2002

An Adaptive neural network for understanding website usage patterns

Victor Perotti

Raj Kiran

An Adaptive Neural Network for Understanding Website Usage Patterns

Victor Perotti, Assistant Professor of Management Information Systems
Raj Kiran, Graduate Research Assistant
Rochester Institute of Technology College of Business
107 Lomb Memorial Drive
Rochester, New York 14623
Phone: 716-476-7753
Fax: 716-475-7055
Email: vic@mail.rit.edu

As the importance of the Internet rises, the need to create more adaptive and more usable web sites also grows. Most improvements to a web site require some knowledge of the site's users and how they are interacting with the pages. However, web professionals today have relatively few good options for capturing this information. Certainly, there are software and services to help summarize the basic information from the web site logs. This could mean keeping track of the frequency of visits for the individual web pages that make up a site, counting how many times the overall web site is visited from a specific web location, or other basic statistics.

At the IRMA 2001 conference, Perotti and Burke presented a technique and visualization that offers web developers an opportunity to easily see the pattern of usage at a website. Unlike earlier depictions, their Web Usage Plot emphasizes the relationship between the various pages at a web site by displaying them in a topographic organization: sites that are visited together frequently appear close together, while those that are seldom visited together in the same session appear far apart. Their process to create the Web Usage Plot visualization has several steps, as depicted below:

1. Cleaning and organizing the web server logs
2. Creating an aggregate representation of all users web page visits, the co-occurrence matrix
3. Visualizing this representation.

Table 1: A simple process for visualizing web usage (adapted from Perotti and Burke, 2000).

The final visualization step relies on a multivariate statistical technique called Multidimensional Scaling (MDS). This technique allows the reduction of the high dimensional data into lower dimensional coordinates that can be more easily visualized. The Web Usage Plot created with MDS does have many advantages over earlier representations of web site usage patterns.

Unfortunately, using MDS for web usage visualization can be tedious because the algorithms for reducing the data dimensionality are computationally expensive. For example, the authors used the SPSS software package, which limits the user to visualizing no more than 100 web pages. Clearly, many web sites have more web pages than this arbitrary limit. The present research explores an alternative and potentially

superior approach using a Neural Network to capture the usage patterns at a web site. In this technique, a neural network would be trained with the patterns of usage at a web site, and then would automatically organize a low dimensional representation of these patterns.

Kohonen Self Organizing Map

Kohonen's Self Organizing Map (SOM) is a well-known neural network technique to do data dimensionality reduction. In this technique, a neural network is created in the desired low dimensionality, say two dimensions for the sake of explanation. This network is then trained with a set of input patterns that correspond to the high dimensional data to be reduced. As the network adapts, 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).

The goal for the present research is to create and visualize a self-organizing map neural network representation of web site usage patterns. As in the Web Usage Plot, the self-organizing map visualization should be useful for web page developers to identify clusters of web pages that are visited together frequently. However, the new techniques go well beyond a simple substitution of the SOM for the Multidimensional scaling in the procedure outlined above.

One of the key issues in using a SOM is how the data is represented for training. We have found that the co-occurrence matrix (in Table 1 above) is not well suited for training a neural network. To understand why, consider the structure of the co-occurrence matrix. For every web page at the given web site, both a row and a column are created. So, if there were n total web pages at the web site, then the resulting co-occurrence matrix would be of size n^2. Inside a specific cell in the matrix is the number of times that the two pages (represented by the row and column) were visited together in the same session. So, for example, if webpage 16 were visited frequently with webpage 42, then we would see a high number in the cell for column 16 and row 42. Of course, only half of the matrix is really needed, since the usage of two pages in the same session is symmetrical.

To use the co-occurrence matrix as input to the SOM simply requires the treatment of each row in the matrix as an input pattern, since each row is a vector that describes the aggregate usage for one web page with all other web pages. The problem with this is that the goal for the SOM is to have pages that are visited together frequently map to nearby nodes in the two dimensional network. Unfortunately, the vectors representing two highly associated pages may be very different. Consider the example given above: the row for

web page 16 will have a high number in column 42, while the row for web page 42 will have a high number in column 16. These two vectors are thus very different!

A potentially superior representation of the same information for input to the SOM could be called the **session membership matrix**. As before, each row corresponds to a specific web page. However, each column now corresponds to a particular user session that was recovered from the web log file. For a given row, each column will have a one (1) in it if the web page represented by the row is visited the session corresponding to the column, and a zero (0) otherwise. Thus, to continue the example above, because web pages 16 and 42 are visited together in the same sessions, they should have a similar pattern of ones and zeros along their corresponding rows. Because web pages that are visited together in the same session will have similar vectors, the session membership matrix is more appropriate to train the SOM than the earlier co-occurrence matrix.

Visualization of the Self Organizing Map

Another unique contribution of the present research is in the visualization of the SOM. While there are several existing techniques to create a depiction from the self organizing map, the resulting pictures are often much more difficult to interpret than the simple map-like presentation in the Web Usage Plot. For example, a common SOM depiction requires the viewer to infer the presented relationships from a complex image of gray-scale or color levels. Since the goal of the research is to make an effective tool for web administrators and developers, a simpler image is desirable.

One existing way to visualize an SOM is to simply note which node in the matrix responds the most when presented with a given input pattern. The matrix can then be visualized by plotting a point at every grid location whose node responded the most during the presentation of the input. In our case, each web page would be represented by a point located at a grid location. However, this approach has two problems. For one, multiple pages frequently map to the same network node. So, the viewer would only see one point, when in fact several associated web pages may be represented there. A second problem is that the distance between points is somewhat arbitrary, since it simply corresponds to the regular distance between the SOM nodes in their grid.

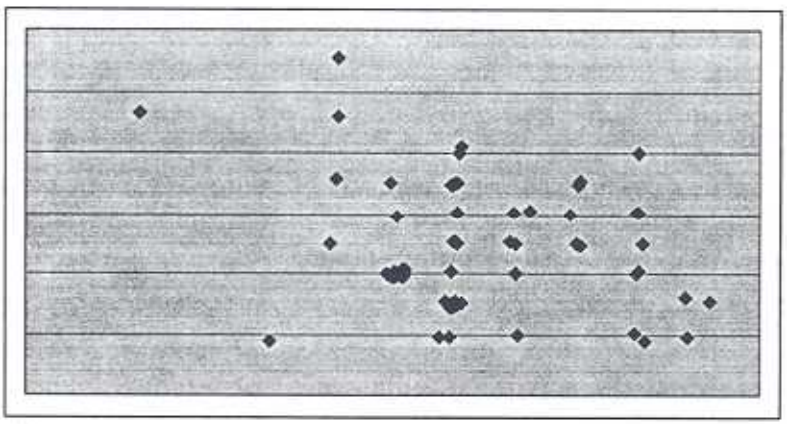

Figure 1: a jittered grid depicting clusters of web pages

Figure 1 demonstrates a novel "jittered" visualization of the self organizing map neural network, which overcomes the two problems mentioned above. Using this procedure, the visualized location of each web page is jittered by a small amount to displace it from the regular grid location. The displacement of each point is proportional to the error reported by the SOM network when responding to the specific input pattern. Thus, multiple web pages can be visualized at the same node location, and the association between the two of them can be seen as a cluster by the viewer. Also, the distance of any point from the regular grid location is a measure of how well that grid location's node succeeded in distinguishing that input pattern from the rest. Such a depiction is easy to interpret, and a viewer can quickly get a sense for the primary usage patterns that are present at the web site.

Discussion

The Self Organizing Map has great potential as a tool for creating useful visualizations of web sites. The present research has begun to develop the techniques necessary to get meaningful and useful results from the SOM neural network. In doing so, a visualization can be created that is as useful as the Web Usage Plot, but more robust in its computation. The SOM technique is relatively fast to compute, and has no restriction on the number of web pages that can be considered.

However, there is still much more research necessary to successfully use the SOM as part of a professional visualization system. One difficulty in using the SOM is a problem with dealing with sparse data sets. In a small web site sample, it is quite possible that most of the web pages are never visited, or are visited in one session. This means that of the hundreds of sessions available, a given web page will have only a single column active. The SOM network will frequently overlook the subtle difference between such pages, in considering the vast similarity in their pattern of not being accessed in so many sessions. Finding a set of parameters and appropriate training regimen for dealing with this problem can be quite time consuming. At present, a variety of different parameters must be experimented with using trial and error in order to find a useful visualization.

Perhaps even more important than the enabling of intuitive visualizations, capturing a web site's usage pattern in a neural network could provide a remarkably versatile component in new web-based applications. Recommendations could be made to the user on the fly, since when a user goes to a particular web page, it will be clear which other pages the other users have visited from there. Also, it would be possible to recreate some user behaviors from the network itself, so that novel web site structures can be readily evaluated or compared.

References

Kohonen, T., Kaski, S., Lagus, K. Salojärvi, J., Paatero, V. and Saarela,A. (2000) *Self Organization of a Massive Document Collection.* IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery, volume 11, number 3, pages 574-585. May 2000

Perotti, V. and Burke, A., (2000) *The Visualization of Usage Patterns for Web Customization*, presented at the IRMA conference, Toronto, CA

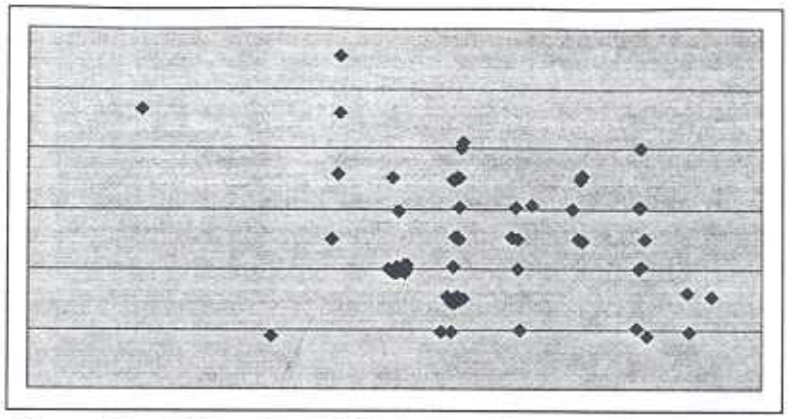

Figure 1: a jittered grid depicting clusters of web pages