Reconocimiento flexible de formatos de archivos para el descubrimiento de evidencia digital

Gabriel Casullo[(1)], Erika Givaudant[(1)], Leandro Tami [(1)], Leopoldo Sebastián Gómez [(2)]

[(1)] Facultad de Ingeniería
Universidad FASTA
info@casu-net.com.ar – erikalaura@gmail.com – leandro.tami@gmail.com

[(2)] Laboratorio Pericial Informático
Poder Judicial del Neuquén
sebastian.gomez@jusneuquen.gov.ar

CONTEXTO

Esta línea de investigación es definida a partir de una necesidad detectada en el área de especialidad de informática forense. Se pretende contribuir a profundizar el análisis forense para el descubrimiento de evidencia digital mediante el reconocimiento flexible del formato de archivos a partir de sus características estructurales y de contenido.
El proyecto en curso contribuye a fortalecer los vínculos de cooperación entre la universidad y el ámbito profesional. Concretamente se apoya un trabajo académico de grado impulsado desde la Facultad de Ingeniería de la Universidad FASTA en colaboración con el Laboratorio Pericial Informático del Poder Judicial de la Provincia del Neuquén.

RESUMEN

Se presentan avances en el estudio de un método computacional robusto para la identificación del formato de archivos, con capacidades especiales para la detección de aquellos formatos que puedan tener daños mínimos por fallas en el medio de almacenamiento o transmisión, o bien cuya estructura sea modificada intencionadamente con el objeto de obstruir una investigación digital. En una primera etapa se compara el encabezado del archivo a identificar con una base de datos de firmas de formatos de archivo conocidos utilizando una métrica de distancia y confeccionando un ranking mediante una lista de formatos candidatos con aquellos que obtuvieron las menores distancias. La segunda etapa tiene por objeto profundizar el análisis del contenido de archivos, extrayendo metadatos y características para una posterior aplicación de técnicas de inteligencia computacional que permitan aumentar la precisión de la identificación. Los métodos utilizados son implementados y evaluados mediante experimentación sobre diferentes corpus digitales.

Palabras clave: *magic numbers, header obfuscation, transmogrify, file types, file fragments, content-based file type identification.*

1. INTRODUCCION

Los archivos digitales tienen el potencial de servir de evidencia inculpatoria o exculpatoria y pueden aportar elementos de prueba relevantes para el esclarecimiento de un hecho delictivo. Resulta de vital importancia poder identificar el formato de los archivos que se someten a análisis forense para conocer preliminarmente mayores detalles sobre su estructura y contenido. Por ello, durante el desarrollo de actividades periciales esta tarea es uno de los primeros exámenes que el investigador realiza sobre el material probatorio.

En condiciones normales la identificación consiste simplemente en observar la extensión de un archivo para obtener a priori una primera descripción general de su contenido, pero dichos atributos son poco fiables ya que

pueden ser fácilmente modificados. Las herramientas forenses actuales permiten la identificación automática del formato de un archivo mediante un procedimiento habitualmente reseñado como "análisis de firmas". Este método computacional es independiente de la extensión del archivo, y consiste en verificar si un conjunto de bytes pertenecientes al encabezado del mismo está presente en una base de datos de firmas ya conocidas para diversos formatos de archivo.

A pesar de ser de gran ayuda para la detección de ocultamientos de archivos mediante modificaciones en la extensión, las firmas de archivo son también vulnerables a cambios accidentales o intencionales en el encabezado, y en estos casos dichas herramientas forenses no tolerantes a errores fracasan en su identificación. El "pattern matching" puede fallar en aquellos casos que el encabezado haya sido alterado aunque sea por pequeños cambios a nivel de bit. Por ello resulta necesario el desarrollo de un método de reconocimiento más flexible que permita obtener un espectro de formatos "candidatos", posibilitando al investigador intentar el acceso a la información digital contenida en el archivo a través de la aplicación que sea más apropiada para la manipulación de los mismos.

2. LINEAS DE INVESTIGACION y DESARROLLO

Actualmente existen diferentes enfoques científicos para el reconocimiento automático de formatos de archivos. Se han estudiado métodos de identificación basados en el contenido [1], incluso cuando el repositorio digital está corrupto, utilizando funciones de extracción jerárquicas que sirvan al PCA (Principal Component Analysis) juntamente con un entrenamiento de redes neuronales auto-asociativas, obteniendo porcentajes muy favorables en el reconocimiento.

Otros investigadores [2] han optado abordar esta problemática mediante un método computacional de clasificación de fragmentos de archivos, considerando aspectos de su estructura interna para una identificación efectiva. En algunas situaciones esto demuestra no ser una tarea sencilla, ya que el reconocimiento suele basarse en la aparición de patrones pero a veces los archivos carecen de dichas señas identificadoras por la aplicación de métodos de compresión o cifrado sobre su contenido. En el recorrido que llevan a cabo de los métodos computacionales disponibles se destaca la librería *libmagic*, base del comando *file* de UNIX. Dicha utilidad es efectiva para la identificación de archivos completos, pero no tiene los resultados esperados en el tratamiento de fragmentos. Una aproximación a la solución es el análisis de distribución de secuencias de bytes, también conocido como “análisis de 1-gramo”. Este método consiste en construir un histograma con cada carácter ASCII posible detectado en el archivo. Estas frecuencias constituyen una huella o “fingerprint”, y la pertenencia a un tipo específico de archivo se determina según la distancia entre la huella generada y las de una base de datos de huellas conocidas.

Desde otra perspectiva de análisis estadístico se han propuesto dos algoritmos para predecir el tipo de archivo a partir del análisis de un fragmento perteneciente a dicho contenedor digital [3]: el primero está basado en el discriminante lineal de Fisher mientras que el segundo utiliza las subcadenas y subsecuencias comunes de caracteres de un fragmento. El discriminante lineal de Fisher es un método estadístico utilizado para clasificar individuos en grupos. Hay una fase de entrenamiento en la cual los datos de individuos pertenecientes a grupos conocidos se utilizan para armar un modelo de clasificación. Este modelo está compuesto por un conjunto de funciones lineales, una para cada grupo. A partir de esta información, el modelo es capaz de predecir el grupo al que pertenece un individuo de acuerdo a la función de clasificación que devuelva el valor más alto. El algoritmo basado en subcadenas y subsecuencias de caracteres comunes presupone que dos archivos del mismo tipo probablemente tengan en común cadenas de caracteres más largas comparados con otros archivos de diferente estructura o contenido. Los resultados empíricos corroboran tales

afirmaciones aún en archivos con ausencia de metadatos.

Se han presentado trabajos científicos [4] que promueven el uso de la métrica de similaridad del coseno como evaluador de distancia en las funciones de extracción de patrones de bytes y el uso de un esquema que aplica pasos recursivos en clusters para la identificación de archivos, obteniendo resultados satisfactorios.

3. RESULTADOS OBTENIDOS/ESPERADOS

Como inicio de las actividades de investigación se intentó un abordaje inicial a través del reconocimiento del *header* de los archivos para obtener una mejora sobre el ya conocido "análisis de firmas". En una primera instancia se procuró la localización de bases de datos de firmas identificadoras de archivos. Se utilizó como estructura básica de información la base de firmas *TrID* [5] que cuenta con 4000 firmas de distintos formatos de archivo. Se hicieron pruebas de medición de distancias entre los *headers* de archivos de prueba y las contenidas en la base de firmas anterior utilizando la distancia de Levenshtein. Esta métrica es de gran utilidad ya que permite medir la distancia de entradas de longitudes desiguales. Pese a ello, existe una gran proporción de falsos positivos debido a que la base de datos de *TrID* contiene muchos tipos de archivos comparten los mismos encabezados.

FIRMA	TIPO	EXTENSION
21 42 44 4E	Outlook Exchange Offline Storage	OST
21 42 44 4E	Microsoft Personal Address Book	PAB

Luego de una depuración se obtuvieron 2734 firmas únicas y se implementó un analizador básico para el reconocimiento del formato de archivos. Los *headers* repetidos corroboran que dicho análisis no es suficiente para la identificación del formato de un archivo desconocido y que son necesarias nuevas fuentes de información, como por ejemplo la presencia o ausencia de metadatos de formatos de archivo conocidos. La identificación de archivos con la herramienta forense se comportó satisfactoriamente aunque es intolerante a errores, equivalentemente a lo que sucede con la utilidad *file* de sistemas UNIX.

Para las tareas de experimentación se generaron lotes de prueba mediante la implementación de una aplicación informática complementaria. Dicha aplicación utiliza un diccionario y realiza una selección de palabras en forma aleatoria para luego realizar consultas automáticas sobre los buscadores *Google*, *Bing* y *FindThatFile* con un formato requerido pasado como parámetro. Para cada resultado de la consulta la aplicación accede desde el enlace al sitio donde se aloja el archivo y lo descarga. Finalmente se genera una lista de las extensiones de los archivos obtenidos. Para contar con un corpus digital sin falsos positivos, se corroboró que la extensión de los archivos descargados se correspondiera con sus tipos verdaderos. Para ello, se usó como valor de referencia la salida retornada por la herramienta *file* y se estableció una paridad entre la salida de dicha herramienta y los tipos MIME correspondiente a cada formato de archivo.

La experiencia aportada desde la práctica profesional ha determinado la necesidad de concentrar la efectividad de la aplicación forense para la identificación de formatos de archivos hacia aquellos que habitualmente son más proclives a contener evidencia digital. Se han priorizado formatos específicos de imagen (png, jpg, bmp), de video (avi, mpg), audio (wav , wma, mp3), de procesamiento de texto y planillas de cálculo (pptx, ppt, doc, docx, xls, xlsx), de base de datos (mdb) y otros compuestos en formatos propietarios (pdf, psd).

A partir del corpus digital de archivos a identificar y la lista depurada de firmas de formato de archivos que han sido determinadas como prioridad para el análisis forense, se desarrollaron herramientas para medir la distancia entre los n primeros bytes de un archivo y cada una de las firmas de la base. Si A y B son dos firmas de longitudes diferentes

y A es un prefijo de B, se considerará más representativa a la firma de mayor longitud. Para ello, se define el 'ratio' como la relación entre distancia y longitud, y se utiliza este criterio para el ordenamiento de los resultados de la identificación.

Complementariamente se implementó una herramienta forense que efectúa dicho reconocimiento sobre grandes lotes de archivos, incorporando errores aleatorios en los encabezados. Cada identificación produce un tipo MIME que es comparado con el que se obtiene al utilizar el comando *file*, y se determina automáticamente si el resultado ha sido correcto o si no se logra el reconocimiento del formato.

De los resultados obtenidos de la experimentación se verifica que: a) Un daño de 1 byte en una firma corta, puede ser demasiado significativo como para permitir la identificación de un formato. Por ejemplo, un daño de 1 byte en la firma de un archivo BMP, representa un 50% de daño, lo cual invalida la identificación con éste método. b) Existen tipos de archivo que comparten la misma firma, por ejemplo ZIP y JAR. El método utilizado por el analizador básico solo acota el conjunto de posibles formatos a la hora de identificarlos.

Ambas aseveraciones hacen necesaria una segunda etapa de identificación que tendrá que basarse en el contenido del archivo, considerando la extracción de características particulares y la búsqueda de presencia de metadatos de formatos conocidos.

Utilizando la herramienta *exiftool* [6] es posible acceder a los metadatos disponibles en una amplia variedad de tipos de archivos. Si bien esta herramienta se especializa en EXIF (EXchangeable Image Format), un estándar para especificación de formatos de archivos de imagen, es viable su utilización para extraer metadatos de otros tipos de archivo que no son imágenes. Se trabaja para integrar un filtro adicional en el analizador básico a través de metadatos de formatos de archivos conocidos. Asimismo se prevé verificar la presencia de datos en posiciones específicas dentro de cada archivo, comprobando que estos se encuentran dentro del rango de valores válidos y tengan el tipo adecuado para cada formato soportado.

Como corolario de las primeras experimentaciones ha de mencionarse que el análisis del *header* como única métrica de evaluación es insuficiente para lograr resultados satisfactorios en la identificación del formato de archivos. Se ha comprobado mediante el uso de una base de datos de firmas utilizada como esquema básico de validación que un mismo *header* puede repetirse en distintos tipos de archivos. Para subsanar los problemas de ambigüedad sobre algunos tipos de archivos mediante evaluaciones sobre del *header* se ha decidido hacer uso de MIME como valor estándar de comparación.

La investigación en curso se ha planteado en etapas que conllevan la implementación de métodos computacionales que actúan como clasificadores combinados para la identificación del formato de archivos. La estrategia propuesta consiste en integrar las siguientes líneas de análisis forense: a) Identificación mediante el *header,* b) Identificación mediante metadatos, c) Identificación mediante la tasa de compresión. Finalmente se prevé integrar dichos predictores mediante métodos de inferencia con lógica difusa. Se procura que el analizador reconozca el formato de un archivo con un porcentaje de verosimilitud, y los métodos implementados en las etapas subsiguientes de la presente investigación reafirmen ese reconocimiento.

El objetivo principal de este proyecto de investigación es el estudio de un método computacional flexible capaz de poder indicar con un grado de verosimilitud el formato de un archivo desconocido o sospechoso, eventualmente sin extensión y/o con alguna pequeña modificación en su estructura interna, a fin de poder acceder a su contenido mediante la aplicación nativa que lo generó.

Concretamente se desea implementar una herramienta que asista a un profesional forense en el reconocimiento del formato de un archivo, para aquellos casos en que el mismo

tenga pequeños cambios internos o en la extensión, como producto de un daño lógico eventual o bien provocado intencionalmente para intentar obstruir una investigación digital.

Para realizar pruebas de performance en el reconocimiento de formatos de archivos, preliminarmente se planteó utilizar un corpus digital generado a partir de archivos extraídos aleatoriamente desde Internet con consultas automáticas sobre diversos buscadores. Se pretende complementariamente realizar una experimentación sobre archivos extraídos desde discos rígidos en uso, en aras de simular escenarios más próximos a la realidad de trabajo de una pericia informática.

4. FORMACION DE RECURSOS HUMANOS

La línea de investigación presentada sirve de apoyo a una tesis de grado de tres alumnos pertenecientes a la carrera Ingeniería en Informática de la Facultad de Ingeniería de la Universidad FASTA bajo la dirección funcional de un especialista en informática forense del Laboratorio Pericial Informático del Poder Judicial de la Provincia de Neuquén

5. REFERENCIAS

[1]. Amirani, M.C.; Toorani, M.; Beheshti, A., "A new approach to content-based file type detection", *Computers and Communications, 2008. ISCC 2008. IEEE Symposium on* , pp.1103,1108, 6-9 July 2008.
[2]. Roussev, V.; Garfinkel, S.L., "File Fragment Classification - The Case for Specialized Approaches", *Systematic Approaches to Digital Forensic Engineering, 2009. SADFE '09. Fourth International IEEE Workshop on* , pp.3,14, 21-21 May 2009.
[3]. Calhoun, W., Coles, D, "Predicting the types of file fragments", *Digital Investigation: The International Journal of Digital Forensics & Incident Response*, vol. 5, pp.14-20, September 2008.
[4]. Amhed, I., Lhee, K., Shin, H. and Hong, M., "On Improving the Accuracy and Performance of Content-Based File Type Identification", *Information Security and Privacy, Lecture Notes in Computer Science*, vol. 5594, pp 44-59, 2009.
[5]. TrID by Marco Pontello, http://mark0.net/soft-trid-deflist.html (Último acceso: 09/03/2013)
[6]. ExifTOOL by Phil Harvey, http://www.sno.phy.queensu.ca/~phil/exiftool/ (Último acceso: 09/03/2013)