$S(j) = \{ r_j(i), t_j(i) \;\; i \in I, k \} \quad j = 1, \ldots\ldots, N$$

where

$r_j(i)$ = number of revisits to page i
$t_j(i)$ = time spent on page i
j = tag of the session vector
I = set of web pages
k = cluster number
N = total number of sessions derived from the web log data.

For example, for a website having 10 web pages, each session vector will be assigned a tag, which is a unique identifier for the particular session. An example of a session vector, with tag number 8, is given as follows

$$S(8) = \{(2,0,1,0,6,1,9,4,7,4),(1,3,2,2,3,1,3,2,3,2),1\}$$

The first set of 10 values represent the number of visits to the ten pages of the website and the next 10 values are the durations spent on each these pages. The last value represents the cluster it belongs. This value is obtained as a result of the clustering process. The cluster number will be set as default to 0 before performing the clustering process. At the completion of the clustering process, a number of clusters, each with its own similarities will be obtained. Members of these clusters are session vectors with similar contents interest as well as browsing patterns.

Once the clustering is complete, the GA will be employed to optimally match the attributes derived from the navigation pattern of an active user to a cluster based on a similarity measure of the session vector.

We describe the characteristics of our GA as follows:

1. GA Population will consist of session vectors, randomly selected from the clusters, for example, the following might be two individuals of the population:

   $S(3) = \{(2,0,1,0,6,1,9,4,7,4),(1,3,2,2,3,1,3,2,3,2),4\}$

   $S(10) = \{(2,0,1,0,6,1,9,4,7,4),(2,3,3,2,3,1,3,2,3,2),5\}$

2. Chromosomes are represented by the session tag number in binary format, for example,

Individual	Chromosome
S(3)	0000011
S(10)	0001010

The size of the chromosome (i.e. number of bits) corresponds to the number of bits required to represent *N*, the total number of session vectors obtained from the web logs

3. Different similarity measure functions can be used to define the fitness function, depending on the user characteristics used for personalization. In this work, we use the cosine similarity measure as the fitness function.

Suppose the session vector of a current active user is obtained as follows:

$$S(m) = \{ r_a(i), t_a(i) \ \ i \in I, 0\}.$$

Then, the fitness function value associated with *S(j)* is given by

$$F(j) = \frac{\sum_{i=1}^{|I|}\{r_j(i)r_a(i)+t_j(i)t_a(i)\}}{\left(\sqrt{\sum_{i=1}^{|I|}\{r_j^2(i)+t_j^2(i)\}}\right)\left(\sqrt{\left(\sum_{i=1}^{|I|}\{r_a^2(i)+t_a^2(i)\}\right)}\right)} \quad (1)$$

The GA is terminated when the fitness value exceeds a predetermined bound or when it reaches the maximum number of generations allowed. At the conclusion of its run, the GA assigns to the active user, the cluster number of the individual with highest fitness value.

4. The crossover operation randomly selects 2 individuals from the population and performs a simple one point crossover.

5. The mutation operation randomly selects an individual from the population and performs a random swapping of 2 alleles in the chromosome.

The straightforward crossover and mutation operations can be performed very quickly, and so together with the simple chromosome encoding, the GA is expected to be a very fast algorithm. Our proposed system would be especially suitable if extended over a framework in personalization for mobile web applications.

C. Conceptual Framework

In our framework, we integrated both components discussed above to form a conceptual framework for a web personalization system as depicted in Figure 1. This framework will adopt a two-level architecture where the clustering process will be done off-line and the GA will be performed online to be able to make recommendations on-the-fly.

For an active user who is currently browsing, we can obtain an active session vector as the user is browsing. Using this active session vector as input to the GA, we can then find the optimized session vector as the output of the GA process. This session vector produce from the GA is then identified from the clusters, where the cluster from which the session vector is identified will contain the similar preferences as the current user. Therefore, the recommender process can then make the necessary suggestions to the user as the user is browsing. This process is a dynamic process that changes every time the user visits a new page.

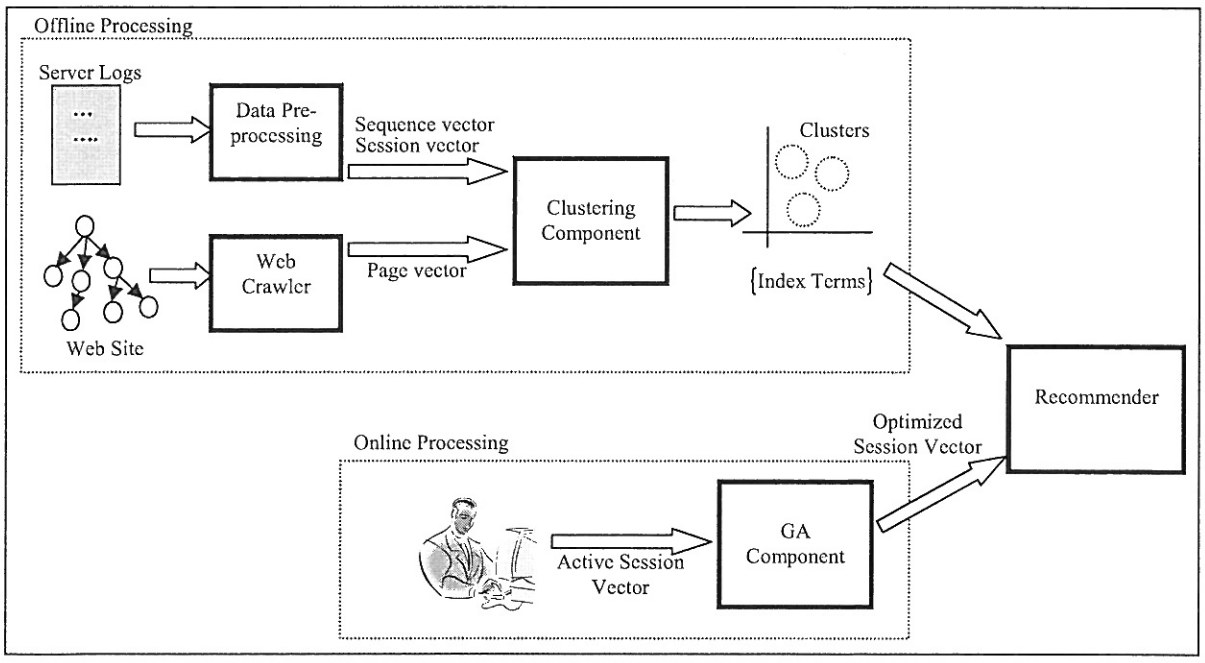

Fig. 1 Proposed Conceptual Framework for Web Personalization System

V. DISCUSSION AND CONCLUSION

As we are at the initial stages of this research, we have yet to implement the system and are working towards implementing this system in the near future. Given that the GA that we have proposed is expected to produce results quickly, we will further explore the possibility of extending this framework (see Figure 1) for use in the mobile web personalization context.

VI. REFERENCES

[1] Anderson, C. R. and Horvitz, E., "Web Montage: A Dynamic Personalized Start Page", in *Proceedings of the 11th International Conference on WWW 2002*, ACM, pp. 704-712

[2] Anderson, C. R., Domingos, P. and Weld, D. S., "Adaptive Web Navigation for Wireless Devices", in *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01) 2001*, pp. 879-884

[3] Baeza-Yates, R. and Ribeiro-Neto, R. *Modern Information Retrieval*, ACM Press, New York, 1999

[4] Baraglia, R. and Palmerini, P., "SUGGEST: A Web Usage Mining System", in Proceedings of 2002 International Symposium on Information Technology (ITCC) 2002, IEEE, pp. 282-287

[5] Chen, Y. and Shahabi, C., "Improving User Profiles for E-Commerce by Genetic Algorithms", in *E-Commerce and Intelligent Methods Studies in Fuzziness and Soft Computing*, Kluwer Academic Publishers, 2002

[6] Davis, L., *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York, 1991

[7] Cooley, R., Mobasher, B. and Srivastava, J., "Data Preparation for Mining World Wide Web Browsing Patterns", in *Journal of Knowledge & Information Systems*, vol. 1, 1999.

[8] Falkenauer, E., *Genetic Algorithms and Grouping Problems*, John Wiley and Sons, 1998.

[9] Fu, Y., Sandhu, K. and Shih, M. Y., "Clustering of Web Users Based on Access Patterns", in *Proceedings of the 1999 KDD Workshop on Web Mining*, Springer-Verlag, San Diego, CA, 1999.

[10] Goecks, J. and Shavlik, J., "Learning users' interests by unobtrusively observing their normal behaviour", *in Proceedings of the 2000 International Conference on Intelligent User Interfaces*, ACM, pp. 129-132

[11] Heer, J. & Chi, E. H., "Separating the Swarm: Categorization Methods for User Sessions on the Web", in *Proceedings of CHI 2002*, vol. 4, No. 1, ACM, pp. 243-250

[12] Labroche, N., Monmarche, N. and Venturini, G., "AntClust: Ant Clustering and Web Usage Mining", in *Proceedings of Genetic and Evolutionary Computing Conference 2003*, LNCS 2723, Springer-Verlag, pp. 25-36

[13] Mobasher, B., "Web Usage Mining and Personalization", Draft Chapter in *Practical Handbook of Internet Computing*, Munindar P. Singh (ed.), CRC Press. *To appear in 2004.*

[14] Morita, M. and Shinoda, Y., "Information filtering based on user behaviour analysis and best match retrieval", in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development*, pp. 272-281

[15] Perkowitz, M. and Etzioni, O., "Towards Adaptive Web Sites: Conceptual Framework and Case Study", in *Proceedings of International Joint Conference on Artificial Intelligence 1999*, pp. 264-269

[16] Rafter, R., Smyth, B. and Bradley, K., "Inferring Relevance Feedback from Server Logs: A Case Study in Online Recruitment", *in Proceedings of the 11th Irish Conference on Artificial Intelligence and Cognitive Sciences* (AICS 2000)

[17] Wu, K., Aggarwal, C. C. and Yu P. S., "Personalization with Dynamic Profiler", in *Proceedings of Third International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems 2001*, IEEE, pp. 12-20

[16] Wu, Y. H., Chen, Y. C. and Chen, A. P. L., "Enabling Personalized Recommendation on the web based on user interests and behaviours", in *Proceedings of the 11th International Workshop on Research Issues in Data Engineering 2001*, pp. 17-24.