Research of Multiple-type Files Carving Method Based on Entropy

Jun GUO[1, a], Jingsha HE[2, b,], Na HUANG[*]

[1]Department of Software Engineering, Beijing University of Technology, Beijing, 100124, China

[2] Department of Software Engineering, Beijing University of Technology, America, 100124, China

[a]email: guojun801@163.com, [b]email: jhe@bjut.edu.cn, [*]605051989@qq.com

Keywords: File Carving, File Fragments, Entropy, Bloom filter, BFD, SVM, PUP

Abstract. File carving is a technique of recovering data from disk without depending on the File System, and the key step is the extraction and reassembly of file fragments. Efficient recognition and extraction of file fragments is not only the prerequisite of recovering file, but also the guarantee of a low false positive rate and high accuracy Digital Forensics. In this paper, when the entropy of file fragments is low, the validation algorithms I used for the extraction are header/footer validation and entropy feature extraction validation, but when the entropy of file fragments is high, besides the previous two algorithms I introduced Bloom filter feature extraction validation, byte frequency distribution (BFD) feature extraction validation and support vector machine (SVM) with supervised learning ability to detect the type of file fragments. After the extraction, I used Parallel Unique Path (PUP) for the reassembly of file fragments. I used DFRWS 2007 carving challenge data set to test my method and the result is better than only using entropy to classify multiple-type files especially in the case of high entropy.

Introduction

Traditional data recovery technology mainly depends on metadata information of the file system. Recovery technique based on metadata information of file system is often only restore continuous storage of files on disk. It has low accuracy and high rate of false positives in the face of data fragments, disorder, and part of the data damage or loss. File carving technology is further supplement and development of traditional recovery technique. This technique does not rely on the file system of original disk image, it attempt to restore and refactor file from the surface without the structure of the binary data stream (i.e., the original disk image), having important significance to the development of the digital forensics.

File carving has three steps [1]: Firstly, the original disk image was obtained from the investigation target; secondly, identifying whether a file is intact and handle file fragments; finally, checking validities of the files and reproduce the complete file from the image. An important concept of file carving is handling file fragments. File fragments processing including identification and extraction of the fragments as well as reassembly of fragments. Nowadays, most file carving techniques based on fragments have three shortcomings: Firstly, it is difficult for people to classify the fragments of different files; secondly, it is hard to reassemble the fragments of same file; thirdly, their results have high “false positives”. This paper focuses on finding out an efficient classification method of file fragments, that is, as much as possible to distinguish between fragments of different file types and different fragments of same file type, which is premise of increasing the accuracy and reducing “false positives” on digital forensics technology. According to Shannon's information theory, entropy can be used to measure unit density of information or state of data compression, which can be applied to the field of pattern recognition and classification. File type of lower entropy value is relatively easy to classify and detect, and the poor for the file type of higher entropy value is corresponding classification and testing results, the degree of confusion between all various of file type of high entropy value is relatively large. This paper describes and uses information entropy, Bloom filter, BFD, and SVM to classify file fragments, and uses Parallel Unique Path to classify and resemble multiple-type files at the same time. This method can improve the classification accuracy of file fragmentation and result of file carving has a low rate of “false positives”.

Related Work

File carving exist for some time and has been more attention in 2006 and 2007. File carving technology was successively issued a challenge problem in 2006 and 2007 two years as digital forensics research group (DFRWS). In 2004, Olivier de Vel used colored the concept of generalized suffix tree to calculate the kernel matrix of broadband K spectrum of data flow, and according to the various file types use support vector machine classification algorithm to categorize documents collection; In 2005, Li wei-jen, Wang ke et al. [2] in the international conference described 1-gram method to calculate all kinds of different file types of ``file fingerprint''; In 2006, Martin Karresand and Nahid Shahmehri [3] proposed a method called Oscar to classify the disk, memory, and swap partition of binary data, this method can well distinguish files fragmented data in the RAM and memory image; In 2007, Cor J.Veenman [4] studied a method that using entropy, complexity, entropy-complexity and histogram as a feature vector to classify data on disk cluster. In 2008, Pal et al. [5] proposed a method of matching keywords and signature to recover fragment file and detect fragment points. In 2008, Calhoun and Coles [6] proposed using the Fisher linear discriminant to a set of different statistics including sum of the 4 highest byte frequencies and correlation of the frequencies of byte values as well as combinations of the statistics.

As file carving method is put forward, file carving tools also gradually developed. Foremost is probably the ancestor of all file carving process, by Kris Kendall and Jesse Kornblum [7] jointly developed an open-source carving tools in the United States air force special research institute. In order to improve performance and reduce memory footprint, Golden g. Richard and Vassil Roussev developed Scalpel in 2005. Scalpe is an extension of Foremost, both of them are based on the principle of combining flag information of file header and footer to carve all kinds of format file. In the case of file system damages or partial files damage, PhotoRec is widely used to recover more types of file data. In 2006, Joachim Metz and Robert-Jan Mora jointly designed a called Revit carving tools. Revit not only search flag information of the file header and footer, but also deal with other file characteristics, such as entropy value, the probability statistical distribution, embedded file structure, etc., which have better recovery effect.

Proposed Method

3.1 Header/Footer Validation

Header refers to the beginning of the file, and footer refers to the termination of the file. Headers and footers of different file have different characteristics, beginning and termination of the file were expressed by two different and fixed sequences of characters. For example, the header of JPEG files is “FFD8” [8], the footer of JPEG files is “FFD9”; The header of ZIP files is “50 4B 03 04”, the footer of ZIP is fixed character sequence of data block, which is “50 4B 05 06”. File header can determine the file type, and include some attribute information, attribute information refer to file type, file length, file size and so on. These attributes play an important role to confirm the connection of fragments.

3.2 Entropy Feature Extraction Validation

The German physicist Clausius (Rudolf Julius Emanuel Clausius) in 1865 proposed the concept of entropy, it is used to measure degree of disorder of a thermodynamic system in thermodynamics. Shannon (Claude Elwood Shannon) in 1948 proposed the concept of "information entropy", using to solve the problem of information quantitative measure, so it is often called the Shannon entropy or information entropy. According to the principle of information entropy, the entropy is a measurement of uncertainty information and can be used to identify different file type. The higher the entropy, the more information can be transmitted; Conversely, then it means fewer transmission of information. Some research shows that entropy value of plain text file is small, compressed file and encrypt files is larger entropy value.

Entropy is used to measure the probability of occurrence of a random variable in information theory, Formula [9] can be expressed as follows:

$$H(x) = -\sum_{i=1}^{n} p(i) \log_2 p(i) \quad (1)$$

Where, $H(x)$ represents the entropy value of the random variable x, $p(i)$ represents the probability of i event, entropy unit is bit, n can be seen as limited sequences of consisting of different byte for file fragments on the logic, matching ASCII/UTF-8 collection of different elements, $n = 256$.

Low entropy fragments can easily detect and extract, the probability of cross entropy is very big among high entropy fragments and the “false positives” is very high, therefore, entropy feature is applied to classify low entropy fragments.

3.3 Bloom filter Feature Extraction Validation

Bloom filter [10] is very high random data structure with space efficiency and time efficiency, which uses a bit vector to express a set and can detect whether an element belongs to this set. The basic principle of Bloom filter is a bit vector joint hash function. Firstly, an empty Bloom filter is a bit vector of m bits, all set to 0. Secondly, using k different hash functions, each of which maps a key value to one of the m positions in the vector. To insert an element into the Bloom filter, we compute the k hash function values and set the bits at the corresponding k positions to 1. Thirdly, to test whether an element was inserted, we hash the element with these k hash functions and check if all corresponding bits are set to 1, in which case we say the element is in the filter. Shown in Figure 1.

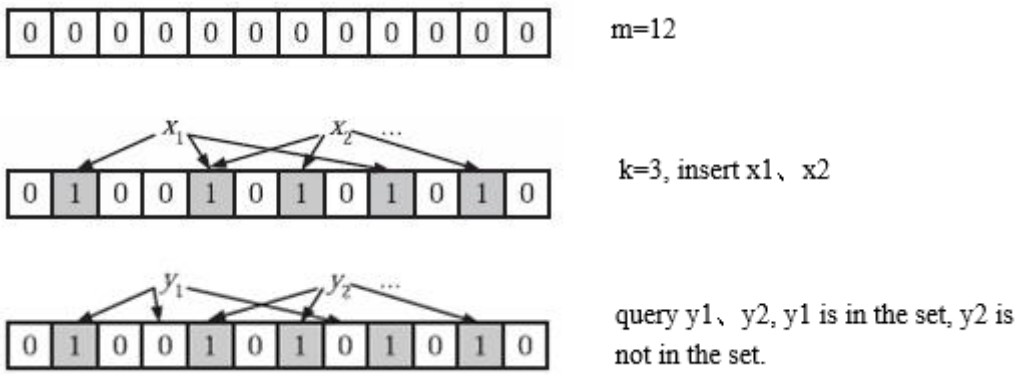

Fig.1. Principle of Bloom filter

Using the Bloom filter, when it judges an element does not belong to the set, the element must not belong to this set; When it judges an element in this set, the element does not belong to this set. Using this principle, for fragments of high entropy and cross entropy, all fragments of the same type can be detected, but which include does not belong to fragments of this file type.

3.4 BFD Feature Extraction Validation

McDaniel and Heydari proposed a method using the file fingerprint to identify file type and this method has higher accuracy. They put forward three different algorithms to generate the file type of "fingerprints", one of the algorithm is based on byte frequency distribution (BFD) [11]. Before using the method of SVM, file fragments firstly need to express as a feature vector, and the classification accuracy and selection of feature vector is very relevant. Feature vector includes histogram of byte value, BFD and Rate of Change (ROC), this paper uses the BFD to express file fragments.

BFD has 256 value for each byte, after statistics on a large number of different types of files, calculating the BFD of various file types, that is, fingerprint of file type, which used to identify disk data blocks as a standard. BFD is a method which to gather statistics the number of occurrences of byte value among file fragments and then the data is normalized and displayed in the form of graphs.

3.5 Support Vector Machine

Corinna Cortes and Vapnik et al. proposed SVM to solve small sample, nonlinear and high dimensional pattern recognition, it exhibits many unique advantages and can be applied to the function fitting and other Machine learning problem. It is a Machine learning algorithm based on supervision with stronger robust and better efficiency and is widely used in various types of data classifier model. A SVM builds a hyperplane, or in high or infinite dimensional space, it can be used for classification, regression, or any other task in set the hyperplane [12].

Figure 2 shows optimal classification surface of a SVM, solid circle and hollow circle represent two different samples. L is a straight line for distinguishing two samples, L1 is a the most recent straight line from L samples of the lower left, L2 is a the most recent straight line from L samples of the upper right , and L, L1, L2 parallel each other. Straight line distance between L1 and L2 is interval of classification.

Specifically described as follows:

1. Given a sample collection (the training set),

$$S = \{(r_1, c_1), (r_2, c_2), ..., (r_n, c_n) \mid r_i \in R_n, c_i \in (-1,1)\} \quad (2)$$

SVM will find an optimal hyperplane to conduct data classification.

2. If $\omega \in R_n, d \in (-1,1)$ and $\varepsilon > 0$, when $c_i = 1$, for any r_i, all have

$$(\omega \times r_i) + d \geq \varepsilon \quad (3)$$

then the sample set S linearly separable.

3. In the case of linear inseparable, considering to use nonlinear mapping algorithm to process the sample set S, that is used the mapping algorithm to map sampling from low dimensional space to high dimensional space, then whether calculating the linearly separable.

SVM algorithm is more common in the identification field of file type. The working principle of SVM [13] is as follows:

1. According to known a set of data samples with known classes (the training set) to establish a model, and use the model to predict classes of data samples that are of unknown classes.

2. Given the training set with class labels and features, SVM treats all the features of a data sample as a point in a high dimensional space, and tries to construct hyperplanes to divide the space into partitions.

3. Choosing the appropriate kernel function to establish the model. Common kernel functions include: (1) linear, (2) polynomial, (3) radial basis function (RBF), and (4) sigmoid. Kernel function parameters can be adjusted according to the character of the problem and find the best parameters.

SVM supervised learning process has three steps, as shown in Figure 3: Firstly, the raw data needs to unzip or decipher at preprocessing of the training set. For the training set, we know all extractable type of file fragments, and mark all the type of file fragments, then extract BFD feature vector for every file fragments, and then put them into SVM and train out a model which judges file fragments type, which used to predict the unknown file type of file fragments. Secondly, when an unknown type of file fragments input into the model, using the Bloom filter to group file type of file fragments and calculating BFD feature vector of the file fragments. Thirdly, using the model of the first step to predict the type of the file fragments.

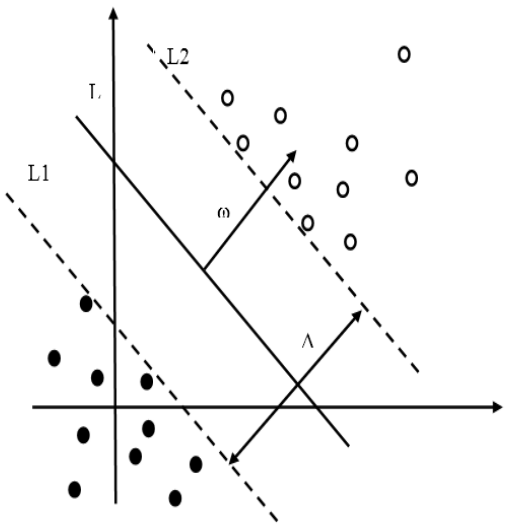

Fig.2. Optimal classification surface of SVM

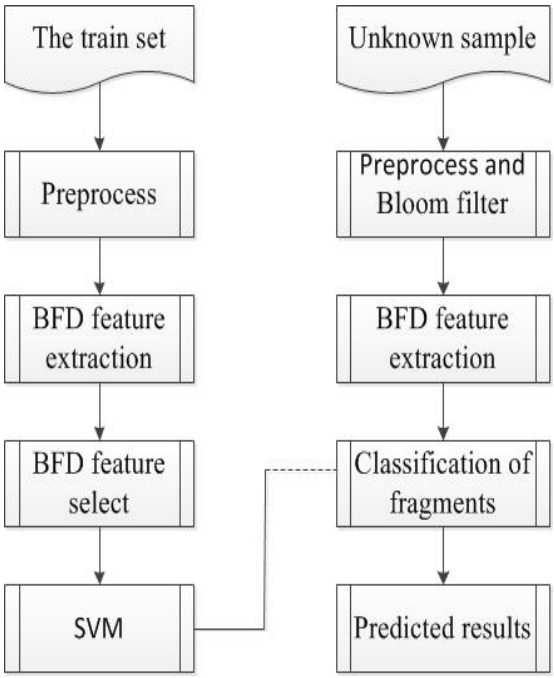

Fig.3. High entropy file fragmentation recognition algorithm framework

3.6 Parallel Unique Path

Multiple-type files contains many file types. In order to achieve a variety of file types to assemble at the same time, we use the method of Parallel Unique Path (PUP) [14], it means multiple-type files to detect their subsequent data cluster at the same time. PUP is described below: By file header signature detected K file header and put them into the beginning of K linked list, each linked list is a reconstruction path stored a file. Then iterate through all the data block and find the best matching of data block. Each pass of the algorithm, fragments that fit best to the current heads of the reassembly paths are determined. As shown in Figure 4, in step (A), H1, H2, H3 represented file header which was confirmed. Search the fragments collection and find fragment 6 fits best H2, so fragment 6 is added to the path 2. In the next step (B), fragment 4 fits best H1, so fragment 4 is added to the path 1. In step (C) fragment 5 becomes the head of reassembly path 1. This procedure is continued until all fragments have been assigned available reassembly paths.

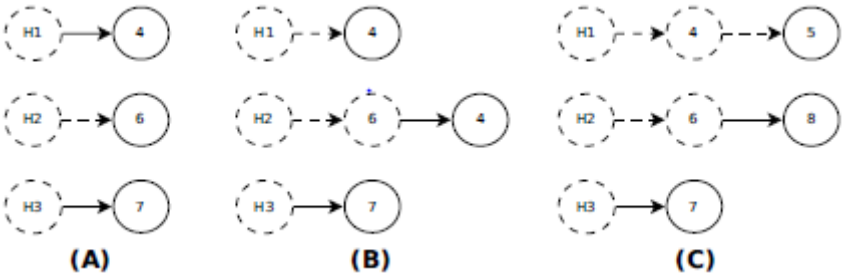

Fig.4. PUP assembly algorithm

3.7 Algorithm Process

Multiple-type files carving method process is as follows, shown in Figure 5:

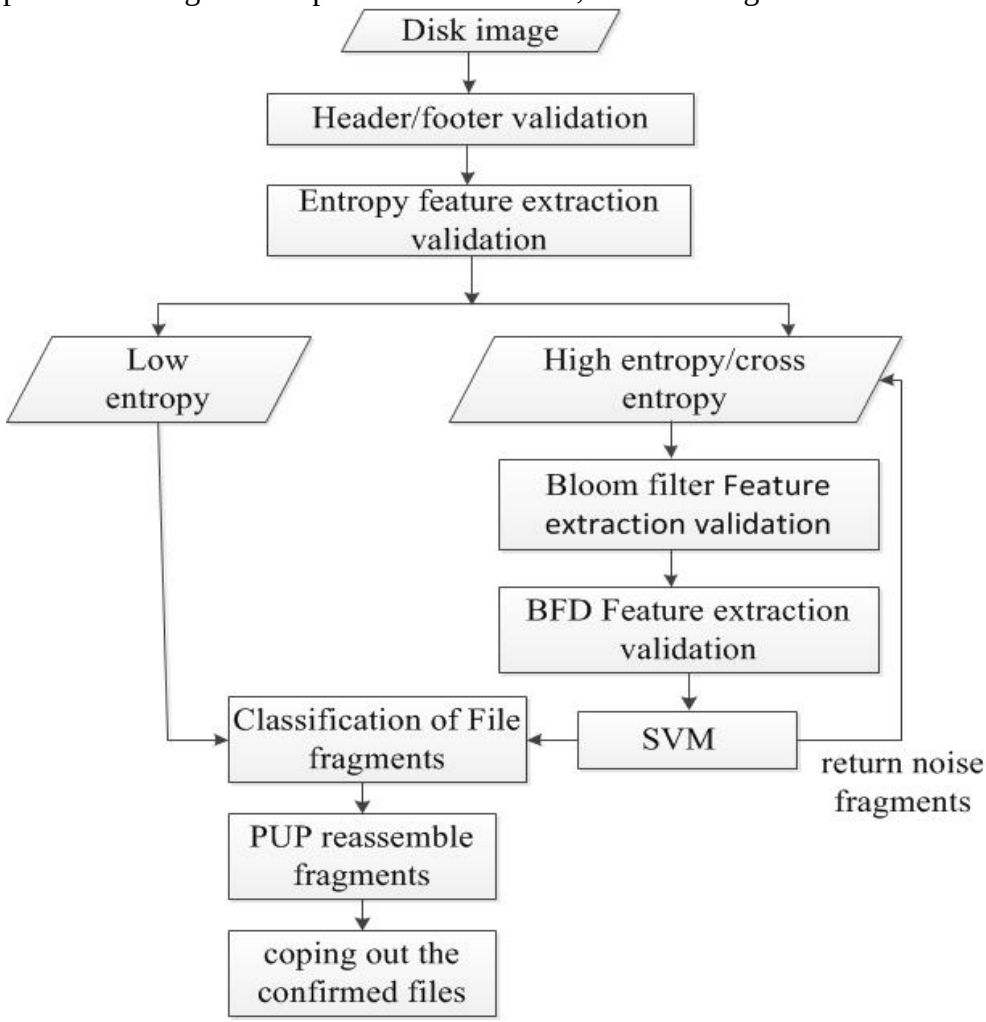

Fig.5. Multiple-type files carving method process

1 Choose unknown disk image as a file under test.

2 According to known file head signature to detect file head, once finding them and put them into the starting position of linked list. Several file headers correspondingly establish several linked lists. No header, ending carving.

3 According to known file footer signature to detect file footer, find them and add them into end of corresponding linked list.

4 For files have headers and footers, calculating entropy of headers and footers to determine average entropy $E_{average}$ of file type. Only file header exist, using known fragments of file types to determine the average entropy of file type. Set their entropy error E_{error}.

5 For each fragment, extracting the entropy features $E_{entropy}$ of the data block under test. For low entropy block, Computing $|E_{entropy} - E_{average}| \leq E_{error}$ and Adding them to corresponding linked list. For noise fragments file $|E_{entropy} - E_{average}| > E_{error}$, classified according to the step 6 and 7.

6 For data block of high entropy and cross entropy, according to file type by signature of file header determined, using the Bloom filter to select one of file fragments type and extracting the BFD feature vector of each data block, through SVM to predict the file type of the data block and add them to corresponding linked list, and return the noise fragments.

7 In another file fragmentation, cycling step 6 and till all the fragments of file types are grouped.

8 PUP deal with multiple-type files simultaneously conduct reassembly fragment. If unidentified fragments exist, for data files, blocks of the same file, its file ID is unchanged and block ID

gradually plus 1 increment. According to the file ID, fragments will be assigned to the most appropriate path. At Each path fragments according to the order of the data block ID to restructure. Extracting content of file fragments to restore the file.

Result and Discussion

In order to verify the validity of the carving method, we select carving mirror of DFRW 2007 as the experimental data, size of mirror is 256 MB, including common Word, C++ source code, JPEG, and PDF. JPEG and PDF belong to high entropy file, entropy value have overlapping phenomenon. For file of high entropy, we only carve JPEG file type, because the more file types were recovered, the effect of carving will be reduced. We collected 1000 JPEG images and 1000 PDF files as the training set. For a cluster in modern file systems, the size of 4096 bytes is common, so we divided the training set and the testing files into fragments of 4096 bytes for each file, removing their header file and footer file. Then according to our method to conduct experiment. In order to compare, we used software tools of Foremost and FTK to analyze the file fragments of the same data, the result is shown in Figure 6.

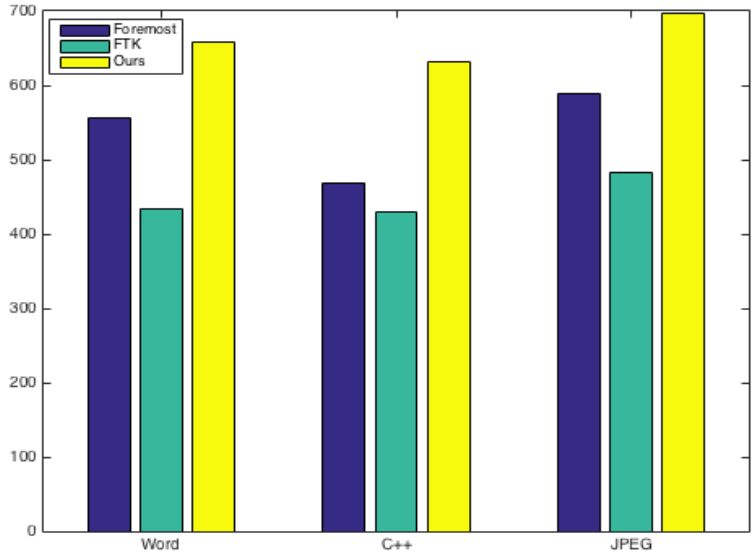

Fig.6. The number of three types of file fragments carving successful statistics

As can be seen from the Figure 6, for low entropy files and high entropy files, the method is effective. For type of low entropy file, their scope of entropy value is small and the probability of entropy overlap is very low, we only use the information entropy to classify and result can achieve high precision. For type of high entropy file, their scope of entropy value is bigger and cross entropy phenomenon will occur. For files of cross entropy, combining Bloom filter, BFD and SVM to classify file fragments, improving classification accuracy of file fragments. This paper solved the problem, for cross entropy and high entropy, comparing with only using entropy, effect has the obvious improvement on files classification.

Conclusion and Outlook

In this paper, we introduced the carving method of multiple-type files to recover file, the method includes four kinds of validation, SVM and PUP. For type of low entropy file, we only use the information entropy to classify and result can achieve high precision. For type of high entropy file, there are cross entropy between different file fragments, so it is difficult to determine the type of file fragments, we are based on information entropy, combining Bloom filter, BFD and SVM to classify file fragments of high entropy. Through the experimental results show that this method can effectively classify different file types and improve the carving success rate of the file.

Future works include two factors. Firstly, in this paper, experiments only involve two high

entropy files, in future work, we will study and restore a few low entropy file and several high entropy file at the same time, further develop the method. Secondly, we will study the method of neural network and k neighbor to achieve a higher precision for low entropy file.

Acknowledgement

We would like to thank all anonymous referees for their valuable comments. The work in this paper has been supported by National High-tech R&D Program (863 Program) (2015AA017204), Beijing Natural Science Foundation (4142008), National Nature Science Foundation of China (61272500) and Shandong National Science Foundation (ZR2013FQ024).

References

[1] Chen, M., Zheng, N., Xu, M., Lou, Y., Wang, X.:Validation algorithms based on content characters and internal structure: The pdf file carving method. In: Information Science and Engineering, 2008. ISISE'08. International Symposium on. Volume 1, IEEE (2008) 168-172

[2] Li, W.J., Wang, K., Stolfo, S.J., Herzog, B. Fileprints: Identifying file types by n-gram analysis. In: Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC, IEEE (2005) 64-71

[3] Karresand, M., Shahmehri, N., Oscar file type identification of binary data in disk clusters and ram pages. In: Security and privacy in dynamic environments. Springer (2006) 413-424

[4] Conti, G., Bratus, S., Shubina, A., Sangster, B., Ragsdale, R., Supan, M., Lichtenberg, A., Perez-Alemany, R., Automated mapping of large binary objects using primitive fragment type classification. digital investigation 7 (2010) S3-S12

[5] Pal, A., Sencar, H.T., Memon, N., Detecting file fragmentation point using sequential hypothesis testing. digital investigation 5 (2008) S2-S13

[6] Calhoun, W.C., Coles, D., Predicting the types of file fragments. Digital Investigation 5 (2008) S14-S20

[7] Kloet, S., et al., Measuring and improving the quality of file carving methods. Almere, Niederlande: Eindhoven University of Technology (2007) 4-79

[8] Mikus, N., An analysis of disc carving techniques. Technical report, DTIC Document (2005)

[9] Shannon, C., A mathematical theory of communication, bell system technical journal 27: 379-423 and 623{656. Mathematical Reviews (MathSciNet): MR10,133e (1948)

[10] Ponec, M.,et al., New payload attribution methods for network forensic investigations. ACM Trans.Info.Syst.Sec.13,2,Article 15(February 2010)

[11] McDaniel, M., Heydari, M.H., Content based file type detection algorithms. In: System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on, IEEE (2003) 10-pp

[12] Li, B., Wang, Q., Luo, J.: Forensic analysis of document fragment based on svm. In: Intelligent Information Hiding and Multimedia Signal Processing, 2006. IIH-MSP'06. International Conference on, IEEE (2006) 236-239

[13] Li, Q., Ong, A., Suganthan, P., Thing, V., A novel support vector machine approach to high entropy data fragment classification. SAISMC (2011) 236-247

[14] Poisel, R., Tjoa, S., A comprehensive literature review of file carving. In: Availability, Reliability and Security (ARES), 2013 Eighth International Conference on, IEEE (2013) 475-484