Chapter 13
Hypothesis Generation and Testing in Event Profiling for Digital Forensic Investigations

Lynn Batten
Faculty of Science and Technology, School of IT, Deakin University, Burwood, Melbourne, VIC, Australia

Lei Pan
Faculty of Science and Technology, School of IT, Deakin University, Burwood, Melbourne, VIC, Australia

Nisar Khan
Faculty of Science and Technology, School of IT, Deakin University, Burwood, Melbourne, VIC, Australia

ABSTRACT

The need for an automated approach to forensic digital investigation has been recognized for some years, and several authors have developed frameworks in this direction. The aim of this paper is to assist the forensic investigator with the generation and testing of hypotheses in the analysis phase. In doing so, the authors present a new architecture which facilitates the move to automation of the investigative process; this new architecture draws together several important components of the literature on question and answer methodologies including the concept of 'pivot' word and sentence ranking. Their architecture is supported by a detailed case study demonstrating its practicality.

INTRODUCTION

In practice, digital forensics is carried out with the aim of extracting evidence which will be tenable in a court of law (Carrier, 2006; Willassen, 2008). A stream of research work in the last decade has attempted to assist the forensic investigator in moving from the historically manual approach towards an automated, and therefore also reproducible, approach to the discovery of digital evidence (Batten & Pan, 2011; Jankun-Kelly, Wilson, Stamps, Franck, Carver, & Swan, 2009; Marrington, Mohay, Morarji, & Clark, 2010; Pan, Khan, & Batten, 2012). Carrier (2006) and

DOI: 10.4018/978-1-4666-4006-1.ch013

Marrington (2009) both developed automated methods of describing a computer system and its activity over a fixed period of time; the former focused on the raw data while the latter focused on events surrounding a crime. Both authors look for relationships between the objects they are examining. The work of Batten and Pan (2011) and Pan, Khan, and Batten (2012) extends the work of both Carrier (2006) and Marrington (2009) by demonstrating how relationships between the objects of investigation can be used to reduce the size of the data set needing analysis and so speed up the investigation time.

All of Batten and Pan (2011), Carrier (2006), Marrington (2009), and Pan, Khan, and Batten (2012) develop extensive methodologies for relationship building. Carrier (2006) gives examples of hypotheses which can be formulated and tested; however, he does not attempt to define the word hypothesis. The authors of Al-Zaidy, Fung, Youssef, and Fortin (2012) use a similar method of relationship building and develop 'hypotheses' in the form of relationships between people and data; however, again, the authors do not define formally what they mean by a hypothesis.

An important contribution of Pan, Khan, and Batten (2012) is a formal definition of hypothesis in the context of digital forensic investigation and an illustration of how the theoretical formulation is able to find relationships from which hypotheses can be developed and examined. In this paper, we move to a new level in investigating the relevance of hypotheses to the situations at hand. We continue to automate the analysis as much as possible in order to apply rigor to the methodology and to provide the ability to replicate the methodology as needed for the court.

First, we describe the relevant literature. The section afterwards contains formal definitions and notations needed to illustrate our subsequent work and we discuss the hypothesis generation and testing methods in detail. A case study is presented and analyzed next; this case study is a continuation of that used in Batten and Pan (2011) and Pan, Khan, and Batten (2012). Finally, we summarize the implications of our work and consider its impact on the future research literature in this area.

RELATED WORK

The paper by Radev, Prager, and Samn (2000) deals with answering natural language questions. In this paper the authors introduce a method called 'predictive annotation' which highlights phrases in the text in advance by assigning labels to make these phrases the targets of a particular question. When dealing with Natural Language Processing, there are many questions that contain words which, when searched for in a corpus may not be returned or answered as they may not exist in the corpus. The concept of 'predictive annotation' was introduced in Radev, Prager, and Samn (2000) precisely to deal with such situations. This method works in two steps. In the first step, the question under consideration is enhanced by assigning labels (which the authors call QA-Tokens) to a set of 'recognized' objects such as places or persons; then, the text in the corpus is labelled with the same QA- Tokens for all recognized objects. Finally, the QA-Tokens in the question are searched for in the corpus to locate matching passages. The system architecture for this step has two components: the Information Retrieval component returns a list of 10 short passages containing a large number of potential answers for each query, and the Answer Selection component which ranks the potential answers using the two algorithms AnSel and Werlec. Both algorithms will return five text passages per query that contain the possible answers. This means that, in the initial stage, no single answer will be returned but rather a list of possible or potential answers.

In the second step, the answer selection process inputs the matching passages identified in the first step and ranks them. In ranking, a weighting