

Chapter 1

SPECIALIZING CRISP-DM FOR EVIDENCE MINING

Jacobus P Venter, Alta de Waal and Cornelius J Willers

Abstract The use of all forms of computer and communication devices is changing human interaction and thinking. Electronic traces of actions and activities are continually being left behind most often unknowingly so. This situation creates opportunities for criminal investigators to make use of these traces and marks to uncover evidence. In this evidentiary discovery process several problems are experienced including the linking of unstructured pieces of data to an evidence trail.

Analysis is a crucial aspect of the overall Cyber Forensic process for which adequate support is not provided. In this article it is argued that in order to alleviate the situation around analysis and further the foundations of cyber forensics it is necessary to learn from another field that also needs to deal with vast amounts of information and develop methods for automated interpretation thereof; the field of Knowledge Discovery and Data Mining (KDD). A specialization of a well known KDD process (CRISP-DM) is developed and named CRISP-EM.

The process of specialization is described and some of the results are shown. It is further shown that the CRISP-EM methodology supports a structured approach in defining the research gaps in evidence mining.

Keywords: Evidence Mining, CRISP-DM, CRISP-EM, Digital Investigation. Data Mining Process

1. INTRODUCTION

"Searching for traces is not, as much as one could believe it, an innovation of modern criminal jurists. It is an occupation probably as old as humanity. The principle is this one. Any action of an individual, and obviously, the violent action constituting a crime, cannot occur without leaving a mark. What is admirable is the variety of these marks. Sometimes they will be prints, sometimes simple traces, and sometimes stains." Professor Edmond Locard[4]

Electronic traces of actions and activities are continually being left behind [14][18], most often unknowingly so. In this electronic and information rich age Locard's Exchange Principle, quoted above, can be extended to include electronic "marks". This situation creates opportunities for criminal investigators to make use of these traces and marks to uncover evidence. In this evidentiary discovery process several problems are experienced, including those of dealing with the ever growing volumes of data and the linking of unstructured pieces of data to an evidence trail [15].

The overall Cyber Forensic process consists out of four major phases Acquisition, Examination, Analysis and Presentation [15]. Software and Hardware tools that support the Cyber Forensic process do exist. Software tools are available for further examination of the data collected. These tools are mostly based on keyword searches and unless very specific knowledge regarding the information to be retrieved is available, the process of retrieving valuable information is complex, manual and extremely time consuming. Some progress has been made in providing automated support to forensic analysis [8]. These tools however do not cover the full spectrum of forensic analysis activities and the functionality of existing tools currently do not adequately reduce the volume of data that are still to be analyzed manually. It is also important to find information that investigators did not know existed [15]. Current cyber forensics processes and tools do not adequately address the specific requirements of the analysis phase of the overall Cyber Forensics process.

Looking at available publications the specific area of forensic analysis (as a subset of the overall process) has to date not received substantial research focus. For example, an analysis performed on 77 articles published in the Digital Investigation journal (all the articles from volume 1 in 1994 until volume 3 supplement 1 in 2006) reveals that only 26% (20/77) of all the articles deals with examination/analysis of forensic data. Of the 20 articles 18 actually deals with the further preparation of the data for manual interpretation. Only 2 out of 77 articles dealt with automating the process of finding electronic evidence, one published in

2004 [9] and the other presented at the Digital Forensics Research Workshop in 2006 [16]. As Garfinkel [8] indicates, forensic examiners have become victims of their own success and that they do not have time to analyze all the data provided by the earlier parts of the forensic process.

Forensic analysis requires a keen detective human mind, but the human mind does not have the capability (or time) to process the millions of words on a computer hard disk. Where methods, sometimes tied into the ability of a single investigator, exist they do not scale very well and, therefore, do not adapt to large data sets [15]. This indicates that not only is more tool/technique support required for the analysis phase but also that such tools should automate some of the tasks currently performed manually.

In this article it is argued that in order to alleviate the situation around analysis and further the foundations of cyber forensics it is necessary to learn from another field that also needs to deal with vast amounts of information and develop methods for automated interpretation thereof; the field of Knowledge Discovery and Data Mining (KDD). KDD is the process of identifying valid, novel, potentially useful and understandable patterns from large volumes of data; It is a multi-disciplinary topic, drawing from several fields including expert systems, machine learning, intelligent databases, knowledge acquisition, case-based reasoning, pattern recognition and statistics [1].

Existing KDD research related to criminal investigations focus on mining data from case databases [3], [14], to find general trends mostly to enable crime prevention. What is required is rather to find specific data elements linked in to a specific case. Although existing data mining knowledge do not adequately address this need it should not be discounted because of this. Building on the foundations of data mining will aid in developing sound practices for data mining in cyber forensics. Pollitt support this argument by stating that research should focus on forensic applications of data mining tools and on developing knowledge management strategies specific to the context of criminal investigations [15]. The term evidence mining is used to indicate the specific application of data mining and knowledge discovery principles in the field of cyber forensics for the purpose of supporting the analysis phase of the Cyber Forensics process.

The rest of this paper will further define evidence mining, indicate the requirement for a process, and suggest a way forward (sections 2 and 3). The specialization of the CRISP-DM methodology for evidence mining is described and a summary of the specialized process (CRISP-EM) is provided (sections 4 and 5). The last part of the paper, sections 6 and 7, discusses research gaps and provides a conclusion.

2. EVIDENCE MINING

The term evidence mining needs further clarification: Evidence is something that validates facts and can be used as testimony in a court or formal hearing. In this context the interest is not general trends that can assist in the prevention of crime. The focus is on the finding of proof in order to testify in court regarding facts.

Mena [14] indicates that criminal analysis uses historical observations to come up with solutions, unlike criminology, which re-enacts a crime in order to solve it. In this sense evidence mining is more like criminology. Evidence mining aims to "re-enact" the crime by analyzing the electronic evidence pieces left behind by a person's everyday actions. Evidence Mining aims to uncover, through the application of KDD principles and techniques, electronic artifacts that can form part of the evidence set to assist in the development of crime scenarios.

Evidence Mining is a new term, or at least sparsely used term. A search of the SCOPUS [22], ACM [20] and IEEE [21] digital/citation libraries returned no relevant results for the term "Evidence Mining".

3. IN SEARCH OF A PROCESS

So far this paper described the gap in support for Cyber Forensic Analysis and Evidence Mining was shown as a way to address this gap. This section argues the necessity for a proper process to support Evidence Mining and suggest the specialization of a popular data mining process for this purpose.

The CRISP-DM consortium developed a Cross-Industry Standard Process for Data Mining (CRISP-DM[2]). The KDNuggets April 2004 poll indicated that CRISP-DM was the most popular methodology amongst the respondents [10]. Clifton [6] also indicates CRISP-DM as a notable effort in process standardization. Therefore attention was focused on this methodology to determine it's appropriateness for application in the evidence mining environment.

The CRISP-DM has characteristics that are useful for applying in the evidence mining environment. In the first instance the CRISP-DM provides for a generic process model that holds the overarching structure and dimensions of the methodology. The methodology then provides for specialization according to a pre-defined context. In the CRISP-DM terminology this is indicated as a specialized process. It is this specialization that was used to conceive CRISP-EM (Cross-Industry Standard Process for Evidence Mining). Mena [14] indicated the use of CRISP-DM for detecting crimes without providing much detail or establishing a specialization of CRISP-DM. This paper offers the specialization of

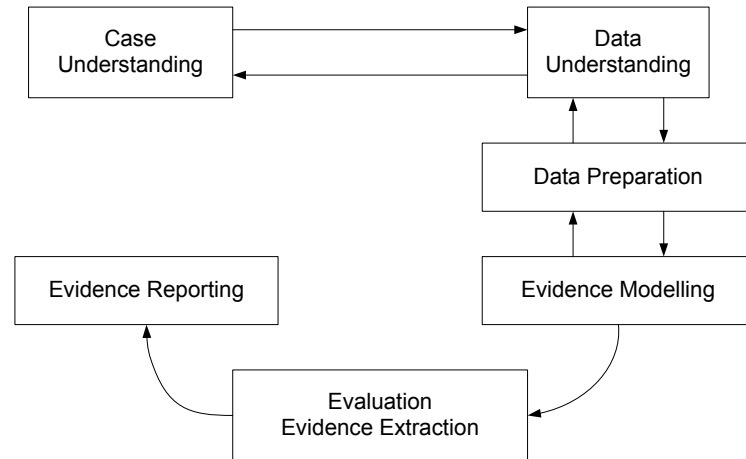


Figure 1. The Main Phases of CRISP-EM.

CRISP-DM as a way forward in meeting the requirement of a process to support evidence mining. The aim of CRISP-EM is not to provide a new Cyber Forensic process but to support the analysis phases of existing Cyber Forensics processes.

4. CRISP-DM SPECIALISATION

Accepting that CRISP-DM is a good basis to start from and that specialization will provide applicability in a cyber forensic context this section describes the specialization process.

The main phases of CRISP-EM are shown in Figure 1. Although CRISP-EM follows the basic structure of CRISP-DM some of the major phases were renamed (see Table 1) to fit the context of digital investigations.

4.1 Specialization Strategy

The CRISP-DM guide [2] indicates that the basic strategy for a specialized methodology is to:

- Analyze your specific context.
- Remove any details not applicable to your content.
- Add any details specific to your content.
- Specialize (or instantiate) generic contents according to concrete characteristics of the context.

- Possibly rename generic contents to provide more explicit meanings in the context for the sake of clarity.

In the rest of this section the application of this strategy is discussed. The need for further work is also indicated.

4.2 Context Analyses

The first aspect of the specialization strategy applied was to analyze the specific context. CRISP-EM is placed within the context of a specific criminal case. It is designed to provide support to an investigator or prosecutor on a specific case. It is not designed to be used to mine for general trends in case databases. Such a project is close enough to normal data mining that CRISP-DM in its original format will suffice.

Cyber Forensics consists out of four major phases namely Acquisition, Examination, Analysis and Reporting. The acquisition phase collects data in a manner that conserves the integrity of the data. Normally a copy of the original data (called a mirror) is made and any further processing is done using the mirror (or even a copy of the mirror). The second phase in the cyber forensic process is examination. In this phase rudimentary processing of the data is performed. This normally consist of either keyword searches or looking in operating system specific important locations (e.g. for a user name). The third phase is to analyze the data in more detail. In this phase the information provided by the examination phase is placed in context with the case and further processed to uncover facts that will stipulate to events, actions, etc. relevant to the case. The fourth phase is to present the evidence to the concerned parties in and out of court.

The context for evidence mining is phases three (analyses) and four (presentation). It is important to note that the data gathering aspects of the CRISP-DM methodology (a part of the data preparation phase) and the cyber forensics acquisition phase is not within the same context. It is due to this context difference that it is believed to be better to speak about data collation in CRISP-EM rather than data collection to eliminate confusion.

4.3 Renaming Generic Content

The original CRISP-DM and renamed CRISP-EM phases are shown in Table 1. As each evidence mining project will be associated with a specific case the renaming of the first phase is obvious. The names for the data understanding and preparation phases remains the same as the intent of these phases for evidence mining is the same as for data mining. The biggest difference is for the last three phases where the

Table 1. Original and Renamed Phases

Original from CRISP-DM	Renamed in CRISP-EM
Business Understanding	Case Understanding
Data Understanding	Data Understanding
Data Preparation	Data Preparation
Modeling	Event Modeling
Evaluation	Evaluation and Evidence Extraction
Deployment	Evidence Reporting

intent is different. A specific evidence mining project is likely to span only one case. The intent is therefore to produce specific evidence for the case at hand rather than to build a model that can be deployed for future use.

The modeling phase is replaced by an Event Modeling or Scenario development phase. This phase creates plausible scenarios from the electronic evidence available in the data set. This is then presented to the investigator and/or prosecutor in the next phase that evaluates the scenarios presented, selects the relevant scenarios and then extract the relevant evidence (hence Evaluation and Evidence Extraction for the phase name). The deployment for evidence mining is where the evidence is reported on, either to an investigator, a prosecutor or in court (therefore Evidence Reporting).

The renaming of various aspects continues within the details of the methodology. One notable example is:

The generic task to "Collect Initial Data" (generic task 1.2 in CRISP-DM) is renamed to "Collate Initial Data" in CRISP-EM (Collate: to collect, compare carefully in order to verify, and often integrate or arrange in order: Merriam-Webster online dictionary). This is to ensure that the distinction between forensic data acquisition (data collection) and the putting together of the data for analysis (data collation) is clear.

4.4 Specialize Generic Content

The specialization phase of the strategy is applied during the detail development of the CRISP-EM methodology. In order to maintain the context of an investigation it is necessary to not only develop specialized tasks but also to specialize the phases and generic tasks. This section describes some of the specialization that was done.

4.4.1 Generic Process Descriptions. The original CRISP-DM descriptions for the generic process phases were adapted to fit within the evidence mining context. The adapted descriptions are shown below. The major changes from the original descriptions are shown in boldface.

It is important to remember that these process phases fit within the analysis phase of the larger Cyber Forensic process and is not meant to replace the overall process.

Case Understanding. This initial phase focuses on understanding the **investigation** objectives and requirements from a case perspective, then converting this knowledge into an **evidence** mining problem definition and a preliminary plan designed to achieve the objectives.

Data Understanding. The data understanding phase starts with an initial data collation and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation. The data preparation phase covers all activities to construct the final dataset (data that will be fed into the **event modeling** tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection, **entity recognition and co-reference resolution** as well as transformation and cleaning of data for **event modeling** tools.

Event Modelling. In this phase, various **evidence modeling and event reconstruction** techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same **evidence** mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

Evaluation and Evidence Extraction. At this stage in the project a **set of scenarios or event lines have been built** that potentially have high quality from a data analysis perspective. Before proceeding to final reporting of the evidence, it is important to more thoroughly evaluate the **scenarios/event lines** and review the steps executed to construct and extract **the relevant ones** to be certain it properly achieves the **case** objectives. A key objective is to determine if there is some important **case** aspect that has not been sufficiently considered. At the end of this phase, a decision on the use of the **evidence** mining results should be reached.

Evidence Reporting. Creation of event lines and the extraction of evidence is generally not the end of the project. Even if the purpose of the evidence mining is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the **investigator** can use it for **evidentiary purposes**. It may involve **augmenting chosen event lines with other data pertinent to the investigation at hand**. In many cases it is the investigator, not the data analyst, who carries out the **reporting** steps. However, even if the analyst will not carry out the reporting effort it is important for the **investigator** to understand up front what actions need to be carried out in order to actually make use of the **extracted event lines and evidence**.

4.5 Add/Remove content

Some new content must be added to address the requirements of evidence mining whereas other content does not make sense within the context of evidence mining. Examples of the changes made to the original guide are discussed next:

Initial Data Mining The development of the event lines is a complex task. It requires more advanced pre-processing and preparation of the data than "traditional" data mining may require. In order to facilitate the additional inputs to the event modeling phase, a task was added to the data preparation phase. This task is named Initial Data Mining. The output this task will produce is a richer dataset that will already include classified and categorized data. Understanding the crime triangle of Willing Offender, Enabling Environment, and Vulnerable target [5] will help in developing the pre-processed data as all three of these aspects is present in every crime instance and as such will also be present in the storyboarding. Therefore identifying entities and classifying them as potential offender, environment or victim indicators will be extremely useful in the next phase.

Develop Event Scenarios The primary purpose of the event modeling phase is to support the investigator through the development of hypotheses regarding the crimes that took place and how it happened based on electronic artifacts found in the evidence set. In this context a hypotheses is an answer to a question about what crime took place and can be wrong or right (adapted from [5]). The set of hypotheses (scenarios) are like a roadmap supporting the investigator in conducting an effective and efficient investigation. The original CRISP-DM task of 'Build Model' was replaced by the Develop Event Scenarios task. The

model that the evidence mining process will build is actually the scenarios thus the replacement.

5. CRISP-EM Summary

In the previous section the development of a specialized process for evidence mining was discussed, this section summarizes the CRISP-EM process and indicates some of the detail in the full process. The major phases of evidence mining are shown above in Figure 1. A next level of the CRISP-EM process, in mind map format, is shown in Figure 2. Further detail in the model identify specific aspects of each task and can also, where appropriate, indicate specific tools that may be necessary to perform the tasks. The detail for the data preparation phase is shown in Figure 3. Substantial further research is required to complete all the aspects of the process and to achieve full implementation.

6. RESEARCH GAPS

The framework provided by CRISP-EM also enables a structured approach to defining research gaps. CRISP-EM provides a level of granularity that allows easier indication of where existing techniques, mostly originating from the KDD knowledge base, would suffice and where new techniques will be required due to the differences in the tasks and outputs of CRISP-DM and CRISP-EM. The biggest difference between CRISP-DM and CRISP-EM lies in the "Event Modeling" and "Evaluation and Evidence Extraction" phases. It is therefore obvious that the biggest research gaps will be in these two areas. Three examples of identified research gaps are described below:

Example Case files: The development of new evidence mining techniques requires example data sets. These data sets, called example case files in this context, must contain known event lines in various forms in order to test the effectiveness of the techniques. Sufficiently large data sets must also be developed mixed within other data in order to test the efficiency of the algorithms. No such example case files currently exist. The research in this area must include the development of plausible crime "stories" and ways to mix this within other data sets. Manually developing all the required examples will be time consuming. Ways of automatically generating examples cases, and plausible electronic evidence associated therewith, must also be researched.

Coping with uncertainty: In the process of developing event lines, uncertainty is a major challenge. Available data is most of the times incomplete leading to beliefs that fall short of evidence, with fallible conclusions and the need to recover from error. This is called non-

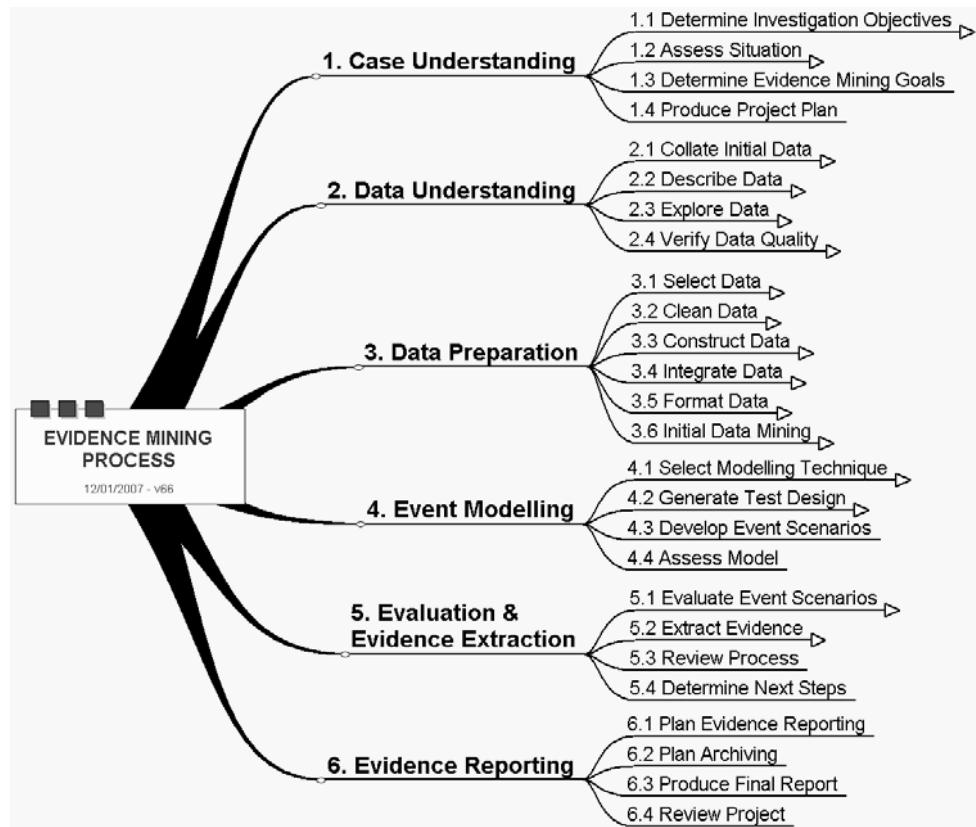


Figure 2. CRISP-EM Second Level.

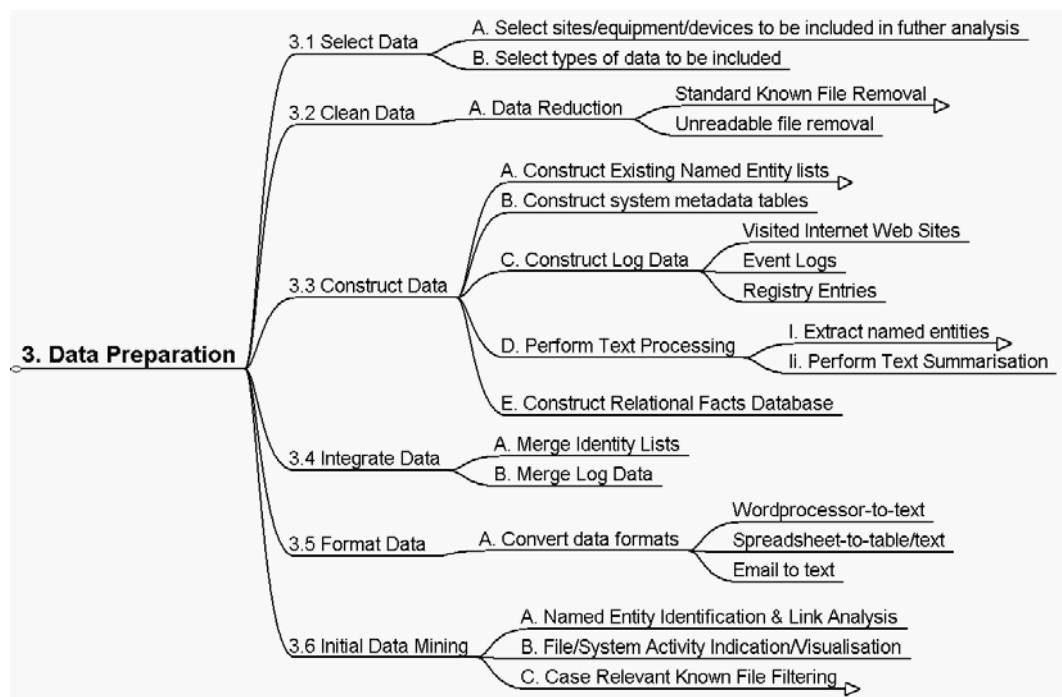


Figure 3. Detailed Data Preparation Level.

monotonic reasoning. Probabilistic reasoning is needed to enable the building of scenarios. The Bayesian paradigm and the artificial intelligence associated with it [12] provide an elegant approach to do that. It also addresses specific problems such as co-reference resolution, record linkage and theme extraction. The probabilistic outcomes of this modeling method enable the association of a probability value to an event line which will facilitate prioritization of evidence extraction.

Automated Investigator Experience: Human investigators have special skills and experience in extracting evidence from unstructured information. However, the human investigation process is a slow laborious process, the number of human investigators is limited, and human concentration diminishes with fatigue. Knowledge and data discovery processes are automated and can be parallelized to handle large volumes of information efficiently. These KDD processes unfortunately do not currently have the skill of the human investigator. What is required are computer automated processes combined with the skill of human investigator. Intelligent Multi-Agent techniques [19][7][17] hold promise in the development of automated investigators.

7. CONCLUSION

Information and communication is changing society forever. It is also changing the amount of electronic traces left behind. This creates opportunities for law enforcement to make use of more electronic evidence. A problem is however created due to the volumes of data that need to be sorted in order to find useful evidence. The wealth of knowledge discovery and data mining science can be used in this process. Due to the specific requirements in the criminal investigative environment, some of the processes, methods and techniques must be adapted or new ones must be developed. In this paper a specialization of the CRISP-DM process for the criminal investigative environment, namely CRISP-EM, is introduced. Related work, research gaps, and future planned work is also indicated.

It can be concluded that evidence mining and the CRISP-EM process provides a route towards better knowledge discovery and data mining support for the criminal investigative environment. CRISP-EM is not yet a proven process; it currently provides a good framework for the initial, mostly manual, application of evidence mining. Apart from this it also provides a framework for researching new methods and techniques required for the enhancement of evidence mining and the automated implementation of the CRISP-EM process.

References

- [1] Bandyopadhyay, S., Maulik, U., Holder, L.B., *Advanced Methods for Knowledge Discovery from Complex Data*, Springer, 2005.
- [2] Chapman, P., et al, *CRISP-DM 1.0 Step-by-step data mining guide*, SPSS, 1999.
- [3] Chen, H., Chung, W., Xu, J.L., Yi Quin, G. W., Chau, M., *Crime Data Mining: A General Framework and Some Examples*, IEEE Computer, April 2004.
- [4] Chisum W.J., Turvey, B.E., *Crime Reconstruction*, Elsevier, 2007.
- [5] Clarke, R.V., Eck, J., *Become a Problem-Solving Crime Analyst*, Jill Dando Institute of Crime Science, 2003.
- [6] Clifton, C., Thuraisingham, B., *Emerging standards for data mining*, *Computer Standards & Interfaces* 23, 2001.
- [7] Dickinson, L., Wooldridge, M., *Towards practical reasoning agents for the semantic web*, Hewlett-Packard Laboratories Filton Road, Stoke Gifford Bristol BS34 8QZ, U.K., 2003.
- [8] Garfinkel, S.L., *Forensic feature extraction and cross-drive analysis*, *Digital Investigation Volume 3, Supplement 1*, Elsevier, 2006.
- [9] Gladyshev, P., Patel, A., *Finite state machine approach to digital event reconstruction*, *Digital Investigation Volume 1*, Elsevier, 2004.
- [10] KDNuggets, *Data Mining Methodology Poll*, April 2004, http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm, last accessed 27 February 2006.
- [11] Keppens, J., Schafer, B., *Knowledge based crime scenario modeling*, *Expert Systems with Applications*, Volume 30, Issue 2, Feb 2006.
- [12] Korb, K. B., Nicholson, A. E., *Bayesian Artificial Intelligence*, Chapman & Hall, ISBN 1-58488-387-1, 2004
- [13] Louis, A., De Waal, A., Venter, J.P., *Named Entity Recognition in a South African Context*, SAICSIT, 2006 [To be published].
- [14] Mena, J., *Investigative Data Mining for Security and Criminal Detection*, Butterworth Heinemann, 2003, ISBN 0-7506-7613-2.
- [15] Pollitt, M., Whitley, A., *Exploring Big Haystacks Data Mining and Knowledge Management*, *International Federation for Information Processing*, Volume 222, *Advances in Digital Forensics II*, eds. Olivier, M., Sheno, S., Boston:Springer, pp. 67-76, 2006.
- [16] Rogers, M.K., Seigfried, K., Tidke, K., *Self-reported computer criminal behaviour: A psychological analysis*, *Digital Investigation Volume 3, Supplement 1*, Elsevier, 2006.

- [17] Van der Hoek, W., Jamroga, W., Wooldridge, M., A logic for strategic reasoning, AAMAS , 2005.
- [18] Witten, Ian. H., Frank, Eibe., Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, 2005.
- [19] Wooldridge, M., An Introduction to MultiAgent Systems, John Wiley and Sons, 2002.
- [20] www.acm.org, last accessed 11 Jan 2007.
- [21] www.ieee.org, last accessed 11 Jan 2007.
- [22] www.scopus.com, last accessed 11 Jan 2007.