C6 Testing Digital Forensic String Search Tools

James R. Lyle, PhD, National Institute of Standards & Technology, 100 Bureau Drive, MS 8970, Gaithersburg, MD 20899; and Barbara Guttman, BA, National Institute of Standards & Technology, Mail Stop 8970, Gaithersburg, MD 20899-8970*

After attending the presentation, attendees will be aware of test data for string search tools and of some of the limitations and constraints of testing computer forensic string search tools. Test data and test method documentation is available via the Computer Forensics Tool Testing (CFTT) and Computer Forensic Reference Data Sets (CFReDS) websites (www.cftt.nist.gov and www.cfreds.nist.gov).

This presentation will impact the forensic science community by increasing awareness of tool test strategies and the ability of tool testing to reveal anomalies in string search tool behavior. This presentation will aid the forensic practitioner in recognizing the limitations of testing string search forensic tools and in being aware of the implications of choices of what to test or not test. The goal of testing forensic tools by a forensic laboratory is not to prove the software is always correct but to show evidence that the software is appropriate for the task at hand.

The CFTT project at the National Institute of Standards and Technology (NIST) develops methodologies for testing digital forensic tools. Currently, there are CFTT methodologies for testing the following: disk imaging, write blocking, deleted file recovery, file carving, forensic media preparation, and mobile devices.

A variety of forensic tools in each of these categories have been tested and observed flaws have been documented and reported by the Department of Homeland Security (DHS) and the National Institute of Justice (NIJ). These results can be used as a basis for identifying the types of likely failures that occur in forensic tools.

At an abstract level, string searching involves the following: (1) something to search with (i.e., a search engine); (2) someplace to search (i.e., an image file or a digital storage device); (3) something to search for; and, (4) search results (presented in a useful way).

A search engine implements a search algorithm that performs the search. A digital forensic string search tool provides an interface between a user and a search engine. The search tool interfaces with at least one search engine, but may interface with additional search engines. Most tools make a pass over the data and construct an index of strings that might be searched for or scan the entire data set for each search. Other tools may allow a user to select the search algorithm to use.

A search tool uses text strings to identify files relevant to an investigation. In addition to active files, a search tool may also need to search deleted files and unallocated space.

In the simplest case, the user is looking for a match to a target search string. Sometimes the tool user has a case-specific list of search terms. In other cases, the user wants the tool to find social security numbers (i.e., groups of nine digits). This can be specified as a regular expression (i.e., a pattern) such as [0-9]{9,9} (a string of nine digits with no separators). In addition, the user might need to search for text that is not represented in ASCII, such as searching for a Chinese word. There are multiple possible encodings for the characters (e.g., UNICODE, GB, Big 5, SHIFT JIS, etc.).

Forensic search tools often have a rich set of search parameters that could be tested. In the design of this test method and test data, this study focused on what seemed to be the most useful features in general. For an individual laboratory, other selections might be a better fit. Some common search parameters addressed in the NIST search data sets include: whole word versus substring; match case versus ignore case; character representation — ASCII versus UNICODE; active file versus deleted file versus unallocated space; exact match versus pattern match; clear text (.txt) versus formatted (.doc or .html); and, indexed search versus live search.

In summary, this presentation will describe a publicly available data set for testing forensic string search tools, including search features that can be tested, how to create test data, and the results of applying the test data set to several commonly used forensic tools.

Digital Forensics, Tool Testing, String Search

*Presenting Author