Using Hypothesis Generation in Event Profiling for Digital Forensic Investigations

L. Pan, N. Khan and L. Batten

School of IT, Deakin University, Melbourne, Australia
e-mail: {l.pan; nakh; lmbatten}@deakin.edu.au

Abstract

The traditional manual approach to the investigation of digital data is no longer feasible as the amount of data which can be saved on hard drives grows out of control. In addition, it is usually necessary to consider data across extensive networks of devices in order to obtain a realistic picture of an investigation and ensure that no evidence is overlooked. The need for an automated approach to forensic digital investigation has therefore been recognized for some years, and several authors have developed frameworks in this direction. The aim of this paper is to enhance and move beyond current work by focusing on hypothesis generation in the later part of the analysis phase. In doing so, we present, for the first time in this context, a formal definition of the word 'hypothesis' and also present an extensive case study to illustrate its usefulness and the method of hypothesis generation and analysis. The scientific approach taken here to hypothesis generation directly supports the investigation procedure and also promotes its acceptance by a court of law.

Keywords

Digital forensics, Hypothesis generation, Confidence level.

1 Introduction

Digital forensics is the investigation of an event based on digital information and is often undertaken with the aim of extracting evidence which will be tenable in a court of law (Carrier 2006) and (Willassen 2008). The last decade has seen an influx of research work designed to assist the forensic investigator to move from the historically manual approach towards an automated, and therefore also reproducible, approach to the discovery of digital evidence (Batten and Pan 2011), Marrington et al. 2010), and (Jankun-Kelly et al. 2009). Automated methods rely on a logical and consistent analysis from which conclusions can be drawn. (Carrier 2006) and (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) extended the work of both Carrier and Marrington by demonstrating how relationships between the objects of investigation could be used to reduce the size of the data set needing analysis (and so speed up the investigation time) and also by developing a visualization technique of the analysis to assist the investigation team.

All of (Carrier 2006), (Marrington 2009) and (Batten and Pan 2011) 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 in the digital forensic context. The authors of (Al-Zaidy et al. 2012) use a similar method of relationship building to find communities of criminals by examining documents seized from suspect computers. Their Algorithm 2 develops 'hypotheses' in the form of relationships between people and data; however, the authors do not define formally what they mean by a hypothesis.

The contributions of the current paper are: 1) to formally present a definition of hypothesis in the context of digital forensic investigation and 2) to illustrate, via an extensive case study, how our theoretical formulation is able to find relationships from which hypotheses can be developed and examined. The impact of our work is first of all to support the investigation by formalizing a procedure which results in the generation of hypotheses which are most likely to be relevant, and secondly to provide the investigator with a formal process presentable to a court of law on the basis of its scientific approach and reproducibility.

In Section 2, we describe the relevant literature. Section 3 contains formal definitions and notations needed to describe our subsequent work. A case study is presented and then analyzed in Section 4. Finally, in Section 5, we summarize the implications of our work on the future research literature in this area.

2 Related Work

Using the hypothesis generation concept, (Chen 1996) proposes a business intelligence framework to improve fraud detection and intelligence decision analysis. In this framework, each potential hypothesis is subjected to three tests: there is first an adjustment to eliminate anomalies; abductive reasoning is then applied taking the context into consideration and results in adaptation yet again; finally, a conflict resolution model checks for contradictions. Chen's framework works well on a business information system, but is not easily adaptable to a digital forensic investigation involving multiple and variable sources of information.

In medical diagnosis, a medical doctor attempts to determine the actual cause of some symptoms. Much attention has been paid to this in the recent research literature where, for instance, Bayesian network models were applied in (Yang 2010), conformal predictors framework in (Lambrou et al., 2011), and clustering discrimination in (Chang et al., 2010). In addition, (Webster et al. 2010) propose a hybrid model which combines multiple information sources and improves the quality of the hypotheses generated. However, the above models are not suitable for our research problem for several reasons. Firstly there is usually no tight deadline for results while speedy solutions are often required in a forensic investigation. Secondly, medical models require input of a uniformed set of data in a standardized format, such as blood test data, patient history and so on, while such standardized data is not available in digital forensics. Thirdly, the above models assume that medical data is from reliable sources and can be taken at face value, but digital forensic investigators often encounter data which has been deliberately manipulated to avoid detection or usefulness.

A popular application of hypothesis generation is found in question/answering (QA) games. QA games may ask a question and request an answer, or provide an answer and ask for a question which leads to that answer. For example, IBM's ***Watson***, is a sophisticated QA system which generates hypotheses of high quality to pose as questions for a given answer (Ferrucci et al. 2009). Watson parses a given answer into a search tree, generates alternative questions as hypotheses, and looks for possible solutions from a given corpus of data. When Watson finds a candidate question, it assigns a confidence level and retains the question as a possible solution if the confidence level is over a certain threshold (Ferrucci et al. 2010). Similar to Watson, (Chen and Garcia 2010) generates semantically valid hypotheses by assessing the semantic quality of a dataset while (Di Lecce and Calabrese 2012) generates hypotheses based on the observations of automated learning results of applying syntactic pattern recognition. These natural language processing approaches successfully rely on data redundancy and explicit rules in natural languages but less ably handle digital evidence of a great variety of complexity and uncertainty.

Utilizing the finite-state machine concept, (Carrier 2006) proposes that investigators should formulate hypotheses to answer questions about the states of events and verify that observations match actual data. Carrier's approach requires all information and resources relevant to the investigation to be collected and observed. The limitation of this approach is that finite-state machines do not scale well and become error-prone as the volume of evidence increases. In this paper, Carrier does not mention the concept of relationship and its application to hypothesis generation.

In (Marrington 2009), the author proposes an automated process to describe a computer system and its activity for computer forensic investigation. He explains that a computer profile consists of finite sets of objects, relationships, the times in the history of the computer system and the events. This work complements existing activities in digital investigations by producing a formal description of a computer system and facilitates the formulation of hypotheses by the investigator about the computer system's activity (Marrington et al. 2010). The authors of (Batten and Pan 2011) expand on Marrington's work by introducing dynamic object sets and relations which is effective in reducing the otherwise fast growing number of objects. A framework proposed to identify relationships between the members of an email network is presented in (Haggerty et al. 2011) and graph theory is applied to a case study.

While informal use of hypotheses is made in all of these approaches, and each paper argues that hypotheses are necessary, no formal definition of hypothesis generation has been proposed. In Section 3, we provide such a definition.

3 Hypothesis Generation

The starting point for the formal approach to hypothesis generation is based on the work in (Marrington 2009) and (Batten and Pan 2011) and the reader is referred to those papers for more details. We begin with a set of objects **O** which have been collected in the preliminary stage of the investigation; relationships are then established between some of these objects. For instance, given objects *printer* and *computer*, we might establish the relationship that '*the computer is connected to the printer*'. Given those relationships, inferred relationships can then be constructed as in (Marrington 2009).

In our context, **O** is the set of items perceived to be in the vicinity of, or connected to, a forensic investigation. The definitions below are standard definitions used in set theory or the theory of binary relations and can be found in (Herstein 1975).

Definition 1. *A relation **R** on* **O** *is a subset of ordered pairs of* **O**×**O**.
Definition 2. *A relation **R** on* **O** *is reflexive if* (a,a) ∈ ***R*** *for all* a *in* **O**.

We can assume without any loss of generality that any relation on **O** in our context is reflexive since this property neither adds nor deletes information in a forensic investigative sense.

Definition 3. *A relation **R** on* **O** *is symmetric if* (a,b) ∈ ***R*** *implies* (b,a) ∈ ***R*** *for all objects* a *and* b *in* **O**.

Again, without loss of generality, in our context we assume that any relation on **O** is symmetric. This assumption is based on an understanding of how objects in **O** are related. So for instance, a printer and PC are related bi-directionally in the sense that they are connected to each other.

Definition 4. *Given a reflexive and symmetric relation **R** on* **O**, *for each element* a ∈ **O**, *we define a relational class for* a *by* (a) ={b| (a,b) ∈ ***R***, b ∈ **O**}.

Note that, because of reflexivity, a∈**O** is always an element of the relational class (a). Given a set **O** of objects and a relation ***R*** on **O**, we now define what we mean by a hypothesis on **O** and ***R***.

Definition 5. *A* ***hypothesis*** *h about* **O** *and* ***R*** *is a statement involving a non-empty subset* $\mathbf{O}_h$ *of* **O** *such that for all* a ∈ $\mathbf{O}_h$, *if* $|\mathbf{O}_h|>1$, *then there is an element* b ∈ $\mathbf{O}_h$ *with* b≠a *such that* (a,b) ∈ ***R***.

Thus, in an investigation, hypotheses are generated using objects from a non-empty subset of the object set; if there is more than one object present, then it must be related to some different object also used in the hypothesis. This latter condition forces the investigator to use the defined relationships to generate hypotheses. We need the condition b≠a since we allow the case of a hypothesis about a single object from **O**, and since each object is related to itself because of the reflexive property, Definition 5 could be trivially satisfied for any set of objects whether related or not.

Thus we have introduced an important forensic constraint that hypotheses should be about objects which are related. However, the word 'statement' in the above definition is quite vague; in our context, we mean an English language statement and we give an example below to illustrate this. Note that, for each subset $\mathbf{O}_h$ chosen, the set of all hypotheses can be generated algorithmically in finite time since the number of words in English is finite; however, the number is combinatorially large and so impractical to generate in real time.

Notation. Let **R** be a set of relations on the object set **O**. Then $\mathbf{H}(\mathbf{O}, \mathbf{R})$ is the set of all hypotheses which can be generated from **O** and **R** for all non-empty subsets of **O**.

Example. Suppose that {Alice, printer, document} is a subset of **O** and that **R** contains (Alice, printer) and (printer, document). Then the statement: 'Alice printed the document on the printer' satisfies definition 5. The statement 'Alice did not print the document on the printer' is also a valid hypothesis.

Analyzing the hypotheses In order to analyze the hypotheses generated from the sets of objects and relations, the investigator assigns confidence levels to them. While a binary logic would assign only one of two values (1 for true or 0 for false) to a hypothesis, in our case, it is generally not possible to be so precise. The investigator, however, may be able to convince herself, based on the evidence and circumstances, that a particular hypothesis may be 'more true than false' and in this case wish to assign it a value between 0 and 1 but closer to 1, say 0.6 or 0.7 on a spectrum of [0,1]. (See (Gerla, 2001) or (Krantz and Kunreuther, 2007.) We formalize this definition now.

Definition 6. *A* ***characteristic function*** *on a set S is a map* $c:S\rightarrow[0,1]$. *For* $x\in S$, *we call c(x) the* ***confidence level*** *of x with respect to S.*

In case S is the set of all hypotheses generated in an investigation, we shall insist on the following: if $h\in\mathbf{H}$ is known to be true, then $c(h)=1$; if $h\in\mathbf{H}$ is known to be false, then $c(h)=0$. Otherwise, the investigator assigns a confidence level strictly between 0 and 1 to **H**. For practical purposes, we limit the choice of confidence value by using only the values 0, 0.1, 0.2, 0.3,…,1.

Continuing from the ***Example***, if Alice shares the printer in an open office and no evidence suggests that Alice prints the document, then an investigator might assign 0.5 to both hypotheses because she feels there is an equal chance that Alice printed the document. But if Alice confesses to printing the document, then the investigator assigns 1 to the hypothesis "Alice printed the document" and 0 to the complement.

In applying our methodology, based on time and resources, the investigator determines a number of rounds to be run, and in each round will generate a set of hypotheses. We use $H_i = H_i(\mathbf{O}_i, \mathbf{R}_i)$ to refer to the set of hypotheses generated in round i based on object set $\mathbf{O}_i$ and relation set $\mathbf{R}_i$. Clearly, $H_i \subseteq \mathbf{H}(\mathbf{O}_i,\mathbf{R}_i)$. Let $U=\cup_i H_i$, the union of the set of hypotheses generated in a fixed set of rounds.

At this point, the investigator needs to make some decisions which we list here.

D1. The investigator assigns a confidence level *c(h)* to each member *h* of *U*.

Let $U^{+}=\{h \in U$ such that $c(h)>0.5\}$.

D2. (a) If $U^{+}\neq$ □□, then consider extensively the elements of U^{+} in order to resolve the investigation. (b) If $U^{+}=$□□, choose a new (larger) bound on the number of rounds and continue until either the bound is reached or $U^{+}\neq$ □□.

D3. If $U^{+}=$ □□ after the rounds and time available have been exhausted, then make changes to the original sets **O** and **R** and repeat the procedure.

In practice, for a large investigation, different sets **O** and **R** can be established initially and the scheme above run in parallel. Our goal is in fact to isolate only those hypotheses in U^{+} of specific investigative interest and of high confidence level. We tackle this by re-interpreting the relationship on **O** in a different way from (Marrington et al. 2010).

4 Case Study and Analysis

In order to illustrate our hypothesis generation methodology, we now present a case study which requires all aspects of our theoretical setup. This case study is taken from the earlier work of (Batten and Pan 2011) where it was used to develop the object set and relations developed during a forensic investigation. In this paper, we modify the object and relational development to show how hypotheses can be generated and examined in assisting the investigator to draw some conclusions about the case in reasonable time and with some certainty about the final decisions.

The case is copied here in its original form as follows:

“Joe operates a secret business to traffic illegal substances to several customers. One of his regular customers, Wong, sent Joe an email to request a phone conversation. The following events happened chronologically —

2009-05-01 07:30 Joe entered his office and switched on his laptop.
2009-05-01 07:31 Joe successfully connected to the Internet and retrieved his emails.
2009-05-01 07:35 Joe read Wong’s email and called Wong’s land-line number.
2009-05-01 07:40 Joe started the conversation with Wong. Wong gave Joe a new private phone number and requested continuation of their business conversations through the new number.
2009-05-01 07:50 Joe saved Wong’s new number in a text file named “Where.txt”.
2009-05-01 07:51 Joe saved Wong’s name in a different text file called “Who.txt”.
2009-05-01 08:00 Joe hid these two newly created text files in two graphic files (“1.gif” and “2.gif”) respectively by using S-Tools with password protection.
2009-05-01 08:03 Joe compressed the two new GIF files into a ZIP archive file named “1.zip” which he also encrypted.
2009-05-01 08:04 Joe concatenated the ZIP file to a JPG file named “Cover.jpg”.

2009-05-01 08:05 Joe used Window Washer to erase 2 text files ("Who.txt" and "Where.txt"), 2 GIF files ("1.gif" and "2.gif") and 1 ZIP file ("1.zip"). (Joe did not remove the last generated file "Cover.jpg".)
2009-05-01 08:08 Joe rebooted the laptop so that all cached data in the RAM and free disk space were removed.
Four weeks later, Joe's laptop was seized by the police due to suspicion of drug trafficking. As part of a formal investigation procedure, police officers made a forensic image of the hard disk of Joe's laptop. Moti, a senior officer in the forensic team, is assigned the analysis task." (*End of case in (Batten and Pan 2011).*)

Moti runs Forensic ToolKit to filter out the files of known hash values from a verified forensic image of Joe's laptop. Then he defines **O** = {250 emails, 50 text files, 100 GIF files, 90 JPG files, 10 application programs} as his initial object set. To avoid analyzing all data bit by bit Moti adopts our hypothesis generation approach which works in multiple rounds. The number of rounds is set by the investigator and possibly adjusted as the investigation proceeds. Moti has two working days before a report is due and decides to try for at least three rounds in the first day and in each round generate hypotheses which satisfy Definition 5.

Round 1 Suspecting that Joe uses the installed programs to process other files, Moti establishes his object set and relational classes as follows: $\mathbf{O_1}$=**O** and $\mathbf{R_1}$ ={(a,b) | a ∈{10 application programs}, b ∈{250 emails, 50 text files, 100 GIF, and 90 JPG files}}on the basis that application programs are indicative of user behavior (Bem and Huebner 2007). Moti's first hypothesis set is $H_1(\mathbf{O_1},\mathbf{R_1})$ = {h_1="Joe used the 10 application programs"}. To validate this hypothesis, Moti establishes that all programs were used in a virtual environment and that the programs S-Tools and WinZip were used frequently.

Round 2 Expecting to see that the visible files are clean, Moti establishes his object set and relational classes as follows: $\mathbf{O_2}$= {250 emails, 50 text files, 100 GIF, 90 JPG files} and $\mathbf{R_2}$={(a,b) | a, b ∈{250 emails, 50 text files, 100 GIF, 90 JPG files}}. Moti's second hypothesis set is $H_2(\mathbf{O_2},\mathbf{R_2})$ = {h_2 = "Joe did not hide information in the object files"}. Moti uses the data carving tool Scalpel but discovers 10 ZIP files each of which is concatenated behind a JPG file.

Round 3 With the newly recovered ZIP files, Moti establishes his object set and relational classes as follows: $\mathbf{O_3}$= {10 newly recovered ZIP files, WinZip program} and $\mathbf{R_3}$= ={(a,b) | a ∈{10 newly recovered ZIP files } and b is the WinZip program} so that he can use WinZip to explore the new files. Moti's third hypothesis set is $H_3(\mathbf{O_3},\mathbf{R_3})$ ={h_3= "Joe hid information in the 10 ZIP files"}. Moti attempts to extract the 10 ZIP files, and finds that they are encrypted.

Moti has now spent a full working day on three rounds and assigns the confidence level $c(h_1)$=1. Moti also sets the confidence levels $c(h_2)$=0 and $c(h_3)$=0.5 since concatenating a file behind a JPG file is a popular and practical anti-forensic method. Then, he reviews U^+ which contains only h_1 and does not help in wrapping up the case. Therefore Moti decides to extend the investigation for another 3 rounds during the second working day.

Round 4 Having decided to use the program PRTK to crack the 10 encrypted ZIP files, Moti establishes the object set and relational classes as follows: $\mathbf{O}_4$={10 encrypted ZIP files} and $\mathbf{R}_4$= {(a,a) | a∈{10 encrypted ZIP files}}. Moti reuses his third hypothesis in the next set: $H_4(\mathbf{O}_4,\mathbf{R}_4)$ = {h_4=h_3= "Joe hid information in the 10 ZIP files"}. Moti manages to crack the 10 encrypted ZIP files and extract the contents as 20 new GIF files.

Round 5 During tea break, Moti recalls that the program S-Tools is frequently used on Joe's laptop but has not yet been investigated. Furthermore, S-Tools steganographically embeds small text files into GIF files. Hence, he decides to reset the object set and relational classes as follows: $\mathbf{O}_5$= {20 new GIF files, 100 GIF files from Round 1, 50 text files, S-Tools} and $\mathbf{R}_5$ ={(a,b) | a ∈{20 new GIF, 100 old GIF, 50 text files} and b is S-Tools}. Moti's fifth hypothesis is $H_5(\mathbf{O}_5,\mathbf{R}_5)$ = {h_5= "Joe used S-Tools to hide information as text files in the GIF files"}. Moti tries to manually recover information by using S-Tools, and thus the progress is slow.

Round 6 As an experienced investigator, Moti suspects that Joe might have used some of his personal information to construct passwords and so he adds Joe's personal information to the object set and to relational classes as follows: $\mathbf{O}_6$= {20 new GIF files, 100 GIF files, 50 text files, S-Tools, Joe's personal information} and $\mathbf{R}_6$={(a,b) | a, b ∈{20 new GIF, 100 GIF, 50 text files, S-Tools, Joe's personal information}}. Moti's sixth hypothesis set is $H_6(\mathbf{O}_6,\mathbf{R}_6)$ = {h_6="Joe used S-Tools to hide information as text files in the GIF files encrypted by his personal information"}. After some trial and error, Moti uses Joe's medical card number to recover two text files from two GIF files. One text file contains the word "Wong" and the other the number "0409267531".

After completing rounds 4, 5 and 6, Moti evaluates his confidence levels. He believes that the information found is related to Joe's alleged drug trafficking business. Moti conservatively sets his new confidence levels to $c(h_4)$=1, $c(h_5)$=1 and $c(h_6)$=1 and decides to extend the investigation to a final round.

Round 7 Moti focuses on the hypotheses and information related to h_4, h_5 and h_6. He adds the new text strings to the object set and relational classes as follows: $\mathbf{O}_7$ = {20 new GIF files, 100 GIF files, 50 text files, S-Tools, Joe's personal information, the two recovered text files, the name Wong, the number 0409267531} and $\mathbf{R}_7$ = {(a,b) | a, b ∈ $\mathbf{O}_7$}. His new hypothesis set is: $H_7(\mathbf{O}_7, \mathbf{R}_7)$ ={h_7 ="The mobile phone number 0409267531 belongs to Wong", h_8 = "Joe used S-Tools to hide drug-trafficking information as text in GIF files encrypted by his personal information", h_9= "Joe and Wong have a client/customer relationship in selling/buying drugs"}. Moti searches through mobile phone registration directories and finds that the number "0409267531" is registered under the name "Alex Wong" who has been charged with drug possession. Thus, Moti regards "Wong" as a suspect in Joe's drug trafficking case. His confidence levels are $c(h_7)$=1 because the ownership of the mobile phone is confirmed, $c(h_8)$=0.8 because Moti believes that more information remains in the other GIF pictures, and $c(h_9)$=0.6 because Moti needs to investigate how Wong relates to Joe.

It is almost the end of the second working day and Moti writes a case report illustrating his steps in finding the two text strings from the digital items. He suggests that Wong is a suspect in Joe's drug trafficking case and requests more time to further explore his hypotheses in the set $U^{+}=\{h_1,h_4,h_5,h_6,h_7,h_8,h_9\}$.

In summary, this case study demonstrates the use of hypothesis generation during a digital forensic investigation where many assumptions and decisions are made by investigators. This new approach moves beyond relationship building, already used by several authors, to focus on hypotheses generation which aids investigators to make justified decisions based on identified evidence. The investigator analyzes only those relations and hypotheses about which he is confident and eliminates others.

5 Discussion and Future Work

This paper takes a scientific approach, using formal and systematic methods of building relationships on objects, to the development of hypotheses in a digital forensic investigation. A formal definition of 'hypothesis' is given and applied throughout the procedure which comprises a series of rounds in which relations are built on the objects under investigation from which, in turn, hypotheses are generated. In each round, the object, relation and hypothesis set may change, depending on the results of the previous rounds. Confidence levels are assigned to the hypotheses by the investigator and when the target number of rounds is reached, those hypotheses with sufficiently high confidence levels are analyzed. The case study of Section 4 demonstrates the usefulness of a formal definition of 'hypothesis'.

Because the methodology is highly structured it lends itself to easy adoption by an investigator and also to a semi-automated process which can reduce the investigation time. Since the structure lends itself to reproducibility, the method is particularly appropriate for presentation in a court of law where each round or stage of the process can be examined thoroughly. While the automated nature of our procedure is evident, we have not yet written software to apply to our case study; this is the next step in our work. In addition, we hope to be able to collaborate with a forensic team to test our methodology on a real case.

6 References

Al-Zaidy, R., Fung, B., Youssef A., and Fortin, F. (2012), "Mining criminal networks from unstructured text documents", *Digital Investigation*, vol. 8, no. 3-4, pp. 147-160, Elsevier.

Batten, L.M. and Pan, L. (2011), "Using relationship-building in event profiling for digital forensic investigations", in Lai, X., Gu, D., Jin, B., Wang, Y. and Li, H. (Ed.) *Forensics in Telecommunications, Information, and Multimedia*, vol. 56, pp. 40-52, Springer.

Bem, D. and Huebner, E. (2007), “Computer forensic analysis in a virtual environment”, *International Journal of Digital Evidence*, vol. 6, no. 2, 13 pages, Utica College, New York.

Carrier, B. (2006), “A hypothesis-based approach to digital forensic investigations”, *CERIAS Tech Report 2006-06,* Purdue University, Center for Education and Research in Information Assurance and Security, West Lafayette.

Chang, J.Y., Huang, Z.C. and Shen, Z.Y. (2010), “Medical Diagnostics System Based on Clustering Discrimination”, in *Proceedings of IEEE 2nd International Conference on Information Engineering and Computer Science (ICIECS)*, pp. 1-4.

Chen, F. (1996), *Hypothesis generation for management intelligence*, PhD thesis, Deakin University, Melbourne, Australia.

Chen, P. and Garcia, W. (2010), “Hypothesis generation and data quality assessment through association mining”, in *Proceedings of 9th IEEE International Conference on Cognitive Informatics (ICCI)*, pp. 659-666, Beijing, China.

Di Lecce, V. and Calabrese, M. (2012), “Syntactic pattern recognition from observations: a hybrid technique”, in *Bio-Inspired Computing and Applications Lecture Notes in Computer Science*, vol. 6840, pp. 136-143, Springer.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J., Nyberg, E. and Prager, J. (2010), “Building Watson: an overview of the DeepQA project”, *AI Magazine,* vol. 31, no. 3, pp. 59-79.

Ferrucci, D., Nyberg, E., Allan, J., Barker, K., Brown, E., Chu-Carroll, J., Ciccolo, A., Duboue, P., Fan, J. and Gondek, D. (2009), “Towards the open advancement of question answering systems”, *IBM Research Report. RC24789 (W0904-093)*, IBM Research, NY.

Gerla, G. (2001), *Fuzzy logic: Mathematical tools for approximate reasoning*, Springer.

Haggerty, J., Karran, A.J., Lamb, D.J. and Taylor, M. (2011), “A framework for the forensic investigation of unstructured email relationship data”, *International Journal of Digital Crime and Forensics*, vol. 3, no. 3, pp. 1-18, IGI Global.

Herstein, I. (1975), *Topics in algebra*. Wiley, New York, 2nd edition.

Jankun-Kelly, T., Wilson, D., Stamps, A.S., Franck, J., Carver, J. and Swan, J. (2009), “A visual analytic framework for exploring relationships in textual contents of digital forensics evidence”, in *Proceedings of VizSec, IEEE*, pp. 39-44, Atlantic City, New Jersey, USA.

Krantz, D. and Kunreuther, H. “Goals and plans in decision making,” Judgement and Decision Making 2(3):137-168, 2007.

Lambrou, A., Papadopoulos, H. and Gammerman, A. (2011), “Reliable confidence measures for medical diagnosis with evolutionary algorithms”, *IEEE Transactions on Information Technology in Biomedicine*, vol.15, pp. 93-99, IEEE.

Marrington, A. (2009), *Computer profiling for forensic purposes,* PhD thesis, QUT, Australia.

Marrington, A., Mohay, G., Morarji, H. and Clark, A. (2010), “A model for computer profiling”, in *Proceedings of IEEE International Conference on Availability, Reliability, and Security*, pp. 635-640, Poland.

Webster, Y., Gudivada, R., Dow, E., Koehler, J. and Palakal, M. (2010), "A framework for cross-disciplinary hypothesis generation", in *Proceedings of the 2010 ACM Symposium on Applied Computing*, pp. 1511-1515, Switzerland.

Willassen, S. (2008), "Hypothesis-based investigation of digital timestamps", *International Federation for Information Processing, Advances in Digital Forensics IV*, vol. 285, pp. 75-86.

Yang, Y.P. (2010), "A consistency contribution based Bayesian network model for medical diagnosis", *Journal of Biomedical Science and Engineering*, 3(5), pp. 488-495, Inderscience.