B2 A Narratological Approach to Digital Forensic Analysis Utilizing Natural Language Toolkit

Mark Pollitt, MS, Daytona State College, 1770 Technology Blvd, Daytona Beach, FL 32117*

After attending this presentation, attendees will gain an understanding of how narratology and natural language processing can be applied to the analysis of digital evidence.

This presentation will impact the forensic science community by providing results of initial experimentation with the application of the theory of narratology, coupled with the Natural Language Toolkit (NLTK), to a forensic corpus.

Digital forensics has made tremendous strides in the acquisition and preservation of electronically stored evidence. As stored data has grown ever larger and more complex, the ability to identify files and data that are investigatively significant and legally probative has not kept pace with this growth. The traditional approach to digital forensics has relied upon the use of metadata such as file system dates and times, classification of files by data type, and simple string searches. With the advent of ever larger storage media, the ability of digital forensic examiners to identify information of investigative or probative value has become less efficient. In earlier work the author has described the concept of narrative as relates to the search for meaning in digital evidence.[1]

Narratology has been defined as: "the ensemble of theories of narratives, narrative texts, images, spectacles, events; cultural artifacts that "tell a story."[2] Digital forensics seek to identify and tell the investigative "story" as it relates to the case under investigation. One part of the narratological theory defines the elements that make up a narrative, such as, actors, events, and chronology. If textual material can be processed in such a way as to identify these elements, then it may be possible to extract the "story" from the text and, by extension, approximate the probative "story." This presentation will explore how this theory can be used to develop criteria for the automated processing of textual material to increase its investigative value, and how using natural language processing tools could be used to assist in the identification of these elements.

In order to test this hypothetical approach, a series of experiments, utilizing the Enron email corpus will be conducted, using a number of procedures included in the Natural Language Toolkit.[3,4] This corpus is a collection of the actual emails sent by employees in the criminal trial of several Enron employees that was admitted into evidence during their trial and was subsequently released to the public by the courts. This corpus was selected since the data is in textual format, the metadata from the email headers has been converted to text, and the "story" is known. The results of these experiments will be reported and an assessment of their value to digital forensic examination will be made.

References:

1. Pollitt, Mark. "Digital Forensics as a Surreal Narrative." Advances in Digital Forensics V: Fifth IFIP WG 11.9 International Conference on Digital Forensics, Orlando, Florida, USA, January 26-28, 2009, Revised ... in Information and Communication Technology). Springer, 2009.
2. Bal, Mieke. Narratology: Introduction to the Theory of Narrative, Third Edition. 3rd ed. University of Toronto Press, Scholarly Publishing Division, 2009. 3.
3. "SGI.nu >> Enron Email Corpus." Web. 21 Feb. 2012.
4. Bird, Steven, Ewan Klein, and Edward Loper. Natural Language Processing with Python. 1st ed. O'Reilly Media, 2009.

Narratology, Language Processing, NLTK

 ** Presenting Author*