INFORMATION FORAGING THEORY AND APPLICATION TO COMPUTER FORENSICS

Yusra Shams
B.S., National University of Computer & Emerging Sciences, Pakistan, 2006

PROJECT

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

at

CALIFORNIA STATE UNIVERSITY, SACRAMENTO

FALL
2010
INFORMATION FORAGING THEORY AND APPLICATION TO COMPUTER FORENSICS

A Project

by

Yusra Shams

Approved by:

__________________________, Committee Chair
Isaac Ghansah, Ph.D.

__________________________, Second Reader
Robert A. Buckley

__________________________
Date
Student: Yusra Shams

I certify that this student has met the requirements for format contained in the University format manual, and that this project is suitable for shelving in the Library and credit is to be awarded for the project.

__________________________, Graduate Coordinator  ___________________
Nikrouz Faroughi, Ph.D.  ____________________________

Department of Computer Science
Computer Forensics involves analysis of a digital artifact to look for evidence associated with an illegal activity. With the growing role of electronic media in our daily lives, most crimes involve the use of electronic media in the process and hence Computer Forensics has become a vital aspect of Crime Investigation Units. There is a variety of Computer Forensic tools (e.g. FTK, EnCase etc.) available to help the experts, but these tools only help in categorizing the information. Examples of categorization include, searching for key words, looking for the deleted files, showing the directory structure etc. Thus to help the decision making process Computer Forensics completely depends on the knowledge, skill and expertise of the Forensic Expert. There is no tool available that can help the forensic experts in the decision making process.

The goal of this project is to present a computational cognitive model which simulates the behavior of a computer forensic analyst. Mainly two types of decision making are involved in the computer analysis process:

- Which path to follow when there are a number of options?
- When is it the right time to leave one patch of information and look for evidence somewhere else?
The presented cognitive model simulates both kinds of decision making. This project proposes the use of “Information Foraging Theory” to achieve the purpose. Information Foraging Theory is an approach to understanding how strategies and technologies for information seeking, gathering, and consumption are adapted to the flux of information in the environment. This theory assumes that people, when needed, will modify their foraging strategies or the structure of the environment to maximize their rate of gaining valuable information.

The computational cognitive model of a foraging system (to perform Computer Forensics) presented in this project embodies a built-in cognitive process, which makes use of the information foraging techniques. The presented model forms the basis for future development of an intelligent computer forensic system, which can potentially reduce the dependency of the computer forensic process on the experience and skills of forensic expert, though it cannot be completely removed.

______________________, Committee Chair
Isaac Ghansah, Ph.D.

____________________
Date
DEDICATION

To my better half Salman, whose love is the driving force of my life.
ACKNOWLEDGMENTS

First and for most, I thank God who makes everything possible.

I sincerely thank my advisor, Professor Isaac Ghansah for his helpful comments and feedback. I genuinely appreciate the valuable time and effort Prof. Ghansah dedicated towards this project in spite of his busy schedule. I thank Professor Robert A. Buckley, my second reader, for reading my project report and making beneficial suggestions on content. I also thank Professor Nikrouz Faroughi for his comments on the format of this report.

Finally, words alone cannot express the thanks I owe to my husband Salman, my parents Shams and Sabiha, my sister Madiha and my aunt Nasreen for their endless support and prayers throughout my Master’s Program. I love you very much.
TABLE OF CONTENTS

Dedication ........................................................................................................................................ vi
Acknowledgments .......................................................................................................................... vii
List of tables ........................................................................................................................................ xii
List of figures ....................................................................................................................................... xiii

Chapter

1. INTRODUCTION ..................................................................................................................... 1
   1.1. Introduction to Computer Forensics ................................................................................ 1
       1.1.1. History of Computer Forensics .............................................................................. 2
       1.1.2. Process of Computer Forensics ........................................................................... 3
       1.1.3. Analysis Phase ....................................................................................................... 5
       1.1.4. Problems/Issues in the Analysis Phase of Computer Forensic ......................... 7
   1.2. Background of Study ........................................................................................................ 7
   1.3. Problem Statement ........................................................................................................... 14
   1.4. Overview of Related Research Work in Context of this Research ............................ 14
   1.5. Project Overview .......................................................................................................... 16
       1.5.1. Idea Origination ................................................................................................. 16
       1.5.2. Research Work ................................................................................................. 16
   1.6. Project Objective ......................................................................................................... 17
   1.7. Report Organization ..................................................................................................... 18

2. INFORMATION FORAGING THEORY AND COMPUTER FORENSICS .................. 20
   2.1. History of Information Foraging Theory ...................................................................... 20
   2.2. Introduction to Information Foraging Theory .......................................................... 21
4.4. Future Work and Implementation ................................................................. 73

Appendix A. Sample Questionnaire for CF Task Information Collection Survey Using CIT ..... 77

Glossary .................................................................................................................. 78

References .............................................................................................................. 79
LIST OF TABLES

1. Table 1: Computer forensic process stages and activities ........................................ 4
2. Table 2: Nature of computer forensic tasks given in percentages ............................ 31
3. Table 3: Relevance rating on a four-point scale ..................................................... 39
4. Table 4: Relevance rating & information scent rating on a four-point scale ............... 41
5. Table 5: Identification of Problem Space of Computer Forensics ............................ 42
6. Table 6: Number of nodes visited by the analyst in a search tree [6] ....................... 50
7. Table 7: Production rules for SNIF-ACT-CF model designed for Computer Forensics . 67
# LIST OF FIGURES

1. Figure 1: Phases of Computer Forensics [19] ................................................................. 4
2. Figure 2: Analysis phase of Computer Forensics [20] .................................................... 5
3. Figure 3: User interface of FTK, for a computer forensic case. [4] ................................. 10
4. Figure 4: User interface of FTK, showing “from emails”. [4] ........................................... 11
5. Figure 5: FTK user interface, displaying indexed search results for a forensic case. [4] .... 12
6. Figure 6: Sample FTK case report. [4] .......................................................................... 13
7. Figure 7: Bar-graph, representing the data given in Table 2. ........................................... 32
8. Figure 8: Laboratory setup for experiment [5] ................................................................. 37
9. Figure 9: Web Behavior Graph Representation [5] .......................................................... 40
10. Figure 10: Files and directories arranged in a hierarchy [7] ........................................... 43
11. Figure 11: Directory structure display of operating system. [7] ..................................... 44
12. Figure 12: Information scent cues presented by a search engine in response to a user query 46
13. Figure 13: Information scent cues presented by the directory tree structure. [18] ......... 47
14. Figure 14: Search tree [5] ........................................................................................... 49
15. Figure 15: Cognitive structure showing proximal cues and desired information. [5] ........ 52
16. Figure 16: Activation Equation [18] ............................................................................ 53
17. Figure 17: ACT-R model of cognition [9] ..................................................................... 56
18. Figure 18: Computational cognitive model for SNF-ACT and SNIF-ACT-CF [5] ........... 59
19. Figure 19: A group of inter-associated chunks.[5] ................................................................. 60

20. Figure 20: A subset of the declarative knowledge of the SNIF-ACT-CF model simulation of user working on the above mentioned task. .......................................................................................... 65
Chapter 1

INTRODUCTION

1.1. Introduction to Computer Forensics

Computer Forensics, also called Computer Forensic Science, is a branch of digital forensic science, which pertains to finding legal evidence on computers and related digital storage media. With the exponentially growing use of internet and digital media, most crimes today make use of some sort of electronic media, and hence the field of computer forensics is growing immensely. Computer Forensics is applied to a variety of situations. A few common examples include unauthorized disclosure of corporate information and data, damage assessment, industrial spying, criminal fraud and deception cases, use of organizations network for illegal activities. In some cases criminals involved in other types of felonies (e.g. robberies, smuggling, drug dealing) have their data and information saved on personal computers or digital devices and storage media. There are countless other situations where computer forensics is used. This explains the growing need and importance of Computer Forensic and this is the reason Computer Forensic Unit has become an integral part of any Investigation Agency.

As explained above, Computer Forensics is used on various different situations but the main goal of computer forensics is to analyze and explain the current state of the given computer system or any related digital artifact such as an electronic storage medium (e.g. hard disk, flash drive, CD ROM). Computer forensics is goal oriented. In any given case there are always a few questions that need to be answered after the analysis. So the scope of a computer forensic analysis can vary from simple information search and retrieval like finding an electronic document or an
e-mail to reconstructing a series of events based on various findings on the computer system (e.g. e-mail conversations, deleted items, instant messages, activity logs etc.).

The “Digital Forensic Research Workshop” (DFRWS), in 2001, defined Computer Forensics as follows:

“The use of scientifically derived and proven methods towards the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.”

1.1.1. History of Computer Forensics

The field of Computer Forensics began to evolve more than 30 years ago. It originated in the law enforcement and military investigation units in the United States, when the investigators started seeing criminals using information technology in their criminal endeavors. At the same time the use of electronic media became a common practice to save important, confidential, and certainly secret information. Therefore Government personnel started to conduct forensic examinations in response to potential security breaches. The purpose behind this was two-fold, firstly to investigate the particular breach, and secondly to learn how to prevent future potential breaches. The fields of information security and computer forensics are correlated. The focus of information security is on protecting information and assets, where as computer forensics focuses on the response to technological offenses of any magnitude and kind. Eventually, the two fields of study have converged.
Over the last three decades the two fields have grown immensely. Information security and computer forensics are used at the local, state, and federal level by law enforcement and military. Private organizations started to hire information security and computer forensic professionals, or contracting such professional or firms on need basis.

Even today, the computer forensic field continues to grow on a daily basis. With time, forensic firms and private investigators are gaining knowledge and experience in the field. And law enforcement and the military continue to train more and more of their personnel in the response to crimes involving technology.

1.1.2. Process of Computer Forensics

A pre-defined process is followed in each case of computer forensics. This involves a set of steps. This is done to make sure that the integrity of information is not compromised at any time, so that the collected information can be presented as evidence in a court of law. The steps involved in the process of Computer Forensics include identification, preservation, collection, examination/analysis, documentation/presentation of computer evidence.

The entire process should follow the “United States Department of Justice” (USDOJ) rules of evidence and legal processes. The process also includes the factual reporting of the information found and the ability to provide expert opinion in a court of law or other legal proceeding as to what was found. A chain of custody should be followed in order to make sure that the integrity of information is not compromised at any level. Chain of custody is a means of accountability that shows who obtained the evidence, when and where the evidence was obtained, who secured the evidence, and who had the possession of evidence at any point of time. Handing and taking over of information or digital media is done through paperwork signed by both parties.
Figure 1 shows the phases of Computer Forensics from start to end provides a general overview of the whole process. Table 1 gives an account of the specific activities performed in each stage.

Table 1: Computer forensic process stages and activities

<table>
<thead>
<tr>
<th>Sr. no.</th>
<th>Stages</th>
<th>Activities performed during the stage</th>
</tr>
</thead>
</table>
| 1       | Plan   | • Identify the purpose of investigation  
 |         |        | • Identify resources required         |
| 2       | Acquire| • Identify sources of digital evidence  
 |         |        | • Preserve digital evidence            |
| 3       | Extract| • Make an exact copy of the evidence for analysis  
 |         |        | • Calculate hash of the original media and the copy to ensure the copy is exactly similar to the original digital media. |
| 4       | Analyze| • Identify tools and techniques to use  
 |         |        | • Process data                        |
|         |        | • Interpret analysis results           |
| 5       | Report | • Report findings                      |
|         |        | • Present findings                     |
This project focuses only on the analysis phase and how it can be made more efficient and productive.

1.1.3. Analysis Phase

Analysis phase is the phase where actual examination is performed, to look for the required evidence. The remaining phases and tasks (i.e. identification, acquisition, documentation and maintaining a chain of custody) ensure that the information is not altered or modified during the process. If integrity is not ensured for the information, then it cannot be used as evidence in a court of law, no matter how critical the information is. Figure 2 shows the ongoing process of analysis phase.

Figure 2: Analysis phase of Computer Forensics [20]
The real media analysis is done during analysis phase. The analysis phase usually involves the following steps [2]:

- Verify integrity of the acquired image of the system. This is usually done by taking a hash [1] of both the original media and the image and matching them for integrity. Example of hash functions includes MD5, SHA1.

- Recovering Deleted files and folders

- Understand the goal of analysis process and determining the keyword list according to the goal of the search process.

- Determining the time zone settings of the suspect system, and knowing what time frame is important for the case at hand, to fasten the search process.

- Examine the directory tree for something that looks out of place or suspicious. Look for doubtful or suspicious software or program installations (i.e. Stenographic tools [3])

- Perform keyword searches on the pre-determined keyword list. Search for information in the slack and unallocated space.

- Search for relevant evidence types, for example, documents, graphics, spread sheets etc.

In most cases, forensic analysis is performed to find some sort of evidence associated with an illegal activity. The person committing the crime or illegal activity, sometimes, hides the information or stores it in a manner that is not very easy to find. The key is to look for the obvious information first and then look for tricky or hidden information.
1.1.4. Problems/Issues in the Analysis Phase of Computer Forensic

“Technological progress is like an axe in the hands of a pathological criminal.”- Albert Einstein

With the exponentially growing technology, the job of computer forensic experts is getting tougher and tougher every day, because it is practically impossible to keep up with the advancements of electronic media all the time. Better and more sophisticated digital artifacts are being introduced to the market every day.

One of the biggest problems is dealing with large media, containing enormous amounts of data. For example, if the criminal case happens to involve an organization, then it is likely that the forensic experts have to analyze organization’s servers. Moreover digital storage media with more capacity are also available in the market for personal use, e.g. large capacity (internal or external) hard drives, flash drives and storage disks. Another major problem is to collect and analyze data for evidence, while the system is running.

Another limitation is the time factor, as all computer forensic cases are bound by time. Hence the speed of doing things is very important. The goal of forensic expert is to find as much of evidence, and collect as many facts as possible in the limited time period, dedicated to the analysis phase.

1.2. Background of Study

In Computer Forensics, the tools used are not nearly as important as the person using them. This does not mean that the tools are useless or not helpful at all. There is a variety of Computer Forensic tools (e.g. FTK, EnCase etc.) available to assist the experts but still the
process heavily relies on the expertise and experience of the forensic expert performing the process, because these tools only help in categorizing the information. Examples of categorization includes, searching for key words, looking for the deleted files, showing the directory structure etc. Thus the decision making process of Computer Forensics depends entirely on the knowledge, skill and expertise of the forensic expert. There is no tool available that can help the forensic experts in the decision making process.

There are various open source and commercial tools available to assist the computer forensic experts in the process of computer forensics analysis and investigation. The most well known and used forensic tools are listed below:

- Encase
- FTK
- PTK Forensics
- The Sleuth Kit/Autopsy Browser
- The Coroner's Toolkit
- Computer Online Forensic Evidence Extractor (COFEE) – by Microsoft
- Selective file dumper

Typical forensic analysis includes a manual review of material on the media, reviewing the Windows registry for suspect information, discovering and cracking passwords, keyword searches for topics related to the crime, and extracting e-mail and pictures for review. To better understand what functionality the existing computer forensic tools can provide, FTK is described below.
FTK is one of the famous computer forensic toolkits. It has variety of features to facilitate the process of computer forensics analysis. These features include various approaches to the classification and categorization of information under analysis. Examples of these features include indexed and live searches for key words, bookmarking, password dictionary creation, slack and unassigned memory description, and report generation etc.

Figure 3 shows the user interface of FTK for a computer forensic case. Near the top of the screen shown in the figure, there are sx tabs. Each tab has a different feature associated with it. The “overview tab” (shown in the figure) provides a dashboard view of the data contained on the digital media. It has three different views of the data, namely:

- **Evidence Items**
  
  This view of data items provides a quick look at the status of analysis, for example, the total number of files/items, the number of items checked, and number of flagged items etc.

- **File Status**
  
  This view presents the statuses of data items. For example, files with bad extension, deleted files, encrypted files, and duplicate files etc.

- **File Category**
  
  This view presents the type of the data item/file. For example, documents, spreadsheets, databases, executables, and images etc.
FTK allows the forensic expert to drill down in each category and view the data items/files present in that category. Figure 4 shows the drilled down view in the “From E-mail” category. All the email messages are displayed in the bottom pane of the screen. The selected message (Message0001) contents are displayed in the top right pane of the screen. All the user must do is click on a message at the bottom of the screen and the message will show up in the file viewer on the right.
Another feature of FTK is the key word/phrase search. Figure 5 shows the search feature of FTK. User needs to type in the key word/phrase in the “Search Term” text box and click OK. All the relevant data items will be displayed for analysis.
Figure 5: FTK user interface, displaying indexed search results for a forensic case. [4]

Reports can also be printed on the computer forensic case findings. A sample FTK forensic case report is shown in figure 6.
All the features of FTK explained above, help to categorize the data items. This categorization of information assists the forensic expert in understanding the organization and type of content stored on the digital artifact under analysis, but the forensic expert has to use his own knowledge and decision making skills to decide which items to pick for analysis first. The available forensic tools does not help the forensic experts in the decision making process. For the forensic tool there is an equal probability of finding the required evidence in all the options. The forensic expert has to use his past experience and skills to determine which path/option to pick.
The decision making becomes very time consuming if there is a large number of options available.

Information Foraging Theory (IFT) can be applied to the process of “Computer Forensics” to make it more productive and less time consuming. This project focuses on making the process of computer forensics more efficient, robust and cost effective with the use of IFT – discussed in detail in chapter 2.

1.3. Problem Statement

“Nothing drives basic science better than a good applied problem.” – [17]

Currently available, computer forensic tools (e.g. FTK, Encase etc.), do not provide the functionality of assisting the user in decision making process. This makes the process of computer forensic analysis highly dependent on the knowledge and past experience of the forensic expert. The existing tools only help by categorizing the information. Examples of categorization includes, searching for key words, looking for the deleted files, showing the directory structure etc. Thus to help the decision making process computer forensics completely depends on the knowledge, skill and expertise of the Forensic Expert. As indicated previously, there is no tool available that can help the forensic experts in the decision making process.

1.4. Overview of Related Research Work in Context of this Research

Information Foraging Theory (IFT) is a scientific theory and it has been used to solve many Human Computer Interaction (HCI) problems [5] [10]. There are three categories of the utility of IFT, namely:
• **Understanding** human-information interaction

• **Predicting** human-information interaction

• **Improving** human-information interaction

Many computational cognitive models and software systems and applications have been designed and built to utilize IFT in the above mentioned ways. Peter Pirolli and Wai-Tat Fu form Palo Alto Research Center (PARC) have developed a model of cognition to simulate the behavior of Web user. This model is called SNIF-ACT, which stands for Scent-based Navigation and Information Foraging in the ACT architecture [5]. Pirolli and Fu also created a system called User Tracer to test the functionality of SNIF-ACT. SNIF-ACT is a tested and working model of cognition based on the principles of IFT.

IFT has helped the researchers to structure and organize the website in a manner that the content is more accessible to the users. Mathematical models have benefited from Information Foraging Theory to predict the following [5].

• **User flow**: The pattern of users visiting states throughout the interface

• **Success matrices**: Such as the average number of steps to success or the rate at which users will achieve success

• **Stickiness**: The amount of time that users will spend interacting with a navigation interface before leaving it (e.g., the number of page visits at a Web site)

This project studies the existing models of cognition based on IFT and researches on how the behavior of a forensic expert can be understood, predicted and improved with the help of IFT principles and models.
1.5. Project Overview

With the growing role of electronic media in our daily lives, most crimes involve the use of electronic media in the process and hence Computer Forensics has become a vital division of Crime Investigation Units. This project discusses the process of computer forensics and its growing importance in the present world. The project focuses more on the ‘Analysis Phase’ of the computer forensic process and discusses the problems and issues involved during this phase.

1.5.1. Idea Origination

This project started when my master’s project advisor involved me in the research to determine if Computer Forensics could benefit from Information Foraging Theory (IFT). My study proved that Computer Forensics provides an ideal environment for the application of IFT and hence could benefit from the use of its techniques and models. This finding derived the research work on the main problems of Computer Forensics and how they could be solved with the use of IFT.

1.5.2. Research Work

This project involves in-depth study of “Information Foraging Theory”. This involves research on various techniques and models of Information Foraging to propose a way these techniques could be used to solve the prevailing problems of computer forensic analysis. Information Foraging Theory is an approach to understanding how strategies and technologies for information seeking, gathering, and consumption are adapted to the flux of information in the environment. The theory assumes that people, when possible, will modify their strategies or the structure of the environment to maximize their rate of gaining valuable information. Field studies
inform the theory by illustrating that people do freely structure their environments and their strategies to yield higher gains in information foraging.

This project presents a computational cognitive model of a computer forensic system, which simulates the behavior of a computer forensic expert. The presented model uses the SNIF-ACT cognitive model as a base architecture. SNIF-ACT stands for Scent-based Navigation and Information Foraging in the ACT architecture. It simulates the behavior of web users and it was created by Peter Pirolli and Wai-Tat Fu [5]. The presented model replaces the production rules and memory content of the SNIF-ACT model to make it work for Computer Forensic environment. The new model is named SNIF-ACT-CF, where CF stands for Computer Forensics. SNIF-ACT-CF simulates the decision making process of the forensic expert in major two ways:

- Indicates the path to pick when there are a number of options
- Estimates the right time to leave one patch of information and look for evidence somewhere else

This project also discusses the advantages and limitations of using this model and the future work and implementation aspects.

1.6. Project Objective

The objectives of this project are as follows:

- Identify the problems and issues in the analysis phase of the Computer Forensics Process. It includes the study of existing Computer Forensic tools to identify their short comings.
Study Information Foraging Theory and research on its different models to figure out how information foraging techniques can be used to solve the problems of Computer Forensic Analysis.

Present a computational cognitive model of a system, which simulates the behavior of computer forensic experts during the analysis phase.

1.7. Report Organization

This report is organized as follows:

- Chapter 1: *Introduction*

  This chapter gives an overview of the computer forensic process. It gives the background of this study and explains the problem statement. This chapter gives an account of previous research done related to the project, and compares it to this project. It also defines project’s objective and scope.

- Chapter 2: *Information Foraging Theory and Computer Forensics*

  This chapter gives an outline of information foraging theory. It explains how the computer forensic process can benefit from the use of this theory. It presents the analogies between conventional Information foraging and computer forensics, and explains how computer forensics presents an ideal foraging environment to benefit from information foraging techniques.
• Chapter 3: *Modeling Information Foraging Based Computer Forensic System*

This chapter first presents a detailed protocol analysis of the cognitive data involved in the process of computer forensics. It also explains the methods to collect cognitive data for analysis. Next, the creation and use of reference information, called information scent is described. It also describes some existing cognitive models, namely, ACT-R and SNIF-ACT. Finally, it presents a computational cognitive model that simulates forensic expert’s information foraging behavior. It discusses the advantages and disadvantages of this model. It also talks about the future work and implementation of this model.

• Chapter 4: *Conclusion*

This chapter presents a summary of the outcomes of this research project. It describes the advantages and limitations of the presented idea. It also talks about the future work and implementation aspects of the research.
20

Chapter 2

INFORMATION FORAGING THEORY AND COMPUTER FORENSICS

In order to benefit from information foraging techniques it is important to answer the following questions. What is Information Foraging Theory? Why is it beneficial to use it in Computer Forensics? What factors constitute the environment of computer forensics? And what are basic computer forensic tasks? This chapter explores these questions and provides a detailed account on each of them.

2.1. History of Information Foraging Theory

The anthropologists and ecologists developed a theory in 1970s to explain animal foraging techniques. This theory mainly explained that the animals assess the situation constantly. The animals consider both, the available food and the cost of obtaining it. By doing so they determine whether to stay for a while or move towards the next food location. This is an intelligent process. The human process works in the same way for information searching. In early 1990s, two scientists, Peter Pirolli and Stuart Card, from Palo Alto Research Center (PARC) pointed out the similarities between users' information searching patterns and animal food foraging strategies [21]. Later on, they began to apply what they called “foraging theory” to information hunting and gathering. After many years, scientists tried to develop algorithms and ideas that use these concepts in real interactive systems. One example of such system is the modeling of the user’s behavior in browsing.
2.2. Introduction to Information Foraging Theory

Information Foraging Theory tries to explain the information seeking behavior in human beings. The literal meaning of ‘Forage’ is to-collect but it is generally used to refer to the process of finding food/material/information by humans and animals.

Information Foraging Theory is an approach to understanding and explaining the strategies and technologies used by the living organisms for information seeking, gathering, and consumption. It also explains that all these activities are adapted, by the information seeker, to the flow of information in the environment. This means that the information seeker, when possible, will modify their strategies or the structure of the environment to maximize their rate of gaining valuable information. Therefore, information foraging theory suggests that people do freely structure their search environments and their strategies to yield higher gains in information foraging. [11]

Information foraging is classically explained using the analogy/example of food foraging by living beings (usually birds, for the ease of understanding and simplifying the foraging environment). Imagine a predator, such a bird in search of prey/food. The scope of search of the bird is called its environment. Food or prey is not evenly distributed throughout the environment. Rather, the environment has a patchy structure and the valuables are present in form of clusters. The bird does not have the ability to scan the complete environment at a time. Therefore, it scans the environment cluster by cluster, looking for the valuables, and constantly assessing the cost of search and the probable value of the findings. Based on this assessment it makes a decision whether to continue searching in the current patch or to move to the next patch. Any forager is
expected to look for patches of high utility and to leave patches when diminishing returns are detected.

The massive growth in technology has lead to an explosive increase in recorded information. Therefore it is becoming more and more difficult to provide people with the facilities to search and understand the available information. Since human minds are limited in their ability to keep pace and allocate time, there is latent demand for tools to support and facilitate the search process.

The next section discusses why it is a good idea to apply information foraging techniques to computer forensics. It is done with the help of food foraging analogy and comparing it to the foraging behavior in computer forensics.

2.3. Analogies between Food Foraging and Computer Forensics

The formalism of computer forensics (including the environment in which it is conducted) and knowledge and purpose of the forensic expert can be viewed as an example of information foraging. Consider the example of a predator, such as a bird in search of prey. The bird faces a problem of finding food and deciding what to eat, keeping in mind its resources (energy and time). An analogous situation in computer forensics might be a computer forensic expert or researcher analyzing a digital media storage device for some evidence. Discussed below are the characteristics that form the basis for the idea of applying information foraging theory to computer forensic analysis.
2.3.1. Goal Driven

Computer forensic analysis, like the bird’s search for food, is driven by goal(s). Before starting the analysis, it is already known to the forensic expert what he wants out of the analysis. It can be a single piece of information, such as a document, an e-mail, a picture or some file of interest, or it can be the activity log report of the system. The goal serves as the point of reference throughout the analysis and keeps the forager focused.

2.3.2. Patchy Information Distribution Pattern

Classical information theory suggests a patchy information structure. Computer forensics environment also presents a patchy information organization. Data is organized on digital media in some sort of file system. The directory structure of file systems is hierarchical in nature. Thus the organization of data in directory structure takes form of hierarchical information patches.

2.3.3. Use of Relevant Experience

Past experience of the information forager plays a vital role in the process of foraging and decision making. This is also true in the case of computer forensic analysis, as it also heavily depends on the knowledge and skills of the computer forensic analyst.

2.3.4. Time Constraint

The process of information foraging is bound by time. For example, the purpose of a bird looking for food is to maximize the rate of gain, which means to gain maximum energy (food) per unit time. Same is case in computer forensic analysis. Depending on the case at hand, specific number of work hours is dedicated to the analysis phase. The goal of the forensic expert is to find
as much information/evidence related to the case as possible. This gives birth to the concept of “Optimal Foraging”, discussed in detail in the next section.

2.4. Optimal Foraging Theory

“An Investment in Knowledge always pays the best interest.” – Benjamin Franklin

In psychological terms, increasing the rate at which people can find, make sense of, and use valuable information improves the human capacity to behave intelligently. Conceptually the optimal forager is one that has the best solution to the problem of maximizing the rate of net energy returned per effort expended, given the constraints of the environment in which he lives. [5]

2.4.1. Information Patch Use

Peter Pirolli explains the phenomena of Information Patch use with the help of Marginal Value Theorem (MVT) [21] [5]. Consider the example of a predator, an animal looking for prey. The animal travels through the environment, encountering patches of food. The travel time \( \tau \) is the (mean) time required to travel from one patch to another. Once in a patch, the animal gains energy at a rate that decreases with time in the patch. This is because the animal is using up the resources in that patch. This means that each additional unit of time spent in the patch results in a smaller gain in energy. This proves that gains are subject to diminishing returns.

It is the strategy of animals that specify how long to spend in each patch encountered. The optimal time spent in a patch maximizes the rate of energetic gain. The resulting rate of energetic gain is called the maximum rate of gain. When the animal starts foraging in a patch, its rate of gain will be above this maximum rate. As time in the patch increases, the rate of gain
decreases, and eventually it becomes more profitable for the animal to travel to a new patch. According to MVT the optimal time to leave one patch for another is when the rate of gain on a patch equals the maximum rate of gain. MVT provides the optimal solution because if the animal stays longer than the optimal time, it is getting a rate on the patch that is lower than the rate it could achieve by leaving the patch and following the optimal strategy. A robust qualitative prediction from the MVT is that the time spent in a patch type increases as the travel time increases [21].

2.4.2. Prey or Diet Choice

Prey choice is also explained by Pirolli and Fu in context of information foraging in patches [21] [5]. During the course of foraging an animal encounters various types of prey item. Each type of item is characterized by its energy content $e$ and the time $h$, required to capture and consume it. Let $\lambda$ is the rate at which each prey type is encountered. Whenever a new item is encountered, there are two possible behavioral options. It can accept (capture and consume) or to reject an item of a given type. The limitation is that the animal cannot encounter other items while it is consuming an item. The purpose is to maximize the rate of energetic gain, therefore the items with higher energy content $e$ are always consumed and the items with lower energy content are always rejected. So the choice for consumption of any type of item is either all or none.

Consider two types of available items, Type 1 and Type 2. The profitability for each item will be “$e/h$”, where $e$ is the energy provided by the item and $h$ is the time required to handle the item. Let the profitability of Type 1 item is higher than that of Type 2 item, that is, $e_1/h_1 > e_2/h_2$. Then Type 1 items should always be accepted, and Type 2 items should be accepted if and only if $e_2/h_2$ is greater than the rate of gain when only Type 1 items are taken. This rate depends only
on the parameters for Type 1 items, so whether Type 2 items should be accepted does not depend on their encounter rate.

The optimal policy can be understood with another perceptive. The item with the higher $e/h$ should always be consumed because it provides the best return of energy from time spent in consuming the item. The item with the lower value of $e/h$ should be consumed if the energy that it yields is more than the energy that could have been obtained if the handling time had been devoted to searching for and eating more profitable items. For more information on the optimal policy in terms of the maximum possible rate of gain, or to read the complete technical paper refer to Optimal Foraging Theory by Houston and McNamara (1999) [21].

2.5. Information Foraging Environment in Computer Forensic Analysis

A typical basic information foraging environment consists of two main areas, namely, the information environment and the task environment. These two environments are tightly coupled and organized in an inter-connected fashion.

Information environment is basically how the information is organized and distributed in the environment. It also defines the boundaries of the search for the forager and how he should look for the information that is valuable to him. On the other hand the task environment consists of the purpose of the foraging and the goals the forager wants to achieve.

Therefore, in order to benefit from information foraging techniques, computer forensics should have similar environment structure as well. The following study discusses how computer forensics environment bears similarities with the information foraging environment, and can benefit from its techniques.
2.5.1. Information Environment

In computer, all the electronic data/information is stored in the hard drive. And it is organized in some sort of file system. On the high level the user can see the data/files arranged in various directories. These file systems and directories organize the information into a hierarchical structure. So it's easier to access and analyze the group of information which is located on the same location in the hierarchy. Here I would like to borrow the word “information patch” from the classic information foraging theory, for this group of information, organized at the same locality in the Information environment. The main property of this information patch is that it’s easier for the forager to look for information within the patch than moving to another patch. One information item/file constitutes a basic information patch in case of computer forensics. It can be any kind of file, for example, text file, execution file, log file, image file etc.

A forager is expected to seek out patches of high utility and to leave patches when diminishing returns are detected.

2.5.2. Task Environment

The task environment is shaped by the actions the forager takes to achieve a certain goal. To get a detailed understanding of the task environment of computer forensics we have to figure what kind of tasks are forensic experts more interested in, what kind of information they want to get and what is the purpose of their forensic analysis, and what course of action do they take.

Now the question arises how to collect this information, because this information should be as accurate as possible. A renowned scientist and researcher in the field of information foraging, Peter Pirolli, suggests that this data could be collected by creating a survey. That is how
he accomplished the purpose while researching on the task environment of World Wide Web (www).

The data regarding important tasks can be collected by preparing a survey for the experienced computer forensic experts. The survey should involve sufficient number of computer experts to make the results more valuable, meaningful and helpful.

Critical incident technique (CIT) can be used to collect the survey data. It has many advantages over the other techniques [12].

- Flexible method that can be used to improve multi-user systems.
- Data is collected from the respondent's perspective and in his or her own words.
- Does not force the respondents into any given framework.
- Identifies even rare events that might be missed by other methods which only focus on common and everyday events.
- Useful when problems occur but the cause and severity are not known.
- Inexpensive and provides rich information.
- Emphasizes the features that will make a system particularly vulnerable and can bring major benefits (e.g. safety).
- Can be applied using questionnaires or interviews.

2.6. Detailed Task Information Collection

It is important to know precisely what the users want to achieve from the system, so that the system can be built to facilitate their activity, and help them achieve their goal efficiently. For
collecting this information, the best and the most logical method would be to contact the users of the system (in our case forensic experts) and ask them about their goals and the course of action taken to achieve the goal, in their forensic activity.

Renowned researcher in Information Foraging, Peter Pirolli, used the survey method (involving Critical Incident Technique) followed by the laboratory experiments, to collect data for his research on the task environment in Web/internet search. This project makes use of the same technique to collect the detailed task data for computer forensics. The method and application of these techniques is discussed below.

2.6.1. Survey Method Using Critical Incident Technique (CIT)

The purpose of the survey is to know the key tasks that users want to perform to achieve their objectives. In this technique a survey questionnaire is prepared for computer forensic experts. This survey is designed to collect information about the key goals and activities of the forensic experts. The survey should be prepared very carefully. The questions should be accurate yet descriptive to capture the true picture of the forager’s activity. Refer to Appendix A for a sample survey questionnaire.

This project uses the Critical Incident Technique (CIT). CIT, by definition, is a set of procedures used for collecting direct observations of human behavior, which have critical significance according to predefined criteria. Therefore the questionnaire should ask the computer forensic expert to describe precisely one of the recent forensic activities where he/she successfully found some useful information which was helpful in solving the case in hand.
2.6.1.1. Sample Selection

The sample selection is one of the critical activities, that need to be done keeping in mind a few things. Firstly the sample size should be sufficiently large, so that the results can be generalized over the domain, rather than just being applicable to the specific experiences of a few forensic experts. The results from a small sample size can be misleading and would kill the whole purpose of the system. After agreeing upon a fairly large sample size, the next concern is how to select the sample in a way that gives the best results. Therefore, secondly it is important that the individual computer forensic experts picked in the sample have sufficient experience in computer forensics. Experience also indicates, generally, that they get better and efficient at their work too. So the resulting system can benefit more from their knowledge and expertise.

Thirdly the sample should represent diversity in their experiences. If most of the people on the sample are from a specific domain area, the system would only be able to solve the problems efficiently from that particular domain, and would not be useful for other problems/scenarios.

2.6.1.2. Preparing the Statistics

The whole purpose of conducting this survey is to collect the real-world examples of tasks. In order to analyze the tasks, we have to prepare the statistics. The statistics should include various aspects of the task environment.

- Nature of the task

Table 2 shows the distribution of the hypothetical response to the significant question categorized by the nature of task. This is just a sample table to convey the idea;
the percentages are not based on any set of data. They are mere approximations of general computer forensic activity. Figure 7 shows the Bar-Graph of the data from table 2 for better understanding.

Table 2: Nature of computer forensic tasks given in percentages.

<table>
<thead>
<tr>
<th>Nature of the task</th>
<th>Percentage of the respondents (Computer Forensic experts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find password information on the system</td>
<td>25%</td>
</tr>
<tr>
<td>Find a file</td>
<td>18%</td>
</tr>
<tr>
<td>Find information about some person</td>
<td>17%</td>
</tr>
<tr>
<td>Find information about some topic</td>
<td>13%</td>
</tr>
<tr>
<td>Find a fact</td>
<td>12%</td>
</tr>
<tr>
<td>Find an email</td>
<td>8%</td>
</tr>
<tr>
<td>Find a software</td>
<td>3%</td>
</tr>
<tr>
<td>Detect stenography</td>
<td>2%</td>
</tr>
<tr>
<td>Find whatever I can get</td>
<td>2%</td>
</tr>
</tbody>
</table>

Similarly many other stats can be prepared to make more sense out of the collected data. Examples of such statistics include the following:

- Type of the System involved for Computer Forensics
- Comparison of “the estimated time for the activity” and “the actual time taken in the activity”
• Type of file system used by the system involved in Computer Forensics

• Type of Operating System used by the system used in Computer Forensics

• Tools used to perform the forensic activity

Figure 7: Bar-graph, representing the data given in Table 2.

2.6.1.3. Analysis of Survey Results

The results of the survey should be analyzed to see the common trends and identify the common and important tasks that the computer forensic experts perform. The purpose of this survey is to get a better understanding of the overall environment and tasks performed during computer forensic analysis. By arranging the survey results (e.g. figure 7 above), specific trends
and tasks can be identified. This helps in understanding the needs and the goals that derive the process of computer forensic analysis.

2.7. Conclusion

This chapter presented an outline of information foraging theory. It explained how the computer forensics process can benefit from the use of this theory. It also presented an introduction to the Optimal Foraging Theory. It presents the analogies between conventional Information foraging and computer forensics and proves that computer forensics presents an ideal foraging environment to benefit from information foraging techniques.

In addition, this chapter explored the foraging environment of computer forensics, by getting an understanding about its specific tasks and goals. It suggests the use of the CIT to conduct a survey and gather first-hand knowledge about the nature of computer forensic tasks.
Chapter 3

MODELING INFORMATION FORAGING BASED SYSTEM FOR COMPUTER FORENSIC ANALYSIS

“Why does this applied science, which saves work and makes life easier, bring us so little happiness? The simple answer runs: Because we have not yet learned to make sensible use of it.” - Albert Einstein

This chapter presents a computational cognitive model which is used to simulate the behavior of computer forensic analyst, while looking for evidence. This model is based on an existing cognitive model called SNIF-ACT, which stands for Scent-based Navigation and Information Foraging in the ACT architecture. SNIF-ACT simulates the behavior of web users and it was created by Peter Pirolli and Wai-Tat Fu [5]. This project uses SNIF-ACT as a base architecture and makes necessary changes in order to make it work for simulating the behavior of computer forensic expert. This project calls this extended and modified version of SNIF-ACT, as SNIF-ACT-CF, where CF stands for Computer Forensics. The author hopes that the reader will appreciate the beauty of applied science and how it can be used to complement various other scientific processes.

3.1. Cognitive Study

Detailed protocol analysis studies have often laid foundation for the development of the models of cognition. With the help of survey method, explained above, knowledge is acquired regarding the trends in computer forensic analysis and important tasks performed by the computer forensic experts. In order to model a compute forensic analysis tool which can make intelligent
decisions, it is necessary to understand the cognitive process of the human mind (of the computer forensic expert) involved behind the execution of the tasks. By the detailed protocol analysis of the process, a model of cognition can be built that replicates the process involved in human thinking while solving a given problem. Therefore a study is conducted in order to begin to understand behavior and cognition involved in basic information foraging during computer forensic analysis. The study involves a set of laboratory experiments, performed in controlled environment, to capture comprehensive cognitive data. Laboratory experiments are followed by detailed protocol analysis of the captured data. These analyses are the basis for developing a computational cognitive model for information foraging while performing computer forensic analysis.

3.2. Laboratory Experiments

In laboratory Experiments many users perform the same tasks under same condition. For this purpose a set of tasks is created and a group of computer forensic experts is selected to take part in the experiment. Each forensic expert has to perform the same set of tasks. Upon completion a detailed protocol analysis is conducted on the collected cognitive data to understand the cognition involved in the problem solving process. The detailed process is discussed below.

3.2.1. Sample Computer Forensic Cases

The goal of conducting the survey (discussed in chapter 2) is to bring realistic computer forensic analysis tasks to the table, in order to perform replicate-able experimental studies. Creating sample forensic cases, for the participants (forensic experts) is a significant activity that needs to be performed after an in-depth analysis of the survey results. The forensic cases should
involve various levels of complexity so that the data captured is more useful to study. Similarly the problems should be from various domains so that more cognitive data can be collected for analysis. If the problems are of similar kind and have same level of complexity then the participant (forensic expert) will get used to of it. As a result they may skip the obvious steps and may not follow the complete cognitive process to solve the problem.

Forensic scenarios should be constructed for each case and evidence should be created on the hard-drive accordingly. A mini case description should be written to handover to the forensic expert.

3.2.2. Participant Selection

The sample selection is one of the critical activities that needs to be done keeping in mind a few things. Firstly the forensic experts should have different sort of expertise and experience. If they are from the same school of thought then they may use the same approach to solve the problem and we would end up collecting redundant set(s) of data. Secondly it will be helpful if they have different levels of levels of experience, so that we get better data for analysis.

3.2.3. Laboratory Setup

Setting up the laboratory is a challenging task. Sufficient number of ways should be implemented to capture as much of the user activity as possible, so that it can be studied to understand the cognitive process involved. Peter Pirolli (2007) used the laboratory setup shown in Figure 8, while performing a research to understand user’s behavior on the web.
3.2.3.1. Think Aloud Technique

In order to follow the user thinking process, the key is to enforce the “think aloud” technique. In this technique all the participating forensic experts are asked to keep on saying loudly whatever is going on their minds, and all their verbalizations should be recorded. In place of Audio recorder, a video recorder can also be used to record all user actions, activities and movements. Peter Piroli in 2007 used a video recorder to capture web user’s activity.
3.2.3.2. Activity Log

All the participant activities should be logged on the machine on which he performs the forensic analysis.

3.2.4. Procedure

The selected forensic cases should be set up on the computer systems in the laboratory for analysis. The systems should have the activity log turned on. The video camera should be fixed to record all the verbalizations revealing the thought process. Forensic cases are presented in random order for each participant. Participants have to start verbalizing when they are presented the question.

3.3. Protocol Analysis

Protocol analysis is the way for studying the cognitive process behind the working of the computer forensic expert’s mind, which later on, becomes the basis for modeling and implementing a system that imitates the human thinking process.

Detailed protocol analysis is very time consuming and should be conducted on the participants who have complete and most intact data on the tasks.

All the data is brought together from the video recordings of the think-aloud protocol and the activity log recordings from the system. This data is then analyzed by a group of forensic experts. For each task under analysis, the forensic experts rate the relevance of every step, taken by the participant during analysis, to the goal of the task. The analysts are presented with every step taken by the participants, then for each step, they decide how helpful/relevant the step is in
order to achieve the goal. Each step can be rated on a four-point scale. Table 3 shows the relevance rating on a four-point scale [5].

Table 3: Relevance rating on a four-point scale

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Relevant</td>
<td>0</td>
</tr>
<tr>
<td>Low Relevant</td>
<td>1</td>
</tr>
<tr>
<td>Medium Relevant</td>
<td>2</td>
</tr>
<tr>
<td>Highly Relevant</td>
<td>3</td>
</tr>
</tbody>
</table>

The participant behavior can be visualized using some sort of graphical representation for each task under analysis. Peter Pirolli (2007) used Web Behavior Graph Representation (WBGR) for a study to understand user’s behavior on the web.

Figure 9 presents a WBGR. In WBCG the final relevance ratings are plotted on a scale of white(0), light grey(1), medium grey(2) and dark grey (3). Each state shown in the figure corresponds to each step taken by the participant while solving the forensic case.
Web Behavior Graph Representation

where

Time runs left to right, then top to bottom

- **State 1** is a node that represents a state
- **State 2** is a distinguished state representing a search result page
- **State 3** represents the application of an operator
- **State 4** indicates a return to a previous state
- Colors surround states at a common Web site
- Grey scaling indicates scent of a state/page
- Surround states within problem spaces

Keyword Link

Figure 9: Web Behavior Graph Representation [5]
The degree of relevance of the available information, to the required information, is called scent of that information [6]. Therefore table 3 can be extended to include corresponding information scent rating. Table 4 shows the direct relationship between the relevance of information and its corresponding scent value.

Table 4: Relevance rating & information scent rating on a four-point scale

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Rate</th>
<th>Information Scent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Relevant</td>
<td>0</td>
<td>No Scent</td>
</tr>
<tr>
<td>Low Relevant</td>
<td>1</td>
<td>Low Scent</td>
</tr>
<tr>
<td>Medium Relevant</td>
<td>2</td>
<td>Medium Scent</td>
</tr>
<tr>
<td>Highly Relevant</td>
<td>3</td>
<td>High Scent</td>
</tr>
</tbody>
</table>

Information scent (discussed in detail in Section 3.6) consists of the proximal cues used to judge the utility of available options, during the analysis. The concept of information scent emerged as the principal component of the heuristics that drive the problem space search process. Two important phenomena observed are: [5]

- Improvement in information scent are related to more efficient information foraging
- Detection of diminishing information scent is involved in decisions to leave an information patch (or information locality)

3.4. Problem Space Analysis

A problem domain analysis should be performed to identify the area of expertise or application that needs to be examined in order to solve the problem at hand. In computer forensic
analysis there are two major problem spaces. The forensic analysts perform their tasks from these problem states to reach their goals. Any of the states encountered can be the goal states, depending on their specific goals.

Table 5 shows the two identified problem spaces in computer forensic analysis. It also presents the states and operators corresponding to each state.

Table 5: Identification of Problem Space of Computer Forensics

<table>
<thead>
<tr>
<th>Sr. no.</th>
<th>Problem Space</th>
<th>States</th>
<th>Operators</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Key words Problem Space</td>
<td>Alternative queries &amp; results</td>
<td>Formulating/editing search queries</td>
</tr>
<tr>
<td>2.</td>
<td>File System Problem Space</td>
<td>Information Patches (data files)</td>
<td>Navigating in the hierarchy (Up, down, back, forward buttons)</td>
</tr>
</tbody>
</table>

3.5. Hierarchical Information Structure Analysis

Data is organized on digital media in some sort of file system. The directory structure of file systems is hierarchical in nature. Thus the organization of data in directory structure takes form of hierarchical information patches. Figure 10 shows a generic hierarchical organization of directories and files containing electronic data.
The types of information patches identified in the process of computer forensic analysis are as follows.

- **Directory**

  Directories, often referred to as folders, are the place holders for data files. The main purpose is to hold related content together. Directories can be at any level of hierarchy.

- **File**

  Files are the leaf nodes of the directory structure. There are numerous types of files. Examples of file types include text files, spread sheets, image files, executable files, log files etc.

- **Data Content**
This data content is the end point of foraging. In most cases this is the data contained inside the files.

- Memory Content

All the files and directories are stored in memory as well, but this category also includes the slack space or unassigned space in the memory. Examples of this category include deleted files.

Figure 11 shows the directory structure as displayed to a user by Windows operating system.

![Figure 11: Directory structure display of operating system.](image)

[7]
3.6. Information Scent

“Information scent is terse representation of content…

whose trail leads to information of interest.” – Peter Pirolli. [5]

Information scent refers to the detection and use of cues, such as bibliographic citations. These cues provide users with concise information about content that is not immediately available [5]. Information scent plays an important role in guiding the users to the information they seek. It also plays a role in providing users with an overall understanding of the available data content.

3.6.1. Examples of Information Scent

When an internet user enters a query in a search engine, the search engine presents him with short descriptions on related content. By going through the presented short descriptions, the user decides which web page the user wants to open to view the whole content. These short descriptions on the web page are an example of information scent. Figure 12 shows the results of a query on Google search engine.
Figure 12: Information scent cues presented by a search engine in response to a user query

Directory structure tree is another example of information scent cues. It gives the user a sense about the overall content organization and helps him/her take the navigation decisions.

Figure 13 shows the information scent presented by the directory tree structure.
Since users take their navigation decisions based on the information scent, therefore, designing user interfaces to have good quality information scent can yield qualitative improvements in usability [5].

3.6.2. Information Scent and Navigation Cost

While searching/foraging for information, there are various times when the forager has to make a decision between two alternatives in order to save time and energy. The better decisions
he make the more efficient would be his search process. This means that high quality information scent reduces navigation cost. It can be safely said that scent quality has an inverse relation with navigation cost. Mathematically, if \( Q \) is the scent quality and \( C \) is the navigation cost then:

\[
Q \propto \frac{1}{C}
\]

In case of complex systems, containing huge amount of data, small perturbation in the accuracy of information scent can cause qualitative increase in the total time consumed during analysis. This increased time affects the overall cost of analysis by making it inefficient. Consider a case of forensic analysis of a computer, where the information is stored in a directory structure. Each directory or file has a name. Directory structure and naming information serves as information scent for the analyst, based on which he decides to pick a certain path. Assume that the analysis takes form of a hierarchically arranged search tree, in which the branches are explored and if unproductive the user returns to the previously searched node of the tree. The search tree may be characterized by a branching factor ‘\( b \)’, which is the average number of alternatives available at each decision point. The desired evidence may occur at various depths ‘\( d \)’ in the search tree. Figure 14 shows a simple search tree where branching factor is three and the depth of the tree is three as well. An exhaustive tree search includes visiting all the nodes in the tree, which is shown by all the lines in figure 14. In the worst case scenario the analyst have to search all the nodes of the tree in order to find the evidence. A random search that terminated upon encountering the evidence at a node would search about half of the tree on average. In the best case, the information scent is perfect, which means that the user will not make any incorrect choices. Figure 14 shows a search tree, representing search path under ideal information scent (thick line), under no information scent (thin line), and the dashed lines show the full tree search.
Consider, ‘d’ is the depth of the search tree, and ‘b’ is the branching factor, then the number of nodes visited ‘n’ in each case can be given as shown in table 6.
3.7. Using Information Scent

"Knowledge, if it does not determine action, is dead to us." — Plotinus

In case of living beings (e.g. birds, humans) their previous experiences guide them to make better decisions. This knowledge is saved in their memory. It is retrieved when they face a situation where they have a choice, in order to help them make better decision. This process is known as pattern recognition. This added knowledge (experiences) is the main difference between intelligent and non-intelligent system. The concept of information scent is utilized in the presented cognitive model in the following ways.

3.7.1. Spreading Activation and Information Scent

Spreading activation is a mechanism to assess the utility of navigational choices. Spreading activation operates on a large associative network. This network represents computer forensic analyst’s total linguistic knowledge. These spreading activation networks are central to the cognitive model presented later in this chapter. Two major characteristics of these networks are as follows: [18]
• Activation networks are general over the universe of tasks

• Activation networks are not estimated from the behavioral data of the users being modeled

These networks can be constructed by using statistical estimates which are obtained from appropriately large and representative samples of the linguistic environment. These networks are used by the presented model to make predictions for users with particular goals. These networks are not made from a specific user data, and are general to all the tasks.

Figure 15 shows an example of a sub-task during information scent assessment. The example is taken from [5]. It assumes that a analyst has the goal of finding evidence about “medical treatments for cancer,” and encounters a document with the text that includes “cell”, “patient”, “dose”, and “beam”. The user’s cognitive task is to predict the likelihood that a distal source of content contains desired information based on the proximal information scent cues available in the information scent. Peter Pirolli, in 2005, presented a rational analysis of the judgment problem presented in Figure. 15. As a result he arrived at the spreading activation model.

The spreading activation model of information scent in SNIF-ACT (presented by Peter Pirolli) assumes that activation spreads from the set of cognitive structures which are the current focus of attention. This activation spreads to the associated cognitive structures in memory. In ACT-R terminology, these cognitive structures are called chunks. Figure 15 shows a set of inter-related chunks. On the right side of figure 15, the chunks represent information scent cues, and on the left side of the figure shows the chunks representing the user’s information need. The associations between the chunks representing information scent cues and the chunks representing
desired distal information is presented by solid line connections. These associations among chunks come from past experience and they may have different strengths. The strength of associations reflects the degree of relevance of proximal information scent cues to the desired information. More relevant cues are supposed to provide better prediction of the occurrence of unobserved features. For instance, the word “medical” and “patient” co-occur quite frequently therefore they would have a high strength of association. If the strength of association is higher it will produce greater amounts of activation flow from one chunk to another.

Figure 15: Cognitive structure showing proximal cues and desired information. [5]

Expressing the spreading activation model in the context of a user evaluating the utility of a data item to his or her information goal, the activation of a chunk ‘i’ in the information goal is \( A_i \) (as shown in figure 16). The activation of a chunk ‘i’ is a function of a base activation plus...
activation spread from associated chunk ‘j’. The activation of any chunk ‘i’ is \( A_i \) and is given by

the equation given in figure 16. It is known as activation equation.

![Activation Equation Image](image.png)

**Figure 16: Activation Equation [18]**

In the activation equation shown in figure 16, Chunk \( i \) is the goal chunk and chunk \( j \) is representing a cue in the data item. \( B_i \) is the base-level activation of chunk \( i \), \( S_{ji} \) is the association strength between chunk \( j \) and chunk \( i \), and \( W_j \) is the attention weight associated with chunk \( j \).

3.7.2. Utility Calculations

The presented model uses spreading activation to calculate the information scent provided by words/names associated with the files and directories, according to the equations specified above. These information scent values are used to evaluate the utility of actions including attending to information items (files or directories), selection of information items, going back to a previous information item within the directory or leaving a directory and searching in another directory. The details of these utility calculations are discussed in the next section.
3.8. ACT-R Model of Cognition

ACT-R is a cognitive architecture. It is a cognitive model for simulating and understanding human cognition. The purpose of ACT-R is to understand how people organize knowledge and produce intelligent behavior. ACT-R was developed to model learning, problem solving and memory.

The basic feature of ACT-R is that it is a production system theory. ACT-R represents each cognitive skill as a conditional statement known as a production rule. A production rule is a statement which pairs an action with a condition such that, if the condition is fulfilled then the action should be executed. The general structure of a production rule is:

\[
\text{IF } \text{<condition>} \text{ THEN } \text{<action>}
\]

For example, the cognitive process, for identifying a shape as square, will be represented in terms of a production rule as follows:

\[
\text{IF } \text{the goal is to classify a shape}
\]

\[
\text{and the shape has four equal sides}
\]

\[
\text{THEN classify the shape as a square}
\]

Cognitive tasks are accomplished by executing a trail of production rules. For executing one task various production rules are linked together. This grouping is referred to as a production. Such a collection of production rules is referred to simply as a production.

There are two different types of long-term memories in ACT-R, namely, declarative memory and procedural memory. Declarative memory consists of facts. The examples of facts include:
• $2 + 5 = 7$

• Islamabad is the capital of Pakistan

• A triangle has three sides and three angles

On the other hand, procedural memory consist production rules. These production rules represent the knowledge of how to do things. For example:

• the ability to hold a pencil and write

• the ability to speak English

• the ability to ride a horse

Figure 17 shows the basic ACT-R model of cognition. The perpetual motor layer interacts with the environment for inputs and outputs. The Cognition layer functions as the decision making component of the system. It replicates the process of human cognition. The ACT-R buffers are used for processing.
3.9. Existing models of cognition based on ACT-R

Initially ACT-R was only used by researchers in cognitive psychology, but later applications were found in fields of Human Computer Interface (HCI) and Intelligent Systems. Many models of cognition have been built using ACT-R as the foundation. Some of these models, related to this study are discussed as follows:
• ACT-IF

ACT-IF, which stands for Information Foraging in the ACT architecture, is a cognitive model which simulates the process of cluster selection in a Scatter/Gather browser [5]. ACT-IF uses a spreading activation model of information scent.

• SNIF-ACT

SNIF-ACT stands for Scent-based Navigation and Information Foraging in the ACT architecture. It is a cognitive model which simulates the user’s foraging behavior on the Web. It also uses spreading activation mechanism for utility calculation.

As the research continues, ACT-R evolves ever closer into a system which can perform the full range of human cognitive tasks: capturing in great detail the way we perceive, think about, and act on the world.


This section presents a computational cognitive model that simulates the behavior of forensic expert while performing computer forensic analysis. This model works on the similar pattern as SNIF-ACT. It uses ACT-R model as a base architecture, and makes use of the cognitive theory of information scent. The only difference between this model and SNIF-ACT is that this model uses a different set of production rules to be followed and different contents of memory and information scent database. This model makes two major predictions related to the following [5]:

• Path selection from the available options

• When to stop following one particular path and try another
The presented model makes changes to the SNIF-ACT model so that it can be used to simulate the behavior of a computer forensic analyst. The basic architecture of the SNIF-ACT is given in figure 18. The working of this architecture is explained in section 3.10.4.

The model presented in figure 18 uses a declarative memory and a procedural memory. Declarative memory contains declarative knowledge and procedural memory contains procedural knowledge. The main difference between the two kinds of knowledge is “knowing that” and “knowing how”. [14]

Declarative knowledge can be thought of as “knowing that”. These are the facts that can be used in a cognitive process, for example comparing two objects. On the other hand procedural knowledge can be thought of as “knowing how”. This is the knowledge that is part of cognitive processing, For example, the knowledge of how to drive a car.
Figure 18: Computational cognitive model for SNF-ACT and SNIF-ACT-CF [5]
3.10.1. Declarative Memory

Declarative memory consists of declarative knowledge. The content of declarative knowledge is represented in the form of chunks in ACT-R. [8]. A chunk can be perceived as one meaningful unit of information. A chunk could refer to pictures, objects, digits, words, chess positions, or people's faces. [15]. Figure 19 shows a group of inter-associated chunks.

![Figure 19: A group of inter-associated chunks.][5]

Declarative memory works on the theory of activation. Activation-based theories of memory predict that more activated chunks of knowledge will be utilized more in the cognitive processing.

This means that chunks with higher activation values take less time to use and have a greater chance to have an impact on behavior. Activation is a way of quantifying the degree of relevance of declarative information to the current focus of attention (mathematically, it represents the posterior probability of how likely each piece of declarative information is needed given the current focus of attention). The user goals and sub-goals are also represented as chunks.
All the goals are organized in a stack. These goals and sub-goals are encodings of user intentions readable by the model. At any point in time, ACT-R is trying to achieve the goal that is on top of that stack and is focused on a single goal.

3.10.2. Procedural Knowledge

Procedural memory contains procedural knowledge. Procedural knowledge is the knowledge or skills that we display in our behavior without conscious awareness. The examples of procedural knowledge include the knowledge of how to climb stairs or how to drive a car. One does not need conscious awareness of the steps to perform these activities. In the presented model the procedural memory is represented as a set of production rules. A production rule is a condition-action pair, such that, if and only if the condition is fulfilled then the specified action will be executed or carried out.

For example, for the task mentioned in the previous section, there will be a production rule in the procedural knowledge that can be expressed in English as follows:

Use-Search-Tool:

IF the goal is to find a key word & there is a key word & there is a key word search tool

THEN

Set a sub-goal Use-Search-Tool

The production “Use-Search-Tool” will be applied (or fired) when the computer forensic analyst has a goal to search for a key word and he/she already has the key words and a search
tool. The condition IF of the production rule is matched with the current goal and the active
chunks in declarative memory. For each possible condition there is at least one production rule in
the procedural memory. The rule is that at any step only one production rule is fired. If more than
one matches are found, then a “conflict set” of production rule is created. A method called
“conflict resolution” is used to decide which production rule should be executed.

The conflict resolution mechanism relies on a utility function. The utility function
calculates the expected utility of applying each of the production rules in the conflict set. The
production rule with the highest utility is selected.

3.10.3. The Role of Information Scent

When a computer forensic analyst browses through different files and directories to look
for evidence, he/she takes guidance from information scent. Information scent consists of the
local cues that provide the analyst with a sense of the content and guide the user towards
information of interest. Examples of information scent include file or directory names and
descriptions.

Each task is represented by a set of chunks in the declarative memory. Whenever there is
a new task/goal, the chunks related to this task are activated. This activation spreads to the related
chunks in the spreading activation network. The amount of activation accumulating on the goal
chunks matched by a production rule is used to evaluate and select the production rules. The total
amount of activation of chunks associated with a production is used to calculate its utility. In case
more than one production is matched with a task, then the production with more utility is
selected.

Consider the production rule “Click-File”: 
Click-File:

IF the goal is Process-element

& there is a task description

& there is a file that has been read

THEN

Click on the link

The production rule “Click-File” will be matched when a file description is read. Consider that other files descriptions are read at the same time too. Then the file description with more utility (combined activation) will be selected and the production is executed. As a result the selected file is clicked. The activation level reflects the degree of relevance of the file description or name to the task description.

3.10.4. Real Computer Forensic Case Example

Consider a computer forensic case where the goal is to find the evidence about a secret document of company ABC. This secret document is a project plan. The corresponding computer forensic task can be defined as “to find a secret project plan document of company ABC”. As soon as the task appears on the screen, all the chunks relevant to the task are activated in the declarative memory. Activation is also spread to their linked chunks in the spreading activation network.

The task is represented in the form of chunks. Figure 20 shows the chunks representing the top goal, task description and the user interface. One task chunk is created with the name “task ABC”. This task chunk is associated with all the chunks created from the keywords of the
task description. These keyword chunks are displayed on the right side of figure 20 and the task chunk is shown in the middle.

When the system starts working in the problem space in order to solve the task, it provides the system with information scent cues in the form directory structure and the file and directory descriptions. When information scent cues are presented their relevant chunks get activated in the declarative memory and their utility is calculated in relation to the goal chunks. Left side of figure 20 shows the chunks from base level information scent. The goal chunk is linked with the related task(s) and information scent chunks provided by the system.
Figure 20: A subset of the declarative knowledge of the SNIF-ACT-CF model simulation of user working on the above mentioned task.

During the process, consider a file description is read saying “Company ABC”. It will be matched with a production rule to determine the next action. There is at least one production rule for every condition. At any time the current state of the system is determined by the values in it
buffers. In the above scenario a file description is read, a task description is given and the goal of the system is to process the data item therefore the production rule “Click-File” will be matched.

Description of the production rule “Click-File” is given as:

Click-File:

IF the goal is Process-element
& there is a task description
& there is a file that has been read

THEN

Click on the link

The production rule “Click-File” will be matched when a file description is read. Consider that other files descriptions may be read at the same time too. In case there are more data items/files, the file description with more utility (combined activation) will be selected, and the matched production is executed. As a result the selected file is clicked. The activation level reflects the degree of relevance of the file description or name to the task description.

There can be more than one productions matching with each condition. The amount of activation accumulating on the goal chunks matched by a production is used to evaluate and select the production rules. The total amount of activation of chunks associated with a production is used to calculate its utility. In case more than one production is matched with a task, then the production with more utility is selected.
3.10.5. Comparing the Model with Actual User Behavior

SNIF-ACT model is tested with the help of a user tracing architecture, also called user tracer. It is a software system that supports a method for developing and testing cognitive models. The goal of user tracer is to trace each and every activity of the user, then compare it to the activity generated by the behavior simulator program to verify if the simulation is working fine. Actual user activity data (also called user trace) is collected through the laboratory experiments explained in section 3.2.

Figure 18 (page 60) shows the user tracing architecture of the model. The top part of the figure shows the memory modules, buffers, and the pattern matching structures. The lower part of figure 18 presents the simulation and compares it to the user trace data.

3.10.6. Production Rules for the presented model – SNIF-ACT-CF

Table 7 presents the production rules designed for the proposed version of SNIF-ACT, called SNIF-ACT-CF, which is designed to simulate the behavior of computer forensic analyst. The production rules are written in simple English for the purpose of understanding.

Table 7: Production rules for SNIF-ACT-CF model designed for Computer Forensics

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Production Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Start-Process-Directory</td>
</tr>
<tr>
<td></td>
<td>IF the goal is Goal*Start-Next-patch</td>
</tr>
<tr>
<td></td>
<td>&amp; there is a task description</td>
</tr>
<tr>
<td></td>
<td>&amp; there is an un-processed file</td>
</tr>
<tr>
<td></td>
<td>THEN Set and push a sub-goal Goal*Process-File to the goal stack</td>
</tr>
<tr>
<td></td>
<td>Process-Files-in-Directory</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------</td>
</tr>
<tr>
<td>2.</td>
<td>IF the goal is Goal* &amp; there is a task description &amp; there is an un-processed file THEN Set and push a sub-goal Goal*Process-File</td>
</tr>
<tr>
<td></td>
<td>Attend-to-File</td>
</tr>
<tr>
<td>3.</td>
<td>IF the goal is Goal*Process-file &amp; there is a task description &amp; there is an un-attended file THEN Choose an unattended file and attend to it</td>
</tr>
<tr>
<td></td>
<td>Read-and-Evaluate-File</td>
</tr>
<tr>
<td>4.</td>
<td>IF the goal is Goal*Process-file &amp; there is a task description &amp; the current attention is on a file THEN Read and Evaluate the file</td>
</tr>
<tr>
<td></td>
<td>Click-File</td>
</tr>
<tr>
<td>5.</td>
<td>IF the goal is Goal*Process-file &amp; there is a task description &amp; there is an evaluated file &amp; the file has the highest activation</td>
</tr>
<tr>
<td></td>
<td>THEN</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>6.</td>
<td>Leave-Directory</td>
</tr>
<tr>
<td>7.</td>
<td>Back-up-a-file</td>
</tr>
</tbody>
</table>
Chapter 4

CONCLUSION

4.1. Project Outcomes

This research provided background information associated with the computer forensic process. It discussed the ‘analysis phase’ of computer forensics and some of the problems and issues involved in it. It explained the background of this study and presented the problem statement. This chapter gave an account of previous research done related to this study, and compared it to this project. It also defined the project’s objective and scope.

This study provided an outline of information foraging theory. It explained how the computer forensics process can benefit from the use of this theory. It presented the analogies between conventional information foraging and computer forensics, and explained how computer forensics presents an ideal foraging environment and hence, can benefit from information foraging techniques.

This research included detailed protocol analysis of the cognitive data involved in the process of computer forensics and also explained the methods to collect cognitive data for analysis. It also discussed creation and use of reference information, called information scent. It also discussed some existing cognitive models, namely, ACT-R and SNIF-ACT. Finally, it presented a computational cognitive model that simulates forensic expert’s information foraging behavior. The presented model is completely based on SNIF-ACT. This study replaced the existing production rules of SNIF-ACT with new production rules (given in table 7, section 3.10.5) to make it work for computer forensic analysis. This study briefly discussed the
advantages and limitations of the presented approach. The study suggested future work and implementation aspects of the research that is needed. This research forms a basis for the implementation of an intelligent computer forensic system, which can provide built-in decision making capabilities, such as, choosing one path to follow from the available options, and deciding when is the right time to stop looking for evidence in one patch and move to another location.

4.2. Advantages of the Model

The advantages of the presented cognitive model for computer forensic analysis are as follows:

- The presented model for computer forensic analysis would help to understand the intelligent cognitive process involved in computer forensic analysis. This model forms a basis for the implementation of an intelligent computer forensic tool. This would lower the dependence of the computer forensic analysis process on the computer forensic expert.

- The presented model can be compared with the actual user behavior to access the accuracy of its findings. The traceability of the system is very convenient, because it records complete information about every step in a process chain.

4.3. Limitations of the Model

The limitations of the presented cognitive model for computer forensic analysis are as follows:
• The most significant limitation of the presented model is that it is not very efficient for un-obvious or disguised information. Sometimes people/criminals store information with totally unrelated and misleading names, which does not make the actual content obvious to anyone. Since names of files and directories are one of the factors that serve as information scent in the presented model. The utility of the information item (file) is calculated on the bases of the relevance of information-scent chunks to the goal chunks. Misleading information scent would cause the system to work inefficiently.

• The second most significant limitation of the presented model is the difficulty involved in applying it. Using this model requires collection of the real word data. The data collection should be performed with extreme care so that intact and complete data can be collected for analysis. Sophisticated and reliable data collection mechanisms should be used. There is simply too much effort involved in developing a model for a specific application. This is the reason cognitive models are confined to research studies.

• The protocol analysis phase which is required to construct the model is very tedious and time consuming. It requires a group of computer forensic experts to analyze the collected cognitive data.

• Finally, some in the cognitive psychology community, question whether human cognitive processes can be modeled as a production system. Time and again, there have been debates that production rules are simply descriptive entities and are not psychologically real. Although many systems of Artificial Intelligence
(AI) and Knowledge Engineering (KE) work on models based on production systems and probabilities. Anderson, in his book “Rules of the Mind” discusses the psychological reality of production rules. [16].

4.4. Future Work and Implementation

The presented model is based on the calculation of probability of occurrence and utility of each option. From implementation point of view, the presented model can be completely implemented and executed. It can be represented as a computer program in the LISP programming language and can be executed. The SNIF-ACT model, which is the base architecture for the presented model, has been implemented and tested by the Information Foraging Research Group Scientists at the Palo Alto Research Center (PARC) [5].

In LISP programming language, the contents of declarative memory and procedural memory have syntax specific representations. Chunks have a schema-like representation. This representation consists of a field called "is-a". This field specifies the category of knowledge. Additional fields are present to encode the knowledge. For example the encoding of the fact "2*5=10" would be represented as follows. [8]

```
fact2*5
  isa multiplication-fact
  multiplicand1 two
  multiplicand2 five
  product ten
```
Production rules also have a specific representation. Below is an encoding of the production rules for counting from one number to another. [8]

( P increment

=goal>

ISA count-from

number =num1

=retrieval>

ISA count-order

first =num1

second =num2

==> 

=goal>

number =num2

+retrieval>

ISA count-order

first =num2

)
The above mentioned notations would be readable and understandable for a person familiar with LISP programming syntax.

This model presented a picture of the cognitive process involved behind solving a case of computer forensic analysis. This project organized the rules in form of production rules, and presented the stepwise functioning of the involved cognitive process, which can be implemented using a programming language. AI techniques can be used to implement the decision making feature, and incorporate them in the architecture of existing computer forensic tools.
APPENDICES
### QUESTIONNAIRE TO COLLECT COMPUTER FORENSIC TASK INFORMATION

**Participant Information:**

- **Name:** ____________________________
- **Organization:** ______________________
- **Job Title:** __________________________
- **Area of Expertise:** ____________________
- **Years of Experience:** ________________

**Significant Question:**

Explain one of your recent successful computer forensic cases where you were able to find the desired information, which was useful in solving the case. Explain your experience in sufficient detail so that we can visualize the event.

**Answer:**

- ______________________________________
- ______________________________________
- ______________________________________
- ______________________________________
- ______________________________________
- ______________________________________

**Time Estimated for the activity (in hrs):** __________________

**Time taken in the activity (in hrs):** __________________
# GLOSSARY

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>CF</td>
<td>Computer Forensics</td>
</tr>
<tr>
<td>CIT</td>
<td>Critical Incident Technique</td>
</tr>
<tr>
<td>COFEE</td>
<td>Computer Online Forensic Evidence Extractor</td>
</tr>
<tr>
<td>CSUS</td>
<td>California State University, Sacramento</td>
</tr>
<tr>
<td>DFRWS</td>
<td>Digital Forensic Research Workshop</td>
</tr>
<tr>
<td>FTK</td>
<td>Forensic Toolkit</td>
</tr>
<tr>
<td>IFT</td>
<td>Information Foraging Theory</td>
</tr>
<tr>
<td>ISD</td>
<td>Information Scent Database</td>
</tr>
<tr>
<td>KE</td>
<td>Knowledge Engineering</td>
</tr>
<tr>
<td>MVT</td>
<td>Marginal Value Theorem</td>
</tr>
<tr>
<td>PARC</td>
<td>Palo Alto Research Center</td>
</tr>
<tr>
<td>SNIF-ACT</td>
<td>Scent-based Navigation &amp; Information Foraging in the ACT architecture</td>
</tr>
<tr>
<td>SNIF-ACT-CF</td>
<td>SNIF-ACT for Computer Forensics</td>
</tr>
<tr>
<td>USDOJ</td>
<td>United States Department of Justice</td>
</tr>
<tr>
<td>WBGR</td>
<td>Web Behavior Graph Representation</td>
</tr>
</tbody>
</table>
REFERENCES


[2] Marc Roger’s Computer Forensics Notes, Department of Computer Science, Purdue University


[21] Houston and McNamara, 1999, optimal foraging theory.
