IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP4707

# Automating Web Personalization with a Self-Organizing Neural Network

Victor Perotti, Ph.d.
Management Information Systems
Rochester Institute of Technology College of Business

The Internet's continuing transformation of business has created new and unique demands for information management. While these demands are multifaceted, perhaps none are more important than pursuing an understanding of website users, and leveraging this information to create site structure and content that is appropriate for users. Indeed, personalization and customization of web services are now commonplace features on many e-Business sites. Certainly, many sites use "cookies" or other technologies to track each individual as they come to the site. Consumers, however, are increasingly wary when their personal information is requested. A less invasive approach is to look at the aggregate behavior of all users, and to try to identify trends therein. Once these trends are identified, a user can be classified as a member of a particular group, and customized web content can be delivered.

Recommendation engines that identify user trends, and deliver user-appropriate content are an active field of research today. For example, Perkowitz & Etzioni (1997) challenged the Artificial Intelligence community to develop truly adaptive web sites that respond to the behavior of their users. Their more recent work (1998) introduces conceptual analysis as part of a "cluster mining" process which identifies groups of users that have common usage profiles. Similarly, Mobasher, Cooley and Srivastava. (2000) have developed an entire process for clustering web users, and have shown its performance relative to other algorithms. While the recovery of usage profiles by these two groups is fairly advanced, the degree of adaptation offered is fairly small.

### A Complex Systems Approach to Web Personalization

The recommendation systems from the aforementioned research create "index pages" or recommended links so that users can access content that will be especially relevant to them. To accomplish this, they use data mining and statistical techniques to identify web usage clusters. An alternative and potentially superior approach is to utilize a complex system that continuously adapts to the user patterns at the web site. Perotti and Kiran (2002) employed this idea to visualize the usage patterns at a web site. In this work, a neural network called a Self-Organizing Map (SOM) was trained with website usage logs, and all subsequent user accesses could then be mapped to a particular cluster of web pages. The authors demonstrated a scatter plot that succinctly summarized the usage patterns at the studied web site.

The present research seeks to build on the work of Perotti and Kiran (2002) by automating personalized website content based on the up-to-the-minute adaptations in a Self-Organizing Map. Figure 1 shows a data flow diagram for the proposed personalization system. As users visit the site their usage patterns are continuously fed to the SOM as training information. In addition, when a user requests a web page, a CGI (common gateway interface) program is used to invoke and interpret the adapted SOM to identify structure and content that is appropriate user. A brief explanation of the Self-Organizing Map neural network will help to explain the recommendation system.

### Kohonen Self Organizing Map

Kohonen's Self Organizing Map (SOM) is a well-known neural network technique to do data dimensionality reduction and clustering. In this technique, a neural network is created in the desired low dimensionality, say two dimensions for the sake of explanation. The network is then trained with a set of input patterns represented as a set of input vectors that correspond to the high dimensional data to be reduced. Each node in the network includes a model vector, which is that node's representation of the input vectors. The process of training involves comparing each input pattern with the model vector at every node. When a node's model vector is the closest to a given input vector, that node and all of its neighboring nodes are altered to be more like that input vector. Thus, through repeated training, one of the network nodes becomes highly associated with each input pattern, so that when the correct input pattern is presented, it will be the most highly active node in the network. After training, the neural network represents a simple (two dimensional) map with nearby nodes representing similar input patterns in the multidimensional input data.

Self organizing maps have been already used for a great variety of problems, including browsing a picture database, data exploration, representing large text collections and classifying web documents based on their textual content (Kohonen et al, 2000).

### Self Organizing Maps as part of a recommendation engine

The advantages to using a complex system like a Self organizing map for website recommendations are many. For one, the neural network generalizes from its training items to any new items. Thus, any new data can be analyzed and compared to earlier items without having to go through training. Thus, a fully trained neural network can make almost instantaneous "decisions" about which cluster a given web site visit belongs in. This is not the case for other recommendation systems, which may have to go through the very large web log in order to include the latest information.

Furthermore, the training of the neural network can happen in parallel with making recommendations, and may in fact go on continually. Continuous training means that if desired, a website can make recommendations based on short term fluctuations in website usage. For example, if a web page deep within the website's structure suddenly becomes popular, the web site can recommend that page from the home page to simplify access for subsequent users.

Another advantage to the self organizing map approach is the generality of the training process. In its most generic form, training is a distance comparison between input vectors and model vectors. What is represented in the input vectors, and how it is represented as a vector are not constrained. In addition, the distance function itself is configurable, allowing certain elements in the vectors to have more weight than others. For the recommendation system, this means that web users can be clustered based not only on the specific web pages they visit, but also on any information available about them in the website log. This flexibility could allow a self organizing map system to recognize when, for example, different geographic regions are using the web site in unique ways. Thus a user from the Southwest United States accessing a page could be given recommendations that are popular in that location, while visitors from Thailand are given a different set of recommended pages. Similarly, website visitors could be differentiated based on their time of access. If the recommendation engine is supporting an electronic commerce site, early morning visitors could thus be directed toward products such as coffee or alarm clocks that are associated with the morning. Late night visitors, conversely, could see after dinner mints, or sleep enhancing medication as their recommendations.

### Directions for future research

While the manipulation of the distance function allows the clustering of website visitors to happen in many different ways, it also raises some challenging issues. How can the weights be assigned to the different sources of user information to best serve the users? Given information about which web pages are being visited, when the visit is happening and where the visit is arriving from, many different clusters can be derived based on the relative weights of these factors. A suitable way to balance these factors to derive the best possible clusters has yet to be developed.

Another challenge for future research is the opportunity to tie in website content data to the process of making recommendations. In the basic form presented here, each webpage is distinguished based only on its name, and location in the site. Recent research on “Information Scent” (Pirolli and Pitkow, 2000; Pirolli, Chen & Pitkow, 2001) has shown the promise of including content information as part of the analysis of website usage.

**References**

Mobasher, B. Cooley, R. Srivastava, J. (1999) *Automatic Personalization Through Web Usage Mining*, Technical Report TR99-010, Department of Computer Science, Depaul University, 1999.

Mobasher, B. Cooley, R. Srivastava, J. (2000) *Automatic Personalization Based on Web Usage Mining*, Communications of the ACM, 43-8.

Perkowitz, M. and Etzioni, O. (1998). *Adaptive Web Sites: Automatically Synthesizing Web pages*, Presented at the American Association of Artificial Intelligence conference.

Perkowitz, M. and Etzioni, O. (1997). *Adaptive Web Sites: An AI Challenge*, Presented at International Joint Conferences on Artificial Intelligence.

Perotti, V. and Kiran, R., (2002) The Visualization of Usage Patterns for Web Customization, presented at the IRMA conference, Seattle, WA

Chi, E. Pirolli, P., Chen, K. And Pitkow, J. (2001). *Using Information Scent to Model User Information Needs and Actions on the Web* In Proceedings of the Conference on Human Factors in Computing Systems, CHI ‘2001 Seattle, WA. 490-497.

Chi, E. Pirolli, P., And Pitkow, J. (2000). *The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site.* In Proceedings of the Conference on Human Factors in Computing Systems, CHI ‘2000 The Hague, Amsterdam. 161-168.