Chapter 10

EXPLORING FORENSIC DATA WITH SELF-ORGANIZING MAPS

B. Fei, J. Eloff, H. Venter and M. Olivier

Abstract This paper discusses the application of a self-organizing map (SOM), an unsupervised learning neural network model, to support decision making by computer forensic investigators and assist them in conducting data analysis in a more efficient manner. A SOM is used to search for patterns in data sets and produce visual displays of the similarities in the data. The paper explores how a SOM can be used as a basis for further analysis. Also, it demonstrates how SOM visualization can provide investigators with greater abilities to interpret and explore data generated by computer forensic tools.

Keywords: Computer forensics, self-organizing map, data visualization

1. Introduction

Dramatic increases in computer-related crimes have led to the development of a slew of forensic tools. These tools ensure that digital evidence is acquired and preserved properly and that the accuracy of results regarding the processing of digital evidence is maintained [9].

Computer forensic investigators are finding it increasingly difficult to use current tools to locate vital evidence in massive volumes of data. In addition, many tools do not present the data in a convenient format for analysis; sometimes, the data presented may actually result in misinforming investigators. In any case, the process of analyzing large volumes of evidence is extremely arduous and time-consuming.

Having an overview of the entire data set obtained directly from a hard drive can be crucial to an investigation. Patterns in the data set could help forensic investigators to locate information, and guide them to the next step in their search.

This paper shows how a self-organizing map (SOM) [6, 7], an unsupervised learning neural network model, can support decision making by computer forensic investigators and assist them in conducting data analysis in a more efficient manner. The technique is used to create graphical representations of large data sets that offer investigators a fresh perspective from which to study the data. In particular, a SOM reveals interesting patterns in data and also serves as a basis for further analysis.

The next section provides background information on computer forensics. The following two sections discuss the SOM technique and its application in computer forensic investigations. The final section, Section 5, presents the conclusions and directions for future work.

2. Background

Computer forensics deals with the preservation, identification, extraction and documentation of digital evidence [9]. Child pornography, threatening e-mails, fraud, and intellectual property theft are all crimes that leave digital tracks [8].

Numerous forensic tools have been developed to collect and/or analyze electronic evidence. Examples include EnCase [5], Forensic Toolkit [1] and ProDiscover [11]. Some tools are designed with a single purpose in mind. Others offer a whole range of functionalities, e.g., advanced searching capabilities, hashing verification and report generation.

A typical computer investigation involves making an exact copy of all the data on a storage medium (e.g., hard drive, compact disk, floppy disk or flash disk). The copy is called an image and the process of making an image is referred to as "imaging." Once the imaging process has been completed, it is essential to have a mechanism or procedure to ensure the integrity [4] of the evidence. Next, it is necessary to analyze the evidence, e.g., performing keyword searches [10] or analyzing signatures and hash values [2].

Computer forensic tools have advanced from using command-line environments to providing sophisticated graphical user interfaces that significantly enhance investigative activities. One useful feature is the presentation of files in a spreadsheet-style format. This ability allows investigators to view all the files on a particular storage medium as well as information regarding each file. The details include file name, file creation date and time, logical size, etc. However, when working with large data sets, the process of scrolling through many rows of data can be extremely tedious. Also, it can be difficult to locate specific information of interest to the investigation.

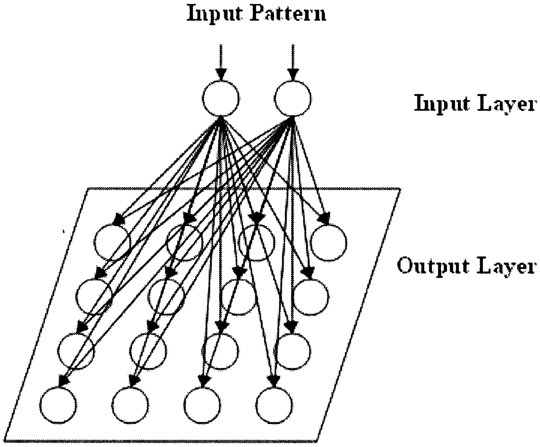


Figure 1. Self-organizing map.

The following section provides a brief overview of a self-organizing map (SOM). The SOM technique is used to enable investigators to visualize all the files on a storage medium and assist them in locating information of interest, both quickly and efficiently.

3. Self-Organizing Map

A self-organizing map (SOM) [6, 7] is a neural network model that has been successfully applied to clustering and visualizing high-dimensional data. It is used to map high-dimensional data onto a low-dimensional space (typically two dimensions). A SOM consists of two layers of neurons or nodes, the input layer and the output layer (Figure 1). The input layer is fully connected with neurons in the output layer and each neuron in the input layer receives an input signal. The output layer generally forms a two-dimensional grid of neurons where each neuron represents a node in the final structure. The connections between neuronal layers are represented by weights whose values represent the strengths of the connections. A SOM is based on unsupervised, competitive learning, which means that the learning process is entirely data driven and that neurons in the output layer compete with one another.

During the learning process, when an input pattern is presented to the input layer, the neurons in the output layer compete with one another. The winning neuron is the one whose weights are the closest to the input pattern in terms of its Euclidian distance [3]. Once the winning neuron has been determined, the weights of the winning neuron and its neighboring neurons are updated, i.e., they are shifted in the direction of the input pattern. After the learning process, a SOM configures the output neurons into a topological representation of the original data using a self-organization process [6].

The effect of the learning process is to cluster similar patterns while preserving the topology of the input space. However, in order to visualize the different clusters, an additional step is required to determine the cluster boundaries. Once the cluster boundaries have been determined, a SOM can be referred to as a cluster map. The size of a cluster is the number of nodes allocated to the cluster. One way to determine and visualize the cluster boundaries is to calculate the unified distance matrix (U-matrix) [3]. The U-matrix is a representation of a SOM that visualizes the distances between neurons. Large values in the U-matrix indicate the positions of the cluster boundaries.

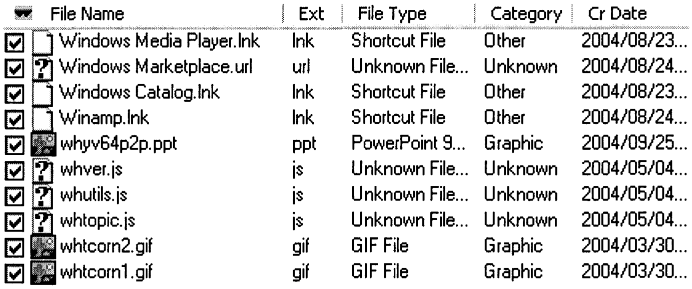
A SOM is useful for inspecting possible correlations between dimensions in the input data [12]. This can be achieved via component maps. Each component map visualizes the spread of values of a particular component (or dimension). By comparing component maps with one another, possible correlations are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

4. Applying SOM to Forensic Data

A SOM application employs an unsupervised neural network which is trained using forensic data. Two-dimensional maps, i.e., the cluster map and the different component maps, are displayed as hexagonal grids, each grid being referred to as a unit. The discussion of a SOM implementation is outside the scope of this work.

The requirements of computer investigations differ. For example, in the case of child pornography, an investigation involves examining all the graphical images on the suspect's computer system. In most cases, the data presented by forensic tools still requires investigators to manually examine the presented data and draw conclusions.

Figure 2 presents an example of what a computer forensic tool may present to an investigator – a spreadsheet-style display of all the files on the storage medium. This allows investigators to view all the files on the storage medium and to see the details of each file. The process of scrolling through the many rows of data for a large data set can be extremely tedious. However, by applying a SOM, the data set can be mapped to a two-dimensional space for convenient visualization and analysis.



	File Name	Ext	File Type	Category	Cr Date
☑	Windows Media Player.lnk	lnk	Shortcut File	Other	2004/08/23...
☑	Windows Marketplace.url	url	Unknown File...	Unknown	2004/08/24...
☑	Windows Catalog.lnk	lnk	Shortcut File	Other	2004/08/23...
☑	Winamp.lnk	lnk	Shortcut File	Other	2004/08/24...
☑	whyv64p2p.ppt	ppt	PowerPoint 9...	Graphic	2004/09/25...
☑	whver.js	js	Unknown File...	Unknown	2004/05/04...
☑	whutils.js	js	Unknown File...	Unknown	2004/05/04...
☑	whtopic.js	js	Unknown File...	Unknown	2004/05/04...
☑	whtcorn2.gif	gif	GIF File	Graphic	2004/03/30...
☑	whtcorn1.gif	gif	GIF File	Graphic	2004/03/30...

Figure 2. Table view of Forensic Toolkit.

4.1 Child Pornography

This section focuses on the application of a SOM to child pornography investigations, in particular, the analysis of temporary Internet files found on a seized hard drive. Most of the temporary Internet files are "image captures" of sites that the user has visited [9]. Obviously, these files may constitute evidence of illegal activity. In an investigation, a law enforcement agent must locate and examine all the images, discover possible patterns and study the suspect's Internet browsing patterns.

We illustrate the application of the SOM technique on an experimental data set containing 2,640 graphical images. The data set generated by Forensic Toolkit [1], a popular forensic tool, contains the four fields listed below. Note that even if file extensions are modified by the user, the tool is able to detect the correct format of each file.

- File name (used only for file identification).
- File extension.
- File creation time.
- File creation date.

The data provided by Forensic Toolkit (strings) cannot be processed directly by the SOM application. Consequently, the strings are converted to numerical values (see Table 1). Dates and times are converted to the formats, "yyyymmdd" and "hhmm," respectively.

The next step is to process the data set with the SOM application. The cluster maps and component maps produced after the SOM training phase are useful visual aids to the investigator.

Figure 3 presents a sample cluster map. The cluster map reveals groups of similar data (clusters), each displayed using a different color.

Table 1. Numerical values for file extensions.

File Extension	Numerical Value
bmp	1
gif	2
jpg	3
png	4

Since this paper is printed in black and white, the clusters are labeled for ease of reference. Figure 3a presents the labeled cluster map corresponding to the one displayed in Figure 3. Similarly, the labeled maps corresponding to Figures 4 and 5 are shown in Figures 4a and 5a, respectively. The letter B indicates the area is blue, C is cyan, G is green, and Y is yellow.

The cluster map in Figure 3 reveals two clusters, one red (R) and the other cyan (C). The brightness of the color reveals the distance of the unit to the center of gravity, i.e., the map unit that most closely represents the average of all the units within a cluster. Brighter colors indicate longer distances, while darker colors indicate shorter distances to the center of gravity.

Component maps, on the other hand, reveal the variations in values of components (or attributes). The combination of all these components determines cluster formation. An example component map is shown in Figure 4. Blue (B) indicates small values, red (R) indicates large values, and the other colors represent intermediate values. The component map in Figure 4 reveals that all the data with small values for the current attribute are grouped in the top right-hand corner of the map. This is the reason why the same units formed a cluster in the cluster map shown in Figure 3. The component maps should therefore be analyzed in conjunction with their corresponding cluster maps.

Figures 5.1 to 5.4 and Figures 5.1a to 5.4a present the unlabeled and labeled maps generated from the image data set after training the SOM. Figures 5.1 and 5.1a present the cluster maps, while the remaining three sets of figures display the component maps for the three components (file creation date, file creation time and file extension).

Figure 5.1 shows 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 when the files were created. This information can be very useful to an investigator.

Figure 5.2 reveals variations in the file creation dates. Blue (B) indicates small values (older files with earlier creation dates) while red (R) indicates large values (new files). Therefore, the most recent files are displayed in the upper half of the map (green (G), yellow (Y), red (R),

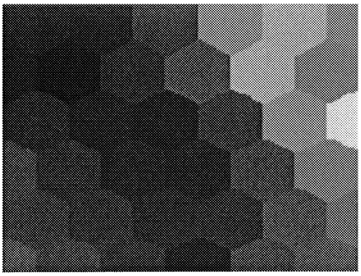


Figure 3. Cluster map.

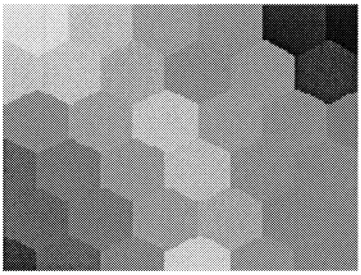


Figure 4. Component map.

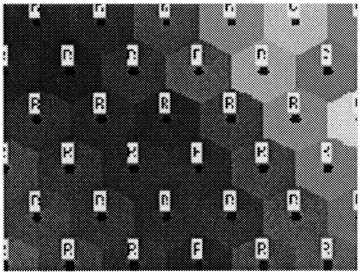


Figure 3a. Labeled cluster map.

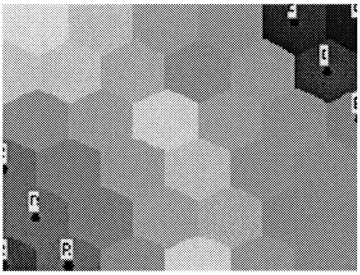


Figure 4a. Labeled component map.

etc.) The bottom half of the map reveals the files that were created earlier (blue (B)). Investigators are thus able to locate the older files by analyzing the bottom portion of the map (see Figure 6).

The older files constitute the area of interest in Figure 6. The specific area being examined by the investigator is marked with a yellow circle. The files were created on 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 6. Comparing Figures 6 and 7 confirms that the top portion of the map indeed reflects the files that were created more recently. According to Figure 7, the dates on which the files were created ranged from 2004/10/03 to 2004/11/21.

Correlations are revealed by comparing component maps. For example, comparing Figures 5.2 and 5.3 shows that a correlation exists between file creation dates and file creation times. Most of the recent files were created between 7:00 and 23:59, meaning that the majority of recent Internet activity took place during this time period. Also, simply by examining Figure 5.3, it is possible to discern a downloading behavior pattern. Specifically, most of the images were created between 7:00 and 23:59, corresponding to normal waking hours.

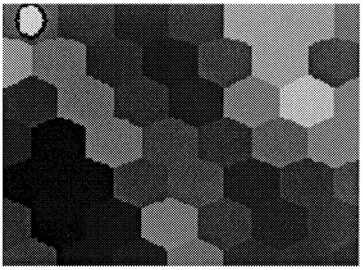


Figure 5.1 Cluster map.

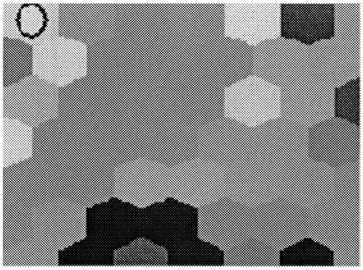


Figure 5.2 Component map (file creation date).

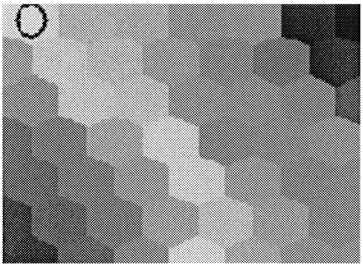


Figure 5.3 Component map (file creation time).

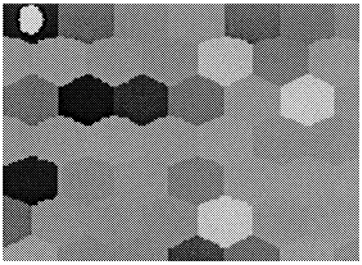


Figure 5.4 Component map (file extension).

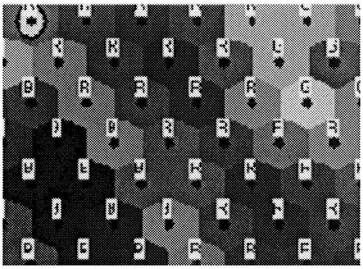


Figure 5.1a Labeled cluster map.

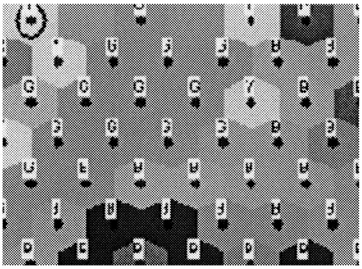


Figure 5.2a Labeled component map (file creation date).

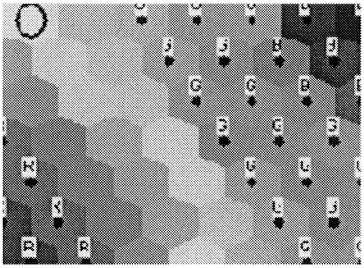


Figure 5.3a Labeled component map (file creation time).

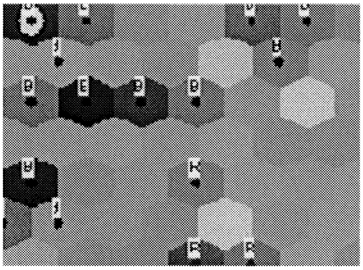
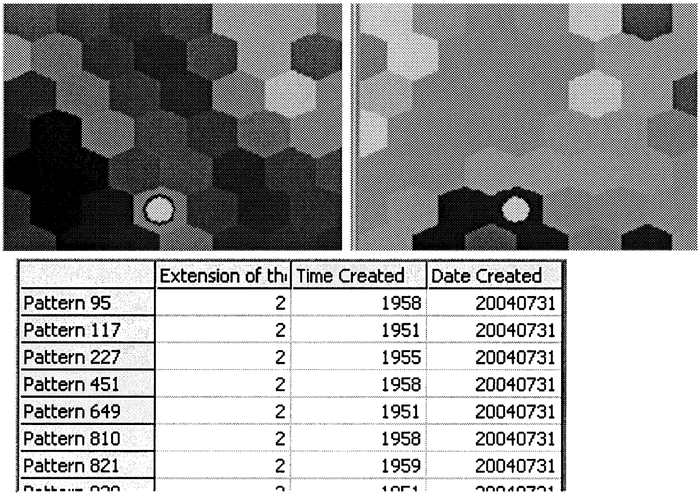
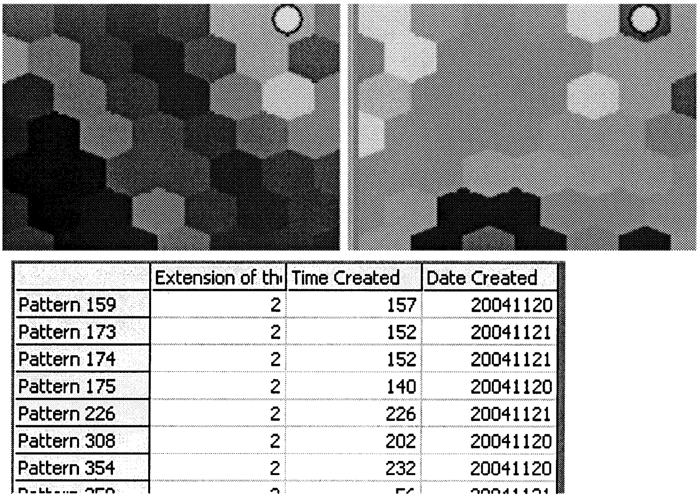


Figure 5.4a Labeled component map (file extension).



	Extension of th	Time Created	Date Created
Pattern 95	2	1958	20040731
Pattern 117	2	1951	20040731
Pattern 227	2	1955	20040731
Pattern 451	2	1958	20040731
Pattern 649	2	1951	20040731
Pattern 810	2	1958	20040731
Pattern 821	2	1959	20040731

Figure 6. Examining the lower portion of the component map.



	Extension of th	Time Created	Date Created
Pattern 159	2	157	20041120
Pattern 173	2	152	20041121
Pattern 174	2	152	20041121
Pattern 175	2	140	20041120
Pattern 226	2	226	20041121
Pattern 308	2	202	20041120
Pattern 354	2	232	20041120

Figure 7. Examining the top portion of the component map.

4.2 MP3 Downloading

Another application of a SOM is the investigation of illegal downloading of MP3 (music) files. Downloading MP3 files may be deemed suspicious when large numbers of MP3 files are downloaded in a short period of time. Investigators can discover illegal MP3 downloading patterns using a SOM. In addition, investigators can identify the MP3 files that were downloaded during a certain period of time. By comparing the different maps generated by a SOM application, investigators can determine exactly when large numbers of MP3 files were downloaded. They can also determine the downloading patterns, e.g., every Friday night or once a month at a certain time.

5. Conclusions

A self-organizing map (SOM) can serve as the basis for further analysis of data generated by computer forensic tools. In particular, maps generated by a SOM application create excellent visualizations of large higher-dimensional data sets. These visualizations enable forensic investigators to locate information that is of interest both rapidly and efficiently.

The SOM technique has several applications in digital forensics. These include identifying correlations (associations) in data, discovering and sorting data into groups based on similarity (classification), locating and visually presenting groups of latent facts (clustering), and discovering patterns in data that may lead to useful predictions (forecasting). By providing new perspectives for viewing data, these applications can facilitate the analysis of large data sets encountered in digital forensic investigations. A major drawback, however, is that the data needs to be transformed manually before it can be processed by a SOM application. One avenue for future research is to develop automated data transformation techniques. Another is to design specialized SOM applications for all the major digital forensic processes.

References

[1] AccessData (www.accessdata.com).

[2] E. Casey, *Handbook of Computer Crime Investigation: Forensic Tools and Technology*, Academic Press, San Diego, California, 2002.

[3] A. Engelbrecht, *Computational Intelligence: An Introduction*, Wiley, New York, 2002.

[4] D. Gollman, *Computer Security*, Wiley, New York, 1999.

[5] Guidance Software (www.guidancesoftware.com).

[6] T. Kohonen, The self-organizing map, *Proceedings of the IEEE*, vol. 78(9), pp. 1464-1480, 1990.

[7] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag, Berlin, Germany, 1995.

[8] W. Kruse and J Heiser, *Computer Forensics: Incident Response Essentials*, Addison-Wesley, Reading, Massachusetts, 2002.

[9] A. Marcella and R. Greenfield (Eds.), *Cyber Forensics: A Field Manual for Collecting, Examining and Preserving Evidence of Computer Crimes*, Auerbach, Boca Raton, Florida, 2002.

[10] D. Schweitzer, *Incident Response: Computer Forensics Toolkit*, Wiley, New York, 2003.

[11] Technology Pathways (www.techpathways.com).

[12] J. Vesanto, SOM-based data visualization methods, *Intelligent Data Analysis*, vol. 3(2), pp. 111-126, 1999.