Chapter 21

SPECIALIZING CRISP-DM FOR EVIDENCE MINING

Jacobus Venter, Alta de Waal and Cornelius Willers

Abstract Forensic analysis requires a keen detective mind, but the human mind has neither the ability nor the time to process the millions of bytes on a typical computer hard disk. Digital forensic investigators need powerful tools that can automate many of the analysis tasks that are currently being performed manually.

This paper argues that forensic analysis can greatly benefit from research in knowledge discovery and data mining, which has developed powerful automated techniques for analyzing massive quantities of data to discern novel, potentially useful patterns. We use the term "evidence mining" to refer to the application of these techniques in the analysis phase of digital forensic investigations. This paper presents a novel approach involving the specialization of CRISP-DM, a cross-industry standard process for data mining, to CRISP-EM, an evidence mining methodology designed specifically for digital forensics. In addition to supporting forensic analysis, the CRISP-EM methodology offers a structured approach for defining the research gaps in evidence mining.

Keywords: Data mining, evidence mining, CRISP-DM, CRISP-EM

1. Introduction

Edmond Locard, a pioneer in forensic science, formulated the Exchange Principle: Every Contact Leaves a Trace [5]:

> "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."

Electronic traces of actions and activities are continually being left behind in the Age of the Internet [16, 21], enabling Locard's Exchange Principle to be extended to include electronic "marks." This situation creates new opportunities for criminal investigators to uncover evidence. However, the electronic evidentiary discovery process is severely limited by the growing volume of data and the linking of unstructured pieces of data to create evidence trails [17].

Digital forensic investigations involve four major phases: evidence acquisition, examination, analysis and presentation [17]. A variety of commercial tools are available for supporting investigations. However, most existing software tools for forensic analysis are based on keyword searches; unless very specific knowledge regarding the information to be retrieved is available, the process of retrieving information is complex, manual and time consuming. Some progress has been made in providing automated support for forensic analysis [9]. However, the tools do not cover the full spectrum of analysis activities, and they do not possess the functionality to adequately reduce the volume of data, let alone find information that investigators did not know existed [17].

Despite its importance, the area of forensic analysis has received relatively limited research attention. For example, our analysis of all 77 research articles published from 1994 through 2006 in the journal *Digital Investigation* revealed that only 26% (20 articles) focused on the examination or analysis of digital evidence. Eighteen of the twenty articles dealt with processing digital evidence to support manual interpretation. In all, only two articles [10, 18] focused on the important task of automating the search for electronic evidence. As Garfinkel [9] indicates, digital forensic examiners have become victims of their own success despite the fact that cannot analyze all the data provided to them by the previous phases of the forensic process.

Forensic analysis requires a keen detective mind, but the human mind does not have the capability (or time) to process the millions of bytes on a computer hard disk. Most analysis methods do not scale very well and, therefore, are unable to cope with large data sets [17]. A new generation of forensic tools is required to support human analysts, at the same time, automating many of the tasks that are currently being performed manually.

This paper argues that digital forensic analysis can greatly benefit from research in knowledge discovery and data mining, which has developed powerful automated techniques for analyzing massive quantities of data to discern novel, potentially useful patterns. The research is multi-disciplinary in nature, drawing from several fields including expert systems, machine learning, intelligent databases, knowledge acquisition, case-based reasoning, pattern recognition and statistics [2].

Previous research in knowledge discovery and data mining related to criminal investigations has focused on mining data from case databases to support crime prevention efforts [4, 16]. However, what investigators really need is support during the analysis phase: automated assistance to find specific data elements in specific cases. This point is underscored by Pollitt and Whitledge [17] who emphasize that research should focus on forensic applications of data mining tools and on developing knowledge management strategies specific to the context of criminal investigations. We use the term "evidence mining" to refer to the application of data mining and knowledge discovery techniques to support the analysis phase of digital forensic investigations.

This paper focuses on the application of evidence mining to support digital forensic investigations. It discusses a novel approach involving the specialization of CRISP-DM, a cross-industry standard process for data mining, to CRISP-EM, an evidence mining methodology designed specifically for digital forensics. In addition to supporting forensic analysis, the CRISP-EM methodology offers a structured approach for defining the research gaps in evidence mining.

2. Evidence Mining

Evidence validates facts and it may be used as testimony in courtroom proceedings or a formal hearing. In this context, the interest is not in general trends that assist in crime prevention. Instead, the focus is on the finding of proof in order to testify about the facts.

Mena [16] observes 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 left behind by the subjects' actions. Evidence mining aims to uncover, through the application of knowledge discovery 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, a scarcely used term. A search of the ACM [1], IEEE [11] and SCOPUS [19] digital libraries returned no relevant results for "evidence mining."

3. Evidence Mining Using CRISP-DM

The CRISP-DM consortium developed the Cross-Industry Standard Process for Data Mining (CRISP-DM) [3]. Clifton and Thuraising-

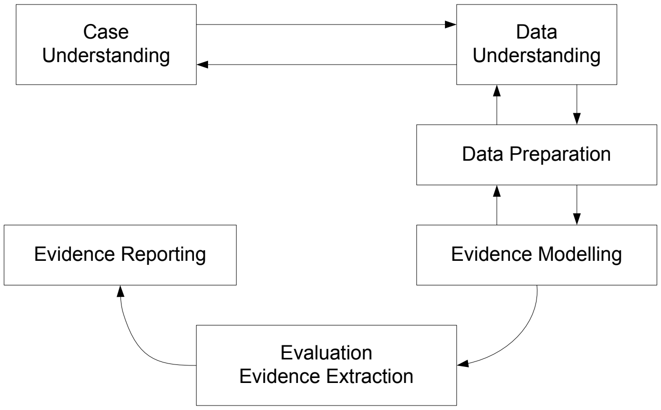

Figure 1. Main phases of CRISP-EM.

ham [7] identified CRISP-DM as a notable effort in process standardization. Moreover, the KDNuggets poll of April 2004 indicated that CRISP-DM was the most popular methodology among the respondents [12].

CRISP-DM has several characteristics that render it useful for evidence mining. It provides 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). We have used such a specialization to create CRISP-EM (Cross-Industry Standard Process for Evidence Mining). Mena [16] proposed using CRISP-DM to detect crimes without providing much detail or establishing a specialization of CRISP-DM. We propose CRISP-EM as an approach for meeting the requirements of a process that supports evidence mining. Our goal is not to create a new digital forensic process but to support the analysis phases of existing forensic processes.

4. CRISP-DM Specialization

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 are renamed to fit the context of digital investigations.

4.1 Specialization Strategy

CRISP-DM [3] proposes the following specialization strategy:

- Analyze the specific context.
- Remove details not applicable to the context.

- Add details specific to the context.

- Specialize (or instantiate) generic content according to concrete characteristics of the context.

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

The next four subsections discuss the application of this specialization strategy.

4.2 Analyzing the Context

The first phase of the specialization strategy is to analyze the context. When CRISP-EM is placed within the context of a specific criminal case, it should be used to provide support to an investigator or prosecutor, not to mine for trends in case databases. Such a project is close enough to normal data mining that CRISP-DM in its original format would suffice.

Digital forensic investigations have four principal phases: evidence acquisition, examination, analysis and reporting. The acquisition phase collects evidence in a manner that preserves its integrity. Normally, a copy of the original evidence (image) is made and all further processing is done using the image (or a copy of the image). The second phase is evidence examination, which involves rudimentary processing of the evidence using keyword searches or examining locations specific to the operating system (e.g. for a user name). The third phase is evidence analysis. During this phase, the information obtained during the evidence examination phase is evaluated in the context of the case and the evidence is further processed to uncover facts about events, actions, etc. that are relevant to the case. The information provided by the examination phase is placed in context with the case and is further processed to uncover facts that stipulate to events, actions, etc. relevant to the case. The fourth and final phase is to present the evidence to the concerned parties in and out of court.

The context for evidence mining is phases three (evidence analysis) and four (evidence presentation). It is important to note that the data gathering aspects of the CRISP-DM methodology (part of the Data Preparation phase) and the digital forensic acquisition phase are not within the same context. Because of this context difference, it is more appropriate to use the term “data collation” instead of “data collection” in CRISP-EM.

Table 1. Original and renamed phases.

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

4.3 Renaming Generic Content

The original CRISP-DM phases and the renamed CRISP-EM phases are shown in Table 1. The first phase is renamed to Case Understanding because each evidence mining project is associated with a specific case. The names of the next two phases (Data Understanding and Data Preparation) remain the same as the intent of these phases for evidence mining is the same as for data mining. The principal differences, however, pertain to the last three phases. A specific evidence mining project is likely to span only one case. Therefore, the intent is to produce specific evidence for the case at hand rather than to build a model that can be used for future cases.

Consequently, the Modeling phase is replaced by the Event Modeling or Scenario Development phase. This phase creates plausible scenarios from the electronic evidence available in the data set. In the next phase, the evidence is presented to the investigator and/or prosecutor who evaluate the scenarios presented, select the relevant scenarios and extract the relevant evidence (hence the phase is renamed to Evaluation and Evidence Extraction). The final phase is Evidence Reporting, where the evidence is reported to an investigator, prosecutor or in court.

The renaming of terms continues within the details of the methodology. A notable example is Generic Task 1.2 in CRISP-DM “Collect Initial Data,” which is renamed to “Collate Initial Data” in CRISP-EM. This is done to distinguish between acquiring forensic evidence (data collection) and putting together evidence for analysis (data collation).

4.4 Specializing Generic Content

To maintain the context of an investigation, it is necessary to not only develop specialized tasks but also to specialize the phases and the generic tasks. In particular, the original CRISP-DM descriptions for the generic process phases must be adapted to fit within the evidence mining context. The adapted descriptions are shown below. The major changes from the original descriptions are shown in italics. It is important to note that these process phases fit within the analysis phase of the larger digital forensic process and are not meant to replace the overall process.

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

- **Data Understanding:** This phase starts with an initial data collation and proceeds with data familiarization, the identification of data quality problems, the discovery of patterns in the data and the detection of interesting subsets that create hypotheses for hidden information.

- **Data Preparation:** This phase covers all the activities involved in converting the initial raw data to the final data set, which is input to *event modeling* tool(s). Data preparation tasks are likely to be performed multiple times and not in any prescribed order. The tasks include table, record and attribute selection, *entity recognition and co-reference resolution*, and the transformation and cleaning of data for *event modeling* tools.

- **Event Modeling:** In this phase, various *evidence modeling and event reconstruction* techniques are selected and applied and their parameters are calibrated to optimal values. Typically, several techniques can be applied to an *evidence* mining problem. Many of these techniques have specific requirements on the form of data. Therefore, it may be necessary to return to the Data Preparation phase.

- **Evaluation and Evidence Extraction:** At this stage in the project, a *set of scenarios or event lines have been built* that are of high quality from a data analysis perspective. Before proceeding to the final reporting of the evidence, it is important to thoroughly evaluate the *scenarios/event lines* and review the steps executed in order to construct and extract *the relevant scenarios/event lines* that achieve the *case* objectives. A key objective is to determine if important *case* aspects have not been considered adequately. A decision on the use of *evidence* mining results should be reached at the end of this phase.

- **Evidence Reporting:** A project generally does not conclude with the creation of event lines and the extraction of evidence. Even if the purpose of evidence mining is to increase knowledge about the data, the knowledge gained should be organized and presented appropriately to enable the *investigator* to use it for *evidentiary purposes.* This *may* involve *augmenting chosen event lines with other data pertinent to the investigation.* In many cases it is the investigator, not the data analyst, who performs the *reporting* steps. However, even if the data analyst is not involved in the reporting effort, it is important for the *investigator* to understand the actions that must be carried out to make use of the *extracted event lines and evidence.*

4.5 Adding/Removing Content

New content has to be added to address evidence mining requirements whereas other content that does not make sense in the evidence mining context has to be removed. Some of the key changes are discussed below.

- **Initial Data Mining:** The development of event lines is a complex task that requires more advanced data pre-processing and preparation than "traditional" data mining. The initial data mining task was added to the Data Preparation phase to facilitate the additional inputs to the Event Modeling phase. The output of this task is a richer data set that includes classified and categorized data. Understanding the crime triangle of "willing offender," "enabling environment" and "vulnerable target" [6] will help in developing the pre-processed data as all three of these aspects are 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 would be very 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 a crime and how it occurred 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 what can be right or wrong (adapted from [6]). The set of hypotheses (scenarios) constitute a roadmap that enables the investigator to conduct an effective and efficient investigation. The original CRISP-DM Build Model task is replaced by the Develop Event Scenarios task. The replacement is necessary because

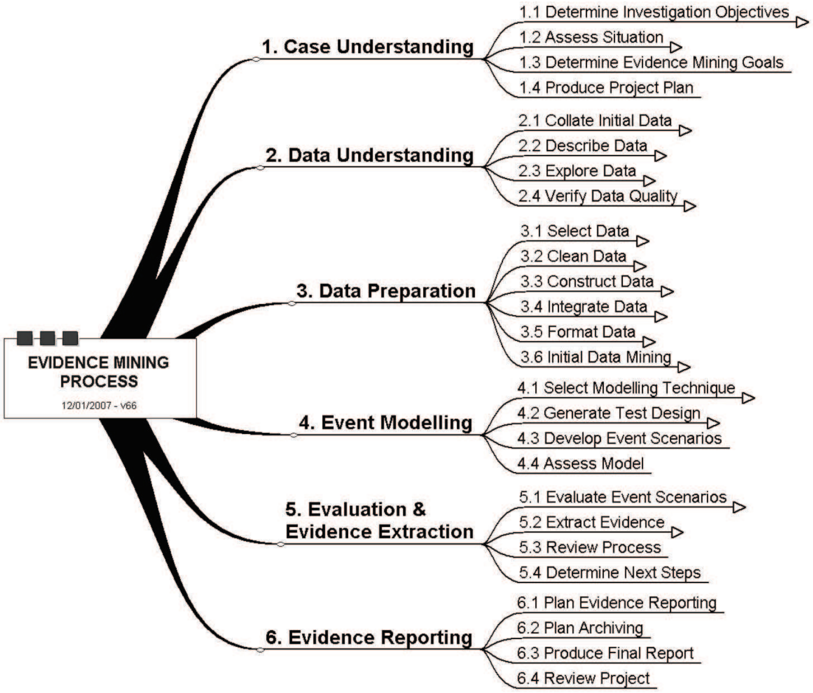

Figure 2. CRISP-EM second level.

the model built by the evidence mining process is in actuality the scenarios.

5. CRISP-EM Summary

The previous section discussed the development of a specialized process for evidence mining. This section summarizes the CRISP-EM process and provides details of the Data Preparation phase.

The major phases of the evidence mining process are shown in Figure 1. Figure 2 presents the next level of the CRISP-EM process in "mind map" format. Details of the Data Preparation phase are shown in Figure 3. Substantial further research is required to complete all the aspects of the process and to implement it completely.

6. Research Gaps

The CRISP-EM framework supports a structured approach for defining research gaps. CRISP-EM provides a level of granularity that makes it easier to identify where existing knowledge discovery and data mining techniques suffice and where new techniques would be required due to

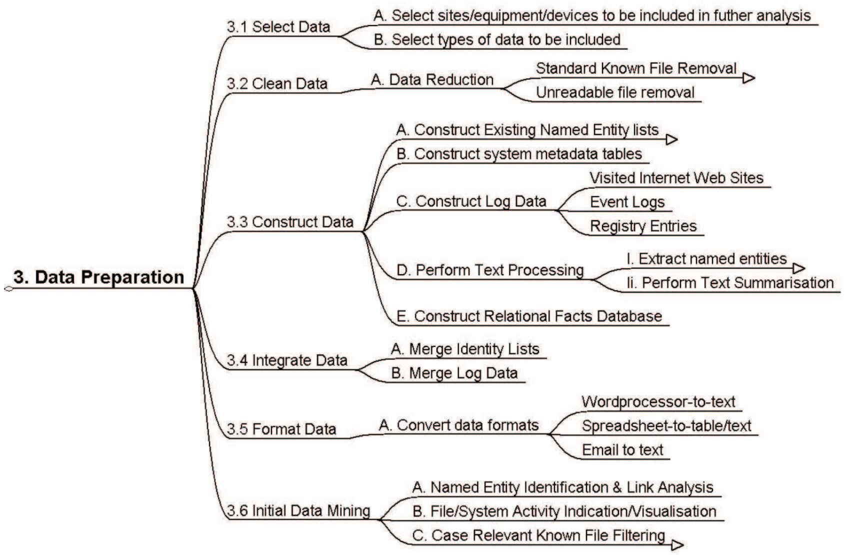
3. Data Preparation
3.1 Select Data
A. Select sites/equipment/devices to be included in futher analysis
B. Select types of data to be included
3.2 Clean Data
A. Data Reduction
Standard Known File Removal
Unreadable file removal
3.3 Construct Data
A. Construct Existing Named Entity lists
B. Construct system metadata tables
C. Construct Log Data
Visited Internet Web Sites
Event Logs
Registry Entries
D. Perform Text Processing
I. Extract named entities
Ii. Perform Text Summarisation
E. Construct Relational Facts Database
3.4 Integrate Data
A. Merge Identity Lists
B. Merge Log Data
3.5 Format Data
A. Convert data formats
Wordprocessor-to-text
Spreadsheet-to-table/text
Email to text
3.6 Initial Data Mining
A. Named Entity Identification & Link Analysis
B. File/System Activity Indication/Visualisation
C. Case Relevant Known File Filtering

Figure 3. Data Preparation phase.

the differences in the tasks and outputs of CRISP-DM and CRISP-EM. The biggest differences between CRISP-DM and CRISP-EM lie in the Event Modeling, and Evaluation and Evidence Extraction phases. It is, therefore, obvious that the largest research gaps would also be in these areas. Three examples of identified research gaps are described below.

- **Example Case Files:** Sample data sets are required for the new evidence mining techniques. 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 that contain mixed data must also be developed to test the efficiency of the algorithms. No such example case files currently exist. Research efforts should focus on developing plausible crime "stories" and ways for mixing them within other data sets. Furthermore, the example case files would have to be created automatically as creating them manually would be extremely time consuming.

- **Coping with Uncertainty:** Uncertainty is a major challenge when developing event lines. The available data is often incomplete, leading to beliefs that fall short of evidence and produce fallible conclusions. Probabilistic reasoning models [14] may be used to build scenarios; they also address problems such as co-reference resolution, record linkage and theme extraction. The association of probability values to event lines would also facilitate prioritization during the Evaluation and Evidence Extraction phase.

- **Automating Investigative Processes:** Human investigators have special skills and experience that enable them to extract evidence from unstructured information. However, the number of human investigators is limited, manual investigative processes are slow and laborious, and human concentration diminishes with fatigue. Automated knowledge discovery techniques can be parallelized to handle large volumes of data efficiently. Unfortunately, these techniques do not exhibit the skill of human investigators. Knowledge discovery techniques involving intelligent agents [8, 20, 22] can be used to automate certain aspects of the investigative process, reducing the burden on human investigators.

7. Conclusions

Forensic investigators are being inundated with massive volumes of electronic evidence, and the situation is only expected to become worse. A new generation of forensic tools are needed to automate analysis tasks that are now being performed manually. Research in knowledge discovery and data mining has developed powerful automated techniques for discovering useful patterns in massive quantities of data. Evidence mining is the application of these techniques in the analysis phase of digital forensic investigations. The CRISP-EM process described in this paper specializes the well-known CRISP-DM data mining methodology to provide sophisticated knowledge discovery and data mining support for digital forensic investigations. CRISP-EM is not yet a proven process; nevertheless, it offers a powerful framework for the initial, mostly manual, application of evidence mining. Also, it provides a basis for researching new methods and techniques for enhancing evidence mining and implementing the underlying processes.

References

[1] Association for Computing Machinery (ACM), (www.acm.org).

[2] S. Bandyopadhyay, U. Maulik, L. Holder and D. Cook, *Advanced Methods for Knowledge Discovery from Complex Data*, Springer-Verlag, Secaucus, New Jersey, 2005.

[3] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartzrysler, C. Shearer and R. Wirth, CRISP-DM 1.0: Step-by-Step Data Mining Guide, The CRISP-DM Consortium, SPSS (www.crisp-dm.org /CRISPWP-0800.pdf), 1999.

[4] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin and M. Chau, Crime data mining: A general framework and some examples, *IEEE Computer*, vol. 37(4), pp. 50–56, 2004.

[5] W. Chisum and B. Turvey, *Crime Reconstruction*, Elsevier, Burlington, Massachusetts, 2007.

[6] R. Clarke and J. Eck, Become a Problem-Solving Crime Analyst, Jill Dando Institute of Crime Science, University College London, London, United Kingdom (www.jdi.ucl.ac.uk/publications/other_publications/55steps), 2003.

[7] C. Clifton and B. Thuraisingham, Emerging standards for data mining, *Computer Standards & Interfaces*, vol. 23(3), pp. 187–193, 2001.

[8] I. Dickinson and M. Wooldridge, Towards practical reasoning agents for the semantic web, *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 827–834, 2003.

[9] S. Garfinkel, Forensic feature extraction and cross-drive analysis, *Digital Investigation*, vol. 3(S1), pp. 71–81, 2006.

[10] P. Gladyshev and A. Patel, Finite state machine approach to digital event reconstruction, *Digital Investigation*, vol. 1(2), pp. 130–149, 2004.

[11] Institute for Electrical and Electronics Engineers (IEEE), (www.ieee.org).

[12] KDNuggets, Data mining methodology poll (www.kdnuggets.com/polls/2004/data_mining_methodology.htm), 2004.

[13] J. Keppens and B. Schafer, Knowledge based crime scenario modeling, *Expert Systems with Applications*, vol. 30(2), pp. 203–222, 2006.

[14] K. Korb and A. Nicholson, *Bayesian Artificial Intelligence*, Chapman and Hall/CRC Press, Boca Raton, Florida, 2004.

[15] A. Louis, A. de Waal and J. Venter, Named entity recognition in a South African context, *Proceedings of the Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*, pp. 170–179, 2006.

[16] J. Mena, *Investigative Data Mining for Security and Criminal Detection*, Elsevier, Burlington, Massachusetts, 2003.

[17] M. Pollitt and A. Whitledge, Exploring big haystacks: Data mining and knowledge management, in *Advances in Digital Forensics II*, M. Olivier and S. Shenoi (Eds.), Springer, New York, pp. 67–76, 2006.

[18] M. Rogers, K. Seigfried and K. Tidke, Self-reported computer criminal behavior: A psychological analysis, *Digital Investigation*, vol. 3(S1), pp. 116–120, 2006.

[19] Scopus, (www.scopus.com).

[20] W. van der Hoek, W. Jamroga and M. Wooldridge, A logic for strategic reasoning, *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 157–164, 2005.

[21] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, California, 2005.

[22] M. Wooldridge, *An Introduction to Multiagent Systems*, John Wiley, Chichester, United Kingdom, 2002.