A Taxonomy of File-Type Identification Techniques

Nasser S. Alamri
Florida Institute of Technology
Melbourne, FL, USA
nalamri2005@my.fit.edu

William H. Allen
Florida Institute of Technology
Melbourne, FL, USA
wallen@fit.edu

ABSTRACT

It has become increasingly important to identify corrupted or malicious files before they are processed by vulnerable applications. In the past, it was sufficient to rely on a filename extension or on information in the file's header, such as a magic number or text string, to reliably identify a file's type. Recent changes in attack patterns show that these approaches are no longer sufficient to avoid misidentifying harmful files.

The authors of this paper have surveyed the existing research on file type identification and determined that a wide range of techniques have been investigated, but that no work has yet been done to organize them into a useful taxonomy. This paper describes our preliminary research on the development of such a taxonomy with the goal of aiding the efforts of other researchers in creating solutions to this problem. In this paper, we present an initial organization that incorporates over 30 different algorithms or approaches.

Keywords

Computer forensics, File systems, Taxonomies.

1. INTRODUCTION

Applications rely on protocol or file format specifications to correctly access the header fields and data in files. However, to process an input file, the application must first be able to determine the correct file type. In the past, it was sufficient to rely on a file's extension or on information in the file's header, such as a magic number or text string, to reliably identify a file's type. Recent changes in attack patterns show that these approaches are no longer sufficient to avoid misidentifying harmful files. Attackers are aware that it is possible to change a file extension to one that is associated with a "safe" file type or to craft a false header that displays the correct magic number or text string, but is not actually indicative of the file's true type or contents [4].

The authors of this paper surveyed the existing research in file-type identification and found that it employs a number of different approaches to classify unknown files according to their file type. Some approaches to file-type identification have found modest results with simple techniques, such as byte frequency, without regard to file structure. Other methods create a 'fingerprint' or signature for each file type and classify an unknown file by matching its signature to one of the known file types.

ACM SE '14, Mar 28-29 2014, Kennesaw, GA, USA
ACM 978-1-4503-2923-1/14/03.
http://dx.doi.org/10.1145/2638404.2638524

However, while research into identifying file types has been underway for over a decade, there has yet to be a systematic organization of this body of work. Therefore, this paper proposes the first such effort by presenting a taxonomy of file type identification techniques. The work presented here is preliminary, but is intended to establish a first step towards creating a formal basis for further research in this field.

2. MOTIVATION

Taxonomies are important to science because they both organize existing work and provide a guide for future research. A taxonomy can be characterized in several ways. Landwehr et al. [20] defines a taxonomy as a way of distinguishing between specimens of recorded data and stated that "it implicitly embodies a theory of the universe from which those specimens are drawn". A taxonomy can also be described as a classification schema that plays a primary role in classifying the major types or domains into subtypes or sub-domains [7]. It should be noted that the goal of creating new taxonomies or developing existing taxonomies is not to make inferences regarding a particular subject; instead, it is to clearly describe, explain, and organize the subject, regardless of the subject's area.

3. THE PROPOSED TAXONOMY

The first draft of the taxonomy is shown in Figure 1 where we classify 27 different research papers by the technique(s) used to identify a file's type. Some papers incorporate hybrid techniques or use two or more complementary approaches to deal with different file types. Thus, the organization of this taxonomy is based on the algorithms or techniques employed to solve the file identification problem. As shown in Figure 1, we use a table format that allows us to show each of the techniques used in a given research paper, while also showing whether the same technique is used by multiple researchers. Each of the papers shown in Figure 1 is listed in Section 6 of this paper and can be identified by the first author's name and the year of publication.

The current broad categories include four major divisions, statistical learning techniques, analysis of frequency distributions and other statistical analysis techniques, along with a category of techniques that focus on detecting and identifying file fragments.

From the papers surveyed so far, we found that a number of researchers have employed statistical learning techniques with several different supervised learning approaches used for classifying file types [1, 2, 5, 6, 9, 13, 15, 25, 29] and two common unsupervised learning algorithms that are used for data reduction [1, 3, 5, 6, 8, 10].

The earliest work on file type identification [21, 22] focused on byte frequency analysis and later research has used byte frequencies as the basis for more complex analysis techniques, such as compressibility and entropy [9, 13, 16, 25].

Several papers performed a range of statistical analysis techniques on file data in an attempt to classify file types by the characteristics of that data [9, 12, 23].

We have separated those techniques that focus specifically on file carving (identifying and extracting file fragments) into a distinct category [10, 11, 14, 24, 26, 28, 29, 30]. In some cases, this research utilizes techniques from other categories, but a number of the papers in this category do not clearly describe how they perform fragment identification.

One of the difficulties in developing a taxonomy of file type identification techniques is the lack of detail in some papers. Another potential problem arises from differences in the terminology some researchers use to describe the technique they employed. We are aware that this may have caused us to categorize the same technique in two different subgroups, but intend to review all of the categories with the goal of correcting duplications such as these in the future.

4. FUTURE WORK

The research papers listed below represent a number of different approaches to the problem of file-type identification. The proposed taxonomy classifies that research according to the algorithms or techniques used for identifying a file's type. Some of the techniques are applicable to the problem of identifying and extracting file fragments (file carving) while others work best on whole, undeleted files.

Future work includes broadening our search for research on file type identification, including scholarly papers, technical reports and theses, and incorporating those sources into the current taxonomy. As new techniques are added, a revision of the organization may be necessary to add new categories or to merge categories that share common features.

5. ACKNOWLEDGEMENT

The authors would like to thank the Institute of Public Administration (IPA) in Saudi Arabia for their support of this research.

6. REFERENCES

[1] Ahmed, Irfan, Kyung-Suk Lhee, Hyun-Jung Shin, and Man-Pyo Hong. "Fast content-based file type identification." In Advances in Digital Forensics VII, pp. 65-75. Springer Berlin Heidelberg, 2011.

[2] Ahmed, Irfan, Lhee Ks, Hyunjung Shin, and M. Hong. "Content-based file-type identification using cosine similarity and a divide-and-conquer approach." IETE Technical Review 27, no. 6 (2010): 465.

[3] Ahmed, Irfan, Kyung-suk Lhee, Hyunjung Shin, and ManPyo Hong. "On Improving the Accuracy and Performance of Content-Based File Type Identification." In Information Security and Privacy: 14th Australasian Conference, ACISP 2009 Brisbane, Australia, July 1-3, Proceedings, vol. 5594, p. 44. Springer, 2009.

[4] Altschaffel, Robert, Stefan Kiltz, and Jana Dittmann. "From the Computer Incident Taxonomy to a Computer Forensic Examination Taxonomy." In Fifth International Conference on IT Security Incident Management and IT Forensics, , pp. 54-68. IEEE, 2009.

[5] Amirani, Mehdi Chehel, Mohsen Toorani, and A. Beheshti. "A new approach to content based file type detection." In Computers and Communications, 2008. ISCC, IEEE Symposium on, pp. 1103-1108. IEEE, 2008

[6] Amirani, Mehdi Chehel, Mohsen Toorani, and Sara Mihandoost. "Feature-based Type Identification of File Fragments." Security and Communication Networks, vol. 6, no. 1 (2013): pgs. 115-128.

[7] Axelsson, Stefan. "Intrusion detection systems: A survey and taxonomy", Vol. 99, No. 15. Technical report, Chalmers University of Technology, 2000

[8] Balakrishnama, Suresh, and Aravind Ganapathiraju. "Linear discriminant analysis-a brief tutorial." Institute for Signal and information Processing (1998).

[9] Beebe, N., L. Maddox, Lishu Liu, and Minghe Sun. "Sceadan: Using Concatenated N-Gram Vectors for Improved File and Data Type Classification." (2013): 1-1.

[10] Calhoun, William C., and Drue Coles. "Predicting the types of file fragments." Digital Investigation 5 (2008): S14-S20.

[11] Cohen, Michael I. "Advanced JPEG carving." In Proceedings of the 1st international Conference on Forensic Applications and techniques, p. 16. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008.

[12] Erbacher, Robert F., and John Mulholland. "Identification and localization of data types within large-scale file systems." In Systematic Approaches to Digital Forensic Engineering, (SADFE). Second International Workshop on, pp. 55-70. IEEE, 2007.

[13] Fitzgerald, Simran, George Mathews, Colin Morris, and Oles Zhulyn. "Using NLP techniques for file fragment classification." Digital Investigation 9 (2012): S44-S49

[14] Garfinkel, Simson L. "Carving contiguous and fragmented files with fast object validation." Digital Investigation 4 (2007): 2-12.

[15] Gopal, Siddharth, Yiming Yang, Konstantin Salomatin, and Jaime Carbonell. "Statistical learning for file-type identification." In Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on, vol. 1, pp. 68-73. IEEE, 2011.

[16] Hall, Gregory, Davis, Wilbon. "Sliding Window Measurement for File Type Identification", Technical Report, ManTech Security, 2006

[17] Hand, Scott, Zhiqiang Lin, Guofei Gu, and Bhavani Thuraisingham. "Bin-Carver: Automatic recovery of binary executable files." Digital Investigation 9 (2012): S108-S117.

[18] Karresand, Martin, and Nahid Shahmehri. "File type identification of data fragments by their binary structure." In Information Assurance Workshop, pp. 140-147. IEEE, 2006.

[19] Karresand, Martin, and Nahid Shahmehri. "Oscar—file type identification of binary data in disk clusters and ram pages." In Security and privacy in dynamic environments, pp. 413-424. Springer US, 2006.

[20] Landwehr, Carl E., Alan R. Bull, John P. McDermott, and William S. Choi. "A taxonomy of computer program security flaws." ACM Computing Surveys (CSUR) 26, no. 3 (1994): 211-254.

[21] Li, Wei-Jen, Ke Wang, Salvatore J. Stolfo, and Benjamin Herzog. "Fileprints: Identifying file types by n-gram analysis." In Information Assurance Workshop, IAW'05. Proceedings from the Sixth Annual IEEE SMC, pp. 64-71. IEEE, 2005.

[22] McDaniel, Mason, and Mohammad Hossain Heydari. "Content based file type detection algorithms." In System Sciences, Proceedings of the 36th Annual Hawaii International Conference on, pp. 10-pp. IEEE, 2003.

[23] Moody, Sarah J., and Robert F. Erbacher. "SADI-statistical analysis for data type identification." In Systematic Approaches to Digital Forensic Engineering, SADFE'08. Third International Workshop on, pp. 41-54. IEEE, 2008.

[24] Pal, Anandabrata, and Nasir Memon. "The evolution of file carving." Signal Processing Magazine, IEEE 26, no. 2 (2009): 59-71.

[25] Penrose, Philip, Richard Macfarlane, and William J. Buchanan. "Approaches to the classification of high entropy file fragments." Digital Investigation (2013).

[26] Povar, Digambar, and V. K. Bhadran. "Forensic data carving." In Digital Forensics and Cyber Crime, pp. 137-148. Springer Berlin Heidelberg, 2011.

[27] Richard III, Golden G., and Vassil Roussev. "Scalpel: A Frugal, High Performance File Carver." In DFRWS. 2005.

[28] Richard III, Golden, Vassil Roussev, and Lodovico Marziale. "In-place file carving." In Advances in Digital Forensics III, pp. 217-230. Springer New York, 2007.

[29] Sportiello, Luigi, and Stefano Zanero. "Context-Based File Block Classification." In Advances in Digital Forensics VIII, pp. 67-82. Springer Berlin Heidelberg, 2012.

[30] Veenman, Cor J. "Statistical disk cluster classification for file carving." In Third International Symposium on Information Assurance and Security, pp. 393-398. IEEE, 2007.

Domains	SubDomains	Techniques	Ahmed (2011)	Ahmed (2010)	Ahmed (2009)	Amirani (2008)	Amirani (2013)	Balakrishnama (1998)	Beebe (2013)	Calhoun (2008)	Cohen (2008)	Erbacher (2007)	Fitzgerald (2012)	Garfinkel (2007)	Gopal (2011)	Hall (2006)	Hand (2012)	Karresand (2006)	Karresand (2006)	Li (2005)	McDaniel (2003)	Moody (2008)	Pal (2009)	Penrose (2013)	Povar (2011)	Richard III (2005)	Richard III (2007)	Sportiello (2012)	Veenman (2007)
			1	2	3	5	6	8	9	10	11	12	13	14	15	16	17	18	19	21	22	23	24	25	26	27	28	29	30
Statistical Learning	Supervised Learning	Support Vector Machine (SVM)	X				X		X				X		X													X	
		k-Nearest Neighbor (kNN)	X												X														
		Neural Network (NN)	X	X		X	X																	X					
	Unsupervised Learning	Principle Component Analysis (PCA)				X	X																						
		Linear Discriminant Analysis (LDA)	X		X			X		X																			
Frequency Distribution		Byte Frequency Distribution (BFD)							X												X								
		Byte Frequency Cross-Correlation Analysis (BFA)																			X								
		N-gram							X											X									
		Oscar																X	X									X	
		Entropy							X				X			X													
		Compressibility							X							X								X					
		Cosine Similarity		X	X																								
		Single-Centroid																		X									
		Multi-Centroid																		X									
		Exemplar Files																		X									
		NIST Statistical Test																						X					
Statistical Analysis		Average							X			X										X							
		Kurtosis							X			X										X							
		Standard Deviation							X			X										X							
		Distribution of Averages										X																	
		Distribution of Standard Deviations										X																	
Detection of File Fragments (File Carving)		Header/Footer												X							X				X				
		Header/embedded Length Carving												X											X					
		File Structure Based Carving																					X		X				
		Bifragment Gap Carving (Object Validation)												X															
		Header/Maximum File Size												X											X				
		Longest Common Subsequences (LCS)								X																			
		File Block Classification																										X	
		Two-Class and Multi-Class																											X
		Bags of Words											X																
		Scalpel (Header/Footer																								X	X		
		JPEG Carving (Image Processing)									X																		
		Bin-Carving															X												

Figure 1: A Taxonomy of File-type Identification and File Carving Techniques