ELSEVIER

Decision Support Systems 35 (2003) 245–256

Decision Support Systems

www.elsevier.com/locate/dsw

Web page clustering using a self-organizing map of user navigation patterns

Kate A. Smith*, Alan Ng

School of Business Systems, Monash University, P.O. Box 63B, Victoria 3800, Australia

Abstract

The continuous growth in the size and use of the Internet is creating difficulties in the search for information. A sophisticated method to organize the layout of the information and assist user navigation is therefore particularly important. In this paper, we evaluate the feasibility of using a self-organizing map (SOM) to mine web log data and provide a visual tool to assist user navigation. We have developed LOGSOM, a system that utilizes Kohonen's self-organizing map to organize web pages into a two-dimensional map. The organization of the web pages is based solely on the users' navigation behavior, rather than the content of the web pages. The resulting map not only provides a meaningful navigation tool (for web users) that is easily incorporated with web browsers, but also serves as a visual analysis tool for webmasters to better understand the characteristics and navigation behaviors of web users visiting their pages.

Keywords: Data mining; Self-organizing maps; Clustering; Web usage mining

1. Introduction

One of the most important functions of the Internet is information retrieval. However, resource discovery on the Internet is still frustrating and inefficient when simple keyword searches can convey hundreds of thousands of documents as results [13]. The continuous growth in the size and use of the Internet thus creates difficulties in the search for information. As a result, when a user enters a keyword in a search engine, the returned result is often a large list of web pages, many of which are irrelevant pages, moved pages, abandoned pages, etc. Therefore, a sophisticated method to organize the layout of the information is important, particularly as the Internet grows in size. The purpose of this study is to assist information retrieval on the Internet by applying data mining techniques. Particularly, we focus on web usage mining, applying data mining techniques to web server logs.

The use of data mining in this domain can be seen as the application of a new technology to an acknowledged problem. Navigational aids were devised not long after the Internet became popular, including technologies such as Supercard that inspected the hypertext of where the user had been, and helped the user navigate through the hypertext based on this prior activity [12]. More recently though, researchers have been using the techniques of data mining to assist information retrieval on the Internet.

One of the most well-known data mining approaches is WEBSOM [4,5,8], which is a system using Kohonen's Self-Organizing Map [6,7] to organise web documents into a two-dimensional map,

* Corresponding author.
E-mail address: kate.smith@infotech.monash.edu.au (K.A. Smith).

0167-9236/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-9236(02)00109-4

according to their document content. Documents which are similar in content are located in similar regions on the map. This method is very effective because the system is able to automatically organize the documents into meaningful clusters according to their content. For example, a group of web pages about data mining may appear in one cluster but it is unlikely that there are any pages on industrial mining (for example, oil or coal) in the cluster, because the self-organizing map (SOM) clusters by content rather than keyword. In addition, the location of the representing node indicates the closeness (similarity) of the documents represented.

There are several advantages to using the SOM to cluster documents, rather than people, due to the objectivity of the process. In addition, the process is automatic (hence the name "self-organizing"). It can thus be done on a large scale and therefore saves labour costs. It also facilitates search by concept instead of search by keyword. However, for the system to accurately reflect the needs of users, the organization of the web documents should also take into account the feedback from users. While it is useful to have a system to organize the web pages in a content-driven manner, it may be more advantageous to organize the web pages in a web-user-oriented manner. After all, the web documents are organized so that *humans* can search in a more effective and efficient manner. The system should realise that users who visit data mining web pages may also like to visit web pages about self-organization, for instance. The *usage patterns* of web users can therefore play a role in assisting other users.

Several researchers have applied data mining techniques to web server logs, attempting to unlock the usage patterns of web users hidden in the log files. Fu et al. [2] have demonstrated that web users can be clustered into meaningful groups, which help webmasters to better understand the users and therefore to provide more suitable, customized services. Perkowitz and Etzioni [13] propose adaptive websites that improve themselves by learning from user access patterns. The main mining subject is the web server logs. Other information such as counts for links are also used to emphasize the importance of the web pages. These systems aim to advise the webmaster regarding the changes that can be made to improve the website. Web server logs have also been analyzed to infer users' demographic facts [11], which are useful for improving the effectiveness of Internet advertising. Lan et al. [9] have presented a strategy to make web servers "pushier", that is, reduce the loading time of the web browser. This is done by mining the web logs to find rules of the form "Document1 → Document2", so that the web server can push "Document2" to the browser if "Document1" is requested.

Perhaps the research most relevant to our study is the Web Personalization system created by Mobasher et al. [10], who propose a system that organizes web usage data into clusters. The system analyzes the web server logs. It takes into consideration what kind of user group the current user belongs to, and then suggests some paths that might interest the current user. The suggestions are formed by looking into the past experience of the user group, and ranked by, for example, association rules. However, this approach does not provide sufficient information to enable the user to search by concept. It may help to provide documents within a concept. However, it does not appear to help much to move from one concept to another. Moreover, the system does not really help the user to see the web area from a broad perspective. With the increasingly large volume of data on the web, it is necessary to provide users with a means to navigate with a broad view in mind. In other words, it does not provide an efficient means for the user to search for information; it merely helps by providing some suggestions about which pages to visit next.

It is obvious that both WEBSOM and Web Personalization are intelligent systems that look at different aspects to add value to the Internet. While both systems attempt to organize web data, they take different approaches. While mining the web server logs (the approach of Web Personalization) yields great benefits, the SOM output format (used by WEBSOM) is very useful in conveying the relationships of the web documents in a user-friendly manner. Therefore, it seems logical to combine the benefit of these systems, to add more value to information retrieval on the Internet. This is the core idea of our study, which will be explored in the later sections.

In this paper we present LOGSOM, a prototype system that organizes web pages on a self-organizing map (SOM) [6,7] according to user navigation patterns. Our approach is different from WEBSOM since we cluster the web pages according to the users'

navigation behaviors, rather than according to the web content [4,5,8]. Instead of organizing the web-pages according to the words contained in the web-pages, we keep track of the interest of the web-users, and organize the web-pages according to their interest. In this way, the SOM provided by our LOGSOM can be updated regularly to reflect the current interest of the web-users.

We select the SOM as the most appropriate technique for the problem of organizing web pages to assist user navigation because of its strength of not only grouping the web pages into clusters, but also graphically representing the relationship among the clusters. It is flexible enough to be adopted into a hierarchical clustering approach to present large amount of data points. Its feasibility in web-page-organization based on content has already been proven [4,5,8]. These strengths make the SOM an ideal technique for resolving the problem of web-page-organization from a web user's perspective.

The scope of this study is to test the feasibility of clustering web pages using a SOM based on inputs derived from user navigation patterns. The SOM is given no information about the content of the web pages, only which users have visited each page. We have tested the approach within a limited domain: a university web server mostly accessed by students in a computer laboratory. Within this domain, we can understand the usefulness of any resulting navigational information. Naturally, this paper aims to lay some foundations for expanding the approach into the broader domain of the Internet.

2. Data preparation

We have developed the prototype of our LOGSOM system based on the access logs for September of 1999 from the Monash University, School of Business Systems web server. There are 170,515 entries in the web log indicating the date, time, and address of the requested web pages, as well as the IP address of the user's machine.

There are several pre-processing tasks to be done before data mining algorithms can be performed on the web server logs. These include data cleansing, user identification, session identification, path completion, and formatting. These pre-processing tasks are the same for any web usage mining problem and are discussed by Cooley et al. [1]. The original server logs are formatted, cleansed, and then grouped into meaningful transactions before being mapped onto the self-organizing map.

2.1. Data formatting

As the web users visit the Business Systems website http://www.bs.monash.edu.au—including all of its linked web pages), they leave some footprints behind. Like many other servers, that of Business Systems saves the footprints as web server logs, which we have reformatted as shown in Fig. 1.

2.2. Data cleansing

The access log is saved to keep a record of every request made by the users. Since the log is to be used as input to organize the web pages to facilitate more effective and efficient navigation, we only want to keep the log entries that carry relevant information. Some log entries which are irrelevant to our study are deleted from the log file as follows:

- We used a computer terminal to check the web pages for experimental purpose. Since this type of checking does not represent a normal user's behavior, we delete all the entries with an IP address corresponding to the first author's terminal.
- Sometimes a user requests a page that does not exist. This results in error entries being recorded. Since we are organizing the existing web URLs, we are not interested in these error entries, which we have therefore deleted.
- A user's request to view a particular page often results in several log entries because that page consists of other material such as graphics. We are only interested in, and only keep, what the users explicitly request because it is intended that the system should be user-oriented.

2.3. Transaction identification

Following Cooley et al. [1], we group the data into meaningful transactions. We define a transaction as a set of web pages requested by a user in a particular session.

date	time	c-ip	cs-method	cs-uri-stem	sc-status	sc-bytes	cs(User-Agent)	cs(Referer)
9/2/99	0:01:41	203.24.205.3	GET	/subjects/wbus1021/w1021.html	304	165	Mozilla/4.06+[en]+(Win95;+I+;Nav)	-
9/2/99	0:01:42	203.24.205.3	GET	/subjects/wbus1021/right.html	304	165	Mozilla/4.06+[en]+(Win95;+I+;Nav)	-
9/2/99	0:01:42	203.24.205.3	GET	/subjects/wbus1021/heading.html	304	165	Mozilla/4.06+[en]+(Win95;+I+;Nav)	-
9/2/99	0:01:44	203.24.205.3	GET	/subjects/wbus1021/smallogo.gif	304	165	Mozilla/4.06+[en]+(Win95;+I+;Nav)	http://www.bs.monash.edu.au/subjects/wbus1021/heading.html
9/2/99	0:01:46	203.24.205.3	GET	/subjects/wbus1021/left.html	304	165	Mozilla/4.06+[en]+(Win95;+I+;Nav)	-
9/2/99	0:01:46	203.24.205.3	GET	/subjects/wbus1021/australia.gif	304	165	Mozilla/4.06+[en]+(Win95;+I+;Nav)	http://www.bs.monash.edu.au/subjects/wbus1021/left.html
9/2/99	0:01:48	203.108.0.58	GET	/Default.htm	304	226	Mozilla/4.05+[en]+(Win95;+I)	http://www.monash.edu.au/dept.html#Administration
9/2/99	0:01:53	203.108.0.58	GET	/webpage/image/choosetheme.jpg	304	165	Mozilla/4.05+[en]+(Win95;+I)	http://www.bsys.monash.edu.au/
9/2/99	0:01:55	203.108.0.58	GET	/webpage/image/msobs.jpg	304	165	Mozilla/4.05+[en]+(Win95;+I)	http://www.bsys.monash.edu.au/

Fig. 1. An extract of a web server access log.

2.3.1. User identification

The task of identifying unique users is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. Therefore, some heuristics are commonly used to help identify unique users. However, since the Business Systems site is mostly accessed by students in the computer laboratories without passing through proxy servers, we simply use the machines' IP addresses to identify unique users.

2.3.2. User-session identification

For logs that span a long period of time, it is very likely that different users will use the same machine to access the server websites. Therefore, we differentiate the entries into different user-sessions through a session timeout. If the time between page requests exceeds a certain limit, it is assumed that there is another user-session, even though the IP address is the same. We use a 30-min timeout because it is the one used by many commercial products [1]. While this measure is fairly arbitrary, we find that 30 min enables us to find a balance between ensuring that the transactions are attributed to the correct user and generating enough web page accesses in one transaction set.

After the entries are grouped according to the user-sessions, the data is converted into a format as shown in Table 1.

We assume that there is a set of n unique URLs appearing in the pre-processed log:

$$U = \{\mathrm{url}_1, \mathrm{url}_2, \ldots, \mathrm{url}_n\}$$

and a set of m user transactions:

$$T = \{t_1, t_2, \ldots, t_m\}$$

We represent the transactions as a bit vector

$$\overrightarrow{t} = \langle u_1^t, u_2^t, \ldots, u_n^t \rangle$$

where

$$u_i^t = \begin{cases} 1, & \text{if } \mathrm{url}_i \in t \\ 0, & \text{otherwise} \end{cases}$$

For our September web log, the number of transactions $m = 8054$ and the number of URLs $n = 235$.

Table 1
Format of data grouped by transactions

	URL 1	URL 2	URL 3	...	URL 235
Txn 1	0	1	0	...	0
Txn 2	1	0	0	...	1
...	...	...	...	...	...
Txn 8054	1	0	1	...	1

3. Dimension reduction

3.1. Clustering transactions into user groups

Even after the above pre-processing procedure, the data is still not ready for effective application of the SOM technique, because the number of transactions is very large (8054 for the September log). The number of inputs of the SOM will need to be equivalent to the number of transactions, and because this number is so large, it will not be feasible with this data. The size of the data set not only consumes large amounts of processing time, but also limits the applicability of the system to data in the real world (for example, web logs collected by a search engine in a month).

To solve this problem, Ref. [10] combines hypergraph partitioning and association rules, to cluster the web documents. However, we cannot use the same technique because we are trying to achieve more than just clustering of documents: we want to map the web documents into a two-dimensional space, where the locations will indicate the similarity between documents, as indicated by the navigation patterns. The pages are deemed similar if they are accessed by similar kinds of people. That is, rather than deeming two pages to be similar because the web-master has identified their relationship based on content or keywords, we are employing a data-driven approach where the similarity of two documents or web pages depends on them being accessed by the same users. If the same collection of

The *K*-means algorithm used in this study comprises the following four steps:

1. Choose K initial cluster centres (representing the K transaction groups) randomly from the centre of the hypercube.

2. Assign all data points (representing the transactions) to their closest cluster (measuring from the cluster centre). This is done by presenting a data point $\boldsymbol{x}$ and calculate the similarity (distance) d of this input to the weights w of each cluster centre j. The closest cluster centre to a data point x is the cluster centre with minimum distance to the data point x.

$$d_j = \| x - w_j \| = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})}$$

3. Recalculate the centre of each cluster as the centroid of all the data points in each cluster. The centroid $\vec{c}$ is calculated as follows:

$$\vec{c} = < w_1^c, w_2^c, \ldots, w_n^c >$$

where

$$w_i^c = \frac{\sum_{j \in c} u_i^j}{N^c}$$

where: N^c is the number of data points in the cluster

$$u_i^t = \begin{cases} 1, & \text{if } \text{url}_i \in t \\ 0, & \text{otherwise} \end{cases}$$

4. If the new centres are different from the previous ones, repeat Steps 2, 3 and 4. Otherwise terminate the algorithm.

Fig. 2. The *K*-means algorithm.

web pages is repeatedly accessed by a group of users, then these pages are defined as being similar (at the very least, they are similar in terms of the interest level of these users at that time).

Due to our focus on user navigation patterns, we have devised another simple yet effective approach to reducing the dimensionality of the problem. By using the *K*-means clustering algorithm [3], we cluster the transactions into nine groups. The number $K=9$ is chosen arbitrarily. In fact, we can choose any number as long as it is small enough for the data to be feasible for SOM processing, and large enough for the data to carry sufficient information to produce a meaningful outcome. In our experiment, each 235-dimensional binary transaction vector is treated as an input vector and clustered into $K=9$ groups. These are essentially similarity groups, that is, collections of transactions that involve similar web page access patterns. The *K*-means algorithm is outlined in Fig. 2.

Within each processed transaction group, we use the column totals as the activity of the transaction group. This is illustrated in Table 2, where the cluster depicted contains 2679 transactions, including transaction numbers 5, 6, and 8012. Instead of having transactions as features, we now have transaction groups as features. The URLs are now described by the relative interests or activities of each of the transaction groups.

Given a transaction group $g \in G$, we can represent the transaction group as a vector:

$$\overrightarrow{g} = \langle w_1^g, w_2^g, \ldots, w_n^g \rangle$$

where

$$w_i^g = \sum_{j \in g} u_i^j$$

After the transactions have been grouped, the number of dimensions has been reduced. The data is now ready to be used for SOM processing. Since the URLs, instead of the transactions, are the subject to be organized, we transpose the data so that each URL is represented as a vector of transaction group activity. This is illustrated by Table 3.

Table 2
Column totals become the vector of activity representing a transaction group

	URL 1	URL 2	URL 3	...	URL 235
Txn 5	1	0	0	...	0
Txn 6	1	0	0	...	1
...	...	...	...	...	...
Txn 8012	0	0	1	...	0
Transaction group 1 (activity)	101	0	3	...	12

Table 3
Representing URLs as vectors of transaction group activity

	Transaction group				
	1	2	3	...	9
URL 1	101	0	0	...	2091
URL 2	0	0	0	...	0
...	...	...	...	...	...
URL 235	1	0	0	...	23

We have successfully reduced the number of dimensions of the data. Before the dimension reduction, our data consisted of 8054 transactions × 235 URLs. After the dimension reduction, it consists of 235 URLs × 9 transaction-groups. The same principle can be applied to web logs collected by search engines to make the data feasible for SOM processing.

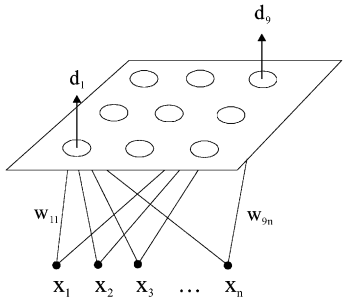
4. Self-organisation of usage patterns

A Self-Organizing Map [6,7], or SOM, is a neural clustering technique. It is more sophisticated than *K*-means in terms of presentation: it not only clusters the data points into groups, but also presents the relationship between the clusters in a two-dimensional space. SOM is also capable of presenting the data points in one- or three-dimensional space, however, two-dimensional space is most commonly used due to the trade-off between information content and ease of visualisation. The SOM concept is outlined in Fig. 3 and the algorithm is presented in Fig. 4.

This study uses Kohonen's Self-Organizing Map [6,7] to organize web pages into a two-dimensional map, according to users' navigation patterns. By doing so, we aim to provide an interactive tool so they can retrieve information more effectively and efficiently. Our inputs consist of a set of 235 URLs (defined by the usage pattern of transaction groups). The desired output is a two-dimensional map of *M* nodes (in this case, a 16 × 16 map of 256 nodes).

The users' navigation patterns are indicated by inputs of URLs showing their usage within nine

The input vectors are connected to an array of neurons (usually 1 dimensional (a row) or 2 dimensional (a rectangular lattice):



When an input is presented, certain regions of the array will "fire", and the weights connecting the inputs to that region will be strengthened

During learning,

- the weights connecting the input space to the winning neuron are strengthened
- the weights of neurons in the "neighbourhood" of the winning neuron are also strengthened (although not as much).

Once learning is complete, "similar" inputs will "fire" the same regions

In this way, similar input patterns can be identified and grouped together or clustered

Fig. 3. The basic concept of SOM.

different transaction groups. The architecture is a SOM consisting 16×16 nodes (a rectangular map). The parameters experimented with are: α={0.1, 0.5, 0.9}; Ω={1, 10, 20, 30, 40, 50}. α represents the learning rate, and Ω determines the number of times each URL is presented within one learning cycle before the neighbourhood size is decreased. In our algorithm, there are 16 cycles of learning to organize the web pages, decreasing the neighbourhood size from its initial value of 16 to 0.

5. Results

In this section we present the results of our SOM approach to web usage mining. The resulting map generated by analysing the data is presented, and its robustness to parameter section is demonstrated.

5.1. A typical SOM by LOGSOM

Fig. 5 shows a typical SOM produced by our LOGSOM system. The map contains 235 URLs. The blank nodes contain no URLs, while those that are numbered indicate the number of URLs contained within each node. For example, the node at the middle of the first row is numbered 17 because it contains 17 URLs.

The distance between nodes on the map indicates the similarity of the web pages, measured according to the user navigation patterns. For example, the 39 web pages contained in the node numbered 39 (bottom right corner of Fig. 5) are grouped in the same cluster because they have been accessed by people with similar interests as indicated by their transaction patterns. The node numbered 2, above the node numbered 39, contains 2 web pages similar to the 39 contained in the

The proposed SOM algorithm for LOGSOM is based on the conventional SOM algorithm developed by Kohonen [3, 9]. An outline of a revised SOM algorithm for our LOGSOM system is summarized below:

1 Initialise:

- weights w_{ij} to small random values. (Seeded for reproduction in all experiments)
- neighbourhood size $N_m(0)$ to be large (but less than the number of nodes in one dimension of the array), in this case 15.
- parameter function $\alpha(t)$ and $\sigma^2(t)$ to be between 0 and 1, in this case, $\alpha(0)$=0.5 and $\sigma^2(0)$=1. In our experiments, we also tried $\alpha(0)$=0.1 and 0.9.

2 Present an input pattern x through the input layer and calculate the similarity (distance) d of this input to the weights w of each node j.

$$d_j = \| x - w_j \| = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})^2}$$

3 Select the node with minimum distance as the winner m.

4 Update the weights connecting the input layer to the winning node and its neighbouring nodes according to the learning rule:

$$w_{ij}(t+1) = w_{ij}(t) + c[x_i - w_{ij}(t)]$$

where $c = \alpha(t)\exp(-\| r_i - r_m \| / \sigma^2(t))$ for all nodes j in $N_m(t)$

where $r_i - r_m$ is the physical distance (number of nodes) between node i and the winning node m.

5 Continue from step 2 for Ω epochs (in this case, 40 epochs); increase t by 1, then decrease the neighbourhood size, $\alpha(t)$ and $\sigma^2(t)$ such as

$$\alpha(t) = \alpha(0)N_m(t)/N_m(0)$$

Repeat until the weights have stabilized. (We also tried Ω=1, 10, 20, 30 and 50.)

6 After the network is trained through repeated presentations of all URLs (each URL is presented for Ω epochs), present unit input vectors of every URL to the trained network and assign the winning node the URL address. Update the number labelling the node as the number of URLs allocated to the node.

Fig. 4. SOM algorithm.

node numbered 39. Similarity here is measured not by the similarity of the content, but by the similarity of usage.

It is difficult to quantitatively measure the effectiveness of the SOM, and comparison with other clustering techniques will not address this issue. Clearly the SOM has generated clear clusters, but until we inspect these clusters we will not be able to ascertain if the approach is useful. We need to convince ourselves that the web pages that are in the same cluster or node are indeed similar according to user navigation patterns, despite the fact that the SOM has been provided with no information about the content of these web pages.

To confirm what we have said about the similarity of web pages within a node, we may look into the web pages inside the node numbered 39. This can be done by

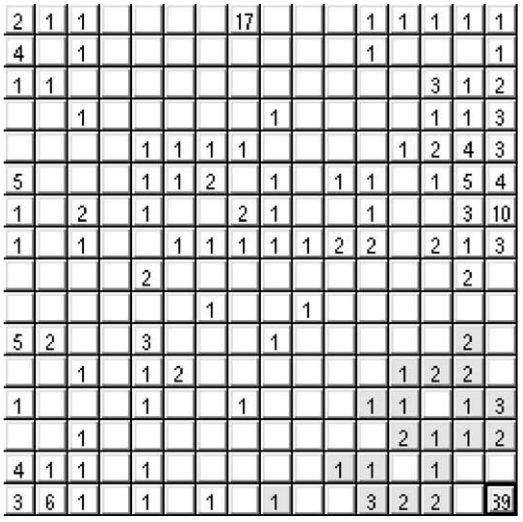


Fig. 5. A typical SOM, with highlighted nodes containing web pages with URL addresses in the form "../wbusx502/...".

clicking the node, which opens a list box showing the title content of the node, as indicated by Fig. 6a. This is a list of the titles of the web pages contained in the node. At the same time, another list box shows the URL addresses of the web pages, as indicated by Fig. 6b.

The list boxes in Fig. 6 are useful for measuring the effectiveness of our system. All the web pages in the selected node have been kept in the "../wbusx502/..." directory by the web author, which indicates that they are treated as web pages in the same group in the eyes of the web author. It is important to note that neither the content of the web pages nor the directory structure is used as inputs to the processing part of our system. The organisation of the system is therefore based solely on the user navigation pattern, as summarised in Table 3. The URL addresses and the titles of the web pages are therefore only used as labels for identifying the web pages. Thus, we have illustrated that when the web pages are organised by user navigation patterns, the resulting SOM is very meaningful, in terms of organising the web pages into clusters based on similarity of usage. Within this familiar domain, we can see that students are indeed likely to have been navigating between web pages within this node, even though the SOM was given no information about the directory structure of the server or the content of the web pages. It has placed these web pages together simply because they are commonly accessed by students in the same transactions.

To investigate the effectiveness of organising the clusters on the two-dimensional map, so that the location reflects the relationship between the clusters, we look at the 'neighbours' of the node numbered 39. The node numbered 2, above the node numbered 39, contains the web page titled "Module 0" and "Laboratory 4" with the URL address "subjects/wbusx502/Plab03/Plab03/Module00.htm" and "subjects/wbusx502/Plab04/ default.htm". The node numbered 2 at the left of the node numbered 39 contains the web page titled "Assessment" and "BUSx502 Lecturer" with the URL address "subjects/wbusx502/assessment.htm" and "subjects/wbusx502/lecturer.htm". These nodes are close to the node numbered 39 and contain web pages related to those contained in the node numbered 39. In fact, all the shaded nodes in Fig. 5 contained web pages with the URL address in the form "../wbusx502/..." despite the SOM not being told of this directory structure.

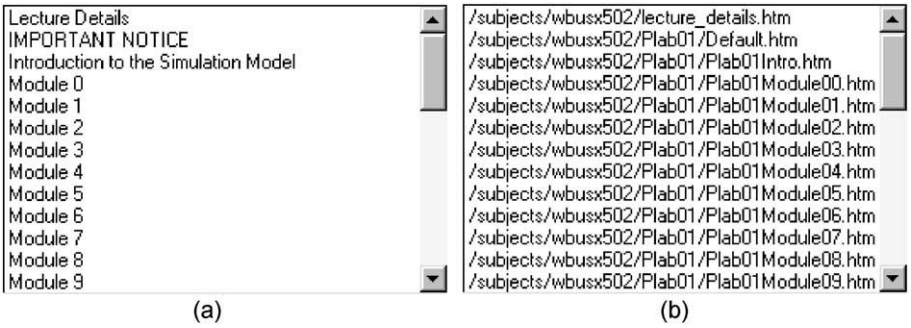


Fig. 6. Inside the current active node: (a) the title contents and (b) the URL addresses.

Therefore, we have illustrated that when the web pages are organised by user navigation patterns, the resulting SOM can be quite meaningful in terms of representing the relationship among the clusters of web pages on a two-dimensional SOM. The system is readily incorporated within a web browser to assist user navigation.

5.2. Effect of changing parameters

Because two important parameters of the self-organising map are α and Ω, this section aims to investigate the effect of changing these parameters on the quality of the SOM produced by our LOGSOM system. α represents the amount of change of weights of the nodes in the process of learning. Ω determines the number of repetitions of the process of learning within one cycle (or the number of times each URL is presented within one cycle before the neighbourhood size is decreased). In our algorithm, there are 16 cycles of learning required to organise the web pages.

We have run 18 experiments for this purpose. Some of the resulting maps are shown in Appendix A. We have changed α from 0.1 to 0.5 and 0.9, while changing Ω from 1 to 10, 20, 30, 40, and 50, respectively. The quality of the resulting maps is measured by the clarity of the resulting clusters, as well as inspection of the web pages within each cluster to ascertain the usefulness given the domain knowledge we have about the data.

We found that while there are some obvious differences in the quality of maps for different α values when the Ω value is fixed at 1, the quality becomes similar when Ω is increased to 40 or above. In other words, at the initial stage of learning, the effect of the amount of learning does make a difference to the quality of the resulting map. However, as we repeat the number of learning cycles, the difference in quality will be reduced. When Ω is at 40, the maps seem fairly stabilised. Further increase in the value of Ω does not seem to significantly increase the quality of the maps. This suggests that the resulting SOM is quite robust and insensitive to narrow parameter selection. Our earlier example is based on the map with $\alpha = 0.9$ and $\alpha = 40$.

Note that at $\Omega = 40$ or above, although the groups of web pages may be located at different corners of the maps, the quality of the maps is similar. For example, in the map $\alpha = 0.1$, $\Omega = 40$, the "../wbusx502/..." pages are scattered around the left top corner; in the map $\alpha = 0.9$, $\Omega = 40$, the "../wbusx502/..." pages are scattered around the lower right corner. The similar pages are still grouped together although they may be located at different corners of the maps.

6. Conclusion

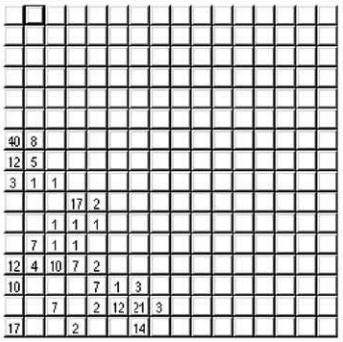
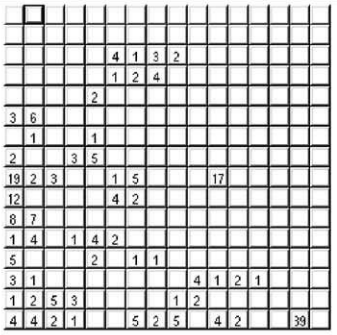
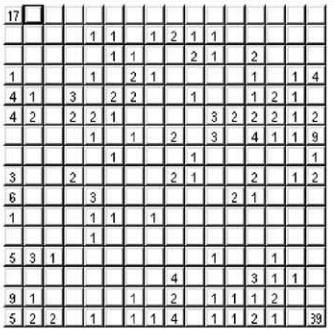
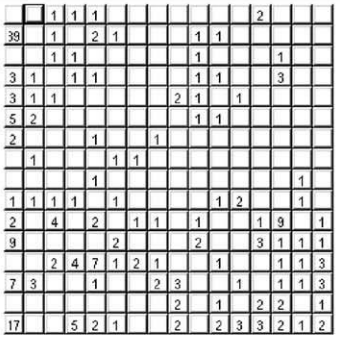
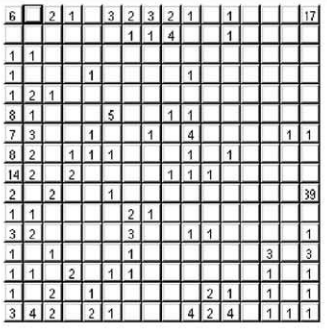
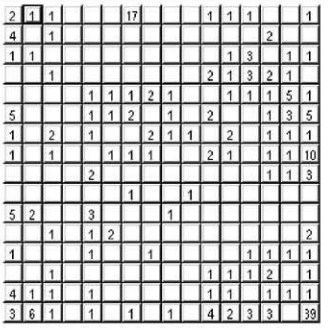
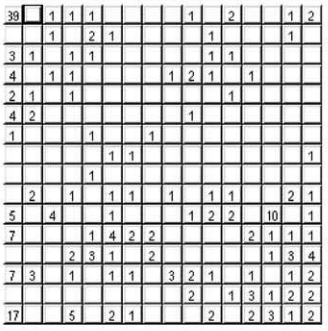
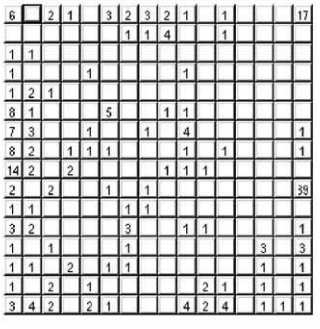
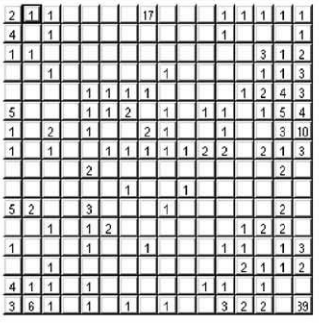
We have presented LOGSOM, a system that utilizes Kohonen's self-organizing map to organize web documents in a domain onto a two-dimensional map. The organization of the web documents is based solely on the users' navigation behavior. It has been demonstrated that the resulting map of this system is very meaningful and can be easily incorporated with a web browser to assist user navigation. LOGSOM provides a visual tool to enable users to see the relationship between web pages based on the usage patterns of web users similar to themselves. LOGSOM also provides an analysis tool for web masters and web authors to better understand the interests of visitors to their pages, and identify potential referring pages.

The limitation of the current prototype LOGSOM system is that it has only been evaluated on a sample of data: the log files for 1 month of access to the School of Business Systems web server at Monash University. This sample has demonstrated that the prototype is effective, and raised issues about pre-processing and dimension reduction needed to tackle larger data sets. Certainly the prototype has been a useful tool for internal purposes, and it is likely to be a useful approach to assist other organisations in better understanding the interests of their web visitors and assisting user navigation on their servers. The bigger challenge, however, lies in the size of the web. Instead of using LOGSOM on a particular server, the next challenge will be to use it on a larger portion of the Internet. It is possible to combine LOGSOM with a search engine so as to organize the data on the Internet. If our LOGSOM system is to be applied to a larger portion of the Internet, the large amount of data would increase the processing time. Therefore, further pre-processing techniques may prove essential. The value of K in the K-means pre-processing will need to be carefully selected to ensure that the interests of all users are represented.

The aim of this study has been to demonstrate the feasibility of the approach within a controlled domain,

and lay the foundations for future research in this area. Our future research will be investigating the scalability of the system, so that it can be incorporated with search engines. We will also investigate the benefits of combining content information with usage patterns, and consider the use of temporal information to enable the SOM to adapt over time to changing user navigation patterns.

Appendix A. SOMs for a sample of experiments conducted with varying Ω and α

$\Omega\backslash\alpha$	0.1	0.5	0.9
1			
20			
40			

References

[1] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining World Wide Web browsing patterns, Journal of Knowledge and Information Systems 1 (1) (1999) 5–32.

[2] Y. Fu, K. Sandhu, M. Shi, Clustering of web users based on access patterns, Lecture Notes in Artificial Intelligence, vol. 1836, Springer-Verlag, Berlin, 2000, pp. 21–38.

[3] J.A. Hartigan, Clustering Algorithms, Wiley, New York, 1975.

[4] K. Lagus, T. Honkela, S. Kaski, T. Kohonen, Self-organizing maps of document collections: a new approach to interactive exploration, Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996, pp. 238–243.

[5] S. Kaski, T. Honkela, K. Lagus, T. Kohonen, WEBSOM—self-organizing maps of document collections, Neurocomputing 21 (1–3) (Oct. 1998) 101–117.

[6] T. Kohonen, Construction of similarity diagrams for phonemes by a self-organizing algorithm, Technical Report TKK-F-A463, Helsinki University of Technology, Espoo, Finland (1981).

[7] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43 (1982) 59–69.

[8] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive document collection, IEEE Transactions on Neural Networks 11 (3) (May 2000) 574–585.

[9] B. Lan, S. Bressan, B.C. Ooi, Making web servers pushier, Lecture Notes in Artificial Intelligence, vol. 1836, Springer-Verlag, Berlin, 2000, pp. 112–125.

[10] B. Mobasher, R. Cooley, J. Srivastava, Automatic personalization based on web usage mining, Technical Report, TR99-010, Department of Computer Science, Depaul University (1999).

[11] D. Murray, K. Durrell, Inferring demographic attributes of anonymous internet users, Lecture Notes in Artificial Intelligence, vol. 1836, Springer-Verlag, Berlin, 2000, pp. 7–20.

[12] J. Nielsen, The art of navigating in hypertext, Communications of the ACM 33 (3) (1990) 296–310.

[13] M. Perkowitz, O. Etzioni, Adaptive sites: automatically on line at www.scope.gmd.de/info/www6/posters/722/index.html learning from user access patterns, Proceedings of WWW6, 1997.

Alan Ng graduated from Monash University, Australia with a Master of Business Systems (2000) and Bachelor of Commerce—Accounting and Finance (1998). His Masters thesis examined data mining techniques for analysing web usage information. In 2001, he joined Singapore Telecommunications Ltd and was appointed as a Systems Analyst. His research interests include data mining, neural networks, database administration, and the application of these skills on the Internet.

Kate A. Smith is an Associate Professor in the School of Business Systems at Monash University, Australia, where she is also Deputy Head and Director of Research. She holds a BSc(Hons) in Mathematics and a PhD in Electrical Engineering, both from the University of Melbourne, Australia. She is also Director of the Data Mining Research Group in the Faculty of Information Technology at Monash University. Dr. Smith has published two books on neural networks in business, and over 80 journal and international conference papers in the areas of neural networks, combinatorial optimization, and data mining. She is a member of the organizing committee for several international data mining and neural network conferences, and regularly acts as a consultant to industry in these areas.