

Exploring Data Generated by Computer Forensic Tools with Self-Organising Maps

B.K.L. Fei, J.H.P. Eloff, H.S. Venter and M.S. Olivier

Information and Computer Security Architectures Research (ICSA) Group

Department of Computer Science

University of Pretoria, Pretoria, 0002, South Africa

Email: benniefei@yahoo.com

Abstract

Computer forensic tools have been developed to assist computer forensic investigators in conducting a proper investigation into computer crimes. In general, the majority of the tools available on the market have the ability to permit investigators to analyse data that was gathered from a computer system. Since storage media are steadily growing in size, the process of analysing large volumes of data consumes an enormous amount of time. Yet, the data on the storage media may contain implicit knowledge that could improve the quality of decisions in a computer investigation.

The focus of this paper is to demonstrate how an unsupervised learning neural network model, the self-organising map (SOM), can aid computer forensic investigators in decision making and assist them in conducting the analysis process more efficiently during a computer investigation. The SOM can be used to search for patterns in data sets and produce visual displays of the similarities in the data. The paper will aim to explore how the SOM can be used to serve as a basis for further analysis. It will demonstrate how the easy visualisation of the SOM provides investigators with greater abilities to interpret and explore the data generated by computer forensic tools.

Keywords

Self-organising map, visualisation, correlations, patterns, computer forensics

1. INTRODUCTION

The dramatic increase in crime relating to the Internet and computers has caused a growing need for computer forensics. Computer forensics identifies evidence when computers are used in the perpetration of crimes [1]. It involves the use of sophisticated technological tools to ensure that the digital evidence is correctly preserved and that the accuracy of results regarding the processing of digital evidence is maintained.

In general, computer forensic tools exist in the form of computer software [1]. Such tools have been developed to assist computer forensic investigators in a computer investigation. However, because storage media are growing in size, investigators may have difficulty in locating their points of interest from a large pool of data. In addition, the format in which the data is presented may result in disinforming the investigators. As a result, the process of analysing large volumes of data may consume a very large amount of time.

Having an overview of the entire data set obtained directly from a hard drive can be crucial in a computer investigation. It reveals the overall patterns of the data set, which guide computer forensic investigators to the next step in their search. In addition, it can aid investigators to locate their points of interest. Data generated by computer forensic tools may be meaningless at times, due to the amount of data that can be stored on a storage medium and the fact that current computer forensic tools are not able to present a visual overview of all the objects (e.g. files) found on the storage medium.

This paper explores how the self-organising map (SOM) [2], an unsupervised learning neural network model, can be used not only to reveal interesting patterns, but also to serve as a basis for further analysis. It will also introduce the main advantage of the SOM – the graphical representation of large sets of data – which is evident from the easy visualisation and interpretation of clusters formed by the map. By presenting data in a visual and graphic manner, the SOM offers investigators a fresh perspective from which to study the data.

The rest of the paper is structured as follows: Section 2 provides some background information on computer forensics. Section 3 gives a brief overview of the SOM. Section 4 contains a SOM application that demonstrates how the SOM can be used in the field of computer forensics. Section 5 concludes the paper with an outlook on future work.

2. BACKGROUND

The concept of computer forensics has been around for a while – actually, since the invention of the computer. Years ago the evidence collected was primarily paper based. Today, the majority of evidence for some crimes resides on a computer. Digital evidence is somewhat unique when compared with different forms of documentary evidence. For example, it may be found in unusual locations that are generally unknown to the computer users, while it is also very mutable and vulnerable to alteration.

Computer forensics deals with the preservation, identification, extraction and documentation of digital evidence [1]. Child pornography, threatening letters, fraud and theft of intellectual property are all crimes that leave digital tracks [3]. The unique needs of computer forensics have resulted in the creation of computer forensic tools in the form of computer software designed to either collect or analyse that data in order to uncover information that is crucial in a computer investigation.

Currently, there are numerous computer forensic tools available on the market. For example, EnCase [4], Forensic Toolkit [5] and ProDiscover [6] are some of the tools that are available. They differ from one another in the sense that some computer forensic tools are designed with only a single purpose in mind, whereas others may offer a whole range of functionalities. Examples of these functionalities are advanced searching capabilities, hashing verification, report generation and many more. Some computer forensic tools do provide similar functionalities, but with a different graphical user interface.

A typical computer investigation involves the making of an exact copy of all the data on a storage medium. Such storage medium may be a hard drive, compact disk, floppy disk or flash disk. The copy is called an image and therefore the process of making an image is frequently referred to as imaging. In most cases, once the imaging process has been completed, it is essential to have a mechanism or procedure to determine whether the evidence has been altered. This is to ensure the integrity [7] of the evidence. After the data has been successfully gathered, it must be analysed to extract the evidence that the investigators would wish to present. Analysis of digital evidence involves a mixture of techniques. For example, one technique is to perform keyword searches [8] on digital evidence. Other techniques can involve signature analysis [9] or hash value analysis [9].

The field of computer forensics has advanced from carrying out analysis within a command-line environment such as DOS, to a graphical environment such as Windows. Due to this transition, computer forensic tools can now assist computer forensic investigators in carrying out their task more efficiently in a computer investigation. A useful feature that several computer forensic tools offer is the ability to display all the files found in a spreadsheet-style format. This ability allows users to view all the files on a particular storage medium as well as the information

regarding each file. These details may include the name of the file, the date the file was created, the logical size of the file and other information about the file. However, when working with a large data set, the process of scrolling through many rows of data can be tedious. The user may also have difficulty when trying to locate his/her points of interest.

The next section will give a brief overview of the self-organising map (SOM) – an approach that allows computer investigators to view (visualise) all the files on the storage medium and assists them in locating their points of interest quickly by greatly reducing overall human investigation time and effort.

3. THE SELF-ORGANISING MAP (SOM)

The self-organising map (SOM) [2] is a neural network model that has been successfully applied in the clustering and visualisation of high-dimensional data. It is used to map high-dimensional data onto a low-dimensional space that is usually two-dimensional. The SOM is based on unsupervised competitive learning, which means that the learning process is entirely data driven and that neurons (or nodes) in the output layer compete with one another. The SOM consists of two layers of neurons namely the input layer and the output layer (see Figure 1). The input layer is fully connected with nodes at the output layer and each neuron in the input layer represents an input signal. The output layer generally forms a two-dimensional grid of neurons where each neuron represents a node of the final structure. The connections between the layers are represented by weights whose values represent the strength of the connection.

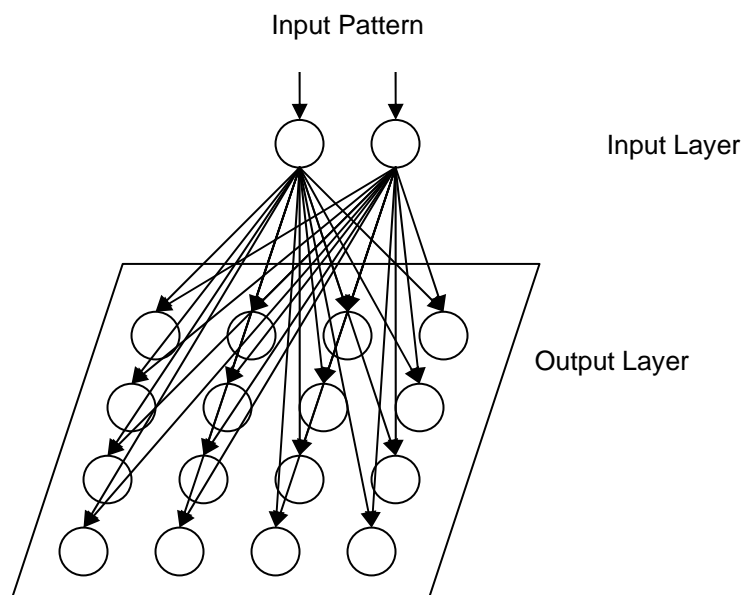


Figure 1. The self-organising map

During the learning process, when an input pattern is presented to the input layer, the neurons in the output layer will compete with one another. The winning neuron will be the one whose weights are the closest to the input pattern in terms of Euclidian distance [11]. Once the winning neuron has been determined, the weights of the winning neuron and its neighbourhood will be updated, i.e. shifted in the direction of the input pattern. After the learning process, the SOM configures the output neurons into a topological representation of the original data, by means of a process called self-organisation [10].

The effect of the learning process is to cluster together similar patterns while preserving the topology of input space. However, in order to visualise the different clusters, an additional step is required to determine the cluster boundaries. Once the cluster boundaries have been determined, the SOM can be referred to as a cluster map. The size of the clusters is the number of nodes allocated to each cluster. One way to determine and visualise the cluster boundaries is to calculate the unified distance matrix (U-matrix) [11]. The U-matrix is a representation of the SOM that visualises the distances between neurons. Large values within the U-matrix indicate the position of cluster boundaries.

The SOM is useful in inspecting the possible correlations between dimensions in the input data [12]. This can be achieved through the visualisation of component maps. Each component map visualises the spread of values of a particular component (or dimension). By comparing component maps with one another, possible correlations are revealed.

4. APPLYING THE SOM TO COMPUTER FORENSIC DATA

This section demonstrates, by means of a SOM application, how the SOM can be used to serve as a basis for further analysis. The SOM application employs an unsupervised neural network that uses the provided data to train, based on the concept of the SOM. Two-dimensional maps, i.e. the cluster map and the different component maps are displayed in the form of hexagonal grids where each hexagonal grid can be referred to as a unit. (Please note that although the SOM application has been demonstrated, it is not the aim of this paper to discuss such implementation.)

The requirements of computer investigations will differ from one another, depending on the unique nature of the investigation. For example, in the case of child pornography the investigation involves the extracting of all graphical images on the suspect's computer system. In most cases, data presented by computer forensic tools still requires users to examine the presented data and draw conclusions.

File Name	Ext	File Type	Category	Cr Date	Acc Date	L-Size
Windows Media Player.lnk	lnk	Shortcut File	Other	2004/08/23...	2004/10/21...	804
Windows Marketplace.url	url	Unknown File...	Unknown	2004/08/24...	2004/08/26...	169
Windows Catalog.lnk	lnk	Shortcut File	Other	2004/08/23...	2004/10/21...	398
Winamp.lnk	lnk	Shortcut File	Other	2004/08/24...	2004/10/21...	672
whyv64p2p.ppt	ppt	PowerPoint 9...	Graphic	2004/09/25...	2004/09/25...	733,184
whver.js	js	Unknown File...	Unknown	2004/05/04...	2004/05/04...	1,136
whutils.js	js	Unknown File...	Unknown	2004/05/04...	2004/05/04...	9,495
whtopic.js	js	Unknown File...	Unknown	2004/05/04...	2004/05/04...	14,769
whicorn2.gif	gif	GIF File	Graphic	2004/03/30...	2004/03/30...	815
whicorn1.gif	gif	GIF File	Graphic	2004/03/30...	2004/03/30...	56
whproxy.js	js	Unknown File...	Unknown	2004/05/04...	2004/05/04...	1,306
whmsg.js	js	Unknown File...	Unknown	2004/05/04...	2004/05/04...	1,666
whatisit-01.gif	gif	GIF File	Graphic	2004/08/23...	2004/08/23...	430
What's New.lnk	lnk	Shortcut File	Other	2004/08/24...	2004/10/21...	652
WG11-9CallForPapers[1].htm	htm	Hypertext Do...	Document	2004/10/15...	2004/10/15...	20,345
wecerr.txt	txt	Plain Text Do...	Document	2004/09/17...	2004/09/17...	105
webex_player.exe	exe	Executable File	Executable	2004/09/06...	2004/09/06...	2,256,784
web_searching_long.gif	gif	GIF File	Graphic	2004/04/23...	2004/04/23...	470
Web Application Forensics...	pdf	Acrobat Porta...	Document	2004/06/09...	2004/06/09...	741,670
wcuevent71.txt	txt	Plain Text Do...	Document	2004/08/24...	2004/08/24...	60,756

Figure 2. The table view of Forensic Toolkit presents numerous fields

An example of what a computer forensic tool presents to the user is shown in Figure 2 above, namely a display of all the files found, in a spreadsheet-style format. This allows users to view all the files on the storage medium and see the details of each file. As mentioned earlier, when working with a large data set, the process of scrolling through many rows of data can be tedious. By applying the SOM, the data set can be mapped onto a two-dimensional space. For example, in Figure 3 below the data depicted in Table 1 (also below) is mapped onto a two-dimensional map (or grid of neurons). Similar input patterns will be clustered together through the learning process of the SOM. The final structure of the map is shown in Figure 3. The data depicted in Table 1 is sorted in ascending order according to the date and time at which the graphical images were created. Therefore, it is expected that the final structure of the map would have the first input pattern positioned at the top right-hand corner and the last input pattern positioned at the bottom left-hand corner as indicated in Figure 3. This is because the time value between the first input pattern and the last input pattern is furthest apart from each other compared to the other input patterns.

Table 1. Details of 9 graphical images

Input Pattern	File Name	Extension	Date and Time Created
1	arrow[1].gif	gif	2004/10/20 09:12
2	auto[1].gif	gif	2004/10/20 09:19
3	bounceout_smalllogo[1].gif	gif	2004/10/20 10:16
4	blko1[1].gif	gif	2004/10/20 10:18
5	Books_03[1].gif	gif	2004/10/20 12:54
6	BRWNSHHHsmall[1].jpg	jpg	2004/10/20 14:16
7	bevel[1].gif	gif	2004/10/20 14:48
8	bg-quicklinks[1].gif	gif </td <td>2004/10/20 14:48</td>	2004/10/20 14:48
9	bishop1[1].jpg	jpg	2004/10/20 15:26

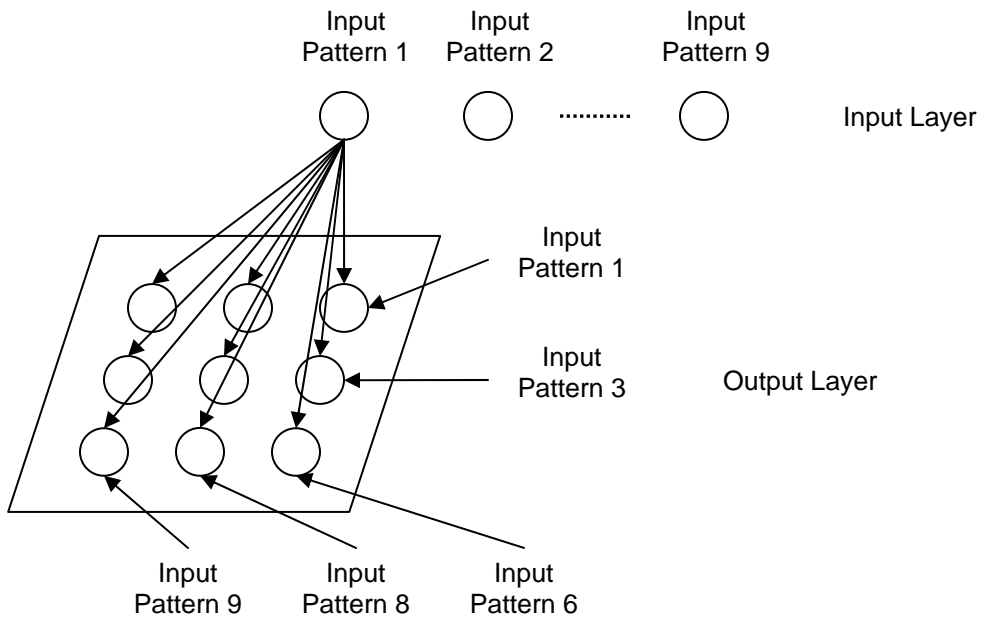


Figure 3. Illustration of the SOM mapping the data onto a two-dimensional grid of neurons

4.1. Dealing with child pornography

The focus of the demonstration in this section will be on the temporary Internet files found on a particular hard drive. Temporary Internet files are those files that are “image captures” of every site that the user visits when accessing the Internet [1]. Looking at temporary Internet files can be very useful when gathering evidence of too much Internet access or inappropriate Internet access.

The demonstration will be based on a scenario dealing with child pornography. In this scenario, it would be the task of the investigator to look for all the graphical images, discover possible patterns, and study the behaviour of the suspect while browsing the Internet. An experimental data set that contains data of all the graphical images – in this case a total of 2 640 – is used. These graphical images are found in the suspect’s temporary Internet files folder, i.e. where all temporary Internet files are stored. Even if the file extensions have been modified, it will not be an issue since the computer forensic tool will be able to detect the correct format of each file. The data set generated by Forensic Toolkit [5] contains the following fields:

- The name of the file. This field will not be used by the SOM application, but only to identify the file.
- The extension of the file.
- The time at which the file was created.
- The date on which the file was created.

It should be noted that the original (or input) data set contains strings that cannot be processed by the SOM application. These strings were consequently converted to numerical values. For example, each file extension was replaced by a numerical value as depicted in Table 2. The date and time were also converted to the format “yyyymmdd” and “hhmm” respectively.

Table 2. Numerical values for the file extension

File extension	Numerical value
bmp	1
gif	2
jpg	3
png	4

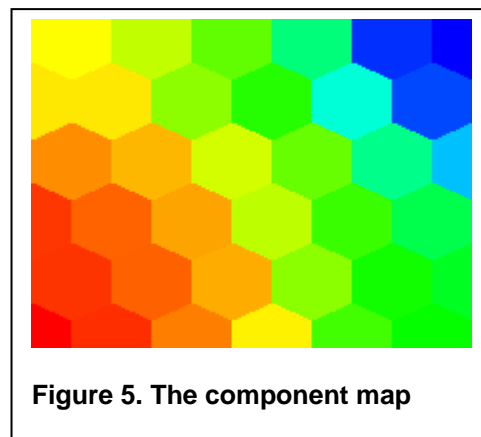
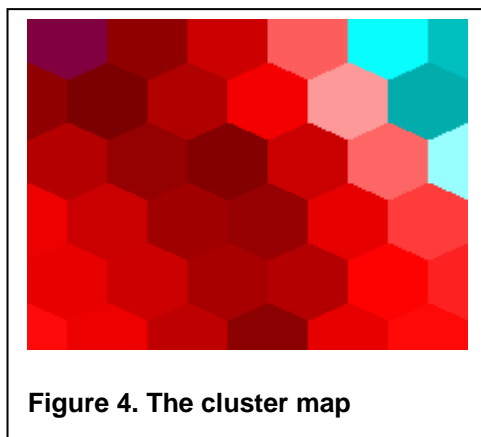
Once the data transformation has been completed, the next step is to process the data set by using the SOM application. After the training phase, the cluster map and the component maps that are subsequently generated can be used as an important visualisation aid as they give a complete picture of the data.

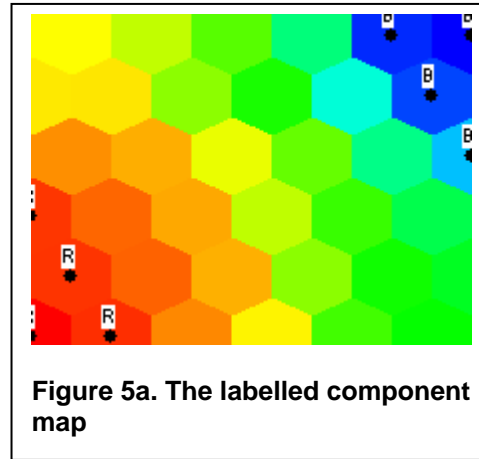
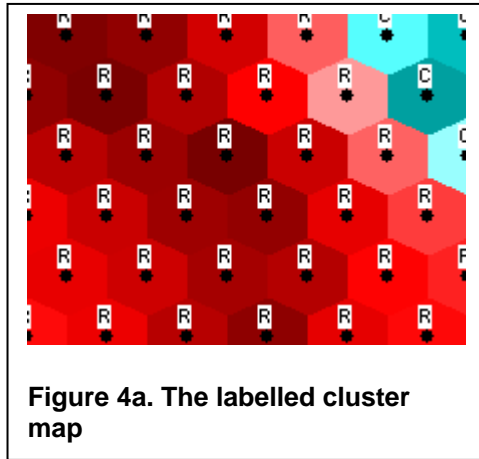
The cluster map reveals groups of similar data. The groups of data are referred to as clusters and each cluster has a distinct colour. An example of a cluster map is shown in Figure 4. Note that for the ease of viewing this paper in black and white, maps have been provided with labels. For example, the map in Figure 4 was labelled and is shown in Figure 4a. Similarly, the labelled maps of Figure 5 and Figure 6 are shown in Figure 5a and Figure 6a respectively. The reader should refer to these figures when reading the paper in black and white. Although the maps have been labelled according to the colour of the specific hexagonal grid, only the necessary portions have been labelled. The following letters represent the colours indicated:

- B indicates the area is blue
- C indicates the area is cyan
- G indicates the area is green
- Y indicates the area is yellow

The cluster map in Figure 4 reveals two clusters – one displayed in red and the other in cyan. The brightness of the colour reveals the distance of a unit to the centre of gravity. The centre of gravity is the map unit that most closely represents the average of all the units within a cluster. Brighter colours indicate a large distance, while the darker colours indicate a smaller distance to the centre of gravity.

The component maps reveal the value variation of components (or attributes) across the map. The combination of all these components determines the formation of clusters. An example of a component map is shown in Figure 5. The blue colour indicates small values, while the red indicates larger values and the other colours represent intermediate values. A close investigation of this map reveals that all the data with small values for the current attribute has been grouped in the top right-hand corner of the map. This is one reason why the exact same units formed a cluster in the cluster map shown in Figure 4. The component maps should therefore be used in conjunction with the cluster map, since together they form the backbone of the analysis.





After the training phase, the cluster map and three component maps – one for each component – were generated (see Figure 6). In Figure 6.1 one can see that three clusters were formed within the data set. By examining the cluster map and the component maps, it is evident that clusters are formed based on the time at which the files were created. These maps therefore provide an overview of the data, thereby making it easier for investigators to locate their points of interest.

For example, the component map that reveals the value variation of the date when the file was created is given in Figure 6.2. By looking at this map, one can see that the files have been grouped according to the period during which they were created. As mentioned earlier, the blue colour indicates small values, while the red indicates large values. Small values depict older files and large values depict newer files. Therefore the most recent files that were created are displayed on the upper half of the map, i.e. the area that contains the colours green, yellow, red, etc. The bottom half of the map reflects the files that were created at an earlier stage, i.e. the area that contains the colour blue. Investigators are consequently able to locate the files that were created further back by analysing the bottom portion of the map (as shown later on in Figure 7).

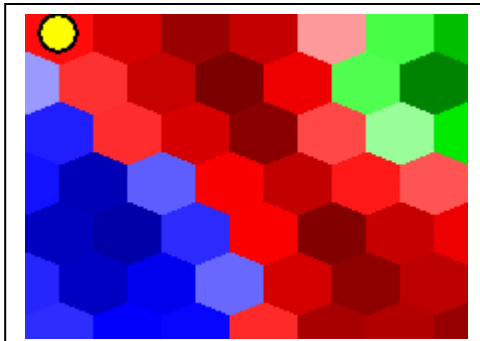


Figure. 6.1. The cluster map

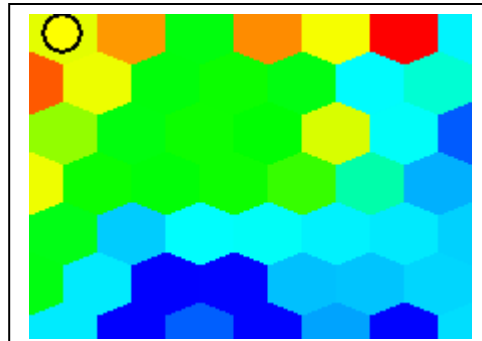


Figure 6.2. The component map of date created

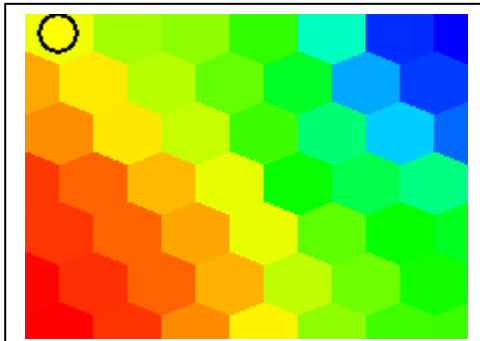


Figure 6.3. The component map of time created

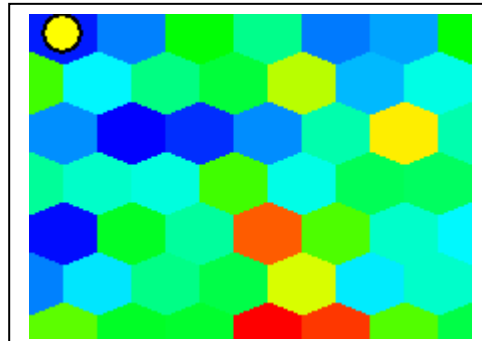


Figure 6.4. The component map of file extension

Figure 6. The cluster map together with the component maps

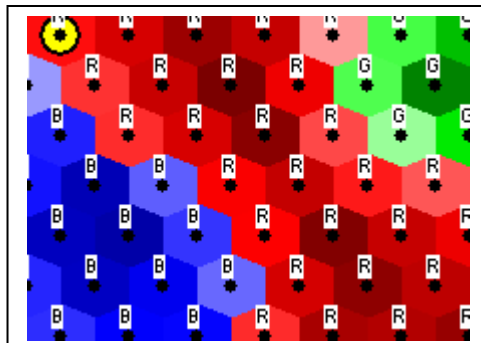


Figure. 6.1a. The labelled cluster map

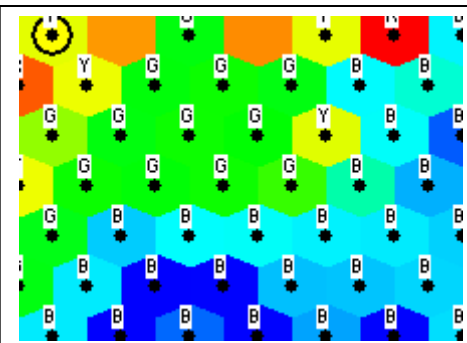


Figure 6.2a. The labelled component map of date created

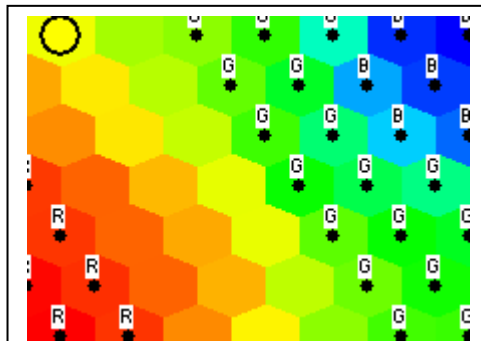


Figure 6.3a. The labelled component map of time created

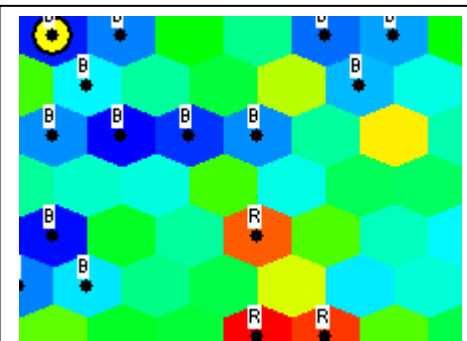


Figure 6.4a. The labelled component map of file extension

Figure 6a. The labelled cluster map together with the labelled component maps

In Figure 7 below, the area of interest is the files that were created further back. The area where the investigator is currently investigating is marked with a yellow circle. The date on which the files were created is 2004/07/31 and each pattern number refers to a particular file or graphical image. This information appears in the bottom right-hand corner of Figure 7. By comparing Figure 7 and Figure 8, one can confirm that the top portion of the map indeed reflects the files that were created more recently and vice versa. According to Figure 8, the dates on which the files were created were between 2004/10/03 and 2004/11/21.

As mentioned earlier, possible correlations are revealed by comparing component maps with one another. For example, a comparison of Figure 6.2 and Figure 6.3 shows that there is a correlation between the dates and the times when the files were created. The majority of the recently created files were created between 7:00am and 23:59pm, meaning that the majority of recent Internet activities took place between 7:00am and 23:59pm. Apart from these correlations, possible patterns can also be identified. For example, just by examining Figure 6.3, one can see a pattern of the graphical images downloaded onto the temporary Internet files folder. This

pattern shows that large portions of the graphical images were created between 7:00am and 23:59pm. This is just as expected, since most people are still asleep until 7:00am.

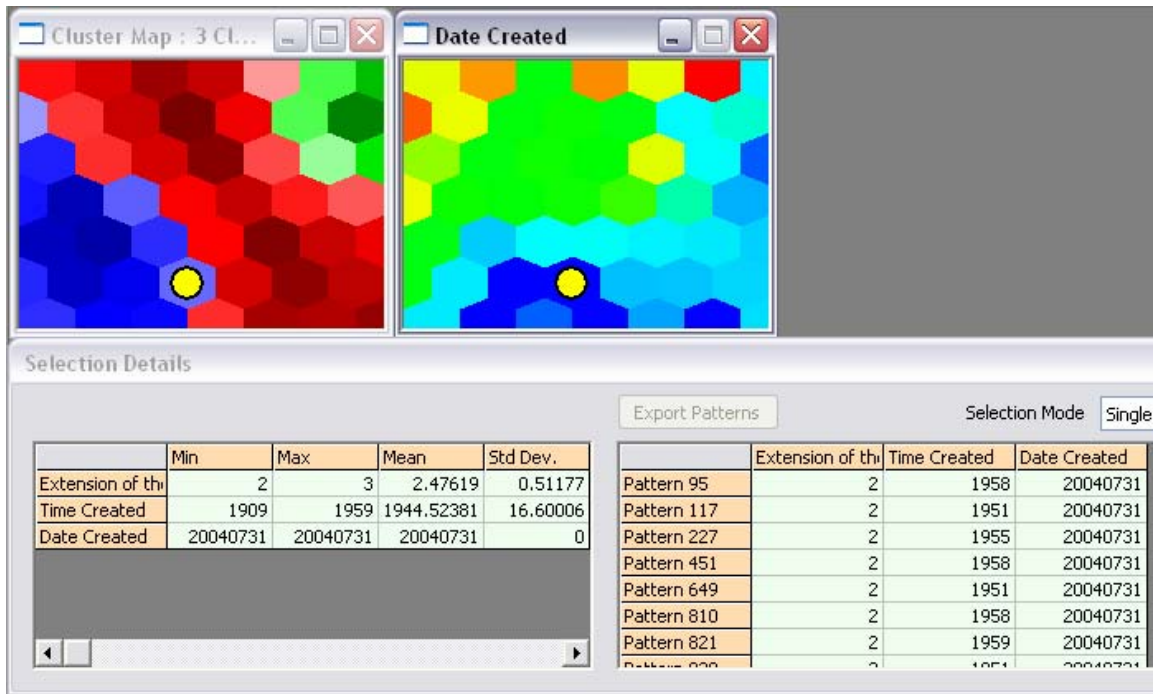


Figure 7. Examining the lower portion of the component map

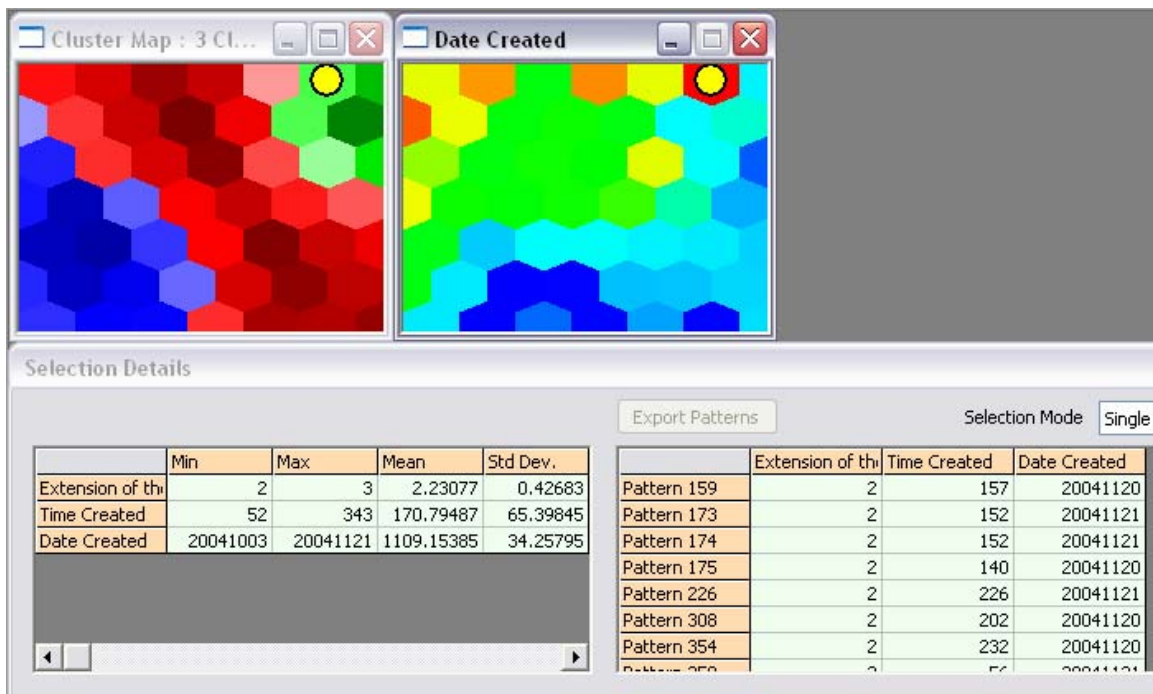


Figure 8. Examining the top portion of the component map

To conclude, possible correlations and patterns were discovered based on the data set. The locating of points of interest was also done quickly and easily by viewing the different maps that were created. The SOM can indeed aid computer forensic investigators to make better decisions and assist them in analysing evidence more efficiently during a computer investigation. In addition, the SOM is used not only to reveal interesting patterns, but also to serve as a basis for further analysis.

4.2. Dealing with the downloading of mp3s

Another application of the SOM involves investigations into users who are suspected of the illegal download of mp3s. Mp3s are used to store sound and often to store illegal copies of music. Downloading mp3s is suspected to be illegal, especially when a large number of mp3s are downloaded within a short period of time. By applying the SOM, investigators can discover possible patterns according to which the mp3s were downloaded or study the behaviour of suspects. In addition, investigators can locate the mp3s that were downloaded over a certain period. By comparing the different maps generated by the SOM application, investigators can immediately determine the date on and time at which such large portions of mp3s were downloaded. They are also able to determine the pattern according to which illegal downloads were made, for example, once a month or once a week at a certain time.

5. CONCLUSION AND FUTURE WORK

This paper has shown that the SOM can serve as a basis for the further analysis of data generated by computer forensic tools. A SOM application was briefly demonstrated, which showed that the easy visualisation provided by the different maps can greatly assist computer forensic investigators in their investigations. This visualization will help investigators to locate their points of interest more quickly since it provides them with an overview of the data.

The SOM is furthermore also ideal for the following tasks:

- Association: identifying correlations among data
- Classification: discovering and sorting data into groups based on similarities of data
- Clustering: finding and visually presenting groups of facts previously unknown or left unnoticed
- Forecasting: discovering patterns and data that may lead to reasonable predictions

The above features of the SOM can be useful when dealing with large volumes of data as might occur in a typical computer forensic investigation. The SOM offers a new perspective from which investigators may view the data.

A current drawback, unfortunately, is that transformation of the data needs to be done manually before the SOM application can be used to process the data. However, future work will be aimed at improving the SOM application or implementing a prototype that will allow the data transformation process to be performed automatically. Furthermore, there is a need for future research to improve or modify the SOM application so that it will be tailor-made for specific use in the field of computer forensics.

REFERENCES

- [1] Marcella, A. & Greenfield, R. 2002. *Cyber forensics: a field manual for collecting, examining and preserving evidence of computer crimes*. Auerbach.
- [2] Kohonen, T. 2001. *Self-organizing maps*. Springer-Verlag.
- [3] Kruse II, W. & Heiser, J. 2002. *Computer forensics: incident response essentials*. Addison-Wesley.
- [4] Guidance Software, Inc. 2004. <http://www.guidancesoftware.com>.
- [5] AccessData Corp. 2004. <http://www.accessdata.com>.
- [6] Technology Pathways, LLC. 2004. <http://www.techpathways.com>.
- [7] Gollmann, D. 1999. *Computer security*. Wiley.
- [8] Schweitzer, D. 2003. *Incident response: computer forensics toolkit*. Wiley.
- [9] Casey, E. 2002. *Handbook of computer crime investigation: forensic tools and technology*. Academic Press.
- [10] Kohonen, T. 1990. *The self-organizing map*. Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480.
- [11] Engelbrecht, A. 2002. *Computational intelligence: an introduction*. Wiley.
- [12] Vesanto, J. 1999. *SOM-based data visualization methods*. Intelligent Data Analysis, vol. 3, no. 2, pp. 111-126.