République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieure et de la Recherche Scientifique

Université A.MIRA-BEJAIA

Faculté des Sciences Exactes Département d'Informatique

Mémoire

Présenté par

BERMAD Nabila

Pour l'obtention du diplôme de Magister

Filière : Informatique

Option : Cloud Computing

Thème

La fouille de données pour l'investigation numérique légale

Soutenu Devant le Jury composé de :

Mr	**TARI Abdelkamel**	MCA	Université de Béjaia	Président
Mr	**KECHADI Mohand Tahar**	Professeur	Université de Dublin, Irland	Rapporteur
Mr	**BOUKERRAM Abdellah**	Professeur	Université de Béjaia	Examinateur
Mr	**MELIT Ali**	Professeur	Université de Jijel	Examinateur
Mr	**OMAR Mawloud**	MCB	Université de Béjaia	Invité

Année Universitaire : 2012-2013

Résumé

Dans le cadre de ce mémoire, nous nous intéressons au problème d'analyse de preuves, que nous proposons d'étudier dans le domaine de l'investigation mobile légale, puisque dans nos jours, l'utilisation des smartphones dans des activités criminelles est de plus en plus fréquente.
L'utilisation des techniques de fouille de données pour extraire des connaissances utiles à partir des preuves stockées pour leur analyse est une approche très prometteuse, notamment dans le domaine de digital forensics, où le volume élevé des ensembles des preuves issus de la phase de collecte sont très complexe et ne peuvent pas être interprétées facilement.
Nous proposons dans ce travail deux stratégies d'analyse scalable. Dans la première, il s'agit de mettre en place une nouvelle approche de reconstruction de time line basée sur la classification non supervisée des événements(SMS et appels). Le deuxième apport concerne l'analyse relationnelle de preuves. Notre algorithme effectue une recherche des motifs fréquents sur la base de preuves afin d'extraire l'ensemble des numéros de téléphones suspects, leurs associations téléphoniques et l'ensemble des contextes des SMS à travers la génération des cliques.
Les résultats d'expérimentation obtenus montrent que notre approche permet de réduire le temps de génération des clusters, de temps de réponse et un meilleur parallélisme.

Mots clés : *Investigation Numérique Légale, Investigation Mobile Légale, Analyse des Preuves, Reconstruction des Événements, Time line, Fouille de Données, Recherche de Motifs Fréquents, Classification non Supervisée, Analyse Relationnelle, Analyse Temporelle et Fonctionnelle, Smartphone.*

Abstract

In this work, we are interested in the problem of proof analysis, that we propose to study in the domain of the mobile forensics, since in ours days, the use of the smartphones in the criminal activities is more and more frequent.
The use a techniques of datamining to extract some useful knowledge from proofs stocked for their analysis is a very promising approach, notably in the domain of digital forensics, where The volume raised of proofs set descended of the collection phase is very complex and cannot be interpreted easily.
We propose in this work two strategies of analysis scalalable. In the first it is about putting a new approach of timeline reconstruction based on the unsupervised classification of events (SMS and calls). The second contribution concerns the relational analysis of proofs. Our algorithm does the frequents tatterns research on the basis of proofs in order to extract the set of the phones numbers suspects, their associations telephonic and the set contexts of the SMS by the generation of clicks.
The results of experimentation gotten showed that our approach permits to reduce the time of clusters generation, of answer time and a better parallelism.

Keywords : *Digital Forensics Investigation, Mobile Forensics Investigation, Analysis of Proofs, Reconstruction of Events, Timeline, Data Mining, Frequent Tatterns Research, Unsupervised Classification, Relational Analysis, Temporal and Functional Analysis, Smartphone.*

Remerciements

Tout d'abord, je tiens à exprimer mes plus vifs remerciements et ma gratitude à mon directeur de mémoire, Mr Mohand Tahar KECHADI Professeur à l'UCD de Dublin, pour leur encadrement continu, pour les remarques constructives qu'il m'a fourni ainsi que pour leurs précieux conseils durant toute la période de mon travail. Je le remercie également pour la confiance qu'il m'a accordée et pour la grande liberté d'idées et de travail qu'il m'a donnée.

Ma grande reconnaissance aussi au responsable de l'école doctorale Mr Abdelkamel TARI, maitre de conférence à l'université de Béjaïa pour votre présence tout au long de ma période de formation, vos encouragements et votre aide pour que le travail se déroule dans des bonnes conditions.

Mes sincères salutations sont cordialement présentées à tous les membres de jury, composés de :

- Mr. Abdelkamel TARI de m'avoir honoré par sa présidence du jury. Je le remercie également pour l'intérêt qu'il porte à ce travail.
- Mrs Abdellah BOUKERRAM, Ali MELIT et Mawloud OMAR d'avoir accepté de faire partie du jury. Leur participation m'honore.

Enfin merci à ceux que je n'ai pu citer mais qui ont toutes mes amitiés et mes remerciements.

Dédicaces

Louange à Dieu le tout puissant qui nous a aidé à achever ce travail de recherche, et que le salut soit sur son prophète Mohamed ;

A ...
La mémoire de mon père ;

A...
Ma mère ;

Et à ceux qui m'ont aidé moralement et concrètement, de près ou de loin, pour réussir notre effort ;

Je dédie cet humble mémoire.

Table des matières

Liste des tableaux

Table des figures

Liste des algorithmes

Liste des abréviations

DF Digital Forensics

SWGDE Scientific Working Group on Digital Evidence

LAI Location Area Identity

ICCID Integrated Circuit Card Identifier

IMSI International Mobile Subscriber Identity

SOM Self Organization Map

SOMAT Self Organization Map Analysis Tool

WAP Wireless Application Protocol

MSISDN Mobile Subscriber Integrated Services Digital Network

MAC time of last Modification (M), time of last Access (A), time of Creation (C)

DOS Disk Operating System

FTK Forensic Toolkit

CD Compact Disc

FBI Federal Bureau of Investigation

DoS Deny of Service

U2R User to root

R2L Remote to Local

IDS Intrusion Detection System

SMS Short Message Service

SIM Subscriber Identity Module

SHA Secure Hash Algorithm

MD5 Message 5 Digest

KDD Knowledge Discovery from Data

BD Base de Données

SVM Support Vector Machine

RBF Radial Basis Function

FP Faux Positifs

BP Backtracking Propagation

TH Targeted Hardware

TS Targeted Software

RD Reported Damage

GA Generic Attacks

AA Additional Actions

IT Investigation Techniques

KMO Kaiser-Meyer-Olkin

FP Profile Factuel

BP Behavioural Profile

LA Link Analysis

Wi-Fi Wireless Fidelity

VoIP Voix Internet Protocol

PDA Personal Digital Assistant

GPS Global Positioning System

UML Unified Modeling Language

MMS Message Multimedia Service

DAG Direct Acyclic Graph

SQL Structured Query Language

PDF Portable Document Format

CATFA Clustering Algorithm For Temporal and Functional Analysis

RAFM Relational Analysis for Frequents Motives

CHAMELEON Hierarchical Clustering Algorithm Using Dynamic Modeling

ROCK Robust Clustering Algorithme for Categorial Attributes

Introduction Générale

En raison de l'augmentation des crimes numériques à cause de l'utilisation croissante de l'Internet et la technologie de la communication mobile dans le monde, le domaine de l'investigation numérique a rapidement émergé afin d'améliorer la qualité et l'efficacité des enquêtes numériques. L'investigation numérique légale (Digital Forensics) développées en tant que champ autonome à la fin des années 1990 et au début des années 2000 lorsque le crime par ordinateur a commencé à grandir avec l'utilisation croissante des ordinateurs et plus encore, l'Internet. Dans les premiers jours, elle a été appelée l'informatique légale puisque la preuve recueillie a été limité à des computers. Cependant, ces dernières années, avec plusieurs avancées technologiques, cette restriction n'est plus vraie. Par conséquent, le processus des enquêtes judiciaires impliquant des preuves numériques est devenu plus difficile, un nouveau terme appelé l'investigation numérique légale a été inventé. Ce nouveau terme se réfère maintenant à enquêter sur tout type des supports capables de stocker des informations numériques dans le cadre d'une enquête légale.

L'investigation numérique légale représente l'ensemble des méthodes qui permettent de collecter, conserver et analyser des preuves issues des supports numériques en vue de les produire dans le cadre d'une action en justice. L'analyse désigne le processus d'organisation et de structuration des données récupérées à la phase d'acquisition. Il est constitué de deux étapes principales : La collecte et la reconstruction de preuves.

Le volume élevé des ensembles de données issus des objets de la scène du crime et aussi la complexité et la dimension des relations entre ces types de données (preuves) ont fait la

phase d'analyse de preuves, l'une des tâches les plus consommatrices du temps et la plus complexe lors d'une investigation.

Les outils actuels de l'investigation numérique légale ne sont tout simplement pas évolutives pour les données volumineuses. L'extraction et l'analyse de données deviennent excessivement lentes et inefficace. Par conséquent y a-t il d'autre façons d'améliorer l'efficacité et la qualité de processus d'analyse légale ?.

Plusieurs travaux ont été réalisés dans la phase d'acquisition et d'examen des preuves numérique [80], [52], [3], [5], [81]. Peu de recherche, ont été faites dans la phase d'analyse. Elle dépend largement de l'expérience et l'intuition de l'enquêteur. La résolution des crimes est une tâche complexe qui exige de l'intelligence humaine et l'expérience. Depuis quelques années, l'idée d'utiliser les techniques de fouille de données (data mining) pour extraire des connaissances caché dans l'ensemble de preuves ou afin de les localisées lors de processus d'investigation est avancée. Cependant, peu de travaux ont été entrepris dans cette optique [24], [50], [79], [27] [30], [1].

Ce mémoire s'attaquera à la problématique d'analyse de preuves collectée à partir des smartphones puisque ils sont devenus une source de preuves dans l'investigation mobile légale. Le but ultime de ce texte est de démontrer comment la preuve numérique peut être utilisée pour la reconstruction des événements qui se sont produits lors de l'incident. C'est une activité fondamentale dans toute enquête. La reconstruction du crime se réfère au processus systématique d'assemblage des éléments de preuves et des informations recueillies lors de l'étape de collecte qui conduit à une image plus complète de l'incident afin de mieux répondre sur les questions *quand ?, qui ?, quoi ?, comment ? et pourquoi ?*. Notre objectif principal consiste alors à fournir des stratégies qui permettent d'analyser les preuves extraites des smartphones dans la phase d'acquisition de processus d'investigation mobile légale. Le cœur de ces stratégies repose sur des techniques de fouille de données. Ces techniques sont employées comme des heuristiques qui aident à réduire la complexité des problèmes de reconstruction de crime. Notre travail s'articule donc autour des deux axes principaux :

◇ Le développement d'une stratégie pour l'analyse temporelle & fonctionnelle.

◇ La proposition d'une stratégie de l'analyse relationnelle.

Nous avons introduit une technique de clustering ascendante basée sur la causalité pour la création d'un time line (chronologie) des événements SMS et appels pour aider un enquêteur à identifier des anomalies et d'informations du crime en question et fournir à l'enquêteur une vision globale de tous les événements à travers toutes les preuves numériques sources qui peuvent être très utiles au cours d'une enquête. En effet, notre stratégie d'analyse relationnelle de preuves quant à elle exploite la recherche des motifs fréquents pour l'ensemble des numéros de téléphones candidats. Notre intuition est que l'utilité d'un numéro de téléphones donné est fortement corrélée avec sa fréquence dans le journal d'appel et SMS, ainsi que son filtrage chez le destinataire.

Il nous a par ailleurs paru intéressant d'adapter ces stratégies au contexte de l'investigation mobile(smartphones) légale puisque l'utilisation de ces dispositifs dans des activités criminels est sans cesse croître. Actuellement, les smartphones sont devenus une partie intégrante de la vie quotidienne d'un nombre croissant de personnes partout dans le monde selon Garner Inc [22], les ventes mondiales de téléphones intelligents atteindront 468 millions d'unités en 2011, augmentation de 58 millions par rapport à 2010. Les smartphones sont essentiellement des ordinateurs de poche, contiennent une grande quantité des données. Les informations stockées sur ces téléphones tels que les détails de communications (appels, messages texte et e-mail), l'emplacement des utilisateurs(GPS), des contacts et vidéos sont pertinents dans tous enquêtes judiciaires. Le reste de ce mémoire est organisé comme suit :

Dans le premier chapitre, nous commençons naturellement de présenter le domaine d'investigation numérique légale, ses notions, les principales sous disciplines. Ensuite, nous présentons le processus de l' investigation légale, avant de se focaliser sur la phase d'analyse. Nous détaillons quelques méthodes et techniques qui existent dans la littérature pour l'analyse de preuves. Enfin nous donnons un aperçu des outils de l'investigation les plus couramment déployés.

Dans le deuxième chapitre, nous rappelons d'abord la notion de data mining et ses techniques qui lui sont associées et on présente un état de l'art portant sur les travaux de

data mining appliqués à l'investigation numérique légale sur la phase d'analyse de processus d'investigation . Nous discutons ensuite les avantages et les insuffisances des solutions proposées. Cette étude nous permettra de proposer une classification selon les deux étapes de la phase d'analyse, la collecte et la reconstruction de preuves.

Dans le troisième chapitre, nous décrivons notre approche d'analyse de preuves pour l'étape de reconstruction des liens entres les évidences et la résolution du crime. Nous proposons deux stratégies, la première est basée sur la technique de clustering pour l'analyse temporelle et fonctionnelle et la deuxième est basée sur la recherche des motifs fréquents afin de répondre sur les questions *qui ? et quoi ?* pour l'analyse relationnelle.

Pour la validation de notre approche, nous déroulons dans le quatrième chapitre quelques testes. Nous commençons par la description de notre prototype d'implémentation. Ensuite nous évaluons notre approche par quelques scénarios de comparaison et expérimentations.

Nous terminerons ce mémoire par une conclusion générale qui discute les apports de notre travail. Ainsi que les perspectives envisagées en vue d'ouvrir de nouvelles directions de recherche.

Chapitre 1

Investigation numérique légale

1.1 Introduction

L investigation numérique légale est le processus d'emploi des méthodes scientifiques d'analyse des informations stockées électroniquement pour déterminer la séquence des événements qui ont mené à un incident particulier. ce processus permet d'analyser la preuve, classer, comparer et individualiser les interactions entre les suspects et les preuves. En d'autres termes, cette investigation permet de relier les éléments de preuves, des suspects, des victimes et de la scène du crime, comme indiqué dans la figure 1.1

FIGURE 1.1 – Liens entre la preuve, le suspect, la victime et la scène du crime [38]

Dans ce chapitre, nous allons présenter les concepts liés au digital forensics, leurs

domaines et quelques outils d'investigation les plus couramment utilisés.

1.2 Concepts et terminologie

Dans ce qui nous allons donner une suite de définitions pour essayer de clarifier certains concepts liés à la science de digital forensic, nous n'allons pas introduire toute la taxonomie des concepts liés à cette science, nous nous restreindrons uniquement aux concepts que nous allons discuter dans les chapitres à suivre.

1.2.1 Légal et la science légale(forensic and forensic Science)

Définition 1.2.1. *Le terme "Forensic" est dérivé du latin "forensis" ou "légal", qui signifie "en audience publique ou public", qui lui-même vient du latin "du forum", se référant à un emplacement réel " un marché public plus carré utilisé pour les affaires judiciaires et autres". En dictionnaires légales, forensics est définie comme le processus d'utilisation des connaissances scientifiques pour la collecte, l'analyse et la présentation de la preuve devant les tribunaux [23]. la science de forensics est "l'application des principes et des techniques scientifiques pour fournir la preuve aux enquêtes juridiques et aux déterminations" [37]*

Nous retrouvons aussi une autre définition de forensic, qui est due au SWGDE (Scientific Working Group on Digital Evidence) [57] et qui stipule que :

Définition 1.2.2. *C'est l'application de connaissance scientifique au droit, et principalement, appelé à l'investigation de crimes [57]*

1.2.2 Digital Forensics

Définition 1.2.3. *C'est une branche de science légale qui permet l'utilisation des méthodes scientifiquement dérivées et éprouvées pour la préservation, la collecte, la validation, l'identification, l'analyse, l'interprétation, la documentation et la présentation de la preuve numérique dérivée de sources digitales dans le but de faciliter la reconstruction des événements en relation avec le crime, ou aider à anticiper des actions non autorisées [58]*

1.2.3 Digital Forensics Investigation

Définition 1.2.4. *L'investigation numérique légale est un processus qui utilise la science et la technologie pour examiner des objets numériques et que se développe et teste des théories, qui peut être entré dans un cour de justice, et répondre à des questions sur les événements qui produite. [17]*

L'objectif d'une investigation numérique est d'exposer et de présenter la vérité, ce qui conduit souvent à des réponses aux questions suivantes relatives à un crime numérique(voir la figure 1.2) [30]

- ◇ Le quand : Se réfère à l'intervalle du temps pendant la scène du crime
- ◇ Le quoi : Concerne les activités exécutées sur le système informatique.
- ◇ Le qui : Concerne la personne responsable du crime.
- ◇ Le où : Se réfère à l'endroit où se trouve la preuve.
- ◇ Le comment : Traite la manière dont les activités ont été réalisées.
- ◇ Le pourquoi : Chercher à savoir les motivations du crime.

FIGURE 1.2 – Résolution d'un crime numérique

L'investigation numérique légale peut être classée dans trois domaines clés, à savoir l'informatique légale, l'investigation de réseau légale, l'investigation mobile légale tels quels sont illustrés à la la figure 1.3

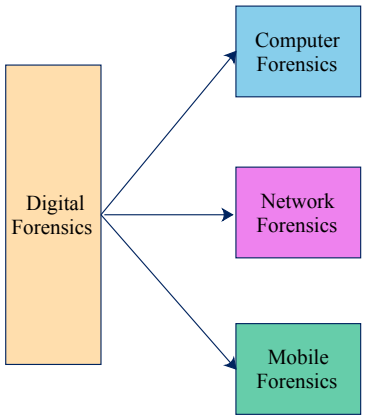


FIGURE 1.3 – Classification des domaines d'investigation numérique légale

Informatique légale(Computer Forensics)

L'informatique légale a été remonté au 1984, lorsque le gouvernement fédéral du bureau d'investigation (FBI), ainsi que d'autres organismes d'application de la loi ont commencé le développement des programmes assistant à l'examen et l'analyse de preuves numériques. Elle sert à identifier la preuve qui peut coexister dans des ordinateurs et ses périphériques qui forment la scène du crime numérique [63], un aperçu général est représenté dans la figure 1.4.

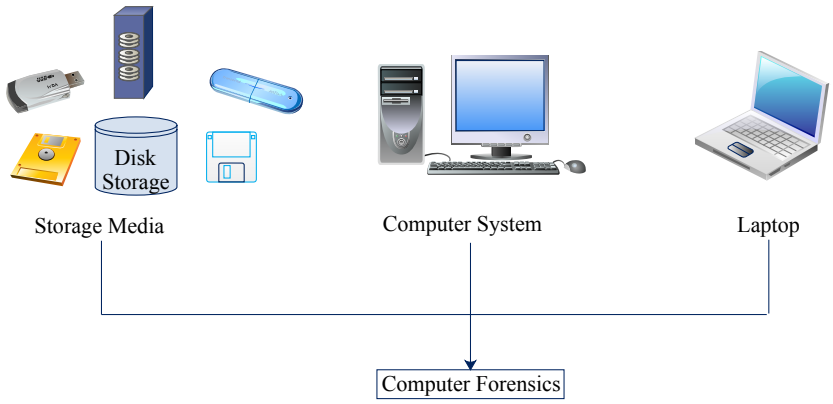


FIGURE 1.4 – Computer Forensics [30]

L'informatique légale traite l'identification, la préservation, l'extraction et la documen-

tation de preuves numériques [51]. Ces phases constituent le cœur du processus d'investigation légale et on va les discuter plus tard dans ce chapitre.

Investigation de réseau légale(Network Forensics)

l'investigation de réseau légale a été introduisé dans le début des années 90, elle sert à la fois comme un moyen de prévention d'attaque dans des systèmes et la recherche des preuves après une attaque ou un incident s'est produit. Un aperçu général est représenté dans la figure 1.5.

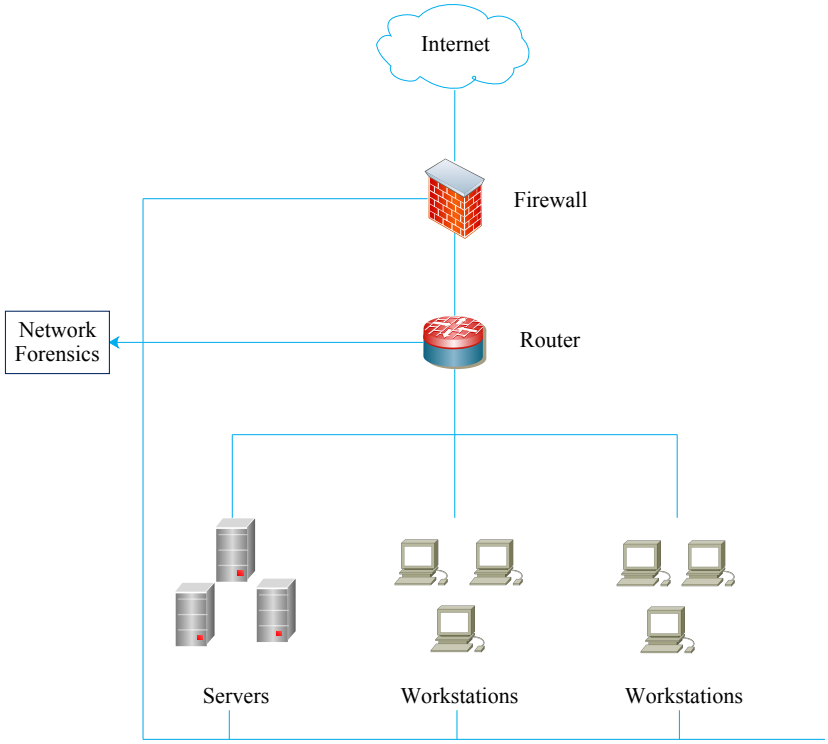


FIGURE 1.5 – Network Forensics [30]

Ces attaques incluent le déni de service (Deny of Service, DOS), utilisateur-à-racine (User to Root, U2R) et télécommande local (Remote to Local,R2L). L'investigation légale de réseau consiste à vérifier, capturer, enregistrer et analyser les pistes de réseau afin de collecter les évidences. Par exemple, l'analyse des fichiers journaux des systèmes de détection d'intrusion(IDS), l'analyse du trafic réseau [20] et l'analyse des périphériques

réseau.

L'informatique légale et l'investigation de réseau légale ont une chose en commune. Les deux domaines peuvent contenir des éléments de preuves quand un incident se produite. Si par exemple, un attaquant attaque un réseau, le trafic de l'attaque passe en général par un routeur. À la suite de cela, les éléments de preuves importants peuvent être collectés en examinant les fichiers journaux de routeur.

Investigation mobile légale(Mobile Forensics)

La popularité des téléphones mobiles intelligents(smartphones) continue de croître. Ils changent la façon du crime au cour des dernières années et les chercheurs ont été confrontés à la difficulté d'admissibilité de la preuve numérique sur les téléphones mobiles. C'est pour ça une science d'investigation légale à été émergé. Cette science permet de collecter des preuves numériques depuis un téléphone mobile dans des conditions juridiquement valides tels que des détails des appels, des SMS(Short Message Service), des e-mails et location (Global Positioning System(GPS)), etc, ainsi que des données supprimées [81]. Un aperçu général est représenté dans la figure1.6.

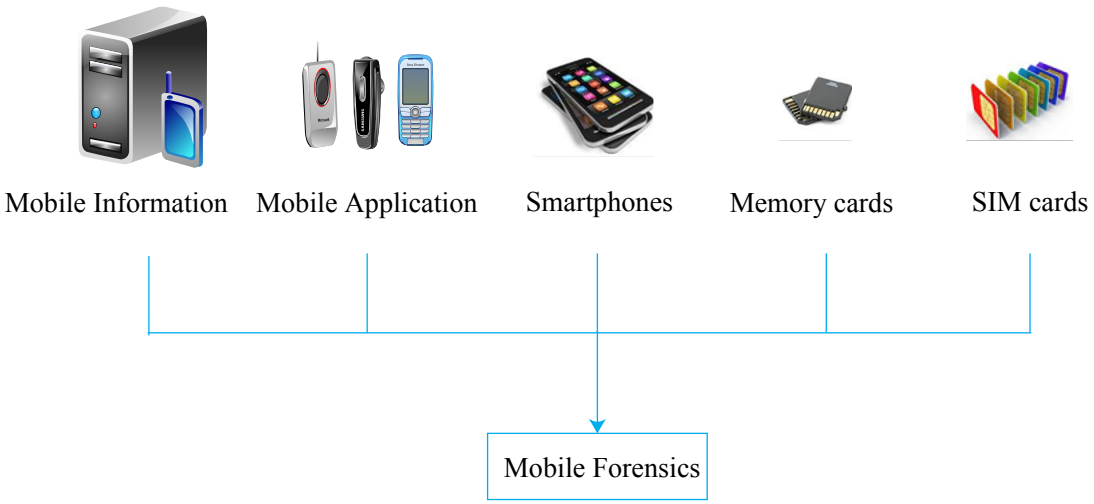


FIGURE 1.6 – Mobile Forensics

Les sources des éléments de preuves dans un téléphone mobile peut inclure(voir le tableau1.1) :

Éléments de preuves	**Source**
Nom du service fournisseur	Imprimé sur l'arrière de la carte SIM (Subscriber Identity Module)
Numéro ID unique	-"-
Emplacement de la zone d'identité (Location Area Identity (LAI))	Stocké à l'intérieur de la carte SIM
Identificateur de carte de circuit intégré (Integrated Circuit Card Identifier (ICCID))	Stocké à l'intérieur de la carte SIM
Identité d'abonné mobile internationale (IMSI)	Stocké à l'intérieur de la carte SIM
Les données de messages texte (SMS)	Stockés sur la carte SIM
Contacts	-"-
Journaux d'appels	-"-
Identité d'équipement mobile internationale(International Mobile Subscriber Identity(IMEI))	Stocké comme imprimé sur le mobile
Messages multimedia	Mémoire du téléphone mobile
Images/Sons/Vidéos	-"-
Navigateur WAP(Wireless Application Protocol)/Historique/Emails	-"-
Points de calendrier / Notes	-"-
Information sur des cartes SIMs précédentes	Les données enregistrées des cartes SIMs précédentes
Abonné mobile des services de réseau intégré(Mobile Subscriber Integrated Services Digital Network(MSISDN))/Numéro de téléphone	Parfois disponible dans la mémoire SIM.

TABLE 1.1 – Éléments de preuves possible sur les téléphones mobiles

Aujourd'hui, les smartphones ont des fonctionnalités similaires aux ordinateurs, mais il existe certaines différences entre l'informatique légale et l'investigation mobile légale qui sont illustrées dans le tableau 1.2 [45]

Aspect	**Informatique légale**	**Investigation mobile légale**
Source des preuves	◇ Disque dur ◇ RAM ◇ Cartes mémoire externe	◇ Mémoire interne ◇ SIM ◇ Cartes mémoire externe
Support de stockage interne	Le disque dur peut être enlevé facilement	Non
Système d'exploitation	Nombre limité des systèmes d'exploitation	Large gamme de systèmes d'exploitations
Mot de passe d'authentification	Non	Ne peut pas contourner le mot de passe d'authentification au cours de l' acquisition logique
Câbles d'alimentation et de données	Alimentation standard des câbles de données	Large gamme des câbles de données puisssants
Système de fichiers	Système de fichiers standard	Large gamme de systèmes de fichiers

TABLE 1.2 – Comparaison entre l'informatique légale et l'investigation mobile légale

D'après cette comparaison, il est clair que l'investigation des smartphones légale est plus complexe que le l'informatique légale.

1.2.4 Preuve numérique(Digital evidence)

Définition 1.2.5. *La preuve numérique est définie comme toute donnée qui peut établir qu'un crime a eu lieu ou fournir un lien entre un crime et sa victime ou un crime et son coupable [19].*

Définition 1.2.6. *C'est toute information numérique de valeur probante stockée ou transmise sous forme numérique. Elle peut être facilement modifiée, reproduite, restaurée ou détruite [30].*

La preuve numérique comprend :
◇ Les données utilisateur

◇ Les métadonnées associées aux données de l'utilisateur
◇ Logs d'activité
◇ Logs du système

Les données utilisateur se rapportent aux données directement crées ou modifiées ou accessibles par un ou plusieurs utilisateurs participant à une enquête. Les métadonnées se rapportent aux données qui fournissent le contexte de *comment, quand, qui* et sous quelle forme les données des utilisateurs ont été crées ou modifiées ou accessibles. Les journaux d'activité sont des enregistrements de l'activité des utilisateurs par un système ou une application et les actions spécifiques menées par un ou plusieurs utilisateurs. Les journaux du système se rapportent à des variations dans le comportement du système fondé sur une ou plusieurs actions menées par les utilisateurs.

Le 3227 RFC (Research Forensics Computer) décrit les considérations juridiques liées à la collecte de preuves. Les règles exigent les preuves numériques d'être [37] :

- ***Admissible :*** Se conformer à certaines règles juridiques avant qu'il peut être mis devant un cour.
- ***Authentique :*** C'est l'intégrité qui permet le suivi de la chaîne des éléments de preuves qui doivent être intactes.
- ***Crédible :*** Les preuves doivent être claires, faciles à comprendre et précises. La version de la preuve présentée au tribunal doit être reliée avec la preuve binaire d'origine dans le cas contraire il n'y a aucun moyen de savoir si la preuve a été fabriquée.
- ***Complete :*** Toutes les preuves soutenir ou contredire une preuve qui incrimine un suspect doivent être examinées et évaluées. Il est également nécessaire de recueillir des preuves qui élimines les autres suspects.
- ***Fiable :*** Les procédures et les outils de la collection des évidences, l'examen, l'analyse, la préservation et la présentation doivent être en mesure de reproduire les mêmes résultats au fil du temps. Les procédures ne doivent pas douter sur l'authenticité de la preuve ou sur les conclusions tirées après l'analyse.

Lors de la recherche de preuves sur des supports numériques, il existe différents types de données à rechercher, parmi ces types on trouve principalement :

- ***Les données actives*** *:* Sont des données qui résident sur des supports de stockage et qui sont facilement visible par le système d'exploitation et accessible aux utilisateurs.

Ces données comprennent les fichiers de traitement de texte, les tableurs, les programmes et les fichiers de système d'exploitation. Celui-ci, y compris les fichiers temporaires, les fichiers internet temporaires, cookies et les métadonnées du système de fichiers, etc [56].

– ***Données résiduelles*** *:* Ce sont des données sembles avoir été supprimées, mais peut encore être récupérées. Les exemples les plus courants sont les fichiers d'échange et des fragments de fichiers récupérés dans l'espace non alloué.[56].

1.2.5 Rapport d'investigation(Chain of evidence)

Le rapport d'investigation est un document qui résume les étapes d'une investigation afin de valider la preuve numérique issue d'une information numérique avec les détails suivants :

◇ Le format de sauvegarde de l'information originale.
◇ L'identification de l'empreinte numérique et les moyens de blocages en écriture
◇ Identifier les opérations réalisées et les logiciels mis en oeuvre.
◇ Les numéros de série des supports d'informations utilisées pour l'enregistrement.
◇ L'ensemble des preuves réunies avec des éventuelles interprétations ainsi que des conclusions.

1.2.6 Rapport de garde(Chain of custody)

Le rapport de garde est un procès-verbal qui commence lors de la réception d'une preuve numérique liée à la scène du crime. Il est définit :

Définition 1.2.7. *La documentation chronologique du mouvement, localisation et la possession d'éléments de preuve [57].*

1.3 Support numérique dans un crime

Un crime ou incident est un événement ou série d'événements qui violent une police et plus spécifiquement, un crime est un événement ou une séquence d'événements qui viole la loi [17]. Un support digital dans un crime implique le numérique et ça englobe les PCs(Personal Computers), les ordinateurs portables, les smartphones, les tablettes, les

outils numérique robot... etc, c'est-à-dire tout ce qui est numérique. Un support numérique peut jouer l'un des trois rôles dans un crime informatisé. Il peut être la cible du crime, il peut être l'instrument du crime, ou il peut servir comme un référentiel de stockage des informations précieuses sur le crime(évidences). Dans certains cas, le support numérique peut avoir plusieurs rôles. Lors d'une investigation sur une affaire, il est important de savoir quels rôles de chaque objet digital a joué dans le crime et puis adapter le processus d'investigation pour cet rôle [72]. Il existe une variété des crimes tel quels sont représentés par la figure 1.7.

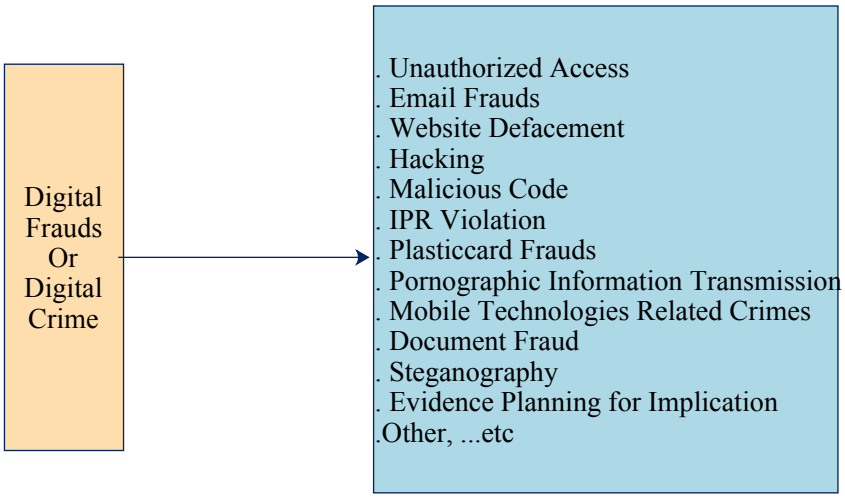


FIGURE 1.7 – Types de crimes digitales ou fraudes [65]

1.4 Processus d'investigation légale(Process of Investigation Forensics

L'investigation numérique dépend de plusieurs facteurs techniques tels que le type de l'information ou le dispositif de communication, en plus le type d'investigation, criminel, civil, commerciale, militaire ou autre contexte et autre facteurs selon le cas d'investigation [21]. Malgré cette variation, il existe des phases communes entre les modèles de processus d'investigation résumé dans le modèle de NIST qui sont [43] :

1.4.1 Collection

Dans cette phase il s'agit d'identifier, labéliser, acquérir, stocker, transporter et conserver des données depuis des différents sources de confiance issue de la scène du crime. C'est un processus long et complexe puisque les données doivent être collectées à temps à cause de la probabilité de perte des données telles que les connexions réseaux ou bien perte de données dues à la décharge des batteries de certains sources (exemple : smartphone).

1.4.2 Examen

Cette phase consiste en une recherche systématique approfondie pour l'évaluation et la localisation des éléments pertinentes à partir de large volumes de données recueillies tout en préservant son intégrité en utilisant une combinaison d'outils automatique, ou manuel. Les résultats de l'examen sont des objets de données. Ils peuvent inclure des fichiers journaux, fichiers de données contenant des phrases spécifiques, des SMS téléphonique, etc.

1.4.3 Analyse

C'est le cœur de processus d'investigation légale. Elle désigne le processus d'organisation et de structuration des résultats des examens, en utilisant des méthodes et des techniques légales et justifiables pour dériver des connaissances utiles qui adressent les questions qui résoudre le cas d'investigation. Cette phase est constituée de deux étapes fondamentales :

La recherche des preuves

Pendant le processus d'une investigation numérique, l'étape de recherche des preuves numériques est la tâche la plus consommatrice du temps. Un des plus grands défis auxquels sont confrontés les enquêteurs numériques est le volume de données qui doivent être recherchées lors de la localisation des preuves numériques, c'est pour ça ils utilisent des méthodes scientifique comme le datamining afin de localiser efficacement la preuve relative au crime. Selon la façon de recherche des preuves électroniques, ce processus sera divisé en deux types de recherches : Statique et dynamique.

- ***La recherche statique :*** C'est de faire la collecte des preuves dans un état passif, c'est-à -dire, la recherche des évidences est faites sur des ordinateurs et autres dispositifs numériques (par exemple smartphones) qui ont quitté la scène du crime(hors marche).
- ***La recherche dynamique :*** C'est la technologie de la collecte des preuves dans le pare-feu, détection d'intrusion et tous les actes possibles d'acquisition en temps réel.

L'analyse de preuves (reconstruction des chaînes des évidences)

L'investigateur développe des hypothèses basée sur des preuves existantes qu'il a collecté à l'étape précédente et teste ces hypothèse en cherchant des preuves supplémentaires indiquant s'ils sont vraies ou fausses [17]. Il existe trois types d'analyse [19] :

- ***L'analyse temporelle :*** Il répond de la question *quand ?*, c'est l'ordonnancement dans le temps des preuves récupérées pour fournir une séquence narrative des événements aider un enquêteur d'identifier les anomalies sur un crime et menant à d'autres sources de données. De nombreux éléments de données numérique légales sont naturellement prêtes à cette séquence par exemple, fichiers MAC(time of last Modification (M) time of last Access (A) time of Creation (C)), les événements journaux avec timestamp, e-mails, etc.
- ***L'analyse relationnelle :*** Il répond des questions *qui ?,quoi ?,où ?* pour montrer les liens entre les entités dans un crime, par exemple l'existence d'un numéro de téléphone dans une base de contacts d'un mobile affiche un lien entre le propriétaire du téléphone et le propriétaire du numéro de téléphone.
- ***L'analyse fonctionnelle :*** Il répond des questions *comment ? et pourquoi ?*. C'est l'acte de déterminer quelles entités pourraient avoir réalisées l'un des événements qui sont liés à un cas d'investigation. Lors de la reconstruction d'un crime, il est souvent utile de se demander quelles conditions étaient nécessaires pour que certains scénarios du crime soit possible. Par exemple, il est parfois utile d'effectuer certains testes fonctionnels sur le matériel utilisé dans le crime pour s'assurer que le système sous investigation était capable d'effectuer des actions. L'analyse fonctionnelle vise à examiner toutes les hypothèses possibles pour un ensemble des circonstances, par exemple, lorsqu'on lui demande si l'ordinateur du défendeur peut télécharger un groupe de fichiers incriminants en une minute, comme il est indiqué par leurs

timestamps, un examinateur légiste peut déterminer que le modem était trop lent pour télécharger ces fichiers rapidement. Cependant, l'examinateur ne doit pas être satisfait de cette réponse et devrait déterminer comment les fichiers ont été placés sur l'ordinateur. Les analyses temporelles, fonctionnelles et relationnelles sont nécessaires pour recréer une image complète d'un crime. Combinant les résultats de ces analyses peut aider les enquêteurs à comprendre le crime et son coupable.

1.4.4 Rapport

C'est la phase dont laquelle les conclusions de la phase d'analyse sont documentées et présentées à l'autorité sous forme d'un rapport d'investigation, il comprend l'enregistrement des détails de chaque étape de l'investigation, telles que les procédures suivies et les méthodes utilisées pour saisir, recueillir, conserver, restaurer, reconstruire, organiser et rechercher des éléments de preuves essentiels.
Les phases de processus d'investigation sont illustrés par la figure 1.8

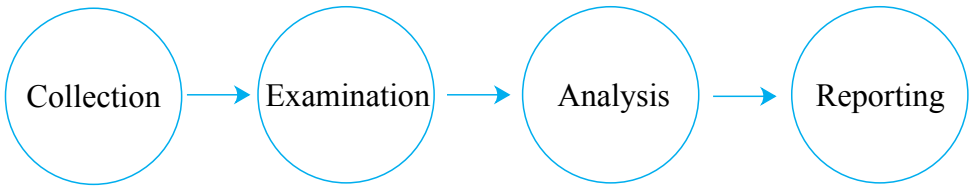


FIGURE 1.8 – Processus d'investigation numérique légale [43]

Dans notre travail, nous allons exploiter les smartphones comme une source de données liée au crime, alors le processus d'investigation est vu comme une transformation des données aux actions comme illustré dans la figure 1.9 :

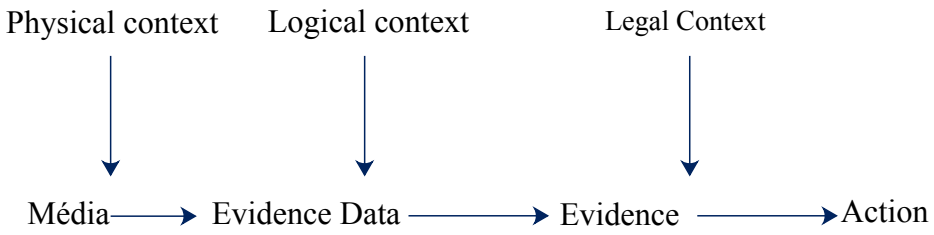


FIGURE 1.9 – Transformation de media en action

	Transformation
Collection	Extraire les données depuis les smartphones en utilisant des outils d'acquisition.
Examen	Les données collectées sont épurées afin de tirer des informations pertinentes
Analyse	La transformation des information en preuves par analogie de transformation des données en connaissances en utilisant des méthodes de datamining afin de répondre sur les questions quoi, qui, ou, comment, quand et pourquoi en faisant des liens entre les preuves
Rapport	La transformation des preuves en action en structurant les résultats de la phase d'analyse

TABLE 1.3 – La signification de processus d'investigation

1.5 Outils d'investigation numérique légale

Une augmentation spectaculaire de crimes numériques a conduit au développement de toute une série d'outils d'investigation. Ces outils assurent que la preuve numérique est acquis et préservée correctement. De tels outils existent sous forme de logiciels mise au point pour aider l'enquêteur numérique lors d'une enquête . L'objectif de cette section est de développer une meilleure compréhension de certains des outils d'investigation légale le plus souvent déployés.

1.5.1 EnCase Forensic

EnCase Forensic, de Guidance Software, Inc (Guidance Software, 2006) est l'un des outils d'investigation numérique légale les plus connus et les plus couramment déployés. C'est un outil informatique basé sur windows destiné à être utilisé pour la collecte et l'examen de preuves numériques à la fois actives et résiduelles. Il est l'un des outils commerciaux les plus chers disponibles sur le marché. Cependant, il offre une gamme complète de fonctionnalités d'interfaces graphiques sophistiquées. Pour faire fonctionner EnCase Forensic, une clé de sécurité est nécessaire. Cette clé de sécurité est un dispositif matériel qui contrôle l'accès à l'application. Une fois l'outil lancé, l'enquêteur doit créer un nouveau cas pour une nouvelle investigation(voir la figure1.10).

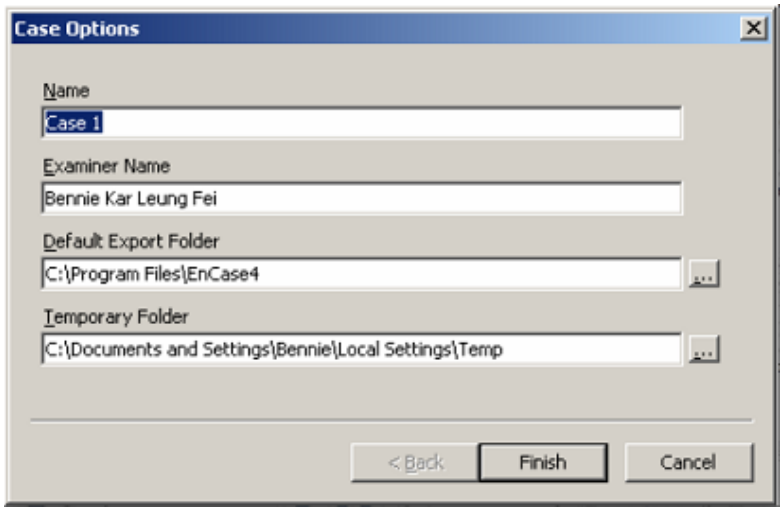
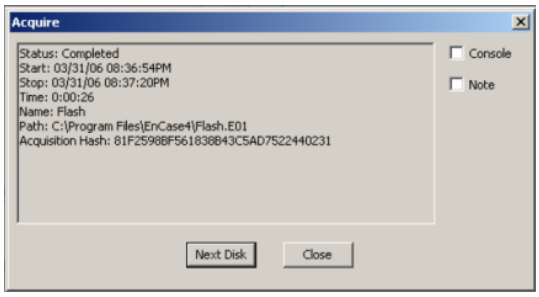


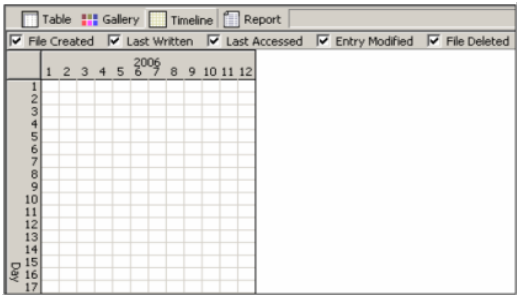
FIGURE 1.10 – Création de nouveau cas dans EnCase Forensic

En outre, il dispose de plusieurs vues qui sont accessibles par un onglet axé sur interface. Ces vues sont les suivantes (voir la figure 1.11) :

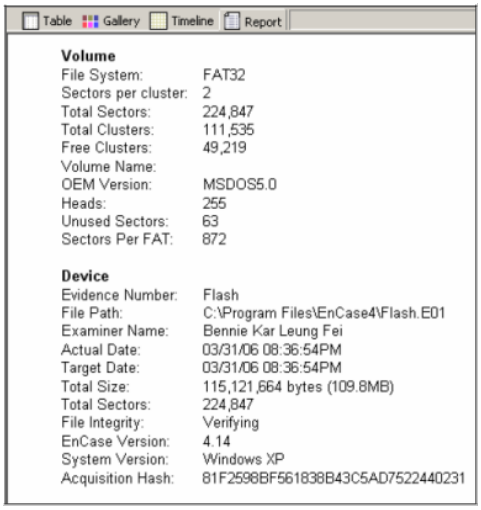
- **Vue de table (Table view) :** La vue table permet à l'investigateur de visualiser les fichiers trouvés sur un périphérique, y compris leurs attributs, dans un format de style tableur.
- **Musée des beaux-vues(Gallery view) :** Gallery view permet à l'enquêteur de voir l'image graphique présentée sur un périphérique.
  Vue de chronologie(Timeline view) : Timeline view fournit une vue d'agenda pour la recherche de modèle de création du fichiers, l'édition et la date de dernier accès.
- **Vue de rapport (Report view) :** Report view présente divers rapports concernant l'enquête. Il peut afficher des informations en ce qui concerne le dispositif, y compris les fichiers et répertoires figurant sur le dispositif.



Vue de table



Vue de timeline



Vue de rapport

FIGURE 1.11 – Fonctionnalités d'EnCase Forensic

1.5.2 Forensic ToolKit

Forensic Toolkit (FTK) de Access Data Corporation (Access Data Corporation, 2006) est un autre outil d'investigation légale bien connu et couramment déployé. Cet outil est reconnu comme l'un des principaux outils pour effectuer l'analyse d'e-mails [61]. De même que pour EnCase Forensic, FTK fonctionne sur la plate forme windows, cependant, en termes du coût, il est moins cher par rapport à EnCase Forensic. FTK est constitué de plusieurs composants, tels que l'imageur FTK, le registre d'affichage(viewer) et le fichier filtre(filter). Une fois FTK est lancé, il propose un choix entre acquérir ou pré

visualiser la source de la preuve (voir la figure 1.12).

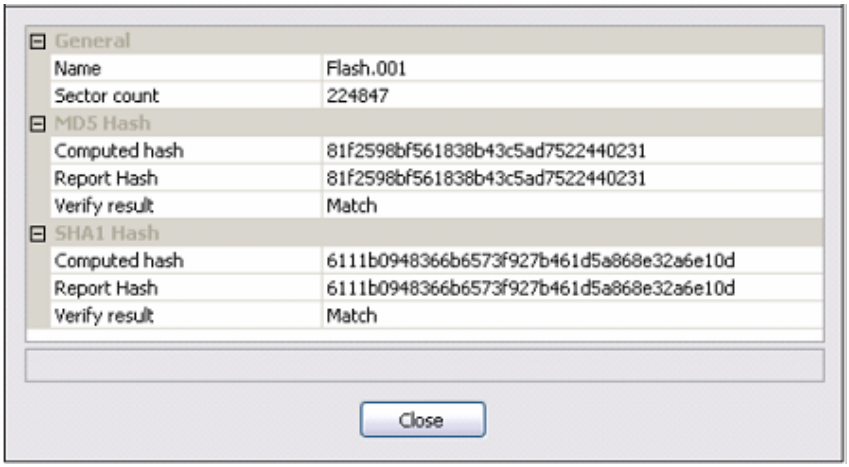


FIGURE 1.12 – Résultat d'acquisition du processus FTK

Après l'acquisition, l'examen peut commencer par le démarrage d'un nouveau cas d'investigation en utilisant l'assistant illustré par la figure 1.13.

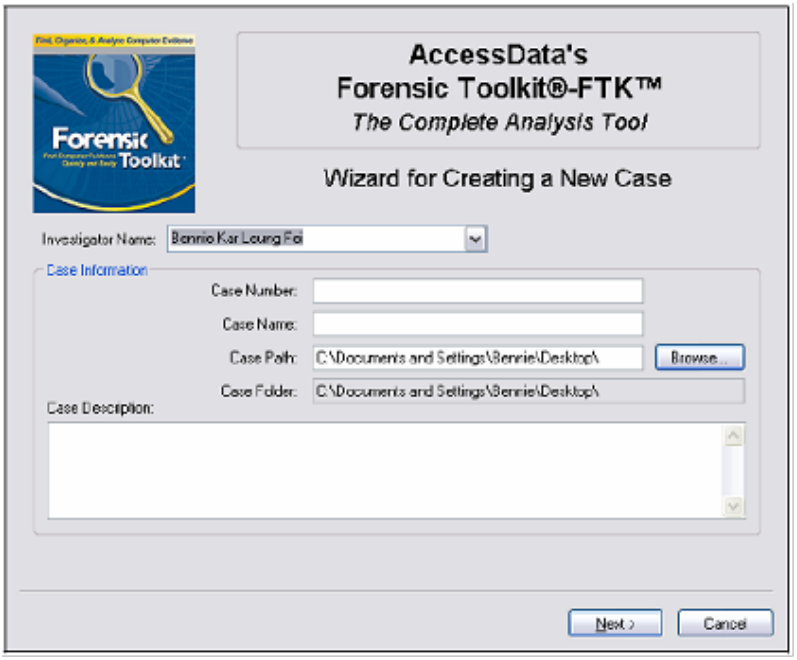


FIGURE 1.13 – Wizard de création du nouveau cas FTK

FTK dispose d'une interface utilisateur graphique sophistiquée, qui comprend six principales fenêtres à onglets pour faire l'analyse et qui sont discutés ci-dessous(voir la figure 1.14) :

- ◇ **Fenêtre de présentation(Overview window) :** Overview window donne un aperçu de cas à étudier. les fichiers sont présentés dans un format de style tableur.
- ◇ **Fenêtre d'exploration (Explore window) :** Dans l'utilisation de la fenêtre explore window, l'enquêteur peut voir la structure hiérarchique des fichiers, des dossiers et des supports de stockage.
- ◇ **Fenêtre graphique(Graphics window) :** La fenêtre graphics window affiche des images graphiques en tant que vignettes qui facilite l'analyse et l'accès rapide.
- ◇ **Fenêtre d'e-mails(E-Mail window) :** La fenêtre e-Mail permet à l'investigateur de consulter la boîte aux lettres e-mails, y compris leurs messages(Outlook, Outlook Express) et les pièces jointes.
- ◇ **Fenêtre de recherche(Search window) :** FTK propose deux modes de recherche distincts : en direct et indexé. La recherche en direct implique une comparaison point

par point avec les termes de recherche spécifié par l'enquêteur, alors que la recherche indexée implique l'utilisation d'un moteur de recherche connu sous le nom dtSearch(dtSearch Corporation, 2006).

◇ **Fenêtre marque de page (Bookmark window) :** Dans cette fenêtre, l'enquêteur peut consulter tous les items qui ont été signée.

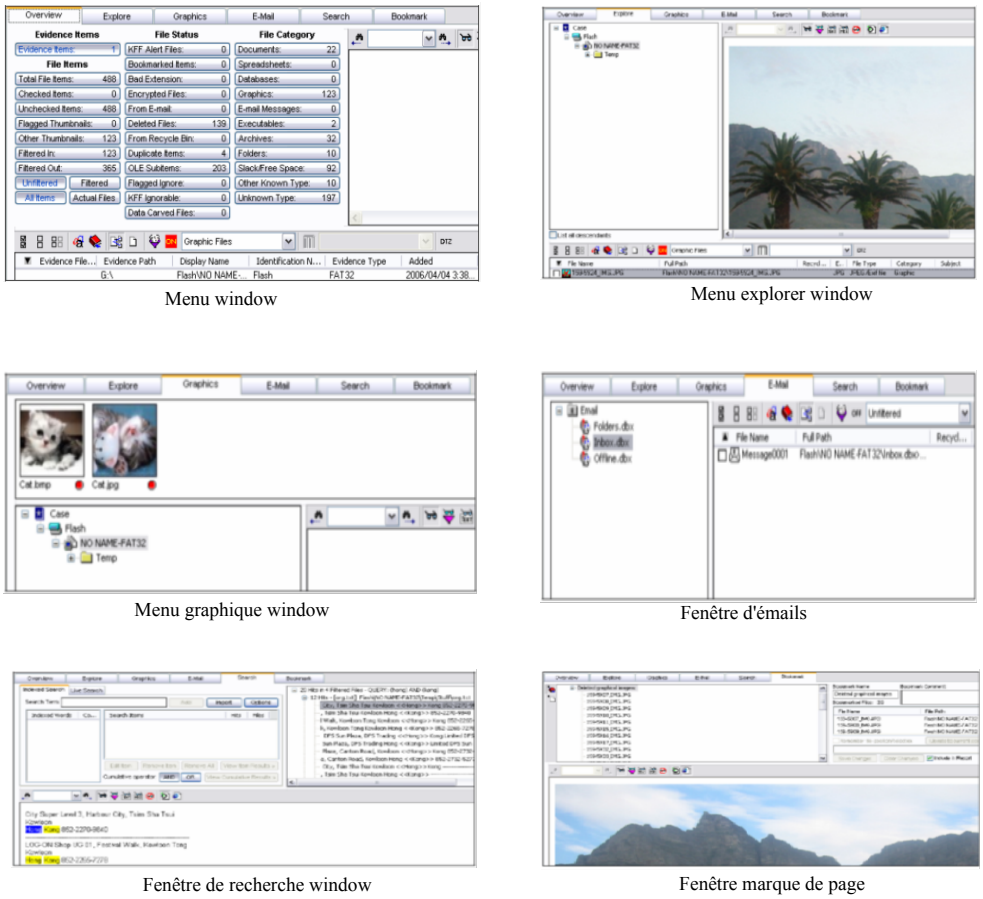


FIGURE 1.14 – Fenêtres Forensic ToolKit

1.5.3 SafeBack

SafeBack est un autre outil d'investigation numérique légale. Il s'agit d'un outil légal basé sur le DOS(Disk Operating System) qui est capable de créer des copies de support

de train de bits (National Institute of Standards and Technology, 2005). Il est principalement utilisé pour l'imagerie des médias et n'est pas équipé de fonctionnalités telles que la recherche et la génération de rapports. SafeBack utilise l'algorithme SHA-256(Secure Hash Algorithm) (Institut national des normes et de Technologie, 2002) pour garantir l'intégrité et l' auto authentification des images. Pour exécuter SafeBack, le disque doit contenir le fichier ¨Master.exe˝ qui est le programme principal de SafeBack. Une fois le support de stockage souhaité et le nom du fichier de sortie ont été sélectionnés, le processus d'acquisition commence.

1.5.4 Support de stockage et récupération d'archives toolkit

Le support de stockage et de récupération d'archives toolkit (SMART) est un outil d'investigation légale fonctionne sous Linux. Il dispose d'une simple interface facile à utiliser. En plus de cela, il propose de nombreuses fonctionnalités similaires à ses homologues telles que EnCase Forensic et FTK. SMART est avant tout un outil d'imagerie qui utilise l'algorithme SHA-1 par défaut pour créer des tables de hachage. En outre, il fournit d'autres algorithmes tels que le MD5(Message 5 Digest) et l'algorithme 32 de redondance cyclique [26] pour assurer l'intégrité des images. Il offre un aperçu, de recherche, rapport et d'acquisition, même à distance. Pour utiliser SMART, une clé de sécurité est nécessaire. Une fois que l'outil est lancé, il donne un aperçu de tous les périphériques explorés. Cette vue d'ensemble comprend des informations telles que la taille de chaque dispositif, les systèmes de fichiers résidant sur chaque dispositif et les données résiduelles disponibles sur chaque support(voir figure 1.15)

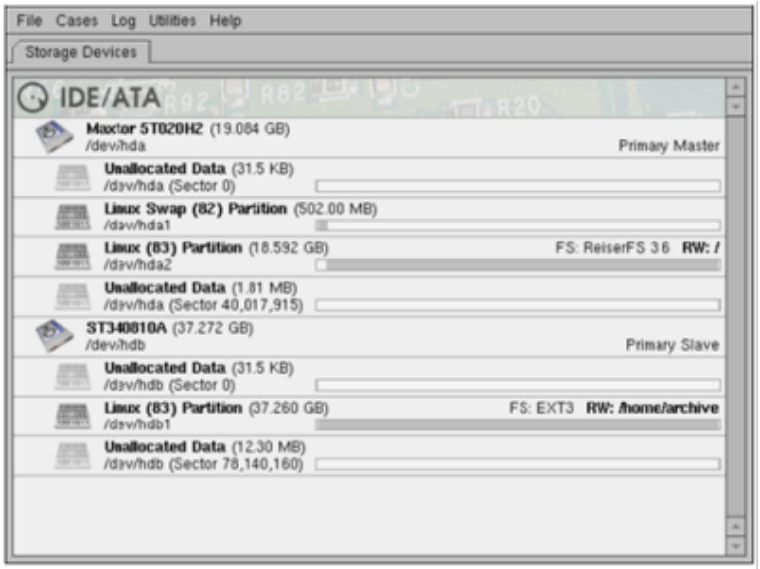


FIGURE 1.15 – Capture d'écran de SMART pour les dispositifs Explorés

1.6 Classification des outils d'investigations

L'étude des outils d'investigation numérique légale dans la section précédente nous a permet d'identifier leurs capacités. Un bref aperçu des différents capacités des outils d'investigation est fourni dans les paragraphes suivants.

1.6.1 Imagerie

L'imagerie est le processus de création d'une image. C'est l'opération de copie des données secteur-par-secteur à partir d'une pièce de support pour créer un flux de bits [42]. Pour acquérir une image, un logiciel spécialisé est utilisé pour lire une pièce de média et créer un fichier image qui contient toutes les données dans le même ordre de lecture. Ce fichier image doit être la réplique exacte des médias, y compris les données actives et des données résiduelles. Une fois la source de données a été acquise, une opération de vérification à la fois des valeurs MD5 et SHA-1 est nécessaire. L'imagerie peut être un processus long en fonction de la taille du support. Une fois terminé, le fichier image peut ensuite être examiné.

1.6.2 Analyse

Une fois que les données ont été recueillies avec succès, une étape d'analyse est nécessaire afin d'extraire tous les éléments de preuves pertinentes. Comme indiqué précédemment, l'analyse est le processus d'interprétation de données et de les placer dans un format logique et utile pour tirer des conclusions [14]. Une partie de la phase d'analyse peut impliquer la récupération des données résiduelles. Celle-ci repose sur un logiciel spécialisé qui a la capacité de récupérer et d'analyser des données à la fois actives et résiduelles. L'analyse de la preuve numérique est un mélange de techniques telles que la recherche par mots clés , l'analyse de hachage, l'analyse des registres, e-mails et filtrage.

1.6.3 Affichage

l'affichage de contenu des pièces de média ou des images est important pour toute enquête numérique. Les enquêteurs numériques sont obligés d'explorer toutes les images graphiques à partir de média source et examiner leurs contenu manuellement [75]. Cela peut être un processus fastidieux, surtout s'il y a des milliers des images graphiques . Toutefois, des outils de visualisation [31], [30], ont été développé récemment pour aider à réduire considérablement le temps de traitement nécessaire au cours de cette partie de l'enquête.

1.6.4 Rapport

Le rapport représente le processus de capture des conclusions d'une enquête. Il doit contenir des détails essentiels de chaque étape de l'enquête, y compris la référence aux procédures suivies et les méthodes utilisés pour saisir, documenter, recueillir, conserver, restaurer, reconstruire, et réorganiser les preuves [19].
La classification figurant dans le tableau 1.4 est basée principalement sur deux caractéristiques :

- ◇ L'environnement d'exploitation, à savoir Windows, DOS et UNIX
- ◇ Les capacités des outils d'investigation légale, à savoir, l'imagerie, l'analyse, l'affichage et les rapports.

	Imagerie	Analyse	Affichage	Rapport
Basés sur Windows				
EnCase Enterprise	√	√	√	√
EnCase Forensic	√	√	√	√
FTK	√	√	√	√
FTK Image	√		√	
Basés sur DOS				
ByteBack	√			
SafeBack	√			
Basés sur Unix				
Smart	√	√	√	√

TABLE 1.4 – Atouts des outils d'investigation numérique légale

1.7 La nécessité des méthodes d'analyse durant le processus d'investigation

Quelque soit le domaine d'investigation numérique légale(mobile, ordinateur, réseaux, bases de données), la phase d'analyse est la plus importante car c'est l'étape qui permet de clarifier les événements de l'incident et de déterminer le scénario le plus approprié du crime.

Comme démontré précédemment, un outil d'investigation légale offre toutes les fonctionalités telles que l'imagerie, l'affichage, le rapport (reporting) et l'analyse. La phase d'analyse est limitée par l'emploi de simple algorithmes d'hachage, indexation simple et des thécniques de recherches utilisées pour répondre à des questions précises sur les objets de la scène du crime, tels que des outils de piratages ou des images qui sont présent dans l'information recueillie. Selon le niveau d'automatisation de recherche, il existe deux techniques : La navigation ou la recherche manuelle et la recherche automatisées.

- ***La recherche manuelle :*** Signifie que l'analyste judiciaire permet d'afficher l'information recueillie et le type d'objets désirés. Le seul outil utilisé en navigation manuelle est un wizard d'affichage . Il prend un objet de données, comme un paquet de réseau, decode l'objet et présente le résultat sous une forme compréhensible par l'utilisateur. la recherche manuelle est lente. La plupart des enquêtes recueillent de grandes quantités d'informations numériques, ce qui rend cette technique inefficace dans l'analyse judiciaire.
- ***La recherche automatisées :*** C'est la recherche automatique de l'information numérique pour les objets de données contenant des mots clés spécifique. C'est la technique la plus répandue pour accélérer la recherche manuelle. Le résultat de la recherche par mot-clé est une liste des objets de données collectées(ou leurs localisations). Les mots-clés ne sont pas suffisant pour spécifier le type d'objets de données souhaitées précisément. En conséquence, le résultat de la recherche par mot clé peut contenir des faux positifs. Un autre problème de recherche avec le mot-clé est le faux négatifs. Ce sont des objets désiré qui sont manquées dans le résultat de la recherche.

Les capacités des outils d'investigation légales sont essentiels pour la conduite des enquêtes numériques. Cependant, les supports numériques (par exemple smartphones) ne cessent de croître en taille, cela a rendu le processus d'analyse des enquêtes numériques de plus en plus complexe et prend du temps en raison de la quantité de données impliqués, et les enquêteurs numériques peuvent se trouver incapables de les conduire d'une manière appropriée et efficace. En plus les résultats offerts par les outils d'investigation peuvent souvent être trompeurs en raison de la dimensionnalité, la complexité et le volume des données légales. Mais compte tenu de la nature légale des enquêtes judiciaires numérique , la nécessité d'avoir une précision des résultats est un objectif critique pour le décideur. Des conclusion d'analyse erronées sont inacceptables dans les systèmes judiciaires. En conséquence, nous pouvons posé la question *y a t-il eu des développements d'approches dans d'autres disciplines qui pourraient bénéficier l'investigation numérique légale ?*, de nombreuses méthodes [33], [71], [39], [44], [31], [74], [8], [7] ont été développées dans la literature pour conduire le processus d'investigation. Cette section décrit les principaux approches utilisées dans le stade d'analyse.

1.7.1 Les approches formelles

Gladyshev et Patel [33] proposent un modèle à états finis pour la reconstruction des événements car de nombreux systèmes numériques réels, y compris les circuits numériques, programmes informatiques et protocoles de réseau peut être décrite mathématiquement comme une machine à état finis. Une machine à états finis peut être présentée sous forme graphique, dont les nœuds représentent l'état du système possible et les flèches représentent les transitions possibles d'un état à état. Tous les calculs possibles conduisant à un état donné peuvent être déterminé par backtracing des transitions menant à cet état en basant sur des formules mathématique pour la génération de scénario de l'incident le plus approprié. La reconstruction des événements d'un système sous investigation est une tâche importante dans la phase d'analyse de processus d'investigation. Ainsi le développement d'une telle théorie qui formalise la reconstruction des événements a plusieurs avantages :

- ◇ L'amilioration de l'efficacité de l'analyse car la formalisation des techniques de reconstruction permet de faciliter leur automatisation.
- ◇ L'amilioration de l'efficience d'analyse car le raisonnement informel employé par les techniques de reconstruction existants augmente la probabilité d'avoir des conclusions erronées.
- ◇ Pour satisfaire les exigences d'admissibilité de preuves.

Dans [71], Stephenson et al ont démontré qu'il est possible de décrire la cause ultime des attaques intranet en formalisant à l'aide de reseau de petri colorés afin de sécuriser et prévenir le système d'information de l'organisation. Cette approche est mieux adaptée pour des enquêtes complexes.

1.7.2 Les approches basées sur le datamining

Les enquêteurs numériques sont confrontés à des défis majeurs créés par le volume le plus important de données disponibles. En vertu de ces circonstances, il est souvent impossible d'effectuer une analyse légale complète en raison de contraintes du temps et de ressources humaines limitées. En conséquence, le datamining est une discipline qui pourraient bénéficier l'investigation numérique légale. Le but de l'exploration de données

est de découvrir des nouvelles connaissances à partir de données où la dimension, la complexité et le volume des données est prohibitif pour une analyse manuelle. Dans ce contexte Jeyaraman et Atallah [39] présentent une étude empirique des systèmes de reconstruction automatique. Ainsi, Khan et al [44] proposent un framwork pour la reconstruction d'un timeline (chronologie) des événements de l'incident à l'aide des réseaux de neurones. Ces deux travaux de recherche utilisent un ensemble de journaux de réseaux pour apprendre un réseau neuronal afin de connaître les attributs des journaux en intégrant les événements dans un seul timeline. Dans un autre coté, Fei et al [31] a introduit SOM(Self Organization Map) qui est un réseau de neurones non supervisé pour détecter les anomalies dans les réseaux contrôlés.

1.7.3 Les approches basées sur la théorie des graphes

Wang et Daniels [74] proposent une approche basée sur les graphes pour générer un graphe de corrélation de preuves en utilisant les captures de réseau.

1.7.4 Les approches basées sur les ontologies

Brinson et al [8] proposent l'ontologie cyber légale qui se concentre sur l'identification des couches correctes de spécialisation, de la certification et de l'éducation au sein de domaine liés au crime. L'ontologie générée a au maximum cinq niveaux d'hiérarchie, elle est déterminée que cette structuration est insuffisante pour l'analyse légale qui est beaucoup plus varié. Bogen et Dampier [7] proposent une approche de modélisation à grande échelle par cas du domaine pour les grandes enquêtes et définir des cas spécifique d'ontologie en utilisant UML(Unified Modeling Language). Le tableau 1.5 illustre les méthodes d'analyse utilisées dans l'investigation numérique légale.

	Approches formelles	Théorie des graphes	Datamining		Ontologies	
	Reconstruction des évenèments	Graphe de preuves	Analyse de log empirique	SOM	Ontologie de cyber légal	Modélisation par cas de domaine
Gladyshev et Patel	X					
Stephenson et al	X					
Wang et Daniels		X				
Jeyaraman et Atallah					X	
Khan et al				X	X	
Fei et al				X		
Brinson et al				X	X	
Bogen et Dampier					X	X

TABLE 1.5 – Méthodes d'analyse utilisées dans le processus d'investigation

1.8 Conclusion

L'investigation numérique légale est un processus compliqué qui commence à la scène du crime, continue aux laboratoires informatiques pour en tirer des éléments de preuves et se termine dans le cour où se fait le jugement final. Dans ce chapitre nous avons passé en revue les différents concepts liés à l'investigation numérique légale, et nous avons présenté les différents domaine liés au DF. Nous avons également présenté certains outils d'investigation légale les plus souvent déployés. Les caractéristiques les plus importantes sont identifiées et classées pour chaque outil. Sur la base de cette classification, nous avons identifié les limites de ces outils qui conduit à proposer des nouvelles méthodes d'analyses durant le processus d'investigation. Parmi ces méthodes on trouve le datamining qui est l'approche la plus adéquate pour traiter le volume de données lors de la phase d'analyse de processus d'investigation.
Dans le chapitre suivant nous passerons en revue les différentes techniques de datamining appliquées à l'investigation numérique légale.

Chapitre 2

La fouille de données pour l'investigation numérique légale

2.1 Introduction

Les crimes numériques sont augmentés avec la croissance remarquable de l'internet. La technologie des smartphones mobiles, les concepts et les idées derrière l'investigation numérique sont bien établis, mais la discipline est encore en naissance. Toutefois, les enquêtes numériques sont de plus en plus complexe et prend du temps en raison de la quantité de données impliquées et les enquêteurs numériques peuvent se trouver incapables de les conduire d'une manière appropriée et efficace. Par conséquent, découvrant des nouvelles méthodes qui permettront d'améliorer la qualité des décisions faites, de réduire le temps de traitement humain nécessaire et de réduire la politique monétaire des enquêtes numériques est d'une importance primordiale. Le datamining est une telle technique potentiel. Il est encore relativement inexploré dans le domaine d'investigation numérique, mais le but de l'exploration de données est de découvrir de nouvelles connaissances à partir des données où la dimension, la complexité et le volume des données est prohibitif pour une analyse manuelle. Le datamining est la synthèse de la modélisation statistique, le stockage de bases de données et la technologie de l'intelligence artificielle [54]. Il a produit de bons résultats lorsqu'il s'agit d'un grand volume de données. Récemment, des recherches récentes ont porté sur l'application des techniques de datamining à l'investigation numérique légale.

L'objectif de ce chapitre est de présenter brièvement le domaine de la fouille de données et de donner un aperçu de ses fonctionnalités et de présenter ensuite la littérature existante limitée relatifs à l'application du datamining à l'investigations numériques légale.

2.2 Le datamining et ses fonctionnalités

La fouille de données (datamining) est le processus d'extraction de connaissances à partir d'énormes quantités de données. Il est une étape essentielle d'un processus appelé KDD (Knowledge Discovery from Data). L 'effort le plus important dans le datamining est de fournir des méthodes efficaces pour les bases de données volumineuses. Les techniques de datamining ont un champ d'application très large, même si elles ne sont pas un remède miracle capable de résoudre toutes les difficultés, elles peuvent être dans la majorité des cas très efficaces pour l'extraction des connaissances utiles à la prise de décision. En appliquant ces techniques, nous ne recherchons pas la meilleure solution prouvée, mais à faire le mieux possible.

Les fonctionnalités de datamining sont utilisées pour spécifier le type de modèles à trouver dans les tâches de datamining. Les principales techniques se répartissent en deux grandes familles : Descriptives et predictives [32] :

- ◇ **Descriptives** : Elles caractérisent les propriétés générales des données de la base.
- ◇ **Prédictives** : Elles consistent à prédire des valeurs non disponibles (données manquantes), ou à prédire des classes pour certaines données.

Les classes et les concepts individuels sont décrits, sachant qu'une classe est un ensemble d'objets ayant des propriétés similaires. Par exemple, une classe dite ″employé″ peut contenir certaines variables qui l'a décrit : (Nom, $Adresse$ et $Date\_naissance$), alors que, un concept est une Représentation générale ou abstraite d'une réalité. Les données sont associées avec des classes et des concepts décrits par des termes précis, concis et résumés. Cette façon de description est appelée : *Description classe/concepts*, elle peut être dérivée via : La caractérisation, la discrimination ou caractérisation et discrimination.

- ◇ **La caractérisation :** Elle permet de faire des résumés sur les caractéristiques des données des classes sous une étude appelées classes cibles et produit des règles dites caractéristiques. Les données pertinentes sont récupérées par des requêtes sur la

base des données et elles passent à travers un modèle de réduction, pour extraire l'essentielle de données à différents niveaux d'abstraction

◇ **La discrimination :** Elle consiste de faire une comparaison entre les classes cibles avec une ou plusieurs classes dites d'opposition. La forme des résultats pour la discrimination est similaire à celle de caractérisation, sauf qu'on doit rajouter des mesures comparatives pour aider à distinguer entre les classes cibles et les classes d'oppositions.

Extraction des modèles fréquents

Définition 2.2.1. *Un modèle fréquent est un item (ou article), une sous séquence ou une structure qui apparaît fréquemment dans un ensemble de données, il joue un rôle important dans les relations intéressantes entre les données comme la corrélation et l'association. Il aide dans le processus de clustering et de classification[34].*

Soit D un ensemble de transactions, chaque transaction (identifiée par TID) est un ensemble de "m" items $I_{TID} = I_1, I_2, ..., I_m$. Une règle d'association est de la forme :

$$A \Longrightarrow B[Support, Confidence] \tag{2.1}$$

Où : le support reflète l'utilité de la règle, c'est le pourcentage des transactions satisfaisant les items sets A et B. Il est calculé comme suit :

$$support(A \Longrightarrow B) \doteq P(A \cup B) \tag{2.2}$$

La confidence reflète la certitude de la règle, c'est le pourcentage des transactions satisfaisant les items sets A sachant B. Elle est calculée comme suit :

$$confidence(A \Longrightarrow B) = support(A/B) = support(A \cup B) \tag{2.3}$$

Le problème d'extraction de règles d'association revient à l'extraction des items sets fréquents, celles qui satisfont le minimum support et le minimum confidence. Lorsque les seuils sont petits, alors il y aura beaucoup d'items à extraire. Pour surmonter cela, les items sets fermés et les items sets maximaux seront introduits.

Classification et Prédiction

La classification et la prédiction sont deux formes d'analyses pour aider dans la prise de décision ou prédire de futures tendances. Par exemple : Un modèle de classification peut catégoriser les emprunts bancaires pour éviter tout risques, ou prédire les dépenses de clients potentiels dans un marché informatique, sachant leur revenus, etc.

◇ **Classification**

C'est le processus qui trouve un modèle (ou une fonction) qui décrit et distingue entre les classes et les concepts, dans le but d'utiliser le modèle pour prédire la classe des objets dont l'etiquette de classe est inconnue. Le modèle dérivé est basé sur l'analyse d'un ensemble de données d'apprentissage[34] . Il se déroule suivant deux étapes :

- **Apprentissage** : L'algorithme de classification construit un ensemble de classes et de concepts prédéterminés à partir d'un ensemble d'apprentissage de tuples de la BD et leurs classes associées. Il s'agit de faire un mapping entre les tuples de données et les classes. Ce mapping est formé par des règles de classification, arbre de décision ou par des formules mathématiques ;
- **Calcul de la précision** : Un ensemble de test est utilisé avec les étiquettes de classes, il est indépendant de l'ensemble d'apprentissage utilisé dans l'étape précédente. la précision représente l'ensemble de tuples correctement classés et si cette estimation est acceptable alors le classificateur peut être utilisé pour prédire des données en ignorant les classes. Parmi les méthodes de classification les plus utilisées, on trouve : L'induction par arbre de décision , le classificateur Baysien, et machine à vecteurs de support [32][34][55].

◇ **Prédiction**

Il s'agit de faire la prédiction des valeurs continues pour des entrées données. Parmi les méthodes utilisées en prédiction on trouve : La régression ou la prédiction numérique avec des variances : Régression linéaire, régression non linéaire, régression multiple [34].

2.3 Datamining appliqué à l'investigation numérique légale

L'investigation numérique légale est de plus en plus couteuse en temps et complexe avec l'accroissement des volumes de données nécessitant une analyse. Les outils d'investigations d'aujourd'hui ne répond pas à tous les besoins surtout concernant le crime utilisant les smartphones.
L'extension de l'application de datamining à l'investigation numérique légale aura tout ou partie des prestations suivantes :

- ◇ Réduire le temps de traitement humain et du système liés à l'analyse de données ;
- ◇ Amélioration de la qualité d'information liée à l'analyse de données ;
- ◇ Réduire les coûts monétaires liés aux enquêtes numériques.

Cette section décrit les différentes techniques de datamining employées à ce jour dans le domaine de l'investigation numérique légale.

2.3.1 La classification supervisée

La classification par SVM(Support Vector Machine)

- ***Travaux de Vel et al***
  Vel et al [24] ont utilisé l'algorithme de machine à vecteur de support (SVM) pour extraire le contenu d'e-mail et identifier son auteur puisque il est devenu une source de plus en plus importante des preuves numériques et qui est maintenant devenu la forme dominante de l'inter et intra communication organisationnelle. Un SVM est un algorithme de classification qui cherche à classifier les données en se basant sur les caractéristiques stylistiques et structurelles de corps de l'e-mail dans l'algorithme d'apprentissage. Les en-têtes d' e-mails sont utilisés pour classer chaque e-mail dans la catégorie appropriée selon son auteur.
- ***Travaux de Malcom et al***
  Selon les travaux de Malcom et al [50], une machine à vecteurs de support a été appliquée pour déterminer le sexe de l'auteur d'un message électronique en se basant sur les caractéristiques des document mails comme les marqueurs du style, les caractéristiques structurelles et le langage préférentiel, c'est-à-dire classifier un

ensemble d'e-mails par genre d'auteur et si possible, obtenir un ensemble des caractéristiques qui restent inchangée avec un grand nombre des e-mails créés par le même genre d'auteurs. Le choix de ces attributs est basé sur la possibilité de distinguer entre les auteurs et les genres d'auteurs au lieu de distinguer entre les sujets des e-mails.

- ***Travaux de Brown et Pharm***
  L'extraction des images est une des nombreuses activités faites durant une investigation numérique. Une machine de vecteur de support peut également être utilisée pour reconnaître certaines tâches ou zones d'une image suspecte [10]. Par conséquent, elle peut être utilisé pour détecter et filtrer les images qui sont non pertinentes ou sans rapport avec l'affaire en cause. D'autres instances où le SVM a été utilisés pour comprendre la récupération de l'image [10] et dans l'exécution de requêtes de recherche sur des images contenant des objets suspects [9]
- ***Travaux de Liu et al*** Liu et al [12] proposent une méthode afin de localiser efficacement la preuve relative au crime informatique, tout en maintenant la précision. C'est une méthode à deux niveaux pour automatiser le processus de localisation, qui emploie d'abord une seule classe de machine à vecteur de support détecteur des outliers pour filtrer les enregistrements insignifiants et puis utiliser un autre groupe de classificateur SVM. Les ensembles de données d'enquête sont d'abord prétraités et les attributs décrivant l'objet cible sont extraites de la base de données par le calculateur d'entité. Après, le détecteur de valeurs aberrantes (outlier detector) est employé afin de rechercher des outliers de certains dossiers suspects aux enquêteurs et éliminer les enregistrements qui ont une haute probabilité d'être normal. Le reste des enregistrements sont transmis à post-processeur qui peut considérer comme un classificateur secondaire afin de déterminer s'il y a des faux positifs en fonction des connaissances existantes composé des connaissances spécialisées et des preuves existantes. Les résultas délivrés par le post-processeur seront étiquetés comme des échantillons de preuves qui peuvent être utilisés comme des connaissances existant dans la prochaine itération et transmirent à l'unité de l'objet cible pour définir ou mettre à jour les objets cibles déjà recherchés. Ces étapes sont répétés jusqu'à ce que la preuve d'un incident réfute ou soutient une hypothèse d'enquête [12](voir la figure 2.1)

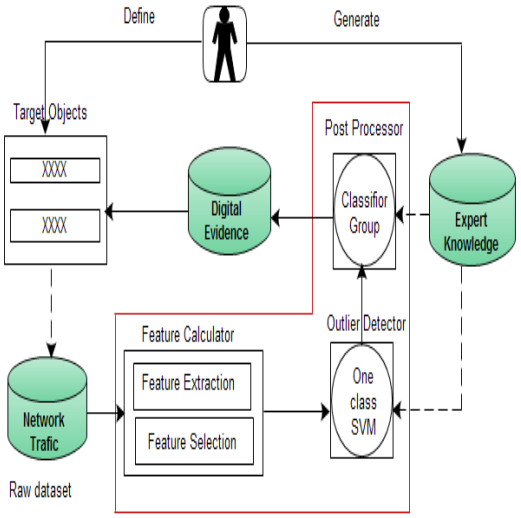


FIGURE 2.1 – Framwork de détection des outliers pour la recherche de preuves numériques [12]

La détection des outliers en utilisant une seule classe SVM avec la fonction noyau RBF(Radial Basis Function) est défini en tant que :

$$K\phi(x,y) = \exp^{-\|D(x,y)\|/\delta^2} \tag{2.4}$$

Où : $D(x,y)$ est la valeur de la différence de fonction hétérogène

$$D(x,y) = (\sum_{i=1}^{m} d_i^2(x_i,y_i))^{1/2} \tag{2.5}$$

Où : m est le nombre d'attributs, d_i est la distance d'ième fonction d'attribut définie par :

$$d_i(x_i,y_i) = \begin{cases} 1 & : x_i \ \ ou \ \ y_i \ \ sont \ \ inconnues; \\ d_{vdm}(x_i,y_i) & : x_i \ \ et \ \ y_i \ \ sont \ \ nominales; \\ d_{diff}(x_i,y_i) & : x_i \ \ et \ \ y_i \ \ sont \ \ numeriques; \end{cases} \tag{2.6}$$

Un SVM d'une seule classe utilise la fonction du noyau 2.6 qui transforme les exemples non étiquetés dans un espace multidimensional afin de séparer au maximum les données "normales" de l'origine par une limite d'hyperplan en séparant

les outliers, la classe SVM doit résoudre le problème de programmation quadratique suivant :

$$\begin{array}{c} Minimiser \quad 1/2\omega^T\omega + C\sum_{i=1}^{l}\xi_i \quad sous \quad contrainte \\ y_i(\omega^T\phi(X_i)) \geq 1 - \xi_i, \xi_i > 0 \end{array} \tag{2.7}$$

Lorsque les vecteurs d'apprentissage x_i sont mappés dans un espace de dimension superieur par la fonction $\phi(\bullet)R^n \longrightarrow R^m$, $m > n$, qui peut être linéaire ou non linéaire, $C > 0$ est le paramètre d'erreur, ω est un vecteur dans l'espace de dimension superieur R^m, le détecteur des outliers est constitue des données non étiquetées. Le reste de l'ensemble de données d'origine est anormale, mais avec des faux positifs (FP) élevés. Pour minimiser le taux FP le groupe de classificateurs dans le post-processeur utilisé pour les filtrer de la sortie du détecteur des outlier. L'algorithme de SVM d'une classe avec un noyau gaussien modifié (RBF) est mettre en œuvre. La sortie de post-processeur sont des preuves et un enquêteur judiciaire peut définir de nouveaux objets cibles ou employer d'autres outils comme l'outil de recherche par mot clé pour obtenir d'autre preuves.

Classification via l'analyse discriminante

– ***Travaux de Carney et Rogers***

Carney et Rogers [13] ont montré comment l'analyse discriminante peut être utilisée pour déterminer la probabilité de l'intention associée au téléchargement des images de trafics (pédopornographie). Leur motivation était de fournir un mécanisme pour la reconstruction des événements avec le calcule de probabilité de précision pour aider les enquêteurs d'étudient la défense de Cheval de Troie. Ils ont examiné sept caractéristiques différentes(variables ou caractéristiques) des données déterminées empiriquement et qu'un modèle unique à deux éléments (le moyen entre le temps de création de fichier et de l'écart type de différence entre les temps de création de fichiers) qui peut être développé pour vérifier l' intentionnalité de l'utilisateur associé à l'incidence de la contrebande stockée sur des supports numériques.

Classification via l'arbre de décision

– ***Travaux de Iu***

Puisque les données forensics sont inconstants, bruyantes et dispersives. Basé sur l'algorithme de classification supervisé ID3, l'auteur [62] propose un nouveau algorithme pour permettre la recherche des preuves utile de façon efficace et pour le rendre plus adaptés aux données forensics. La stratégie de l'algorithme ID3 est simple, si les données d'apprentissage sont dans la même classe, alors l'arbre de décision à un nœud feuille étiquetée avec la classe, sinon, l'algorithme permet de sélectionner les meilleures propriétés de l'échantillon de données comme une propriété de classement, la selection de cette propriété est basée sur le calcule de l'entropie(valeur qui permet de mesurer le désordre dans un système) et le gain d'information(voir les équations 2.8, 2.9, 2.10). L'algorithme est appliqué récursivement jusqu'à ce que tous les échantillons soient classés dans une classe.

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{2.8}$$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + ... + s_{mj}}{s} I(s_{1j}, s_{2j}, ..., s_{mj}) \tag{2.9}$$

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A) \tag{2.10}$$

Malgré que ID3 est un algorithme efficace pour la recherche de preuves mais selon l'auteur, il y a certains inconvénient :

◇ Le calcul de l'entropie dépend relativement à un grand nombre de valeurs d'attributs. Mais l'analyse des données forensics pour la sélection d'attributs dépend de paramètres distincts, plutôt que les données elles mêmes ;

◇ Les valeurs des propriétés ont le même effet sur la classification comme leurs propriétés, mais l'algorithme ID3 sélectionne que les attributs ;

◇ ID3 est sensible au bruit. Dans le processus de l'investigation légale, les exemples de teste devrons être un bruit inévitable ;

◇ Dans le processus de construction de ID3, il détermine la propriété de classement, à base de la valeur du gain d'informations, mais la propriété de ce dernier n'a

pas été prise en compte ;

◇ Le manque de participation des usagers dans le processus de génération de l'arbre de décision.

Une amélioration de l'algorithme ID3 permet [62] :

◇ Introduire un poind α pour chaque attribut des exemples d'apprentissage alors la formule 2.9 devient 2.11 :

$$E(A) = \alpha_A \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + ... + s_{mj}}{s} I(s_{1j}, s_{2j}, ..., s_{mj}) \tag{2.11}$$

◇ A Chaque sélection d'une nouvelle propriété, l'algorithme ne prend pas seulement en compte le gain d'informations apporté à propos de cette propriété, mais prend également en compte les deux niveaux tiers de nœud de l'arbre. L'idée est la suivante : A est la propriété candidat. A possède v valeurs différentes, et les probabilités correspondantes sont $q_1, q_2, \ldots, q_v$ respectivement, en conformité avec le principe de l'entropie minimale, $B_1, B_2, ..., B_V$ sont les v attributs sélectionnés par les sous nœuds. Selon la formule 2.11, l'information d'entropie $E(B_1), E(B_2), E(B_v)$ est calculée, le choix de la propriété du classement de l'algorithme est donné par la formule 2.12 :

$$E\prime(A) = \sum_{j=1}^{v} q_i \cdot E(B_i) \tag{2.12}$$

Les résultats de cette amélioration sont illustrés par la figure 2.2 :

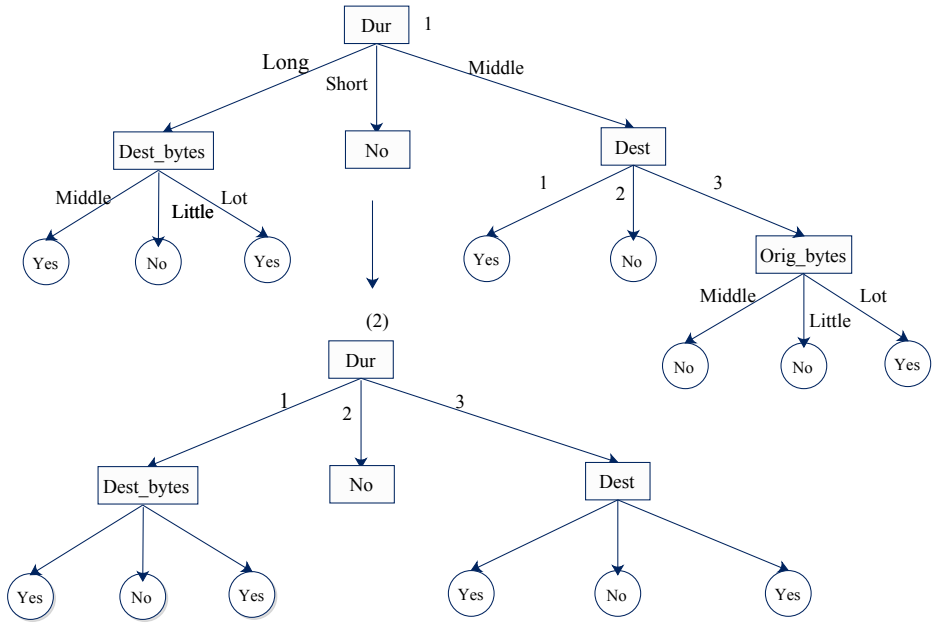


FIGURE 2.2 – Classification via l'arbre de décision avec ID3 (1), avec amélioration de ID3 (2) [62]

Classification via les réseaux de neurones

– ***Travaux de Yan et al***

Yan et al [79] proposent un système dynamique (se réfère au phénomène que les preuves légales seront acquis quand le comportement illégal apparaît automatiquement) pour la collecte des preuves basé sur les réseaux de neurones BP (Backtracking Propagation). Cette technique combine les avantages des réseaux de neurones artificiels dans la reconnaissance de formes qui peuvent identifier des attaques réseau de manière rapide et efficace pour trouver certaines preuves électroniques légales, valides, précis et complètes. L'auteur considère l'étape de découverte de preuves comme un problème de reconnaissance de formes. Le système proposé est illustré par la figure 2.3 :

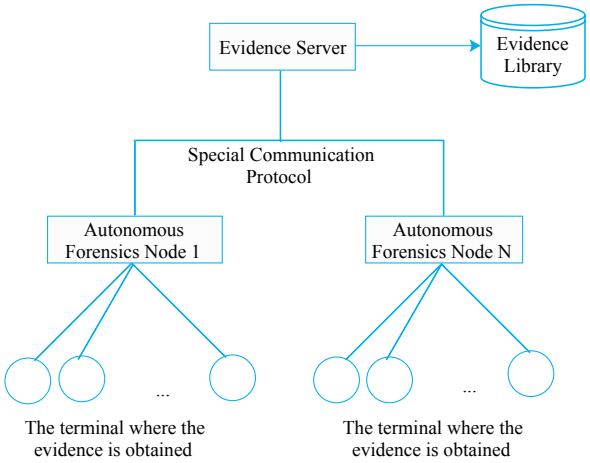


FIGURE 2.3 – Structure topologique globale du système pour l'investigation légale [79]

Le serveur de preuves(evidence server) est le centre de contrôle des trafics, qui est responsable de lancement de la tâche d'investigation et la collection de résultats. La couche de nœud autonome forensics(autonoumous forensics node) est la clé du système qui permet la collecte, l'analyse et la préservation des éléments de preuves dans un temps réel selon le flux de travail illustré par la figure 2.4.

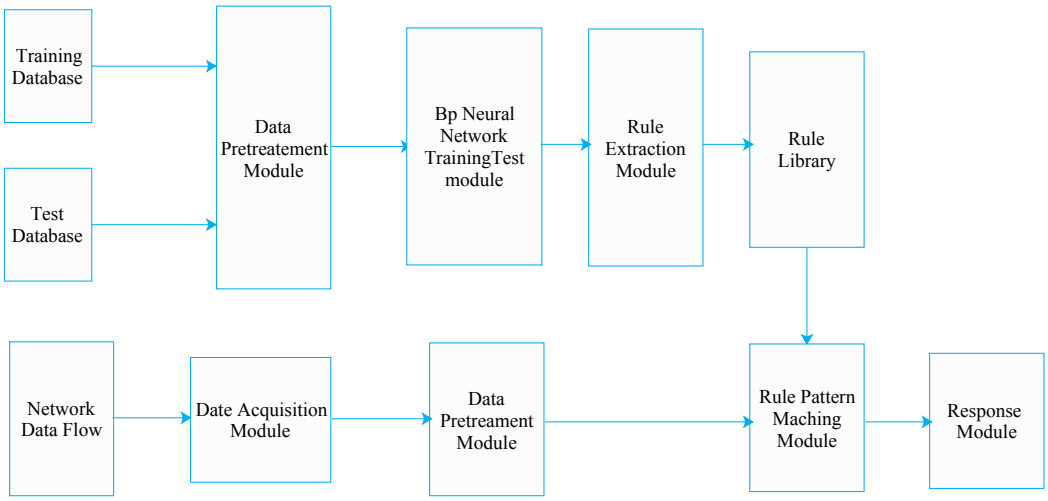


FIGURE 2.4 – Flux de travail d'un noeud d'investigation légale [79]

Au même temps, les résultats de l'expertise pour le cas d'investigation sont soumis au serveur par le biais d'un protocole de communication spéciale et ensuite le serveur stocke les preuves dans la bibliothèque de preuves(evidence library) en temps réel.

– ***Travaux de Khan et al***

La reconstruction d'un timeline post-événement joue un rôle crucial dans les enquêtes légales et sert comme un moyen pour identifier les preuves d'un crime numérique. Une approche basée sur les réseaux de neurones pour la reconstruction de la chronologie post-événement en utilisant les activités des fichiers système [44], elle consiste à surveiller leurs manipulations, capturer les instantanés à des intervalles du temps discrets pour caractériser l'utilisation de différentes applications logiciels et ensuite utiliser ces données collectées pour former un réseau de neurone afin de reconnaître des modèles d'exécution des programmes d'application. Le timeline est utile pour mettre en évidence l'accès des utilisateurs au système cible, l'exécution de certaines applications et l'identification des fichiers de données qui ont été rendus accessibles, modifiés ou supprimés pendant des périodes spécifiques.

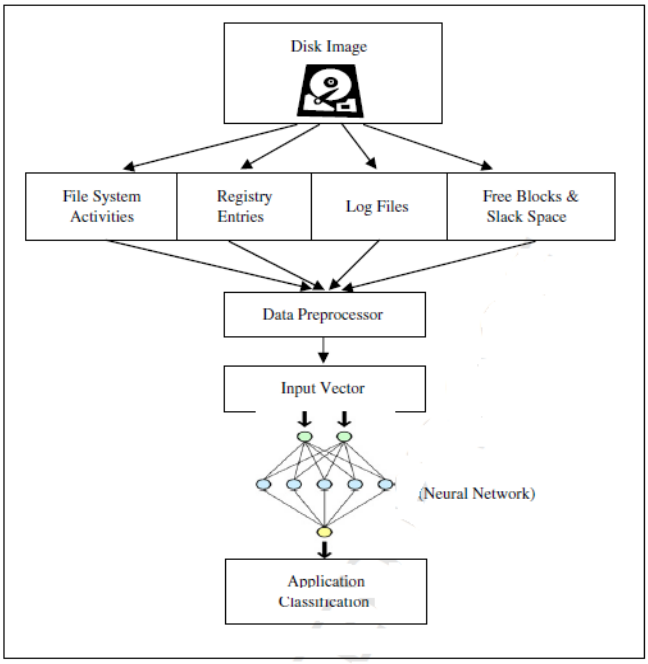


FIGURE 2.5 – Application de modèle conceptuel de classification [44]

Khan et al [44] ont utilisé deux modèles de réseaux neuronaux, perception et réseaux de neurones récurrents, un réseau de neurone multi-couche de perception (feedforward) aussi connu comme de perception [35], constitue la forme générale des réseaux de neurones pour classifier des modèles de données non linéaires. Les interconnexions entre les neurones appelés coefficients synaptiques sont utilisées pour stocker les connaissances acquises par le réseau. Un réseau d'anticipation(recurrent) est un réseau dans lequel les connexions entre les neurones ne forment pas un cycle dirigé [28]. Cette architecture de réseau récurrent est utile pour évaluer les modèles de données de séries chronologiques dans laquelle certains types d'entrées sont répétés périodiquement au cours d'une période du temps.

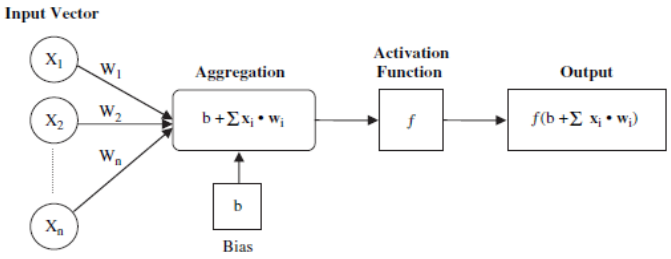


FIGURE 2.6 – Architecture de modèle perception de réseau de neurones [44]

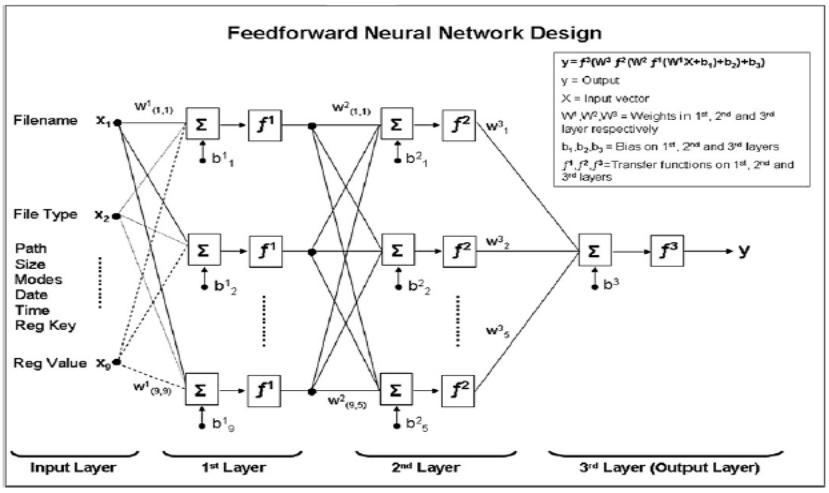


FIGURE 2.7 – Conception du réseau de neurones récurrent [44]

La classification via les réseaux bayésiennes

– ***Travaux Duval et al***

Un réseau bayésien utilise la théorie de probabilité et la théorie des graphes pour construire l'inference probabiliste des modèles de raisonnement. Il est défini comme un graphe orienté acyclique avec des nœuds et des arcs. Les nœuds représentent des variables, des événements ou des éléments de preuves. Un arc entre deux nœuds représente une condition de précédence causale, les arcs sont unidirectionnels et ils n'ont pas de feedback. Grâce à cette caractéristique, il est facile d'identifier la relation parent-enfant ou la dépendance de probabilité entre deux nœuds. la théorème de BAYES est définie :

$$p(E|H) = \frac{P(E)P(H|E)}{p(H)} \tag{2.13}$$

Ou $P(E|H)$ désigne la probabilité conditionnelle de E à H. $P(H|E)$ est la probabilité a posteriori, c'est à dire la probabilité que lorsque E a lieu alors H est effectivement a lieu. $P(H)$ désigne la probabilité a priori de H à un stade où la preuve n'est pas encore présentée. $P(E)$ est la probabilité a priori de E, qui est parfois considéré comme une constante de normalisation.
Duval et al [27] utilisent les réseaux bayésiennes pour modéliser un plan d'investigation d'un système XMeta dans le but d'automatiser les enquêtes légales. Un réseau bayésien est généralement construit en intervention des experts de domaine pour obtenir des informations sur les nœuds, les liens de causalité et les valeurs de probabilité. Le plan d'enquête est crée en entrant la configuration de l'hôte physique(Targeted Hardware(TH)) et logique(Targeted Software(TS)) et le dommage observé(Reported Damage (RD)). Le réseau bayésien utilise la probabilité de pondération approximative d'inference technique pour raisonner sur l'attaques (Generic Attacks(GA)) et les actions supplémentaire(Additional Actions(AA))afin d'énoncer un hôte. D'autre part, les actions (AA) ne sont pas obligatoires, bien que leur présence puisse aider les enquêteurs, XMeta fournit les techniques d'investigation (Investigation Techniques(IT)) qui peuvent être utilisées lors de l'attaque d'un système d'ordinateur et réseaux(voir la figure 2.8)

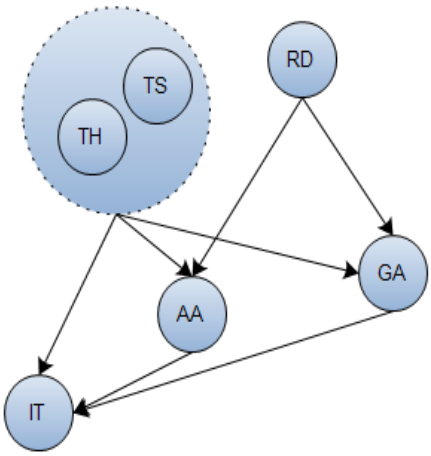


FIGURE 2.8 – Plan d'investigation du système XMeta [27]

– ***Travaux de Kwan et al***

Kwan et al [46] ont utilisé les réseaux bayésiens pour quantifier les preuves qui soutenir des hypothèses formulées lors d'une enquête. Ils ont démontré l'utilité d'un modèle de réseau de croyance bayésien en utilisant des preuves dans une affaire criminelle impliquant le partage illégal de fichiers via BitTorrent. Kwan et ses collègues ont précisé les distributions de probabilité des hypothèses et des éléments de preuve dans le réseau de croyance bayésien. Pour améliorer la fiabilité et la traçabilité de résultats produits par les enquêtes légales. Le modèle, qui est basé sur les distributions de probabilité des hypothèses dans un réseau bayésien signifie aussi que s'il y a des preuves qui supportent l'hypothèse alors elle est valide. La construction d'un modèle de réseau bayésien commence avec l'hypothèse principal afin de prouver l'acte illégale avec les états possibles de l'hypothèse (Oui, Non et Incertain) et affecter des valeurs de probabilité pour ces états. Les valeurs sont également appelées les probabilités a priori de l'hypothèse. Après avoir établi le nœud racine, le processus à explorer des preuves ou événements qui dépendent causalement en produisant un graphique. L'un des principaux défis dans l'application d'un réseau bayésien est d'évaluer les preuves est l'attribution de probabilité à posteriori. La tâche suivante consiste à attribuer des valeurs de probabilités conditionnelles aux événements ou des preuves tels quels sont illustré par la figure 2.9)

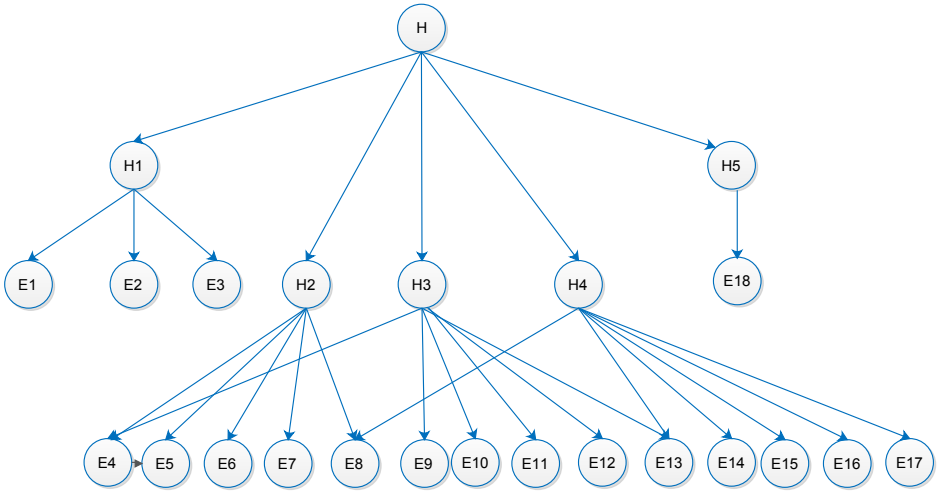


FIGURE 2.9 – Diagramme de réseau bayésien [46]

2.3.2 La classification non supervisée

Extraction du contenu via text mining

– ***Travaux de Decherchi et al***

Les dispositifs numériques contient une quantité d'infirmations non structurées et qui est utile pour une investigation. Decherchi et al [25] proposent un processus d'enquête en deux étapes basé sur l'extraction de l'information textuelle vise à générer une collection de fichiers de texte et puis analyser ces données textuelles via les outils d'analyse de texte(text mining). Ce modèle de clustering dynamique, adaptable pour organiser les documents non structurés en fonction du contenu homogènes de groupes. Le clustering est conçu pour découvrir des groupes dans l'ensemble des documents tels que les documents au sein d'un groupe ont une forte similarité par contre les documents entre groupes ont une forte dissimilarité avec un critère de partitionnement à faible coût. Formellement, l'objectif consiste à calculer une fonction ϕ de mappage $\phi : D \longrightarrow \{1, ..., Z\}$ telle que ϕ minimise le coût de partitionnement de D documents en Z clusters en respectant les similitudes E entre les documents. Un document D est finalement réduit à une séquence de termes représenter par un vecteur, qui se trouve dans un espace enjambé par le dictionnaire(ou vocabulaire) $T = \{t_j; j = 1, .., n_T\}$. Le dictionnaire recueille tous les termes utilisés pour représenter chaque D, le document peut être assemblé empiriquement en recueillant les termes qui se produisent au moins une fois dans la collection du documents, Différents modèles peut être utilisé pour extraire des termes d'indexation et de générer le vecteur qui représente le document d. $\mathbb{D} = \{D_u; u = 1, .., n_D\}$ c'est le corpus(collection de documents). L'ensemble $T = \{t_j; j = 1, .., n_T\}$ est le vocabulaire, qui est l'ensemble des termes qui se produisent au moins une fois dans $\mathbb{D}$. L'auteur propose une similarité entre documents en tenant en compte, la description du contenu de document et la description des propriétés structurelles(style) alors la distance entre les deux documents D_u et D_v peut être définie :

$$\triangle(D_u, D_v) = \alpha\triangle^{(f)}(D_u, D_v) + (1 - \alpha)\triangle^{(s)}(D_u, D_v) \tag{2.14}$$

Un document D est représenté par une paire de vecteurs réels, $v\prime$ et $v\prime\prime$. Le vecteur

$v\prime$ traite de la description d'un document D

$$V\prime_{k,u} = tf_{k,u}/\Sigma_{l=1}^{n_T} tf_{l,u} \tag{2.15}$$

où $tf_{k,u}$ est la fréquence de l'occurrence de k-ème terme dans le document D_u.
Le vecteur $v\prime\prime$ vise à exploiter les propriétés structurelles d'un document D. Ces propriétés sont représentées par un ensemble de lois de probabilités associées aux termes dans le vocabulaire. Chaque terme t ∈ T se produisant dans D_u est associé à la répartition qui donne la fonction de densité de probabilité spatiale pdf (probability density function) de t en D_u. L'auteur utilise une approximation de la fonction gaussienne, il assume que le document est découpé en section, alors un vecteur de dimension s est créé pour chaque terme $t \in T$, chaque élément dans le vecteur estime la probabilité d'occurrence de terme correspond à chaque section dans le document, Alors le vecteur $v\prime\prime$(D) est un tableau de vecteurs n_T dont la dimension S.

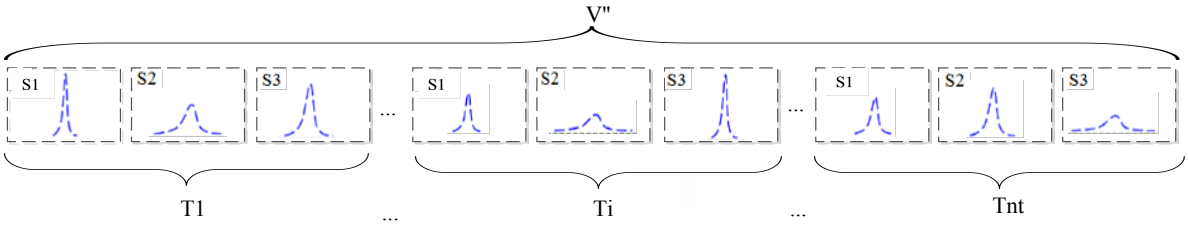


FIGURE 2.10 – Représentation de vecteur $v\prime\prime$ [25]

Les vecteurs $v\prime$ et $v\prime\prime$ prennent en charge le calcul de la distance par fréquence, $\triangle(f)$ et la distance stylistique, $\triangle(s)$, respectivement. Cette mesure de similarité a été employée dans le regroupement de noyau K-Means , le paradigme classique k-means prend en charge un processus de regroupement, qui répartit l'ensemble de corpus, $\mathbb{D}$ = $\{D_u\,;\, u = 1,..,\, n_D\}$, en un ensemble de clusters de Z, C_j (j = 1,..., Z). Dans la pratique, on définit un ˝vecteur de mappage˝ , ou ˝fonction d'appartenance˝ δ_{uj}, (D_u,C_j), qui définit la composition du uème document à la jème cluster : $\delta_{uj} = 1$ si m_u = j et 0 sinon. Par conséquent, le nombre des membres d'un cluster sont :

$$N_j = \sum_{u=1}^{n_D} \delta_{uj}; j = 1, ..., Z \tag{2.16}$$

et le centre de gravité de cluster est donnée par :

$$W_j = 1/N_j \sum_{u=1}^{n_D} X_u \delta_{uj}; j = 1....Z \tag{2.17}$$

Une fonction ϕ peut mapper n'importe quel document D dans un espace éventuellement infinie $\phi(D)$ qui est l'espace de dimension de Hilbert. Dans le nouvel espace mappé les centres de regroupement deviennent :

$$\Psi_j = 1/N_j \sum_{u=1}^{n_D} \phi_u \delta_{uj}; j = 1....Z \tag{2.18}$$

Le clustering de données pour les dispositifs numériques

– ***Travaux de Veena et al***

Veena et al [73] proposent un framework générique pour la génération, le stockage et l'analyse de données extraite de dispositifs numériques physiques et qui se présentent comme preuves pour l'analyse légale. Afin de valider l'approche Veena et al prennent un cas typique de la mémoire fash pour analyser son contenu. Ce framework est composé de six étapes(voir la figure 2.11) :

◇ **Extraction :** Le dispositif numérique focalisé pour l'investigation est la mémoire flash. si elle est protégée par un mot de passe, des logiciels doit être installés afin de cassé le code de sécurisation et décrypter les données encryptées. Les données extraites de dispositif numérique original doivent être protégé contre les autres processus d'écriture ;

◇ **Transformation :** C'est une structuration des données extraites à l'étape précédente, les principales transformations sont : la conversion des dates dans un format standard, la génération des racines de répertoires et l'extraction des extensions du fichiers ;

◇ **Chargement :** Avant que les ensembles des données sont chargé pour la fouille de données, elles ont validé statiquement par le test Bartlett's test of sphericity and Kaiser-Meyer-Olkin(KMO) afin de calculer la mesure de l'adéquation d'échantillonnage ;

◇ **Serveur de datamining :** C'est un système de datamining constitue d'un

ensemble de fonction comme l'association, le clustering, la classification, l'évolution et l'analyse des écarts ;

◇ **Clustering :** Les instances de données sont groupées dans des groupes significatifs par l'algorithme de clustering k-means selon l'intérêt et les auteurs ont évalué les performances du modèle proposé afin de détecter les valeurs aberrantes (outliers).

◇ **Classification :** Les groupes de données sont classifiés, les auteurs ont utilisé l'arbre de décision C4.5 ;

◇ **Validation croisée :** La méthode standard de prédiction du taux d'erreurs d'un arbre de décision est 10 facteurs de validation croisée, cette technique est adoptée surtout lorsque la quantité des données impliquées dans l'analyse est limitée.

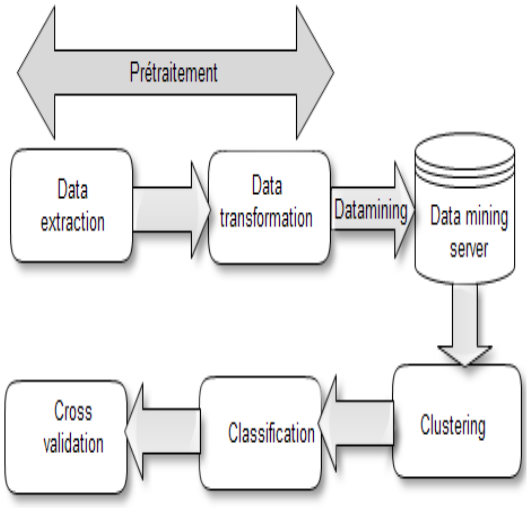


FIGURE 2.11 – Modèle d'analyse des données à partir des mémoires flashs [73]

La visualisation via le clustering

– ***Travaux de Fei et al***

Fei et al [30] exploitent les techniques de web mining pour faire l'investigation des crimes liés à l'internet (Network Forensics). Cette étude est focalisée sur l'extraction d'usage du web (web usage mining). L'objectif de cet étude est de démontrer comment SOM (Self Organisation Map) qu'est un modèle de réseau neuronal peut

aider les enquêteurs numériques pour prendre de meilleures décisions. La capacité de visualisation de SOM peut non seulement être utilisée pour extraire des motifs intéressant, mais peut également servir comme une plate-forme pour l'analyse interactive. La visualisation est le processus d'observation des données sur un support numérique. SOM a l'habilité de mapper des données multidimensionalles à un espace à deux dimensions afin de réduire la complexité coexistée dans les données pour extraire tout les motifs intéressants aux investigateurs, ce qui permet un processus d'analyse efficace et de qualité. L'architecture de l'outil SOMFA (Self Organization Map Forensics Analysis) se compose de trois éléments principaux. Chaque composant exécute l'une des trois phases requises, à savoir, le prétraitement des données, la découverte de motifs et l'analyse du modèle tels quel sont illustrés par figure 2.12

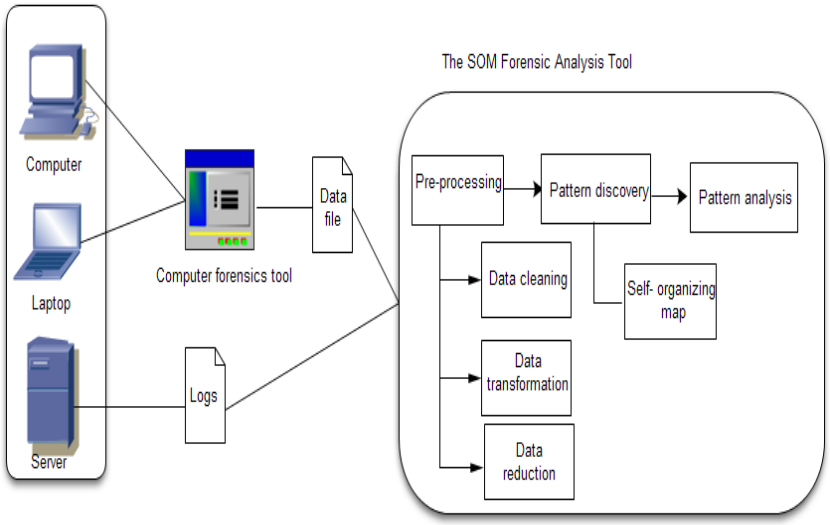


FIGURE 2.12 – Architecture de l'outil SOMFA [30]

◇ Une fois la source de la preuve a été acquise, un fichier de données contenant des informations créés par un outil forensics pour être traité à l'aide de l'outil de SOMFA ;

◇ Les preuves peuvent être également collectées auprès des sources tels que les journaux des serveurs web ou les journaux du web proxy. Néanmoins, une phase de prétraitement doit être effectuée en premier. Elle implique le nettoyage, la transformation et la réduction de données ;

◇ La découverte de motifs pour extraire de nouvelles connaissances, sous forme des anomalies à partir des données légales. Elle implique l'utilisation de SOM pour le mappage de données légales sur un espace à deux dimensions, qui peuvent ensuite être visualisés et analysés au cours de la phase d'analyse des modèles.

◇ L'objectif de la phase d'analyse est d'identifier des éventuelles corrélations, des associations et des anomalies dans les données. Une telle identification est prise en charge par les différentes visualisations des données possibles en utilisant SOM pour fournir aux enquêteurs une vue générale compréhensible de données légales.

2.3.3 L'extraction des règles d'associations

Travaux de Vel et Abraham

L'investigation du profil est une activité importante dans l'investigation légale, elle peut réduire considérablement la recherche de criminel et le motif du crime, Vel a collaboré avec Abraham [1] pour identifier le comportement des utilisateurs ainsi que les irrégularités dans leurs profils. Ces chercheurs ont appliqué les règles d'associations basées sur des événements liés aux rôles de l'utilisateur afin de trouver des preuves qui pourraient être des séquences journaux des ensembles de données. Un profil se compose de deux éléments, à savoir les éléments factuels et de comportement. Le profil factuel(Profile Factuel (FP)) est constitué des connaissances du contexte concernant les suspects tels que leurs noms, statuts, adresses, etc. Le profil comportemental (Behavioural Profile (BP))intègre les connaissances sur le crime et le comportement de son coupable. BP peut être représenté en tant qu'union des hiérarchies multi slot des sous-profiles (PH_j) :

$$BP \quad \leftarrow \bigcup_{j}^{M} PH_j \tag{2.19}$$

BP peut également être modéliser comme un jeu de règles d'association :

$$BP \leftarrow \{R_i / i = 1, 2, ..., N\} \tag{2.20}$$

Ces règles fournissent un moyen intuitif et déclaratif pour décrire le comportement de l'utilisateur.
La figure 2.13 présente le système proposé par les auteurs pour le processus de pro-

fil qui permet de générer des règles d'association en tenant compte de l'hiérarchie des concepts(taxonomie d'attributs) et les données déduites de fichiers journaux. Le concept d'hiérarchies est souvent utilisé dans le processus de génération des règles pour les transmettent comme une généralisation des concepts de nœud à la racine et peut être représenté comme un ensemble des relations parent-enfant dans un fichier de données.

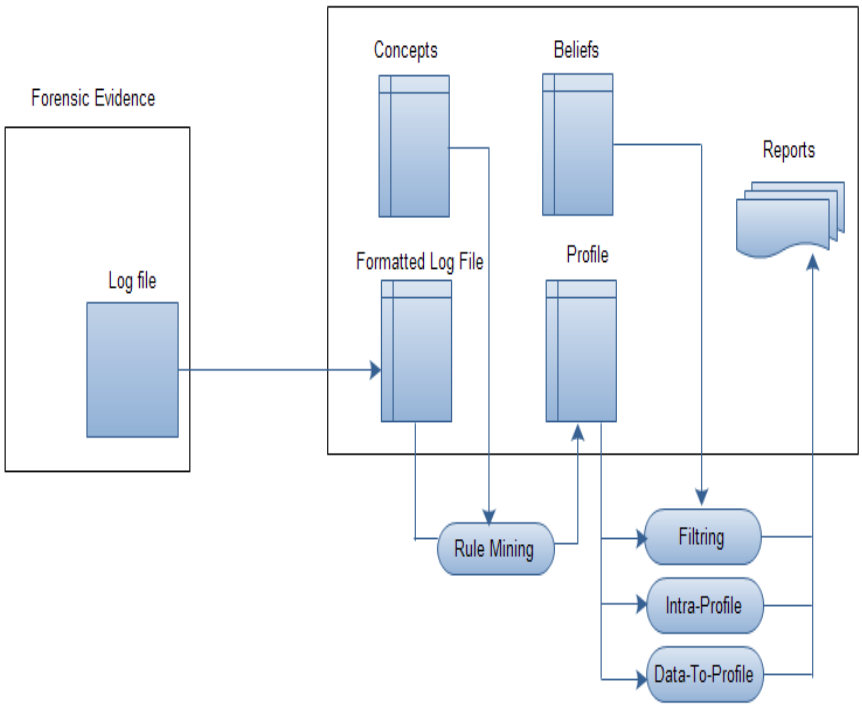


FIGURE 2.13 – Diagramme de flux de données pour le processus de génération de profil [1]

L'algorithme de génération de profil à l'aide des règles d'associations en utilisant la hiérarchie et les croyances de l'expert du domaine (Matrix to Itemsets Using Concepts) est défini :

Algorithme 1 M2IS-c

Require: Matrix M of bits $(n * m)$;
Matrix C of bits for relation concept $(m * m)$;
minsup $\in [0, 1]$;
Ensure: Collection of itemsets I satisfying **minsup** ;
$I = \bigcup_K I^K$
1: Intialize K :=1 ;
2: **For** each column i=1.....m de M
3: Initialize support, $sup_i := \sum_{j=0}^{n} b_{ji}, b_{ij} \in [0, 1]$;
4: **EndFor**
5: Add 1-itemsets I_1^i to I where $sup_i/n >= minsup$
6: Increment k. Stop if $k > l$, otherwise for the current k :
7: Initialize k-itemset count $count_k := 0$
8: Generate potential k-itemset from existing $(k-1)$-itemsets by finding next pair $I_i^{K-1} = \{i_i^1, ..., i_i^{K-1}\}$,$I_j^{K-1} = \{i_j^1, ..., i_j^{K-1}\}$ so that $i_i^o = i_j^o$,$o = 1, ..., K-2$,$i_i^{k-1} < i_j^{k-1}$, and i_i^{k-1} is not in a concept relationship with i_j^{k-1}
9: **For** potential itemset I_{ij}^k=$\{i_i^1, ..., i_i^{K-1}, i_j^{K-1}\}$ calculate support sup in M by counting the rows where all bits appearing in $I_{ij}^K$are set.
10: **EndFor**
11: **If** $sup/n \geq minsup$, add I_{ij}^K into I as a k-itemset and increment $count_k$
12: **Goto Step 7** until all potential k-itemsets are found
13: **Stop** if $count_k = 0$, otherwise **Goto Step 6**.

Après l'étape de génération du profil, l'analyse vient à l'aide de processus de filtrage, ce dernier peut être guidé par un ensemble de croyances. Une collecte séparée des règles est utilisée pour décrire un ensemble de ces croyances. Ce filtrage permet de chercher des irrégularités dans le profil, les règles conformes aux croyances peuvent être automatiquement supprimés, tandis que les règles en contradiction aux croyances peuvent avoir une priorité plus élevée dans le processus d'investigation.
La génération des contradiction intra profil. Cela signifie de trouver des règles en contradiction avec d'autre règles dans le même profil. Pour mesurer la différence entre les règles, une métrique de distance est proposée, l'algorithme (Itemset to itemset distance) permet de la calculer.

Algorithme 2 IS2IS-dist

Require: K-itemsets $I_i = \{b_1^i, ..., b_k^i\}$; $I_j = \{b_1^j, ..., b_k^j\}$;
An $(m * m)$ concept relationship bit-matrix C ;
Attribute function attr(b) ;
Ensure: Distance $d \in [0, ..., k+1]$;
1: Initialize $d := 0$
2: **For** $o := 1$ *to* k
3: **If** attr $(b_i^0) \neq attr(b_j^0)$; set $d := K + 1$ and stop ;
4: **If** attr $(b_i^0 \neq b_j^0) \wedge (b_i^0)$ not in child-parent relationship with (b_j^0),
5: increment d ;
6: **EndFor**

2.3.4 La detection des outliers

Travaux de Carrier et Spafford

Carrier et Spafford [18] ont proposé un framework pour localiser des preuves dans des fichiers et répertoires qui ont été cachés ou supprimés. Afin de localiser les fichiers cachés, les caractéristiques de chaque fichier sont comparées dans un répertoire pour détecter les valeurs aberrantes possibles. Ils ont utilisées des techniques d'exploration de données pour collecter les fichiers et les répertoires créés lors de l'incident. Ce travail décrit deux approches de recherche automatisée de preuves . La première approche suggère de nouvelles recherches basées sur des preuves existantes et la deuxième approche utilise une analyse des valeurs aberrantes pour trouver les fichiers et répertoires qui ont été créés ou modifiés pendant l'incident. Ces approches peuvent aider à faire des enquêtes plus approfondie et plus précise. Le framework proposé constitue de quatre phases qui sont illustrées par la figure 2.14.

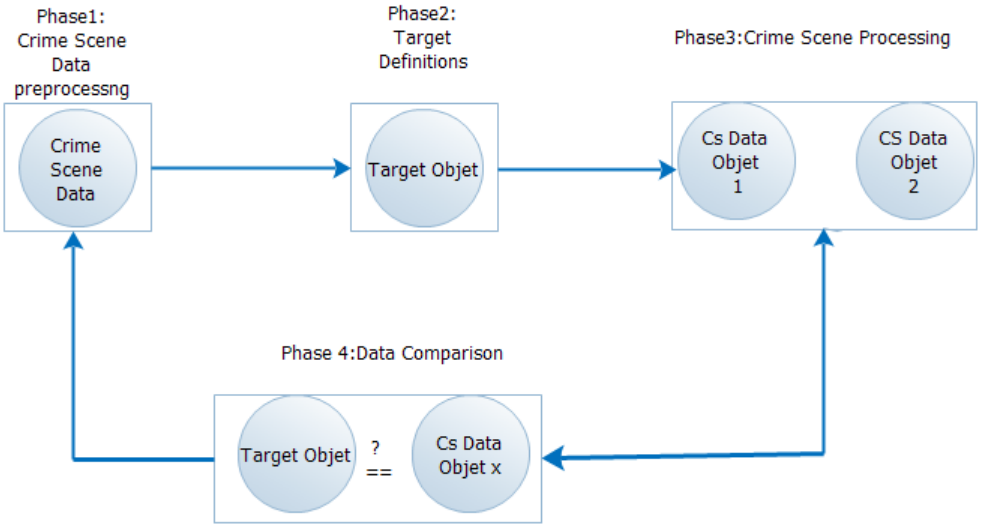


FIGURE 2.14 – Flux des quatres phases de processus de recherche de preuves [18]

phase1 : Le prétraitement du crime

C'est l'endroit où un outil traite les données de la scène du crime pour faire des recherches plus efficaces. L'objectif de cette phase est d'organiser la scène du crimes de sorte que le temps nécessaire au traitement de données sera plus petit.

phase2 : La définition des objets cibles

Cette phase identifier les objets à examiner. Un objet cible est une entité abstraite avec des attributs qui définissent les caractéristiques des preuves à prédire. Les objets cibles sont définies en fonction des hypothèses sur les incidents.

phase3 : Le traitement des données de la scène du crime

Cette étape permet de traiter les objets de données de la scène du crime afin qu'ils soient au format adéquates et de les comparer avec les caractéristiques de l'objet cible.

Phase4 : La comparaison de données

La phase finale dans le processus de recherche compare les caractéristiques de l'objet de données de la scène du crime avec les caractéristiques de l'objet cible. Si elles sont similaires

alors les objets de la scène du crime sont des preuves. Cette phase et la phase précédente sont menée en parallèle où chaque objet de la scène du crime est immédiatement comparé par rapport à l'objet cible. Les comparaisons peuvent être automatisées, manuels ou une combinaison des deux.

La définition des objets cibles est basée sur des preuves existantes, le concept de base est que l'enquêteur utilise un outil d'analyse pour identifier les fichiers comme preuves et fait des suggestions pour des recherches supplémentaires. La deuxième méthode est basée sur la détection des valeurs aberrantes d'un seul attribut du fichier pour trouver les fichiers qui ont été cachés ou qui sont différents des autres. Les objets sont des outliers (aberrants) lorsque ils sont totalement différents ou incompatible avec le reste de l'ensemble de données [34]. Pour cette technique de recherche, l'algorithme itératif spatial aberrant [49] a été utilisé pour la détection des valeurs aberrantes spatiales.

1. Définir chaque fichier comme un point spatial $x_i(pour i = 1, 2, ..n)$ et son voisinage $NN(x_i)$ comme l'ensemble de fichiers appartient au même répertoire que x_i
2. Calculer le résumé de voisinage :

$$g(x_i) = \frac{1}{\mid NN(x_i) \mid} \sum_{x \in NN(x_i)} f(x)$$

   et le comparer pour chaque point x_i :

$$h_i = h(x_i) = f(x_i) - g(x_i)$$

3. Compte tenu de l'ensemble des différences $\{h_1, h_2, ..., h_n\}$, le calcule de la moyenne μ et l'écart type σ normalise les points à l'aide $y_i = \mid \frac{h_i - \mu}{\sigma} \mid$, pour $i = 1, 2, ...n$
4. Soit y_q le plus grand y_i, s'il est supérieure à un seuil θ alors x_q est un outlier.
5. Retirer x_q de données afin d'examiner les autres points. Pour se faire, définir la valeur de l'attribut x_q d'être la moyenne du voisinage, $g(x_q)$. Aller à l'étape 1.

De toute évidence, s'il n'y avait pas des fichiers liés à un incident. Une autre technique de détection des outliers à été proposée. L'algorithme est le suivant :

1. Normaliser chaque valeur d'attribut comme $f_j(x_i) = \frac{f_j(x_i) - \mu f_j}{\sigma f_j}$ pour chaque attribut $0 < j \leq q$. Où μ_{f_j} est la moyenne de l'attribut j et σ_{f_j} est l'écart-type de l'attribut j.

2. Calculer la fonction de voisinage moyenne $g(xi) = \frac{1}{NN(xi)} \sum_{x \in NN(x_i)} f(x)$ et la fonction de comparaison $h(x_i) = f(x_i) - g(x_i)$
3. Calculer la moyenne de l'échantillon du répertoire $\mu_s = \frac{1}{|NN(x_i)|} \sum_{x \in NN(x_i)} h(x)$ et calculer le répertoire de variance-covariance $\sum_s = \frac{1}{|NN(x_i)|-1} \sum_{x \in NN(x_i)} [h(x) - \mu_s][h(x) - \mu_s]^T$
4. Calculer la distance de Mahalanobis de chaque fichier par le calcule de la moyenne de l'échantillon de répertoire utilisant la matrice de variance-covariance. Soit la distance $a(x_i) = a_i = (h(x_i) - \mu_s)^T \sum (h(x_i) - \mu_s)$
5. Identifier les fichiers outliers avec une distance de Mahalanobis en calculant les moyennes μ_m et l'écarts type σ_m pour les distances a_i dans le répertoire. Si $\frac{a_i - \mu_m}{\sigma_m} > \theta$ alors x_i est un outlier.

2.4 Autre recherche en datamining pour l'investigation légale

Les techniques d'exploration de données mentionnées ci-dessus ne sont que quelques unes des techniques appliquées dans l'investigation numérique. D'autres recherches des méthodes d'exploration de données comprennent l'utilisation de transformées en ondelettes comme un modèle statistiques en multi échelle pour analyser les données dans les bases de données de sécurité réseau [48] et afin d'analyser des images pour la détection sténographie [29].

Bien que notre travail focus sur l'investigation des crimes, le datamining à l'aide de ces techniques peut être utilisé pour modéliser les problèmes de détection des crimes.
Selon les travaux précédents étudiés, la catégorie d'analyse semble être la plus prometteuse parmi les différentes catégories de processus d'investigation. Cependant, la littérature n'a que des efforts très concentré et dispersés dans l'étape de la collecte de preuves que l'étape d'analyse (lien entre les preuves). La figure 2.15 illustre notre taxonomie des travaux de datamining pour l'investigation numérique légale.

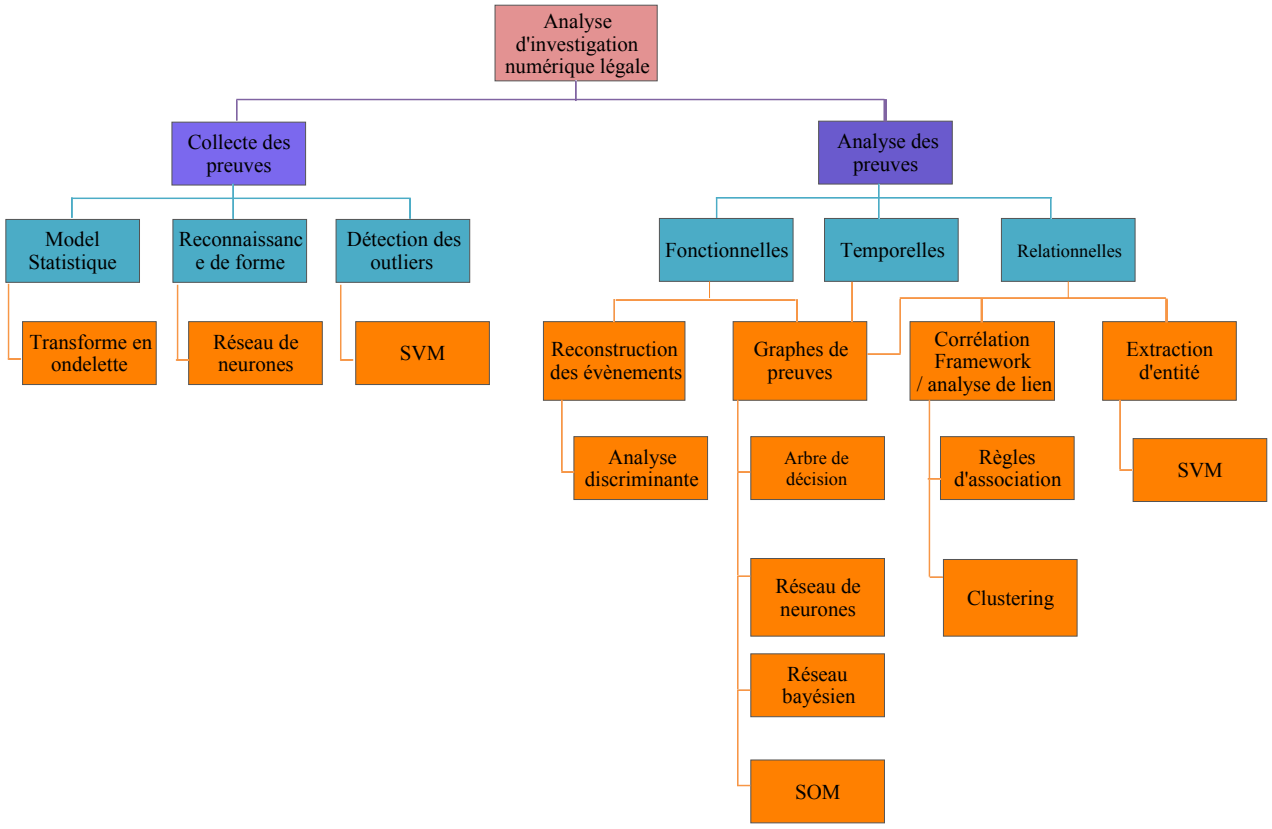


FIGURE 2.15 – Taxonomie des travaux de dataminig au stade d'analyse de processus d'investigation

2.5 Bilan et Discussion

Cette section vise à discuter quelques unes des limites des travaux de datamining en investigation numérique légale qui sont discutées au section précédente. La plus part des travaux de datamining pour l'investigation numérique légale ont été focalisés sur l'étape de collecte de preuves de la phase d'analyse de processus d'investigation par rapport à l'étape d'analyse de preuves (voir le chapitre 1). La classification à base de SVM a été utilisé dans les travaux de Vel et al, Malcom et al [24] [50] pour l'extraction d'entité d'un hauteur d'émail afin de répondre sur la question *qui ?* lors de l'analyse relationnelle, d'un autre coté les travaux de Brown et Pharm [10], Liu et al [12] ont été focalisés pour localiser les preuves de façon automatique dans des repertoires et fichiers cachés. Leurs systèmes permet la re-

cherche de toutes les évidences de manière précis, cependant le processus d'apprentissage peut être très coûteux en terme de temps du calcul. Le passage à l'échelle est donc délicat. Carney et Rogers [13] ont introduit l'analyse discriminante pour la reconstruction des événements afin de répondre sur le *comment ?* Lors d'une analyse fonctionnelle, cependant cette technique souffre d'un pourcentage de certitude. D'autres travaux utilisent l'arbre de décision par amélioration d'ID3 [12] [79] pour localiser des preuves dans des ensembles pertinentes, les algorithmes implémentés sont polynomiale, précis car le taux d'erreurs est faible, compréhensible(moins de branches et de niveaux dans l'arbre), même problème de scalabilité reste posé dans ces travaux qui sont théoriques et ne prennent pas en compte la difficulté de construire des arbres de décisions avec un nombre élevé des valeurs d'attributs, ne permet pas aussi de trouver toutes les évidences en cas des données manquantes ou des attributs de valeurs numériques. Les réseaux de neurones ont été utilisés pour localiser les preuves de façon dynamique [44] lors des attaques réseaux et aussi de construire un timeline afin d'ordonner chronologiquement les événements et répondre sur le *quand ?* lors d'une analyse temporelle. Dès lors que les données sont volumineuse, la recherche de ces preuves ou bien la construction d'une chronologie complètes pour fournir une vision plus large des séquences des événements liés à une activité malicieuse devient une tâche difficile, voir impossible à cause de processus d'apprentissage. D'autres travaux utilisent les réseaux bayésiens [45] pour automatiser les investigations numériques et démontrer la validité des hypothèses fixer d'avance en exploitant des preuves numérique, ils répondent sur la question *comment ?* lors d'une analyse fonctionnelle et *qui ?* lors d'une analyse relationnelle. Le modèle de réseau bayésien est non seulement un outil analytique pour évaluer les preuves, mais aussi un outil de suivi qui permet d'examiner et d'analyser les résultats d'investigation. Cependant, l'affectation des probabilités à priori aux preuves est liées à l'expérience de l'analyste, ce qui donne des résultats un peu critiques. Certains travaux utilisent le texmining pour le clustering des documents non structurés [25]. Ce modèle est efficace pour la recherche d'information sur les preuves à cause de fonction de mappage, exhaustive en termes de la collecte des preuves grâce à la phase de prétraitement. Cependant ce modèle est complexe et n'est pas sacalable en terme de temps du calcul et ne prennent pas des relations sémantiques entre documents qui sont essentielles pour une investigation numérique. Autre que le texmining, d'autres travaux utilisent le clustering [73] afin d'extraire des corrélations entre les preuves à partir de disque, bien que cette approche est

exhaustive, ce modèle n'est pas générique pour tous les dispositifs numériques. Il y a des travaux qui exploite les réseau de neurones (SOM) [30] pour analyser les données liées à l'Internet et extraire des corrélation significative entre les données de bas niveau à travers la visualisation des clusters des données. Cependant, son application immédiate intégrée à l'analyse légale n'est pas claire. D'autres recherche ont appliqué les règles d'association pour identifier des irrégularités comportentale associes aux profils des utilisateurs dans les fichiers journaux du système [1]. Le modèle compréhensible en terme des résultats obtenus n'est pas exhaustifs en terme de collecte des preuves car il prend en compte un seul fichier journal. Certaines travaux exploite la notion d'un outlier [18] afin d'automatiser le processus de collecte des preuves, même pour des données supprimées, leur motivation est de vérifier un rapport d'incident, de trouver des preuves d'un événement particulier, ou pour tester des hypothèses de l'incident. Bien que ce système permet de gagner beaucoup du temps lors de l'étape de collecte des preuves mais il ne permet pas l'analyse, qui est une étape primordiale pour une investigation numérique légale. Farid et liu [29] ont discuté le problème de scalabilité mais uniquement au niveau d'une seule base de données réseaux. En utilisant un modèle statistique multi échelle. Le tableau 2.1 illustre notre classification des travaux de la fouille de données en investigation numérique légale en tenant compte des critères suivants les deux étapes de la phase d'analyse de processus d'investigation(la collecte et l'analyse de preuves). Ces critères sont :

- ◇ La scalabilité de processus d'enquête légale numérique en gardant le temps moyen constant en terme d'investigations avec la croissance de la taille et la diversité des supports cibles.
- ◇ Exactitude des résultats : Mesuré par la précision et le taux du rappel
- ◇ Exhaustivité : La recherche de toutes les preuves liées au crime dans la phase de collecte des évidences
- ◇ Qualité des résultat : C'est la compréhension des conclusions en répondant sur les questions et les hypothèses d'investigation posés à l'avance.

Travaux	**Collecte de preuves**				**Analyse de preuves**		
	Passage à l'échelle (scalabilité)	Précision	Compréhension	Exhaustivité	Passage à l'échelle (scalabilité)	Précision	Compréhension
H. Farid et al (2003)	X						
X. Yan (2011)		X	X	X			
Z. Liu et al (2008)		X	X	X			
B.D. Carrier et E.H. Spafford (2005)		X	X	X			
H.B. Veena et al (2010)		X	X	X			
S. Decherchi et al (2009)				X		X	
B. K. L. Fei(2007)				X		X	X
M.N.A. Khan et al(2007)				X		X	X
M. Carney et M. Rogers (2004)						X	
M. Kwan et al (2008)						X	
O. Vel et al (2001)						X	X
C. Malcom et al (2002)						X	X
R. Brown et B. Pharm (2005)						X	X
Q. Iu (2010)						X	X
T. Abraham et O. Vel (2002)							X
T.Duval(2005)							X

TABLE 2.1 – Classification des travaux au stade d'analyse pour le processus d'investigation

2.6 Conclusion

Nous avons présenté dans ce chapitre un état de l'art sur les principales techniques de datamining définies dans le contexte d'investigation numérique légale. Nous avons constaté que la plus part des travaux sont focalisé sur l'étape de collecte de preuves de la phase d'analyse de processus d'investigation. Nous avons vu aussi que le datamining est une approche très efficace pour traiter le volume de données dans DF.

Très peu de travaux qui ont abordé la problématique de datamining à l'étape d'analyse de preuves, qui est mal traitée et se ramène toujours à l'étape de collecte des évidences . Dans le chapitre suivant nous allons présenter notre approche de datamining à l'étape de l'analyse de preuves qui sont collectées à partir des smartphones.

Chapitre 3

Une approche d'analyse de preuves pour l'investigation mobile légale

3.1 Introduction

Avec le progrès rapides de l'information et la technologie de communication dans le monde, l'accent est mis sur la sécurité sociale de la société, le crime est considérable quand il s'agit de l'utilisation des technologies de téléphones intelligent (smartphones). Aujourd'hui les smartphones devenu un outil indispensable et très populaire dans tous les aspects de nos vies quotidienne et professionnelle. A la suite de l'adoption généralisée de ces dispositifs. Ils sont devenus un choix cible et peuvent être impliqués dans presque n'importe quel crime (traditionnel ou numérique). Par exemples l'attaque terroriste de Mambie 2008 et Londres 2011, l'institut Australien de criminologie a découvert que les téléphones mobiles sont la forme la plus répandue de communication pour les personnes qui achètent le héroïne, le méthamphétamine et de cocaïne, un employé peut voler les informations sensibles de l'entreprise en les téléchargeant sur son mobile, un cas qui s'est tenu à New Zélande concernant un vol de bijouterie, les voleurs utilisent leurs téléphones portables pour organiser et cordonner le crime [53]. En plus de cela, le développement rapide du marché de la téléphonie mobile garanti que les téléphones mobile sont toujours de plus en plus puissant, et offre une augmentation de fonctionnalités. Par conséquent la capacité à utiliser un teléphone mobile à des fins malveillantes et de plus en plus apparente. Les smartphones peut contenir une grande quantité d'information relatives aux

actions de l'utilisateur peut inclure, l'historiques des appels, listes des contacts, l'historique de navigation(recherche Internet), comptes des réseaux sociaux, calendrier/notes, connexions(réseau mobile, Wi-Fi(Wireless Fidelity), bluetooth, mappe(location, favoris), messages textes/emails, médias(photos,vidéos,audios), software (logiciel de traitement de documents, logiciel de VoIP(Voix Internet Protocol), etc. Le but de l'investigation mobile (smartphones) légale est l'utilisation des méthodes d'acquisition pour récupérer ces données pertinentes liées au crime dans des conditions juridiquement valides. Il comprend une analyse à la fois de carte SIM et la carte mémoire ainsi que la mémoire interne de téléphone. Les informations collectées peuvent servir des preuves et qui sont d'intérêt primordial pour un investigateur.

Le processus d'extraction de connaissances à partir de ces preuves pose une difficulté majeure vue l'augmentation exponentielle des données issues des smartphones. Plus on possède de données, plus il est difficile de les traiter et d'en tirer des conclusions. Comme on a vu dans le chapitre précédent que l'utilisation de datamining est une discipline très approprié à la phase d'analyse de processus d'investigation. Donc, on peut l'adopter pour développer une approche d'analyse de preuves afin de reconstruire le scénario le plus approprié au crime.

Notre travail concerne particulièrement la problématique d'analyse de preuves dans la phase d'analyse de processus d'investigation légale. La phase d'analyse a motivé un nombre très important de recherches comme on l'a exposé dans le chapitre précédent. Cependant la plus part des travaux antérieurs se focalisent seulement sur la collecte de preuves pour répondre à la question *où ?* et cela sans prendre en compte l'étape d'analyse de preuves.

Notre objectif est donc de montrer comment les preuves numériques collectés peuvent être utilisées pour reconstruire un crime ou un incident, d'identifier les suspects, d'appréhender le coupable, de défendre les innocents, et de comprendre les motivations criminelles en développant une approche scalable de reconstruction qui se réfère au processus systématique d'assemblage des éléments de preuve pour construire des chaînes des évidences à partir des informations recueillies lors de l'étape de collecte et répondre sur les questions *comment ?, quand ?, qui ?, quoi ?*, et dans quel ordre dans un contexte temporel, fonctionnel, et relationnel. Dans ce qui suit nous allons présenter en détails ces différents aspects de notre approche.

3.2 Système d'Analyse des Preuves pour l'Investigation des Smartphones(SAPIS)

La figure 3.1 introduit les différents composants de l'architecture proposée de notre système d'investigation.

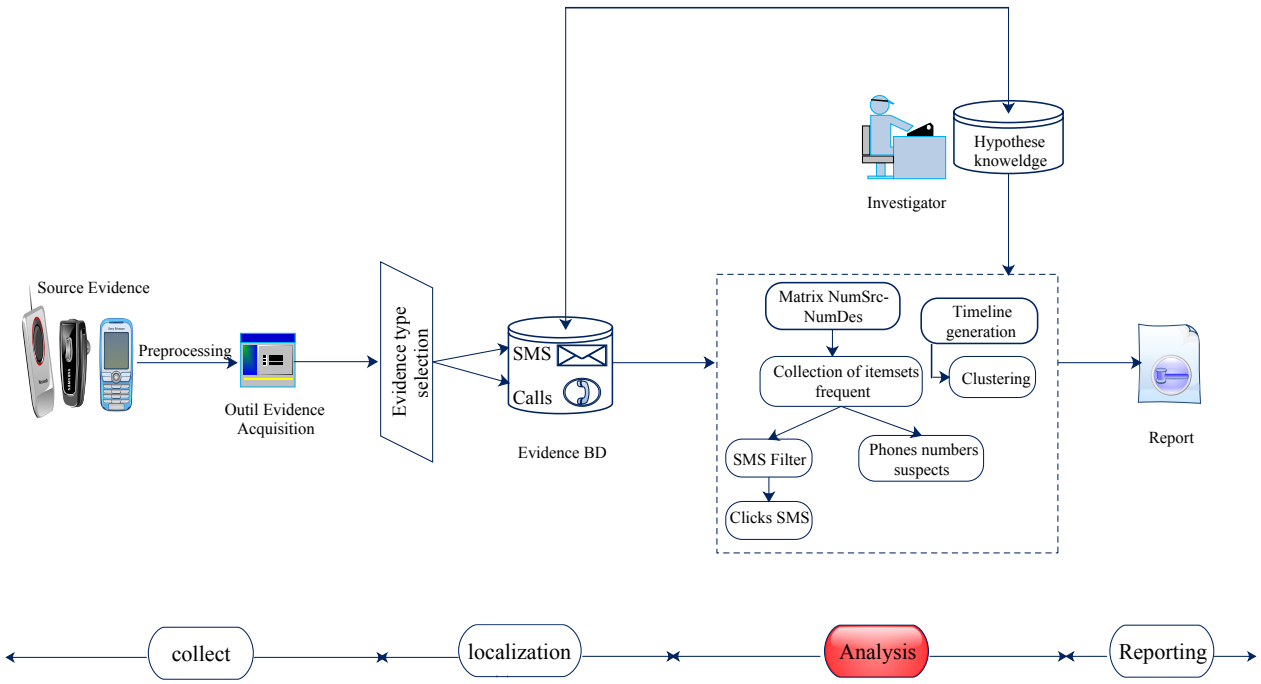


FIGURE 3.1 – Système d'Analyse des Preuves pour l'Investigation des Smartphones(SAPIS)

Chaque composant exécute une des trois phases nécessaires, à savoir, la collecte(avec localisation), l'analyse et la présentation tels quels sont discutés au chapitre 1. Une fois la source de preuves a été acquis, un fichier contenant les données collectées crée par un outil d'investigation légale, il subi un traitement en filtrant les différents appels et SMS issues des smartphones d'acquisition. La preuve peut être recueillis auprès des sources, tels que les mobiles, les smartphones ou les PDAs (Personal Digital Assistant). Néanmoins, un prétraitement qui concerne le nettoyage, la transformation et la réduction de données sera

réalisé. Une fois le fichier de données à la sortie de l'outil d'acquisition soumis à un filtrage des SMS et des appels téléphonique qui seront ensuite charger dans la base des évidences. Après la phase de localisation et de recherche de preuves, la prochaine étape est l'analyse de preuves, elle est le cœur de notre travail. Le but de la phase d'analyse de preuves est d'identifier des éventuels corrélations, des associations et des anomalies dans les données légales en répondant sur les questions *comment ?, quand ?* à l'aide d'un algorithme de clustering et les questions *quoi ? et qui ?* à l'aide d'un algorithme de règles d'associations en se basant sur des hypothèses fixer à l'avance par l'analyste. Après l'analyse de preuves, un rapport est générer qui contient des conclusions sur la phase précédente. L'architecture proposée est constitué de cinq composant majeurs qui sont détaillés aux sections suivantes.

3.2.1 Outil d'acquisition de preuves(Tool of Evidences Acquisition)

Nous exploitons l'outil d'acquisition des évidences de Kechadi et al [4] afin d'extraire les données utiles pour l'investigation, y compris, les messages, les appels téléphoniques, les vidéos, les images, les emails, localisation GPS, etc, en plus les données supprimées (hidden data). Le choix est basé sur ce framwork car il permet une acquisition standard quel que soit le type de smartphones utilisés. les résultats d'extraction sont illustrés par la figure 3.2

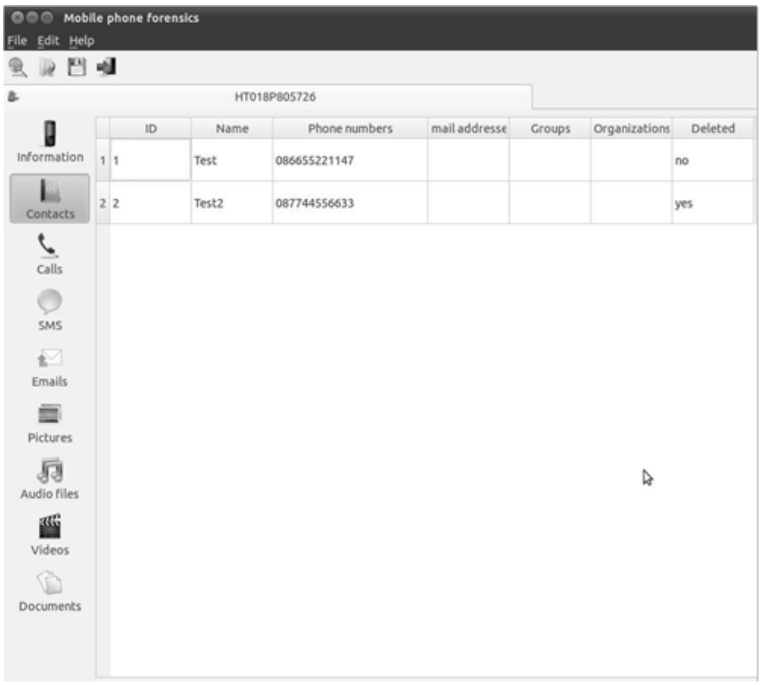


FIGURE 3.2 – Résultats d'extraction [4]

3.2.2 Sélection de type des évidences(Selection of evidence type)

La sélection de type de preuves est une étape qui permet de filtrer et de transformer les données forensics(evidence data). Nous nous seulement intéressons par les messages et les appels téléphoniques.

3.2.3 Base de preuves(Basis of proofs)

Pour pallier les insuffisance de l'organisation des données forensics dans un simple fichier plat utilisé par la plupart des systèmes d'investigation, nous mettons en place une configuration de base de données pour stocker les événements en terme des SMS et des appels et nous formulons notre conception de la base de données selon la norme UML (Unified Modeling Language) standard, le diagramme de classe est illustré par la figure 3.3.

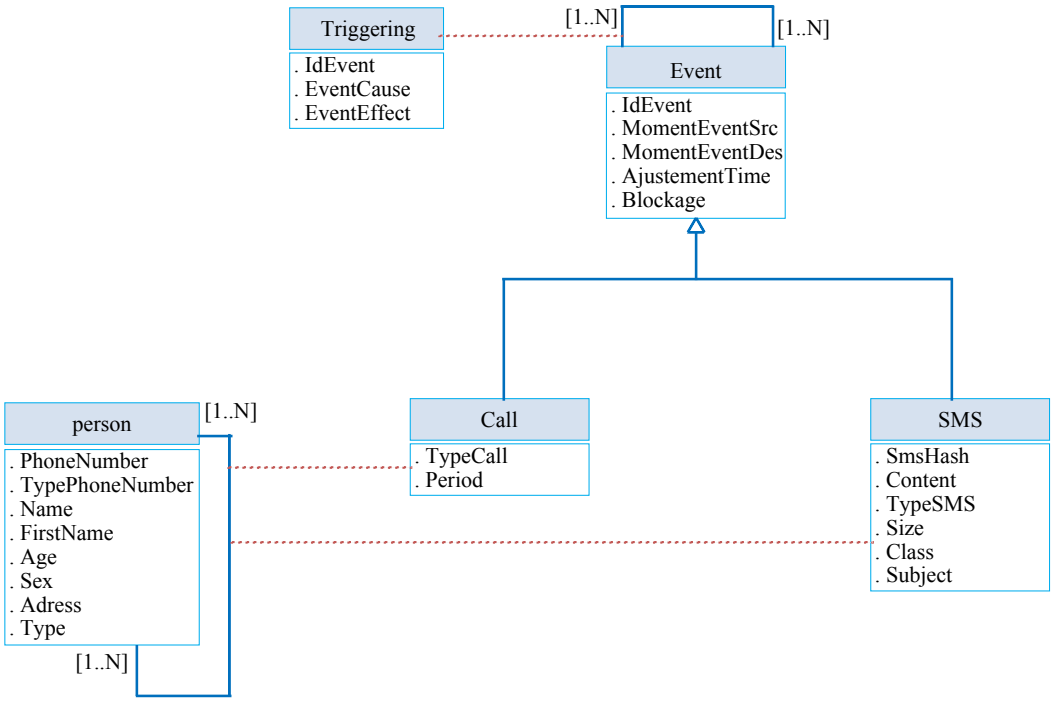


FIGURE 3.3 – Diagramme de classe de la base de preuves

Le schéma de stockage de la figure 3.3 se compose de quatre tables principales provenant des classes suivantes :

- **Classe Event :** C'est la classe mère des sous classes SMS(Short Message Service) et appels, elle permet de stocker les événements en terme générique. Elle contient les attributs suivants :
  - ◇ *IdEvent :* C'est l'identificateur de l'événement.
  - ◇ *MomentEventSrc :* C'est la date de création de l'événement(date d'envoi d'un SMS ou la composition d'un appel), de la forme JJ-MM-AAAA et HH-MM-SS.
  - ◇ *MomentEventDes* : C'est la date de réception d'un SMS ou d'un appel chez le destinataire, de la forme JJ-MM-AAAA et HH-MM-SS.
  - ◇ *AdjustmentTime* : C'est la déviation du temps entre la date d'envoi et de réception des différents événements.
  - ◇ *Blockage* : Il montre la situation des numéros de téléphones sources avec le numéro

de téléphone destinataire (cette propriété se trouve au niveau des smartphones).

- **Classe Triggering :** C'est le rôle joué par un événement dans une communication. Elle contient les attributs suivants :
  - ◇ *EventCause* : C'est le numéro de téléphone qui provoque l'événement en cours.
  - ◇ *EventEffect* : C'est le numéro de téléphone provoquer par l'événement en cours.
- **Classe Call :** C'est la première sous classe de la classe Event, c'est une collection d'informations concernant un appel téléphonique. Elle contient les attributs suivants :
  - ◇ *TypeCall* : Qui désigne soit un appel reçu ou émis(composé).
  - ◇ *Period* : C'est la durée d'un appel émis ou reçu.
- **Classe SMS :** C'est la deuxième sous classe de la classe Event, c'est une collection d'informations concernant des SMS. Elle contient les attributs suivants :
  - ◇ *SmsHash* : Pour crypter le contenu du message à l'aide de MD5 en assurant son authentification.
  - ◇ *Content* : C'est le corp de SMS et de l'attachement.
  - ◇ *TypeSMS* : Qui désigne soit un SMS reçu, envoyé ou répliqué.
  - ◇ *Size* : La taille de message et son attachement s'il existe.
  - ◇ *Class* : C'est la catégorie SMS ou MMS(Message Multimedia Service).
  - ◇ *Subject* : C'est le contexte de SMS.
- **Classe person :** C'est une collection d'informations sur l'émetteur et le récepteur pour les SMS et les appels. Elle contient les attributs suivants :
  - ◇ *PhoneNumber* : C'est le numéro de téléphone du propriétaire de smartphone.
  - ◇ *TypePhoneNumber* : C'est le type de puce, par exemple, en algérie on trouve Nedjema, Mobilis, Djezzy.
  - ◇ *Name, FirstName, Age, Sex, Adress.*
  - ◇ *Type* : Désigne le rôle de la personne qui peut être criminel où victime.

  .

3.2.4 L'analyse de preuves(proofs Analysis)

C'est le bloc le plus important dans notre travail, contrairement à la plus part des travaux d'investigation qui ont basé sur l'étape phase de collecte de processus d'investigation, nous proposons un algorithme de clustering scalable dans le contexte d'analyse temporelle et fonctionnelle et qui va répondre sur les questions *quand ? et comment ?* en

générant un time line d'événements et nous exploitons la recherche des motifs fréquents dans le contexte d'analyse relationnelle pour répandre sur le *qui ? et le quoi ?*.

3.2.5 La base des hypothèses(Hypotheses Knowledges)

Notre algorithme de clustering proposé dans la phase d'analyse est basé sur la notion du temps(timestamp). Le timstamp stocké peut ne pas refléter avec précision le temps de création d'événement(SMS ou bien appel) provoquée par la résolution du format de timestamp et l'horloge qui va le générer. Cette imprécision peut parfois être assez grande pour donner la valeur de timestamp attaché à un événement. Les horloges ne sont pas totalement fiables. Elles peuvent dériver, générer ainsi que des timestamps progressivement plus différents de ceux générés par d'autres horloges. Sur la plupart des systèmes peut être réglée à tout moment par l'utilisateur du système pour afficher une date et une heure différente que le temps en cours. Ce réglage peut être fait intentionnellement par le suspect pour des raisons liées aux crimes. Ces incertitudes sont inquiétantes pour les enquêteurs, c'est pour ça, dans ce travail nous avons utilisé le formalisme de Willassen [77] pour développer les hypothèses sur les horloges des timestamps et tester leurs cohérence.

Hypothèses basées sur l'investigation de timestamp

Définition 3.2.1. *Soit V le domaine de valeurs du temps produites par une horloge. Une fonction d'horloge est défini par : $c(t) : T \rightarrow V$. La définition d'une fonction d'horloge n'impose pas de restrictions sur les valeurs d'horloge en fonction du temps [77]*

Timestamps d'événements

Un événement daté est un événement pour lequel il existe une valeur de timestamp dans le domaine V de valeurs du temps. Un timestamp est créé quand un événement effectue une copie de la valeur fournie par une horloge. Les timestamps associé à un ensemble des événements sont pas nécessairement liés à la même horloge.

Définition 3.2.2. *Soit E un ensemble d'événements datés et V le domaine de valeurs du temps la fonction $\tau_c(e_i) : E \longmapsto V$ est définie de telle sorte que $\tau_c(e_i) = c(t(e_i))$, où $\tau_c(e_i)$ est le timestamp associé à l'événement e_i par rapport à l'horloge c. [77]*

Hypothesis 1. *Les timestamps associés à un ensemble des événements (messages, appels) en terme de date d'envoi(MomentEventSrc) et date de réception(MomentEventDes) sont nécessairement liées à la même horloge c* $MomentEventSrc_c(e_i) = c(MomentEventSrc(e_i))$ $\forall i \in [1..n], e_i = \{SMS, appels\}$ $MomentEventDes_c(e_i) = c(MomentEventDes(e_i))$

◇ Horloges idéal et non idéal

Une horloge idéale est celle qui ne peut aller de l'avant . Une horloge non idéal est une horloge qui n'est pas idéal.

Définition 3.2.3. *Soit I l'ensemble des horloges idéales. Une horloge idéale* $c(t) \in I$ *est une horloge qui satisfait les propriétés suivantes [77] :*

$$\forall i \forall j((t(e_i) < t(e_j) \Rightarrow c(t(e_i)) \leq c(t(e_j))) \forall i \forall j(t(e_i) = t(e_j) \Rightarrow c(t(e_i)) = c(t(e_j))) \quad (3.1)$$

Hypothesis 2. *L'horloge associé à l'ensemble des événements de la base de preuves est idéal c'est à dire deux événements liés causalement par la même horloge idéale ont des timestamps de telle sorte que le timestamp du dernier événement n'est jamais inférieur au timestamp de premier événement.*

$\forall i \forall j(MomentEventSrc(e_i) < MomentEventSrc(e_j) \Rightarrow c(MomentEventSrc(e_i)) \leq c(MomentEventSrc(e_j)))$

$\forall i \forall j(MomentEventSrc(e_i) = MomentEventSrc(e_j) \Rightarrow c(MomentEventSrc(e_i)) = c(MomentEventSrc(e_j)))$

$\forall i \forall j(MomentEventDes(e_i) < MomentEventDes(e_j) \Rightarrow c(MomentEventDes(e_i)) \leq c(MomentEventDes(e_j)))$

$\forall i \forall j(MomentEventDes(e_i) = MomentEventDes(e_j) \Rightarrow c(MomentEventDes(e_i)) = c(MomentEventDes(e_j)))$

tel que l'horloge c est idéal à savoir :

$e_i \longrightarrow e_j \Rightarrow \tau_c(e_i) \leq \tau_c(e_j)$

◇ L'ensemble d'événements observés et leurs précisions

Au cours d'une investigation légale d'un smartphone, l'investigateur peut observer un certain nombre d'événements datés qui sont basées sur la même horloge. Certains de ces événements ont un lien de causalité. L'ensemble d'événements observés datés est appelé le "jeu d'observation"

Définition 3.2.4. *Un ensemble d'observation O est un ensemble d'événements datés qui sont reliées à une même d'horloge $c_o(t)$. Un jeu d'observation a généralement un grand nombre d'événements datés avec un grand nombre de relations causales. Les données de l'ensemble d'observations sont utilisées pour déterminer si une hypothèse d'horloge est titulaire ou non [77]*

Définition 3.2.5. Une hypothèse d'horloge $c_h(t)$ pour un ensemble d'observation O est Correcte si $c_o(t) = c_h(t)$ pour tout t, à savoir :

$$c_o(t) = c_h(t) \Rightarrow \forall e_i(\tau_{co} e_i = c_h(te_i)).[77] \tag{3.2}$$

Hypothesis 3. *L'hypothèse d'horloge choisie dans un ensemble des événements O est correcte, alors toutes les occurrences de timestamps doivent correspondre aux valeurs prédites par l'hypothèse. La précision d'un timestamp reflète à quel point dans le temps le timestamp représente le temps réel lorsque l'événement s'est produit.*
$c_o(MomentEventSrc) = c_h(MomentEventSrc) \Rightarrow \forall e_i(\tau e_i = c_h(MomentEventSrc(e_i))$
$c_o(MomentEventDes) = c_h(MomentEventDes) \Rightarrow \forall e_i(\tau e_i = c_h(MomentEventDes(e_i))$
et $e_i \in \{SMS, Call\}$

◇ **La cohérence d'hypothèses d'horloges**

Les deux théorèmes précédents peuvent infirmer une hypothèse d'horloge, mais ils ne peuvent pas prouver qu'elle est correcte. Cela conduit à la définition suivante d'un horloge d'hypothèse cohérente .

Définition 3.2.6. *Étant donné un ensemble de testes Z, une hypothèse d'horloge est conforme sous Z avec un ensemble d'observations O si aucun teste $z \in Z$ montre que l'hypothèse d'horloge est incorrecte pour O. l'hypothèse d'horloge est inconsistante sous Z avec un ensemble d'observation O si elle n'est pas consistente sous Z avec O [77]*

Hypothesis 4. *L'hypothèse d'horloge interdit certaines choses de se produire pour chercher à détecter les incohérences :*
$\exists e_i \exists e_j (e_i \rightarrow e_j) \wedge (\tau_{co} e_i - d_h(MomentEventSrce_i) > \tau_{co} e_j - d_h(MomentEventSrce_j)) \Rightarrow$
$co(MomentEventSrc) \neq c_h(MomentEventSrc)$
$\exists e_i \exists e_j (e_i \rightarrow e_j) \wedge (\tau_{co} e_i - d_h(MomentEventDese_i) >$
$\tau_{co} e_j - d_h(MomentEventDese_j)) \Rightarrow co(MomentEventDes) \neq c_h(MomentEventDes)$
et $(e_i, e_j) \in \{SMS, Call\}$

3.2.6 Rapport

La présentation est la dernière phase, tous les résultats de la phase d'analyse de preuves sont affichés sous forme d'un rapport PDF(Portable Document Format) destiner au tribunal.

3.3 La démarche d'analyse de preuves

3.3.1 Génération de timeline pour l'analyse temporelle et fonctionnelle

La reconstruction des événements d'un incident à partir des preuves collectées joue un rôle primordiale pour la résolution de crime, c'est pour ça nous exploitons la technique de clustering afin de générer un timeline, c'est à dire un ensembles des événements ordonnées dans le temps. Nous utilisons les concepts suivant pour le développement de notre algorithme de classification non supervisée [15] :

- **Objet digital**
  Un objet numérique est une collection discrète des données, comme un SMS dans un téléphones mobile, appel téléphonique, etc. Chaque objet à des propriétés ou caractéristiques observés dans le temps. L'état de l'objet est définit selon les valeurs de ses attributs.
- **Événement digitale**
  Un événement numérique est un événement qui change l'état d'un ou plusieurs objets numériques.
- **Objet preuve**
  Un objet est la preuve d'un événement si l'événement a changé l'état de l'objet. Il existe deux types d'objets preuves :
  ◇ **Cause** : Un objet joue le rôle d'une cause si ses caractéristiques ont été utilisées dans l'événement. Un test pour ce rôle consiste à identifier si le même effet aura lieu si l'objet n'existe pas.
  ◇ **Effet** : Un objet joue le rôle d'un effet si son état était changé par un autre objet à cause de l'événement.
  Les objets causes peuvent être passifs. Autrement dit, ils sont exploités par

l'événement, mais ils ne sont pas modifiés par l'événement. Si un objet cause est modifié par l'événement, il est à la fois cause et effet et il est considéré comme actif [16]. Dans notre travail nous mettons l'accent seulement sur des objets ou événements actives

– **Chaînes d'événements**
Une chaîne d'événement est une séquence d'événements $(e_0, e_1, e_2, ..., e_k)$ tel que l'effet de l'événement e_i est une cause d'événement e_{i+1} pour $i = 1, 2, ..., k-1$. Une chaînes d'événements peut être représentée par un graphe acyclique (Direct Acyclic Graph(DAG)). Un exemple d'une telle représentation est donné dans la figure 3.4. les événements sont représentés par des carrés. Les relations de précédences entre les événements sont représentées par des arcs ou flèches.

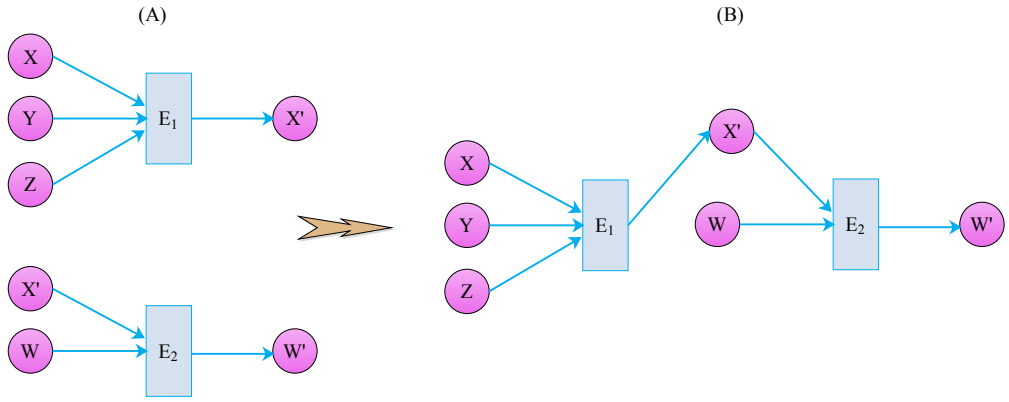


FIGURE 3.4 – Représentation graphique des événements individuelles(A),chaînes d'événements après séquencement(B) [68]

L'algorithme de classification non supervisé des événements

Si un objet preuve contient des informations temporelles fiable et la valeur réel de temps d'un événement est connue, alors il peut être facilement séquencé par rapport aux autres événements dont le temps en se basant sur les deux notion de causalité et du temps (timestamp), nous développons un algorithme de clustering hiérarchique qui permet de générer un time line (un ensemble des clusters d'événements ordonnées dans le temps) .

◇ **Causalité : -(la relation entre la cause et l'effet)-** peut être formellement

exprimée comme une relation mathématique entre les événements. Lamport [47] était le premier à utiliser la relation de précédence ($\rightarrow$) pour commander des événements concernant le processus d'exécution et le passage de messages. Nous avons captons la notion de causalité présentée dans [77]

◇ **Temps** : Nous supposons que chaque événement a un moment dans le temps qui lui est associé et ces moments dans le temps peuvent être commandés en utilisant les relations $<$ et $=$.

Le processus de clustering est formaliser comme un processus de recherche de graphe ordonné des événements. L'algorithme repose sur les hypothèses suivants :

– Chaque événement à plusieurs causes et plusieurs effets
– Nous supposons que la causalité est préservée dans le temps, $MomentEventSrc_c(e_i) < MomentEventSrc_c(e_j)$ et $MomentEventDes_c(e_i) < MomentEventDes_c(e_j)$ n'implique pas que $e_i \rightarrow e_j$, puisque les événements peut se produire à différents moments dans le temps, sans être lié par $\rightarrow$. Cette contraintes n'est pas prise en compte dans notre proposition.

La stratégie de clustering

Nous proposons une stratégie d'ordonnancement des événements exploitant la classification non supervisée. Le problème qui se pose alors est de regrouper les événements avec leurs causes et effets afin d'obtenir des clusters des événements ordonnés dans le temps. Ces clusters sont représentés comme des graphes orientés acyclique, c'est pourquoi, nous utilisons une méthode de classification hiérarchique ascendante. En plus, nous ne connaissons pas a priori le nombre de classes à obtenir car ce nombre dépend de la base de preuves.

– **Contexte de classification**
Un événement sur un smartphone est vu comme une communication faite d'un coté entre l'expéditeur et un seul destinataire (cas d'un appel) et d'un autre coté celui de l'expéditeur et un ou plusieurs destinataires(cas d'un SMS). Alors un événement e_i est vu comme une ligne d'une matrice de similarité composée de cellules qui correspondent aux attributs représentatifs *PhoneNumberSrc* et *PhoneNumberDes* sachant que le *PhoneNumberSrc* est la cause de e_i et le *PhoneNumberDes* est l'effet du même événement. Le terme général e_{ij} de cette matrice est représenté par des valeurs

naturelles attachées aux numéros de chaque événement (numéro de téléphone source et de destination). L'algorithme suivant permet le remplissage de la matrice de similarité selon le principe de causalité définis précédemment.

Algorithme 3 Calcule de la matrice de similarité

Require: Base of proofs ;
Ensure: Matrix of Similarity ;

1: $j = 1$; ▷ Initial value that is to say for the event ;
2: $T(1,1) = j$; $j = j + 1$; $T(1,2) = j$;$j = j + 1$;
3: **If**($MomentEventSrc(e_i) = MomentEventSrc(e_k)$)
And(PhoneNumberSrc(e_i)=PhoneNumberSrc(e_k)) **then**
4: T(i,1)= T(k,1) ; **such as :** k==1. . (i-1)
5: **Goto 13**
6: **Else**
7: **If**((MomentEventSrc(e_i)=MomentEventDes(e_k)+AdjustmentTime)
and(PhoneNumberSrc(e_i) = PhoneNumberDes(e_k)))**Then**
8: T(i,1)= T(k,2) ; **such as :** (k==1. . (i-1))
9: **Goto 13**
10: **Else**
11: T (i,1)= j ;
12: j= j+1 ;
13: **If**(PhoneNumberDes(e_i)=PhoneNumberDes(e_j))**then**
14: T(i,2)= T(k,2) ; **such as :** (k==1. . (i-1)) ; i=i+1 ;
15: **Goto 21** ;
16: **Else**
17: T(i,2)=j ;j=j+1 ; i=i+1 ;**Goto 21** ;
18: **EndIf**
19: **EndIf**
20: **EndIf**
21: **If** ($i \leq$numberEvent)**THEN**
22: **Goto 3** ;
23: **EndIf** ;
24: **End** ;

Le principe de remplissage de la matrice de similarité à partir de contexte d'extraction et l'application de l'algorithme 3 est représenté au tableau 3.1

	Cause	Effect
e1	1	2
e2	1	3
e3	1	4
e4	5	4
e8	4	6
e5	2	7
e6	3	7
e7	3	8
e11	6	9
e10	8	9
e9	7	9

TABLE 3.1 – Matrice de similarité

– **Mesure de similarité**
Soit EN une matrice Événements-Numéros tel que les lignes représentent les événements E= $\{e_i$, i = 1..n$\}$ et les numéros N = PhoneNumberSrc(Cause), PhoneNumberDes(Effect) sont des colonnes. Nous décrivons un événement e_i par un vecteur de deux valeurs naturel : $e_i = [e_{i1}, e_{i2}]$. Ce modèle de description permet de comparer deux événements. Par exemple, nous pouvons considérer les événements e_1 et e_2 comme fortement similaires si leur vecteurs $[e_{11}, e_{12}]$ et $[e_{21}, e_{22}]$ présentent une égalité de valeurs entre le PhoneNumberDes(Effect) de (e_1) et PhoneNumberSrc(Cause)(e_2).

– **Similarité et dissimilarité entre événements**
Nous définissons la relation de similarité et la dissimilarité entre événements par les deux fonctions $\sigma_{Sim(e_i,e_j)}$ et $\sigma_{Dissim(e_i,e_j)}$, qui mesurent respectivement la similarité et la dissimilarité entre les deux événements e_i et e_j suivant les deux attributs Cause et Effect respectivement :

$$\sigma_{Sim(e_i,e_j)} = \begin{cases} 1 & si \quad e_{i2} = e_{j1}, j = i+1, ...n, i = 1...n \\ 0 & sinon \end{cases} \tag{3.3}$$

$$\sigma_{Dissim(e_i,e_j)} = \begin{cases} 0 & si \quad e_{i2} = e_{j1}, j = i+1, ...n, i = 1...n \\ 1 & sinon \quad (e_{i2} \neq e_{j1}) \end{cases} \tag{3.4}$$

La première fonction définit la similarité entre e_i et e_j suivant les deux attributs PhoneNumberSrc(Cause), PhoneNumberDes(Effect), les deux événements e_i, e_j sont considérées similaires s'il y a un numero de téléphone commun entre ces deux événements (Effect=cause dans cet ordre). La deuxième fonction définit la dissimilarité entre deux événements e_i et e_j, les deux événements e_i et e_j sont considérées dissimilaires s' il n'y a pas du numéro de téléphone commun entre les deux événement, sachant qu'il y a un séquencement temporel entre e_i , e_j ($e_i \rightarrow e_j$).

– **Similarité et dissimilarité entre un ensemble d'événements**
Comme nous l'avons défini pour deux événements, nous introduisons deux fonctions qui prennent en compte le degré de similarité et dissimilarité entre deux ensembles des événements.
Soient E_a et E_b deux sous-ensembles de l'ensemble des événements E, rendre compte du niveau de similarité (respectivement, de dissimilarité) entre des ensembles des événements, nous utilisons la fonction Sim(E_a, E_b)(respectivement, Dissim(E_a, E_b) qui détermine le nombre de similarités (respectivement, de dissimilarités) entre les deux ensembles des événements E_a et E_b($E_a \neq E_b$)comme suit :

$$Sim(E_a, E_b) = \begin{cases} 1 \quad si & \sum_{e_i \in E_a, e_j \in E_b} sim(e_i, e_j) = \\ & (card((E_a \bigcup E_b) - 1) \times card((E_a \bigcup E_b)/2)) \\ 0 \quad sinon & \end{cases} \tag{3.5}$$

$$Dissim(E_a, E_b) = \begin{cases} 1 \quad si & \sum_{e_i \in E_a, e_j \in E_b} Dissim(e_i, e_j) \neq \\ & (card((E_a \bigcup E_b) - 1) \times card((E_a \bigcup E_b)/2)) \\ 0 \quad sinon & \end{cases} \tag{3.6}$$

L'algorithme de classification non supervisé que nous avons proposé est basé sur le principe de l'algorithme Chameleon (Hierarchical Clustering Algorithm Using Dynamic Modeling) proposé par Karypis et al [41]. Chameleon est un algorithme hiérarchique de regroupement agglomératif ascendant basé sur un modèle dynamique. Nous choisissons cet algorithme car il permet non seulement de classer les événements d'une manière qui nous intéresse en intégrant facilement notre mesure de similarité pour construire

dynamiquement la plus grand séquence des clusters ordonnés dans le temps. Mais aussi d'intégrer la notion de graphe d'événements obtenue dans le processus de reconstruction au processus de classification lors de l'étape de fusion des sous graphes(chaque cluster est un sous graphe).

Notation 1. ***NotExiste(cluster(i,j))*** *: C'est pour verifier l' existence de cluster(i,j).*

Notation 2. ***EN*** *: La matrice de similarité evenement-Numéros de téléphones.*

Notation 3. ***EventFirstCluster (j,t)*** *: Méthode qui retourne le premier événement du cluster j de taille t.*

Notation 4. ***EventLastCluster (i,t)*** *: Méthode qui retourne le dernier événement du cluster i de taille t.*

Algorithme 4 Classification non supervisé des évènements

Require: Matrix of Similarity **EN** ;
Number of cluster **K** ;
Ensure: A set of representative clusters ordered (time line) ;

```
1: Place each record (event) in its own cluster ;
2: k=0 ;
3: For (i=1...N) Do
4:    GenererCluster() ;
5:    K=k+1 ;
6:    Cluster(k)=e_i ;
7: EndFor
8: ClusterStage_i = 1      ▷ To join the clusters ;
9: i = ClusterStage_i ;
10: Cluster(i) ;      ▷ Method of clustering ;
11: while(i ≤ N) Do
12:     ok=true ; ok1=true ; j=i+1 ;

13:    While(ok) Do
14:       If(EN[EventLastCluster(i,t),2]==EN[EventFirstCluster(j,t),1])
             And( NotExiste(cluster(i,j))) Then
15:          FusionnerCluster(i,j) ; ok1= false ; k= k+1 ;
16:          Cluster(j) ; j++ ;
17:       Else
18:          If(ok1= true) Then j++ ;
19:          Else Cluster(j) ; j++ ;
20:            If(j > N) Then
21:              ok= false ; i++ ;
22:             EndIf ;
23:          EndIf ;
24:       EndIf ;
25:    EndWhile ;
26: EndWhile ;

27: ClusterStage_i = N + 1 ;
28: IF( dedans(intervalAnalysis)) Then
29:       N= N+K ;
30:       Goto(9) ;
31: EndIf ;
32: End.
```

Notons que, la classification par notre algorithme a des propriétés intéressantes :

- ◇ La complexité du calcul est relativement basse car, elle est polynômial O(N^2) suivant le nombre des événements.
- ◇ Les résultats du calcul sont précis car, chaque événement est attaché à sa classe adéquate (le taux d'erreurs est faible).
- ◇ L'algorithme est performant (scalable) pour un nombre élevé des événements puisque il est parallèle. Il permet de construire des clusters de même événement à la fois, il permet aussi de classifier d'autre événement causalement dépendant à cet même événement ou bien plusieurs événements indépendants de celui-ci, en plus il permet de mémoriser les clusters fusionner à chaque étape pour faire le balayage que sur les classes qui sont similaire à l'étape de fusion. Nous illustrons par la figure 3.5, le résultat de la classification non supervisée sur le contexte de classification précédent.

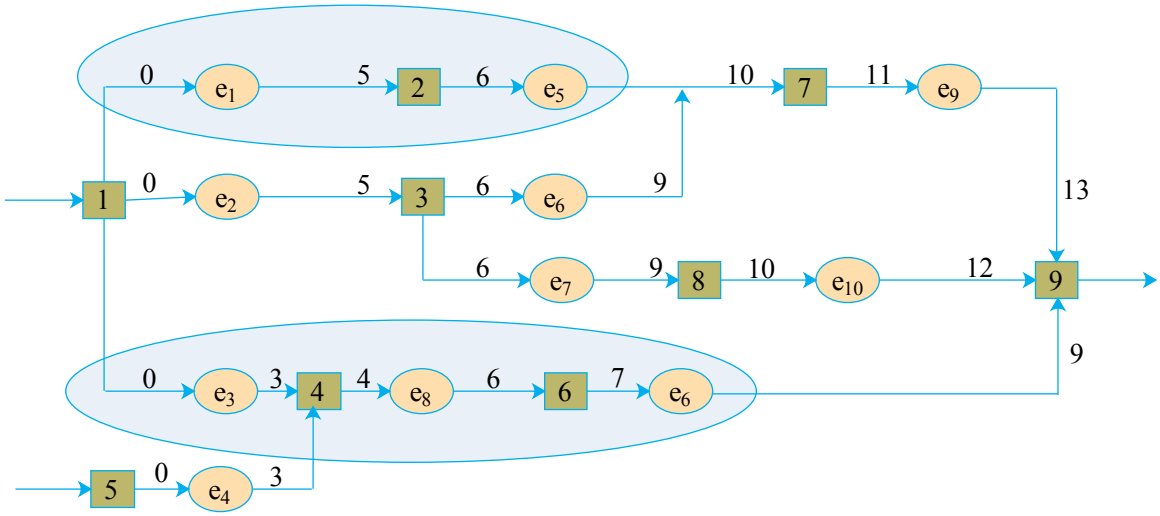


FIGURE 3.5 – Graphe de classification des événements

3.3.2 Recherche des motifs fréquents pour l'analyse relationnelle

Rappelons que nous voulons de concevoir un système d'investigation qui exploite les données stockées dans une base de preuves afin d'en extraire des connaissances utiles au résolution d'un crime numérique en utilisant les smartphones. Nous pensons que la

pertinence d'un numéro de téléphone est fortement corrélée avec la fréquence de son utilisation dans l'ensemble des communications téléphonique et le blocage de ce numéro à chaque fois chez le déstinataire. A priori, la recherche des motifs fréquents est un moyen qui nous semble approprié pour rendre compte de cet corrélation.
L'approche que nous proposons, dont le principe est représenté à la figure 3.1 exploite une vue matérialisé crée à partir de la base de preuves, une instance de cette vue est illustrée par le tableau 3.2.

IdEvent	Numéro Téléphone Source	Numéro Téléphone Destination	Technique de blocage	TypeEvent
1	1	2	false	Appel
2	1	2	false	SMS
3	2	3	true	Appel
4	2	3	true	SMS
5	3	2	true	Mixte
6	3	2	true	SMS
7	2	4	false	Appel
8	2	4	false	SMS
9	4	5	false	Mixte

TABLE 3.2 – Extrait de la base de preuves

Afin d'analyser les données extraites de la base de preuves, nous utilisons deux notions importantes, à savoir :

- **La fréquence** : C'est le nombre d'événements ou communication (SMS, appels) font entre deux numéros de téléphones
- **La technique de blocage** : C'est une technique qui permet de bloquer le numéros de téléphones chez le destinataire. Cette technique est capable d'identifier les irrégularités dans les smartphones, elle permet aussi d'analyser le comportement de propriétaire du téléphone.

La fréquence est un champ calculable à partir des événements de la base de preuves selon les même numéros de téléphones sources et les numéros de téléphones de destination, tandis que le filtrage est un attribut booléen qui permet de tester si l'événement(SMS, Appel) à été bloqué chez le destinataire. Une transformation de vue crée auparavant est illustrée dans le tableau 3.3

Numéro telephone source	Numéro telephone destination	Fréquence	TypeEvent
1	2	5	SMS
2	3	6	Appel
3	2	5	Appel
2	4	7	Mixte
4	5	1	Appel

TABLE 3.3 – Extrait de la base de preuves après transformation

Construction de contexte d'extraction

Pour construire la table de transaction, nous sélectionnons les enregistrements de la vue selon la requête SQL (Structured Query Language) suivante :
Select PhoneNumberSrc, PhoneNumberDes
From vue
Where frequence ≥ 3
and filtrage ='OK'
La table de transaction construite est une matrice (N*N) tel que les linges sont les numéros de téléphones sources et les colonnes sont les numéros de téléphones de destination, le terme général NN_{ij} est défini comme suit :

$$NN_{ij} = \begin{cases} 1 \quad si \quad frequence >= \theta \wedge filtrage = ok \\ \\ 0 \quad sinon \end{cases} \qquad \theta \in [1..n] \tag{3.7}$$

tel que θ est un seuil définit par l'analyste selon son intérêt par rapport à l'investigation en question.
Le tableu 3.4 illustre la table de transaction :

	NumDes1	NumDes2	NumDes3	NumDes4
NumSrc1	0	1	0	0
NumSrc2	0	0	1	1
NumSrc3	0	1	0	0
NumSrc4	0	0	0	0

TABLE 3.4 – Table de transaction

L'algorithme de construction des items sets fréquents

Plusieurs algorithmes traitent le problème de la recherche des motifs fréquents. Nous citons, à titre d'exemple, Apriori [2],[69], AprioriTID [2], Partition [64], Close [59], [60] et Pascal [6]. Ce dernier nous intéresse plus particulièrement pour les raisons suivantes. Dans notre cas d'étude, les objets sont des événements (SMS, appels) et les items sont les numéros de téléphones sources et destinations extraits de ces événements. La base de preuves peut être volumineuse. L'algorithme Pascal s'avère adapté à cette volumétrie. En effet, Pascal est basé sur le comptage par inférence, destinée à réduire le nombre du calculs de supports des motifs lors de l'extraction des motifs fréquents. Avec le comptage par inférence, seuls les supports des motifs clés fréquents (et de certains non fréquents) sont calculés depuis la base de données . Dans Pascal comme dans Apriori, les motifs fréquents sont extraits par niveaux : durant chaque itération, les motifs candidats de taille k sont créés en joignant les motifs fréquents de taille k-1, leur support est déterminé et les motifs non fréquents sont supprimés. Si un motif candidat de taille k est un motif non clé, alors son support est égal au plus petit des supports de ses sous ensembles de taille k-1. Ceci permet de réduire lors de chaque balayage de la base de données le nombre de motifs considérés et encore plus important, de réduire le nombre total de balayages réalisés. En plus, le comptage par inférence permet de calculer les supports d'autant qu'il y a des motifs que possible sans accéder à la base de données en utilisant les informations acquises durant les itérations précédentes.

Exemple 3.3.1. *L'exécution de Pascal sur la base de données D est présentée dans la figure 3.6 qui indique les classes d' équivalences et les motifs clés sur le treillis des parties de D. La première passe de Pascal donne les motifs candidats de taille 1. Tous sont des motifs clés puisque aucun n'a un support de 100 %. Les candidats de la deuxième itération sont donc tous présumés clés. Après lecture de la base de données, il s'avère que*

a,c et b,e ne sont pas des motifs clés : le support de a,c est le même que celui de a, le support de b,e est identique à celui de b (et de e). Par conséquent, aucun candidat de C3 n'est clé, car chacun deux est sur-ensemble de a,c ou de b,e et il est inutile de parcourir la base de données pour connaître les supports. De même à la dernière étape : le support de a,b,c,e est égal à celui de a,b,c (ou a,b,e, ou a,c,e). Pascal recherche le support de 11 motifs dans la base de données en deux passes.

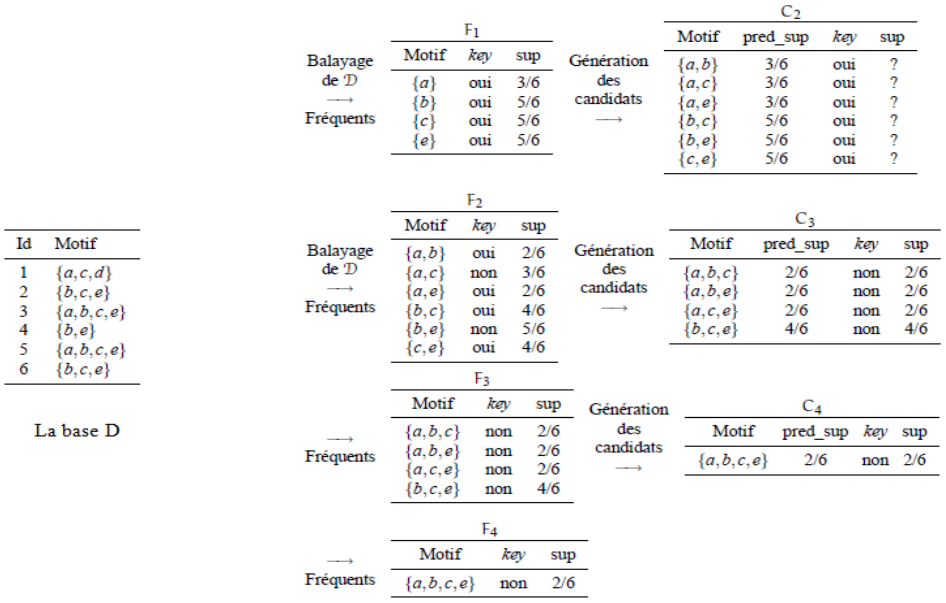


FIGURE 3.6 – Exemple d'exécution de Pascal

L'application de l'algorithme de Pascal sur la table de transaction donne lieu à un ensembles des motifs fréquents(numéros de téléphones suspects), l'algorithme est le suivant :

Algorithme 5 Construction des numéros de téléphones suspect(itemsets fréquents)

Require: Matrix of transaction : **NN[n,n]** ;
Minimum support **Suppmin** ;
Ensure: Items sets frequents **I** such as **I as** $I = U_k I^k$;

1: I_c =null ; ▷ I_c represents the set of items candidates ;
2: **For** all column i=1.. n of the transaction table NN **Do**
3: I_c= all the items sets of size 1 ;
4: $support_i = \sum_{j=0}^{N}(k)NN[i,j]$ such as : $NN[i,j] = 0 \vee 1 \wedge i = 1..n$;
5: **If** $(support_i/n \geq Suppmin)$ **Then**
6: Add I_k^i in I such as $I_k^i = k - itemset \wedge k - itemset \in I_c$;
7: **EndIf** ;
8: **EndFor** ;
9: **If** NotEmpty($I = U_k I^k$)**Then**
10: K=k+1 ;
11: **Goigner(I,k)** ; ▷ To construct the set of items candidates I_c of the size k ;
12: **Else**
13: $I = U_{k-1} I^{k-1}$;
14: **Goto 23 ;**
15: **EndIf ;**
16: **For** (i =1. . card(I_c)) **Do**
17: $support_i = \sum_{j=0}^{N}(k)I_c NN[i,j]$ such as : $NN[i,j] = 0 \vee 1 \wedge i = 1..n$;
18: **If** $(support_i/n \geq Suppmin)$ **Then**
19: Add $I_k^i = k - itemset$ in I ; ▷ k-itemset belongs to the set of candidates items ;
20: **EndIf ;**
21: **EndFor ;**
22: **Goto 9 ;**
23: **End.**

A partir de l'ensemble des items sets fréquents, nous générons toutes les associations téléphoniques suspectes par l'algorithme de génération des règles d'association et de découvrir le criminel :

Algorithme 6 Génération des règles d'associations(Associations Téléphoniques)

Require: Items sets frequents : **I** ;
Matrix of transaction : **NN[n,n]** ;
Minimum confidence : **confMinim** ;
Ensure: A set of rules of association ;

1: **For** each item set frequent L de I **Do**
2: **For** each subset A of L **Do**
3: calculate(support(L)/support(A)) ;
4: **If**(support(L)/support(A)≥confMinim)**Then**
5: $Rule = Rule \bigcup A \rightarrow (L - A)$;
6: **EndIf** ;
7: **EndFor** ;
8: **EndFor** ;
9: **End**.

Chaque motif de l'ensemble des items sets fréquents est analysé afin de générer une collection des SMS filtrés(c'est à dire un ensemble des numéros de téléphones qui communiquent entre eux par messages). Le but de cet algorithme et de répondre sur la question *quoi ?*, la description de l'algorithme est la suivante :

Algorithme 7 Collecte des SMS

Require: Items sets frequents :**I** ;
Ensure: Set of the SMS ;

1: **For** each item set frequent L de I **Do**
2: **For** each item A de L **Do**
3: **If** (A is only communicate by call) **Then**
4: L=L-A ;
5: **EndIf** ;
6: **EndFor** ;
7: **EndFor** ;
8: **End**.

L'algorithme de génération des cliques SMS (Clicks SMS)

Le modèle de clique d'utilisateur est utilisé pour modéliser les communications des SMS sur un smartphone, de résumer ces ensembles et de leur dynamicité dans le temps. Cette

information est utilisée pour classifier les SMS selon leurs contexte qui conduit à déterminer de quoi il s'agit [36].
Formellement, la communication mobile par SMS peut être capturée par un graphe orienté G(V,E) avec l'ensemble de nœuds V étant des numéros de téléphones individuels et E est l'ensemble des événements SMS. Un arc e_{12} existe si v_1 envoie un SMS à v_2. Vu sous cet angle, les cliques sont un certain modèle dans ce graphe que nous essayons de caractériser et d'utiliser comme une norme comportementale de communication. Une clique est un sous ensemble de sommets qui induit un sous graphe complet et symétrique mais ces deux propriétés sont rarement être ensemble dans l'investigation mobile légale .
Dans le cadre d'analyse des SMS, un SMS peut être considéré comme une transaction qui implique plusieurs numéros de téléphone mobile, y compris l'expéditeur et un ou plusieurs destinataires. Nous exploitons la matrice de transaction pour définir la matrice de similarité afin de générer des clusters SMS , chaque classe est considérée comme une clique qui contient une source(numéros de téléphones sources) et plusieurs destinataires(numéros de téléphones de destinations), donc, il s'agit de même contexte échanger dans les extrémités des cliques et d'autre sujets pour d'autre cliques.

La matrice de similarité

Étant donné la matrice de transaction NN. L est un item set (Collecte SMS), A un item tel que A $\notin$ L, B un item tel que B $\in$ L, pour construire la matrice de similarité on peut définit une fonction de similarité Sim(A,B) et de dissimilarité Dissim (A,B) entre A et B comme suit :

$$Sim(A,B) = \begin{cases} 1 & si \qquad NN[A,B] = 1 \\ 0 & sinon \end{cases} \tag{3.8}$$

$$Dissim(A,B) = \begin{cases} 1 & si \qquad NN[A,B] \neq 1 \\ 0 & sinon \end{cases} \tag{3.9}$$

La description de l'algorithme de génération des cliques SMS(Clicks SMS) est la suivante :

Algorithme 8 Génération des cliques SMS

Require: Items sets frequents filtered : **L**.
Matrix of similarity : **M**
Ensure: Set of clustering (clicks) ;

1: **For** each element A of cluster L **Do**
2: **For** each phone number A_i **Do**
3: **If**$(M[A_i, A] = 1)$**Then**
4: click == $click \bigcup A$;
5: **EndIf** ;
6: **EndFor** ;
7: **EndFor** ;
8: **End**.

Notons que, les algorithmes implémentés par notre stratégie d'analyse relationnelle ont des propriétés intéressantes :

- ◇ La complexité du calcul est relativement basse car, elle est polynômial O(N^2) suivant le nombre des événements.
- ◇ Les résultats du calcul sont précis car, le taux d'erreurs est faible.
- ◇ L'algorithme est performant (scalable) avec un nombre élevé des événements puisque il s'appuie sur le principe de comptage par inférence, qui est destinée à réduire le nombre du calculs de supports des motifs lors de l'extraction des items sets fréquents..

3.4 Conclusion

Dans ce chapitre, nous avons présenté notre approche d'analyse de preuves selon le système d'investigation proposé. Cette approche exploite les résultats de la classification non supervisée appliqués sur une base de preuves donnée. Cette classification permet de construire un timeline générique des événements causalement dépendants et qui permet de répondre sur le *quand ?* et le *comment ?* lors de l'investigation d'un crime utilisant des smartphones. Pour effectuer la classification, nous avons défini des mesures de similaritées et de dissimilaritées permettant de construire, à partir de la base de preuves, par une succession de fusions, les classes d'événements. Pour avoir une grande chaîne d'événements , nous avons proposé un algorithme ascendant parallèle qui permet de classifier les événements au même temps. À l'aide de technique de filtrage (blocage) issue

des smartphones et la fréquence des numéros de téléphones durant des communications, nous avons proposé un algorithme pour la générations des règles d'associations afin d'analyser le comportement d'un suspect et de répondre sur la question *qui ?* lors d'une analyse relationnelle du crime. Nous avons filtré les SMS à partir des numéros les plus fréquents afin de les analyser dans des groupes pertinentes et afin de répondre sur la question *Quoi ?*.

Chapitre 4

Mise en œuvre et validation de l'approche d'analyse de preuves

4.1 Introduction

Nous venons de spécifier au cours du précédent chapitre les aspects et les algorithmes constituants notre approche d'analyse de preuves dans la phase d'analyse de processus d'investigation mobile légale. Dans le présent chapitre, nous présentons les outils logiciels utilisés ainsi que les résultats obtenus à l'implémentation de l'approche.
Dans la première partie, nous présentons notre prototype d'implémentation.
Dans la seconde partie, nous évaluons les performance de notre approche en le comparant par d'autres(ROCK, CHAMELEON, AbRVel).

4.2 Conception du notre système d'analyse de preuves

Notre système d'investigation se compose de trois modules principaux. Chaque module exécute une des trois phases nécessaires, à savoir, la collecte, l'analyse de preuves et la génération de rapport. Le cœur de ce dernier est le module d'analyse de preuves.
Une fois les preuves ont été acquises à partir des smartphones à travers l'outil d'acquisition, nous avons stocké les SMS et les appels comme des enregistrements dans une base

de données MySQL.
L'objectif du module d'analyse est d'identifier les corrélations et les associations dans les preuves(SMS, appels). L'analyse temporelle et fonctionnelle permet de reconstruire des évènements ordonnés dans le temps avec la génération d'un timelime. L'analyse relationnelle permet de générer l'ensemble des numéros de téléphones suspects et leurs associations téléphoniques. L'ensemble des numéros de téléphones communiqués par SMS doit être filtrés pour générer l'ensemble des cliques, chaque clique représente un clustering selon le contexte. Enfin l'interprétation des résultats d'analyse est affiché dans un rapport au format PDF.

4.3 Prototype d'implémentation

La figure 4.1 montre la première fenêtre qui apparait lors du lancement du système d'investigation légale des smartphones. Cette fenêtre permet d'identifier les utilisateurs ayants le droit d'utiliser et de manipuler le système.

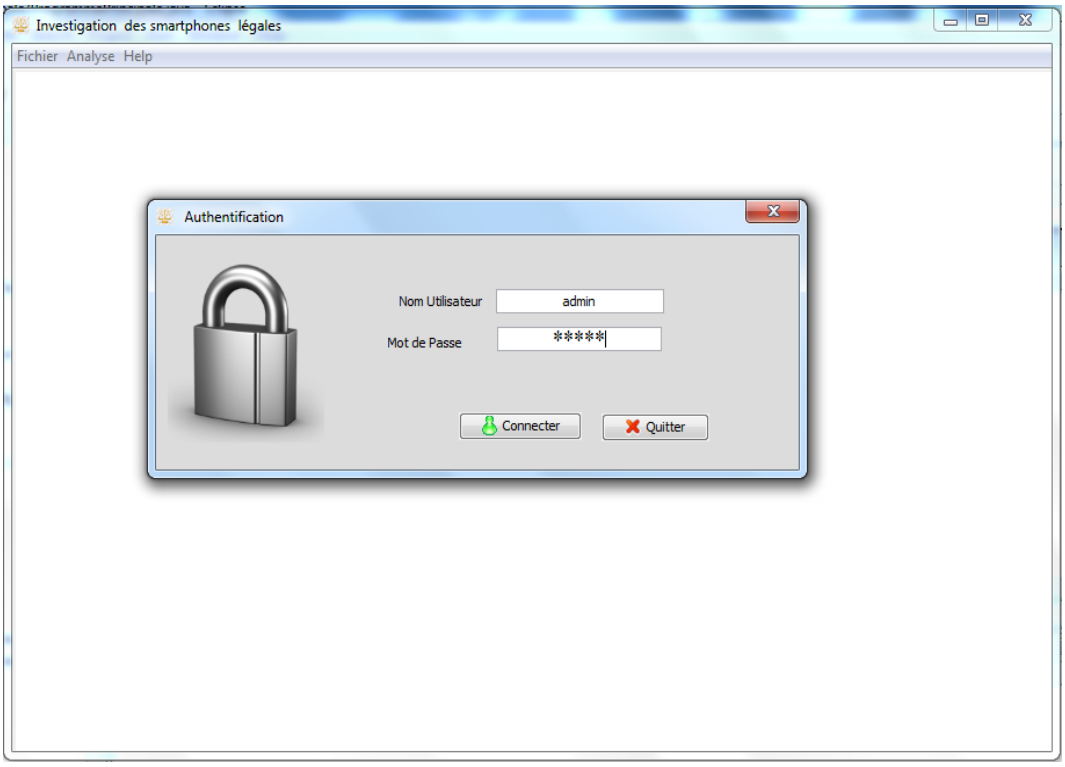


FIGURE 4.1 – Fenêtre d'authentification

Le système est utilisé à travers des sessions d'investigation, nous pouvons ouvrir et fermer la session, modifier le mot de passe et quitter l'application, ces options se trouvent quand on clique sur le menu **Fichier** (voir la figure 4.2)

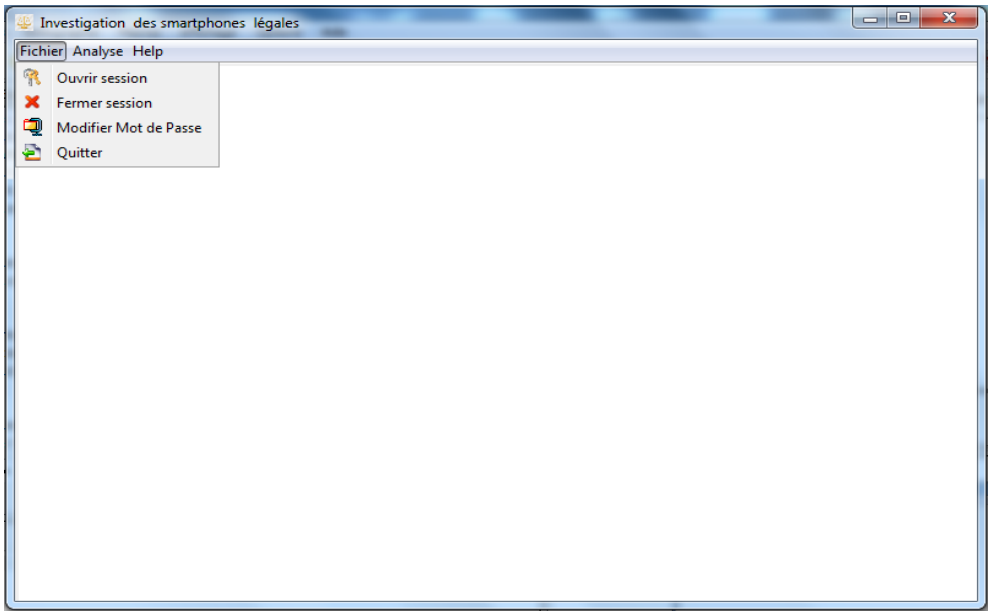


FIGURE 4.2 – Interface principale

Nous avons implémenté le sous-système d'analyse de preuves. Dans ce dernier il y a :

– **Module d'analyse temporelle & fonctionnelle :** Ce type d'analyse va commencer par la génération d'un dendrogramme qui montre la stratégie de parallélisme puis la génération d'un timelime pour la représentation des évènements à chaque intervalle.

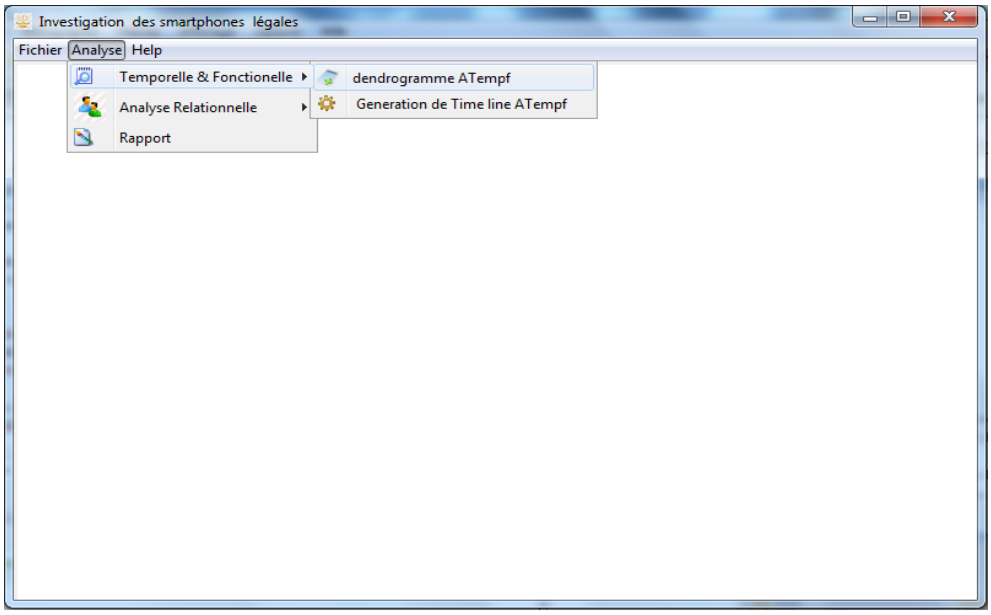


FIGURE 4.3 – Module d'analyse temporelle & fonctionnelle

◇ ***Génération de dendrogramme :*** Représente le déroulement de notre algorithme de clustering pour la génération des clusters illustrés par la figure 4.4 :

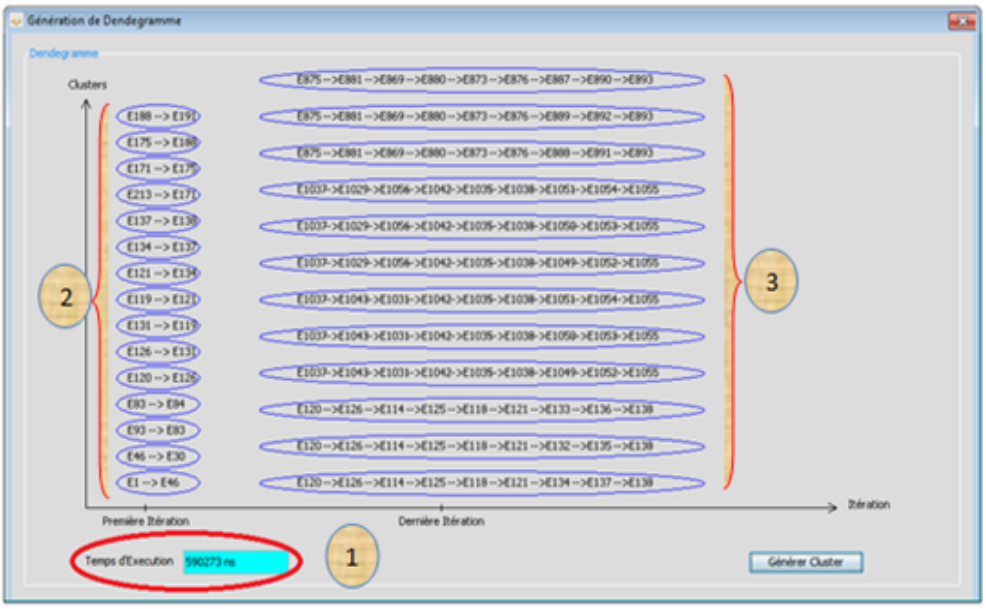


FIGURE 4.4 – Génération de dendrogramme

Au niveau de cette interface, nous présentons le déroulement de notre algorithme de clustering afin de bien illustré la notion de parallélisme pour la construction et la fusion de clusters de la même itération. Sachant que :
1 : Représente le temps d'exécution de tous les clusters générés.
2 : C'est l' ensemble de clusters de la première itération
3 : Représente un ensemble de cluster de la dernière itération.

◇ ***Génération de timelime :*** La génération de timelime est représentée comme un ensemble des clusters des événements ordonnés au niveau de chaque intervalle du temps tel qu'il est illustré dans la figure 4.5 :

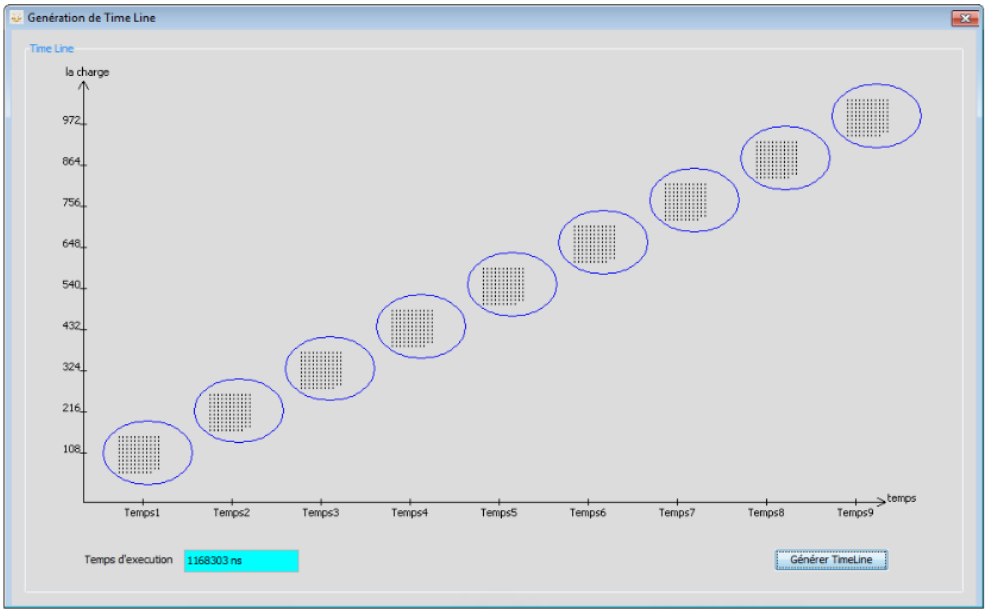


FIGURE 4.5 – Génération de timeline

– **Le module d'analyse relationnelle :** Au niveau de ce type d'analyse, nous allons aborder les items sets fréquents pour la génération des numéros de téléphones suspects et les règles d'associations pour l'extraction des associations téléphoniques et la génération des cliques sous forme d'une source et plusieurs destination et dans ce cas-là, nous pouvons savoir le contexte des SMS transmis au sein de chaque groupe.

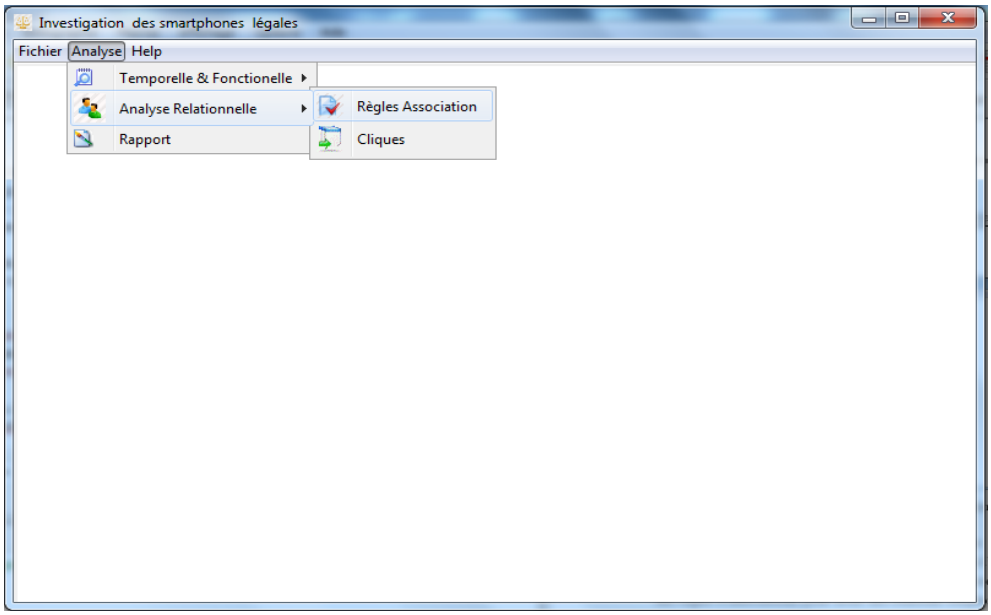


FIGURE 4.6 – Module d'analyse relationnelle

◇ ***Génération des associations téléphoniques*** : Après la génération de l'ensemble des numéros de téléphones suspects(candidats) à l'aide de l'extraction de l'ensemble des items sets fréquent, nous avons appliqué notre algorithme de générations des règles d'associations pour avoir des relations téléphoniques pour préciser le suspect.
Nous avons fait une comparaison avec l'approche d'AbRVel de génération du profil, les résultats sont illustrés dans l'interface suivante.

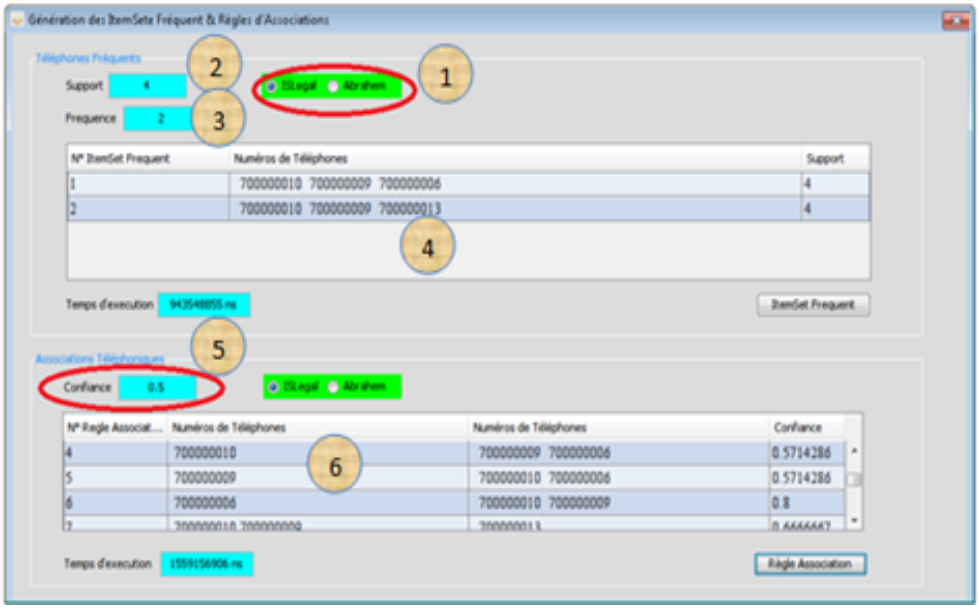


FIGURE 4.7 – Liste des numéros de téléphones suspects et des associations téléphoniques

Au niveau de cette figure nous présentons l'ensemble des items sets fréquents ainsi que les règles d'association, sachant que :
1 : Représente un choix entre notre approche et l'approche d' AbRVel, nous avons sélectionné notre approche(RAFM).
2 : Représente le support seuil.
3 : Représente la fréquence pour prendre en considération que les événements intéressant.
4 : Représente l'ensemble des items sets fréquents selon le support seuil donné.
5 : Représente la confiance minimale.
6 : Représente l'ensemble des règles d'association selon la confiance donnée.

◇ ***Génération des Cliques*** : A partir de l'ensemble des items sets fréquents, nous avons collecté tous les numéros de téléphones qui communiquent à traves des SMS, ces numéros sont regroupés selon un numéro de téléphones sources et plusieurs numéros de téléphones de destination , ces classes permettent de générer des cliques qui permettent d'avoir le contexte d'intéraction (le contenu des SMS) entre les extrémités communiquantes (voir la figure 4.8)

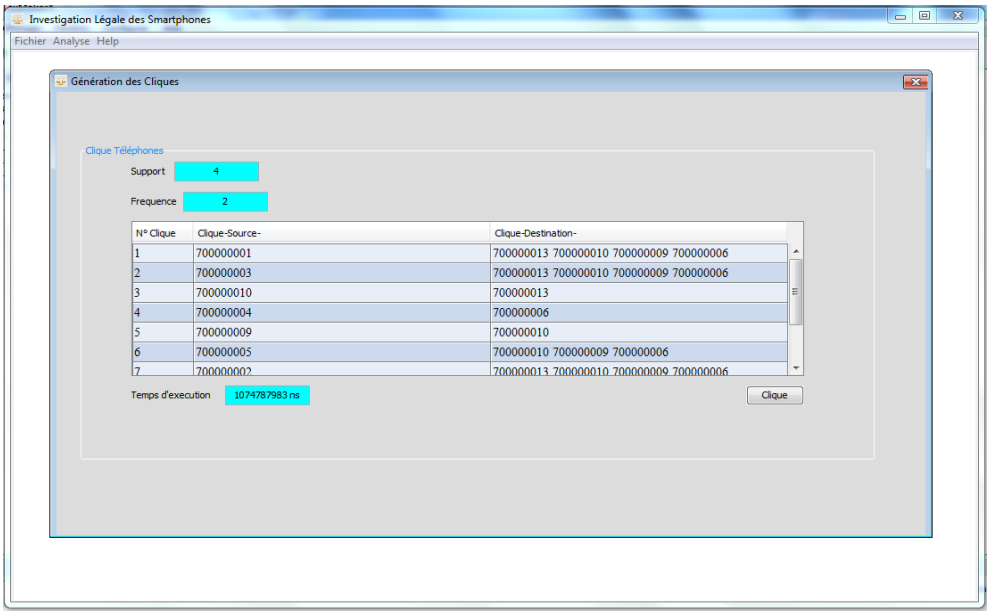


FIGURE 4.8 – Liste des cliques

– **Le module de génération de rapport :** Une fois l'analyse est faite, nous pouvons afficher un rapport sous format PDF qui comporte les résultats du module d'analyse comme les informations personnelles d'un suspect, par exemple tel quel illustré dans la figure 4.9

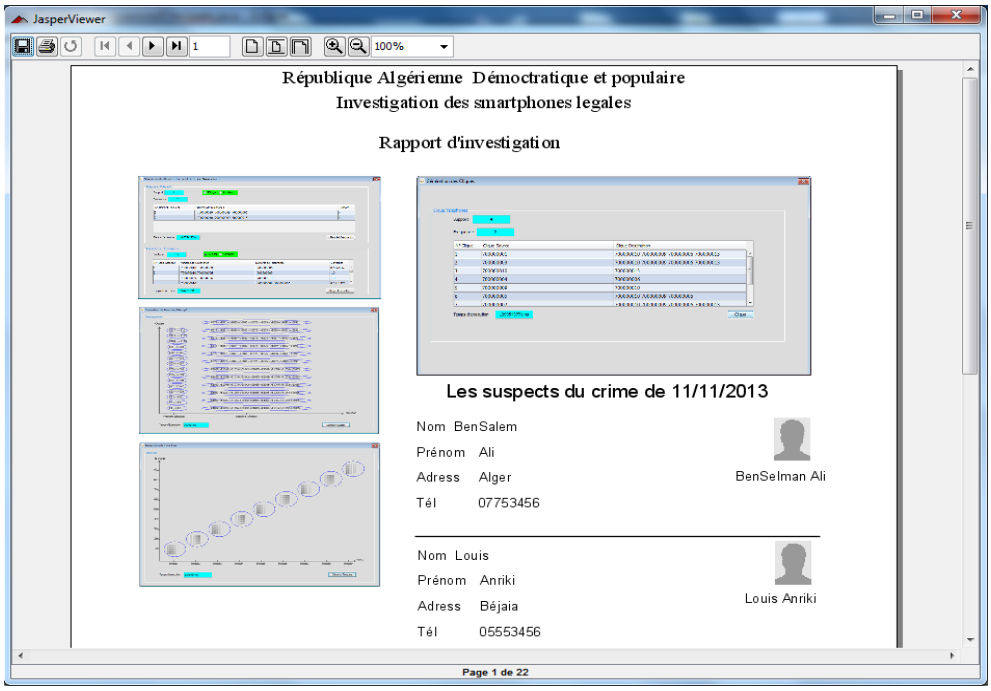


FIGURE 4.9 – Rapport d'investigation

4.4 Évaluation de l'approche d'analyse de preuves

Notre système d'analyse a été implémenté sous JAVA avec l'IDE (Integrated Development Environment) Eclipse (HILIOS, Version : 1.3.0.20100617-0521), plusieurs tests ont été effectués sur notre module d'analyse de preuves. Le but de ces tests est d'évaluer la scalabilité de notre approche, c'est à dire la possibilité de générer des clusters d'un timelime pour un nombre très grand d'évènements. Nous évaluons aussi l'efficacité et la rapidité de l'algorithme de calcule des items sets fréquents pour la génération des numéros de téléphones suspects, ainsi que l'algorithme d'extraction des règles d'associations pour la génération des associations téléphoniques. Tous ces tests sont exécutés sur un PC core (TM)2 Duo CPU avec 3Go RAM, processeur de 2.10 GHZ sous Windows7.

Nous avons mesuré par la suite les performances de l'approche en utilisant deux métriques : temps de génération du cluster et temps de réponse qui comporte le temps de génération des items sets fréquent et le temps d'extraction des règles d'associations. Ces métriques sont affectées par des facteurs comme le nombre des évènements

dans la base de preuves, le seuil de support, nombre d'itérations, etc.

4.5 Discussion des résultats

Dans ce qui suit, nous allons présenter et analyser les résultats d'expérimentation obtenus suivant les métriques de performances décrites précédemment.
Dans tous les scénarios du teste effectué, notre approche (Relational Analysis for Frequents Motives(RAFM)) a été comparée avec l'approche AbRVel pour la génération du profil dans le cadre d'analyse relationnelle, en ce qui concerne l'analyse temporelle et fonctionnelle, notre approche(Clustering Algorithm For Temporal and Functional Analysis (CATFA)) pour la génération de timelime a été comparée avec les algorithmes CHAMELEON et ROCK.

4.5.1 Temps de génération des clusters vs le nombre des évènements

Nous avons mesuré l'évolution du temps de génération de cluster de notre algorithme proposé pour construire le timelime en fonction du nombre des évements SMS et appel de la base de preuves. Après cela nous avons comparé les performances de notre algorithme avec ROCK [66] et CHAMELEON [40]. Les résultats montrent une meilleure scalabilité de notre approche.

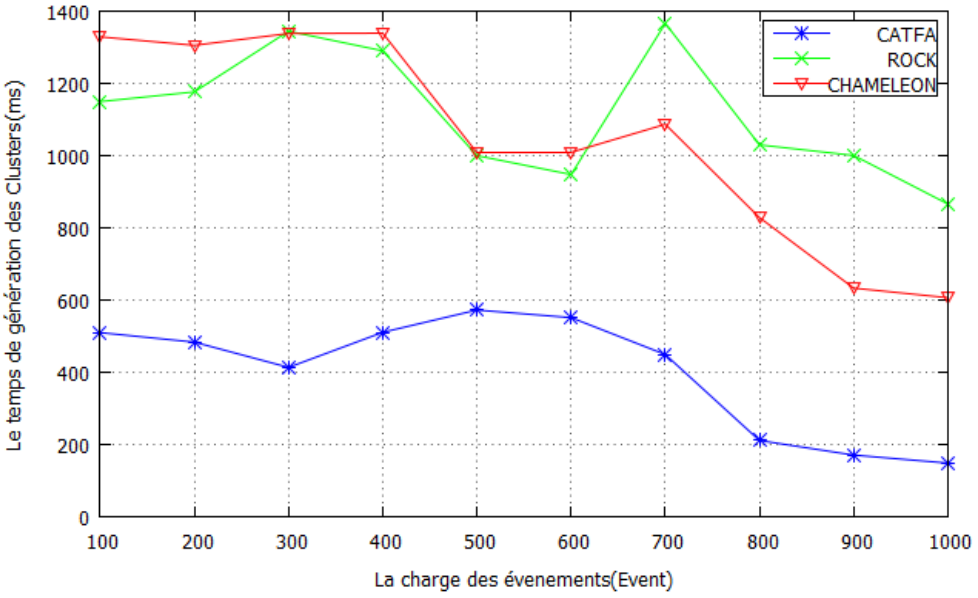


FIGURE 4.10 – Temps de génération des clusters vs le nombre des évènements

Selon les résultats présentés dans la figure précédente, nous pouvons voir que notre approche CATFA, surperforme l'algorithme de ROCK et CHAMELEON en marquant un gain du temps atteint jusqu'à moins de 800 ms de temps (pour 800 évènements par exemple) par rapport a ROCK et 600 ms par rapport à l'algorithme CHAMELEON. Ce gain est grâce aux parallélismes utilisés pour construire tous les clusters ayant des évènements causalement dépendant et les fusionnés dans la même itération lors de la construction de dendrogramme. Alors que les deux algorithmes ROCK et CHAMELEON n'exercent pas aucuns parallélismes, ils ont basé sur un balayage simple de la base de preuves pour construire leurs dendrogramme (construction et fusion des clusters). Egalement, sur la même figure, on peut remarquer que l'accroissement du nombre des évènements dans la base de preuves induit à une diminution dans le temps de génération des trois approches. Pour notre approche cela est raisonnable à cause du parallélisme, par contre dans les deux algorithmes cette diminution est aléatoire induit à un balayage totale à chaque itération pour construire un nombre de clusters. Ce nombre est varié selon la similarité entre les classes générées.

4.5.2 Temps de réponse vs le support minimal

Dans cette section, nous allons expérimenter la variation du temps de réponse par rapport au support minimal. Ce test a pour but de se faire une idée de la complexité de l'algorithme d'extraction des items sets fréquents et l'algorithme de génération des règles d'association en fonction de la charge des événements.

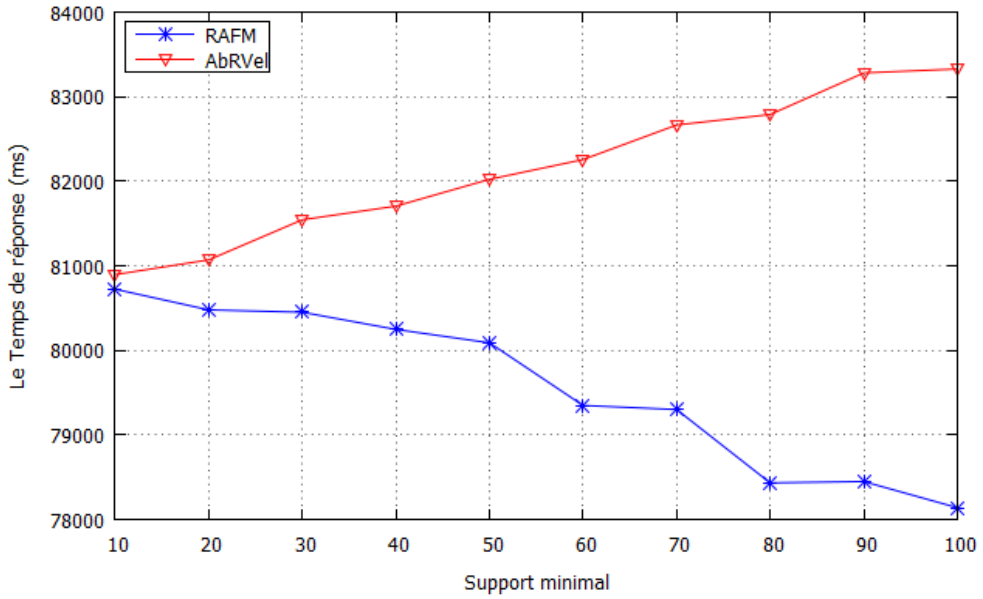


FIGURE 4.11 – Temps de réponse vs le support minimal

Selon les résultats présentés dans la figure précédente, le temps de réponse des approches varie entre 78138 ms et 83328 ms avec un avantage marqué par notre approche par rapport à l'approche AbRVel. On peut remarquer que le temps de réponse le plus bas est celui présenté par notre approche, il atteint à 78138 ms lorsque le support minimal prend la valeur 100 par rapport à l'approche de AbRVel, qu'il dépasse 83328 ms, parce qu' au niveau de l'approche d'AbRVel, il y a un balayage total de la matrice de transaction lors de construction des items sets fréquents, par contre, notre approche fait un balayage partiel de la matrice de transaction (selon les items sets fréquents éxistés au niveau de l'itération précédente). Ces résultats sont très raisonnables et traduisent bien l'efficacité de notre approche.

4.6 Conclusion

Dans ce chapitre, nous avons présenté les résultats de l'implémentation de notre système d'analyse de preuves(CATFA, RAFM). Dans la première partie du chapitre, nous avons présenté la plateforme logicielle que nous avons développé afin de concrétiser nos propositions théoriques dans le cadre de l'investigation mobile légale. Il s'agit d'un outil d'aide à la décision judicaire dédié à l'analyse temporelle, fonctionnelle et relationnelle de preuves issus des smartphones d'acquisition de la scène du crime.

Dans la deuxième partie du chapitre, nous avons présenté une étude comparative entre les performances de notre approche d'analyse et les approches existante de la même famille ROCK, CHAMELEON et AbRVel. Plusieurs tests ont été réalisés, le premier de ces tests portait sur l'effet du nombre des évènements sur le temps de génération du cluster. Le second test porte sur le temps de réponse en variant le support minimal. Les résultats des tests montrent que notre approche améliore le temps de génération du cluster et le temps de réponse. En effet, les performances de notre protocole en termes du temps de réponse ne se dégradent pas avec l'augmentation du nombre de services, ce qui est l'inverse de l'approche AbRVel.

Conclusion Générale et Perspectives

Dans le cadre de ce mémoire, nous avons essayé d'apporter des solutions à la problématique d'analyse de preuves pour la reconstruction du crime qui ont été appliquées au domaine de l'investigation mobile légale. Ces stratégies d'analyses exploitent les techniques de fouille de données en général , la recherche des motifs fréquents et la classification non supervisée en particulier.

Notre travail a été guidé par une étude préalable des travaux existants traitant les problèmes de collecte et d'analyse de preuves qui ont basé sur les techniques de datamining. Chacun de ces travaux a été étudié selon des critères que nous avons jugés pertinents, à savoir, la scalabilite ou le passage à l'échelle de processus d'enquête légale numérique en gardant le temps moyen constant en terme d'investigations avec la croissance de la taille et la diversité des smartphones cibles, l'exactitude des résultats mesurés par la précision et le taux du rappel, l'exhaustivité qu'est la recherche de toutes les preuves liées aux crimes dans la phase de collecte des évidences et enfin la qualité des résultats. Le travail effectué porte essentiellement sur :

- La modularité de nos stratégies qui sont adaptables en modifiant un ou plusieurs modules (par exemple cliques SMS, filtrage SMS) ;
- L'efficacité de nos propositions, ce qui passe par un processus d'expérimentation et de validation impliquant l'implantation de nos solutions.

Dans ce travail, nous avons implémenté un système d'analyse de preuves collectées à partir des smartphones. Ce système est constitué de deux modules : un module d'analyse temporelle et fonctionnelle de preuves et un autre module d'analyse relationnelle. En plus

des module de collectes, de filtrage, de stockage et de reporting qui offrent les moyens et les outils nécessaires pour réaliser l'analyse de preuves.

Les principales contributions concernant notre stratégie d'analyse relationnelle de preuves résident dans les points suivants :

- La recherche des motifs fréquents qu'est une bonne heuristique pour réduire l'espace de recherche de l'ensemble des numéros de téléphones suspects à l'aide des deux notion de blocage et de fréquence.
- L'algorithme de recherche des motifs fréquents peut générer des numéros de téléphones mono et multi-attributs à la volée. Cela évite de passer par un processus à plusieurs itérations
- Notre stratégie d'analyse relationnelle de preuves est modulaire et peut être améliorée ou adaptée en changeant le module adéquat. Par exemple, si nous voulons prendre une nouvelle technique d'analyse, il suffit d'ajouter ce module au module de génération des numéros de téléphones suspects à partir des motifs fréquents.

Notre stratégie d'analyse fonctionnelle présente quant à elle, les caractéristiques novatrices suivantes :

- Elle utilise la classification non supervisée basée sur un algorithme de clustering hiérarchique agglomérative pour construire un timeline (chronologie des événements de différents sources) afin d'analyser le crime et répondre sur les questions *quand ? et comment ?*.
- La matrice de similarité est généré à partir d'un algorithme basé sur la notion de causalité dynamique.
- Les mesures de similarité et de dissimilarité introduites dans le processus de classification permettent de capturer les relations qui peuvent exister entre les évènements(SMS et appels).
- La fusion des évènements réalisée au niveau des classes obtenues à partir de la classification non supervisée est moins coûteuse que la fusion réalisée directement sur la base de preuves.

L'évaluation de notre nouvelle approche d'analyse de preuves dans le cadre de l'investiga-

tion mobile légale a montré son impact positif sur le temps d'exécution par rapport aux autres travaux [1],[66]. Les résultats générés par l'algorithme de clustering sont intéressants et montrent que le timeline généré peut être adapté quel que soit la charge importantes des évènements dans la base de preuves. Les résultats obtenus, en appliquant la stratégie des motifs fréquents pour la génération des numéros de téléphones suspects, ainsi que les associations téléphoniques, en plus l'algorithme de génération de cliques qui permet d'avoir des informations précieuse sur le suspect, ils justifient l'adoption de ces algorithmes pour résoudre le problème d'analyse de preuves.

La série d'expérimentation menée sur notre approche d'analyse nous a permis de prouver sa scalabilité, sa flexibilité et son adaptabilité à l'investigation mobile légale.
Pour les perspectives envisageables à nos travaux :

- L'analyse de preuves en explorant d'autre type de données au niveau des smartphones tels que les MMS, vidéos, photos, emails, comptes facebook, etc.
- L'exploration des autres techniques de datamining pour localiser les preuves dans la phase de collecte et de reporting.
- On peut améliorer notre algorithme qui génère les items sets fréquents afin de diminuer le nombre des items de ce dernier.
- Nous avons expérimenté nos stratégies sur une base de preuves, nous envisageons d'étendre ces tests sur des crimes réels basés sur les smartphones.
- Concernant l'algorithme de génération des cliques, on peut exploiter des techniques de fouille de données pour analyser le contenu des SMS pour chaque clique et essayer d'étudier des corrélations entre eux.
- On peut modifier la structure de la base de preuves pour la prise en compte de la dynamicité des évidences afin d'améliorer les résultats d'analyses.

Bibliographie

[1] T. Abraham, R. Kling, and O. de Vel. Investigative profiling with computer forensic log data and association rules. *in ICDM 02, Proceedings of the 2002 IEEE International Conference on Data Mining, pp. 11-18, Maebashi City, Japan*, (2002).

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. *in VLDB'94 20th international conference on Very Large Data Bases, pp. 478-499, Santiago de Chile, Chile*, (1994).

[3] A.Morum, F.Caus Sicoli, L.Peotta, and R.Timoteo. Acquisition of digital evidence in android smartphones. *In the Proceedings of the 9th Australian Digital Forensics Conference, Vol. 2 , N°. 3*, (2011).

[4] L. Aouad, T. Kechadi, J. Trentesaux, and N.A Le-Khac. An open framwork for smartphone evidence acquisition. *International Conference on Digital Forensics, Vol. 383, pp. 159-166, Pretoria, South Africa*, (2012).

[5] L. M. Aouad and T. Kechadi. Android forensics : A physical approach. *In SAM'12 the 11th International Conference on Security and Management, pp. 311-316*, (2011).

[6] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Pascal : Un algorithme d'extraction des motifs fréquents. *Technique et Science Informatiques, Vol. 21, No. 1, pp. 398-416*, (2002).

[7] AC. Bogen and DA. Dampier. Preparing for large scale investigations with case domain modeling. *in DFRWS'05, Proceedings of the 5th annual digital forensic research workshop, French Quarter, New Orleans*, (2005).

[8] A. Brinson, A. Robinson, and M. Rogers. A cyber-forensics ontology : Creating a new approach to studying cyber forensics. *Digital Investigation, The International Journal of Digital Forensics and Incident Response, Vol. 3, No. 1, pp. 37-43, USA*, (2006).

[9] R. Brown, B. Pharm, and O. de Vel. A grammar for the specification of forensic image mining searches. *EANZIISC'03, Proceedings of the Eighth Australian and New Zealand Intelligent Information Systems Conference, pp. 139-144, Sydney, Australia*, (2003).

[10] R. Brown, B. Pharm, and O. de Vel. Design of a digital forensics image mining. *in IIHMSP'05, Proceedings of the International Workshop on Intelligent Information Hiding and Multimedia Signal Processing, Vol. 3683, pp. 395-404, Melbourne*, (2005).

[11] Z. Brown, D. Lin, and F. Guo. A method for locating digital evidences with outlier detection using support vector machine. *in IJNSEC'08, International Journal of Network Security, Vol. 6, No. 3, pp. 301-308, China*, (2008).

[12] Z. Brown, D. Lin, and F. Guo. A method for locating digital evidences with outlier detection using support vector machine. *in IJNSEC'08, International Journal of Network Security, Vol. 6, No. 3, pp. 301-308, China*, (2008).

[13] M. Carney and M. Rogers. The trojan made me do it : A first step in statistical based computer forensics event reconstruction. *in IJDE'04, International Journal of Digital Evidence, Vol. 2, No. 4, pp. 01-11, India*, (2004).

[14] B. Carrier and E.H. Spafford. Getting physical with the digital investigation. *in IJDE'03,International Journal of Digital Evidence, Vol. 2, N°. 2*, (2003).

[15] B.D. Carrier and E.H Spafford. Defining event reconstruction of digital crime scenes. *in JFS'04 Journal of Forensic Sciences, Vol. 49, No. 6, West Conshohocken*, (2004).

[16] B.D. Carrier and E.H Spafford. Defining searches of digital crime scenes. Under review, (2004).

[17] B.D. Carrier and E.H. Spafford. An event-based digital forensic investigation framework. *In CERIAS'04, Center for Education and Research in Information Assurance and Security Proceedings of the fourth Digital Forensic Research Workshop, pp. 11-13, Baltimore, Maryland*, (2004).

[18] B.D. Carrier and E.H. Spafford. Automated digital evidence target definition using outlier analysis and existing evidence. *in DFRWS'05, Proceedings of the fifth Digital Forensic Research Workshop, New Orleans, LA*, (2005).

[19] E. Casey. *Digital Evidence and Computer Crime : Forensic Science, Computers, and the Internet*. Second Edition, Elseiver Academic Press,Great Britain, (2004).

[20] E. Casey. Network traffic as a source of evidence : Tool strengths, weaknesses,and future needs. *Digital Investigation, Vol. 1, No. 1, pp. 28-43, USA*, (2004).

[21] E. Casey. *Digital Evidence and Computer Crime Forensic Science and Computers and the Internet*. Third edition, Elsevier, cmdLabs, Baltimore, Maryland, USA, (2011).

[22] P. Christy and G. Laurence. Sales of mobile devices in second quarter of 2011 grew 16.5 percent year-on-year ; smartphone sales grew 74 percent. Gartner,UK,Egham, (2011).

[23] M. Clugston and J. Ed. *The New Penguin Dictionary of Science*. (1998).

[24] O. de Vel, A. Anderson, and M. Corney G. Mohay. *E-mail Authorship Attribution for Computer Forensics*. First Edition, Kluwer Academic, Boston, Dordrecht, London, (2002).

[25] S. Decherchi, S. Tacconi, J. Redi, F.Sangiacomo, A. Leoncini, and R. Zunino. Text clustering for digital forensics analysis. *Computational Intelligence in Security for Information Systems, Vol. 63, pp. 29-36, Genova, Italy*, (2009).

[26] P. Deutsch. Gzip file format specification version 4.3. RFC 1952 (Internet Engineering Task Force), (May 1996).

[27] T. Duval, J. Bernard, and R. Laurent. The mitnick case : How bayes could have helped. *in ICDF'05, International Conference on Digital Forensics, Advances in Digital Forensics, Vol. 194, pp. 91-104, Orlando, Florida*, (2005).

[28] JL. Elman. Finding structure in time. *Cognitive Science, Vol. 14, No. 2, pp. 179-211, CA*, (1990).

[29] H. Farid and S. Lyu. Higher-order wavelet statistics and their application to digital forensics. *in CVPRW'03, Proceedings of the IEEE Computer Vision and Pattern Recognition Workchop, Vol. 8, pp. 94-94, Madison, Wisconsin, USA*, (2003).

[30] BKL. Fei. *Data Vizualisation In Digital Forensics , Magister Scientia in computer science*. Department of Computer Science,University of Pretoria , South Africa, (2007).

[31] BKL. Fei, JHP. Eloff, MS. Olivier, and HS. Venter. The use of self-organising maps for anomalous behaviour detection in a digital investigation. *in Forensic Sci Int'05, Forensic science international 17th triennial meeting of the international association of forensic sciences, Vol. 162, No. 1-3, pp. 33-37, Hong Kong*, (2006).

[32] G. Galas. études des principaux algorithmes de data mining. Cours, ecole d'ingénieurs en informatique(EPITA), (2009).

[33] P. Gladyshev. *Formalising Event Reconstruction in Digital Investigations, The PhD thesis of Computer Science*. University College Dublin,Irland, (2004).

[34] J. Han and M. Kamber. *Data Mining : Concepts and Techniques*. Second Edition, Morgan Kaufmann, San Francisco, CA, (2006).

[35] S. Haykin. *Neural networks : A comprehensive foundation*. Second Edition, Prentice-Hall, USA, (1999).

[36] S. Hershkop. *Behavior-based Email Analysis with Application to Spam Detection, The PhD thesis of Computer Science*. School of Arts and Sciences, Columbia university, (2006).

[37] G. Hong, J. Bo, and H. Hang. *Forensics in Telecommunications, Information, and Multimedia*. Second Edition,Springer Berlin Heidelberg ,China, (2011).

[38] K. Hung-Jui, W. Shiuh-Jeng, L. Jonathan, and G. Dushyant. Hash-algorithms output for digital evidence in computer forensics. *In bwcca'11, Proceedings of 2011 International Conference on Broadband and Wireless Computing, Communication and Applications, pp. 399-404, Barcelona*, (2011).

[39] S. Jeyaraman and MJ. Atallah. An empirical study of automatic event reconstruction systems. *in DRFWS'06,Proceedings of the 6th annual digital forensic research workshop, Vol. 3, No. 1, pp. 108-115, Lafayette, Indiana*, (2006).

[40] G. Karypis, E. Han, and V. Kumar. Chameleon : A hierarchical clustering algorithm using dynamic modeling. *To Appear in the IEEE Computer*, (1999).

[41] G. Karypis, E.H Han, and V. Kumar. *CHAMELEON : A Hierarchical Clustering Algorithm Using Dynamic Modeling*. The IEEE Computer, Technical Report n°99-107, USA, (1999).

[42] E.E. Kenneally and Clt. Brown. Risk sensitive digital evidence collection. *Digital Investigation, Vol. 2, No. 2, pp. 101-119, Amsterdam, The Netherlands*, (2005).

[43] K. Kent, S. Chevalier, T. Grance, and H. Dang. Guide to integrating forensic techniques into incident response. *Technical Report n°sp 800-86,National Institut of Standards and Technology, Gaithersburg, MD, United States*, (2006).

[44] M.N.A. Khan, C.R. Chatwin, and R.C.D. Young. A framework for post-event timeline reconstruction using neural networks. *Digital Investigation, Vol. 4, No. 3-4, pp. 146-157, USA*, (2007).

[45] A. Khawla, J. Andrew, and A. Thomas. Guidelines for the digital forensic processing of smartphones. *in ADF'11, the Proceedings of the 9th Australian Digital Forensics Conference, Perth, Australia*, (December2011).

[46] M. Kwan, K.P. Chow, F. Law, and P. Lai. Reasonning about evidence using bayesian network. *in IFIP'08, International Federation for Information Processing, Advances in Digital Forensics IV, Vol. 285, pp. 275-289, Hong Kong, China*, (2008).

[47] L. Lamport. Time, clocks and the ordering of events in a distributed system. *Communications of the ACM, Vol. 21, No. 7, pp. 558-565, New York, NY, USA*, (1978).

[48] W. Liu, H. Duan, P. Ren, X. Li, and J. Wu. Wavelet based data mining and querying in network security databases. *Proceedings of the International Conference on Machine Learning and Cybernetics, Vol. 1, pp. 178-182, Xian, China*, (2003).

[49] C.T. Lu, D. Chen, and Y. Kou. Algorithms for spatial outlier detection. *in ICDM'03, Proceedings of the 3rd IEEE International Conference on Data Mining, pp. 597-600, USA*, (2003).

[50] C. Malcom, O. De Vel, A. Alison, and M. George. Gender-preferential text mining of e-mail discourse. *in ACSAC'02, Proceedings of the 18th Annual Computer Security Applications Conference, pp. 282-289, San Diego, California*, (2002).

[51] AJ. Marcella and RS. Greenfield. *Cyber Forensics : A Field Manual for Collecting,Examining and Preserving Evidence of Computer Crimes*. Second edition, Boca Raton, Auerbach, (2002).

[52] P. McCarthy. *Forensic Analysis of Mobile Phones, A thesis submitted for the Bachelor of Computer and Information Science*. University of South Australia, Mawson Lakes, 2005.

[53] P. McCarthy. *Forensic Analysis of Mobile Phones, The thesis Computer and Information Science*. University of Soutn Australia, (2005).

[54] J. Mena. *Investigative Data mining for Security and Criminal Detection*. Second Edition, Butterworth Heinemann, USA, (2003).

[55] H. Mohamadally and B. Fomani. Machines à vecteurs de support ou séparateurs Ã vastes marges. Versailles St Quentin, (2006).

[56] P. Motion. Hidden evidence. *In JLSS'05, The Journal of the Law Society of Scotland, Vol. 50, No. 2, pp.32-34, Scotland*, (2005).

[57] SWGDE Scientificc Working Group on Digital Evidence. *Swgde and swgit digital multimedia evidence glossary*. Scientificc Working Group on Digital Evidence, Technical Report, (2011).

[58] G. Palmer. *A Road Map for Digital Forensic Research*. DFRWS Technical Report n°DTR/T001-2001,Utica ,New York, (2001).

[59] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. *in ICDT'99 7th International Conference on Database Theory, Vol. 1540, pp. 398-416, Jerusalem, Israel*, (1999).

[60] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems, Vol. 24, No. 1, pp. 398-416*, (1999).

[61] C. Prosise, K. Mandia, and M. Pepe. *Incident Response and Computer Forensics*. Second Edition, Corel Ventura,California,USA, (2003).

[62] Q.IU. Data mining method based on computer forensics-based id3 algorithm. *in ICIME'10, The 2nd IEEE International Conference on Information Management and Engineering, pp. 340 - 343, Chengdu*, (2010).

[63] M. Reith, C. Caar, and G. Gunsch. An examination of digital forensic models. *In IJDE'02, International Journal of Digital Evidence, Vol. 1, No. 3*, (2002).

[64] A. Savasere, E. Omiencinski, and S. Navathe. An efficient algorithms for mining association rules in large databases. *in VLDB'95 21th international conference on Very Large Data Bases, pp. 432-444, Zurich, Switzerland*, (1995).

[65] Y. Seema. Analysis of digital forensic and investigation. *In VSRD-IJCSIT'11, International Journal of Computer Science and Information Technology ,Vol. 1, No. 3, pp. 171-178,India*, (2011).

[66] S.Guha, R.Rasogi, and K.Shim. Rock : A robust clustering algorithme for categorial attributes. *0-7695-0071-4/ 99 $ 10.00 IEEE*, (2000).

[67] K. Sindhu and B.B. Meshram. A digital forensic tool for cyber crime data mining. *In ESTIJ'12, Engineering Science and Technology : An International Journal, Vol. 2, N°1, pp.117-124, India*, (2012).

[68] E.H Spafford and B.D. Carrier. *Defining Event Reconstruction of Digital Crime Scenes.* CERIAS Technical Report n°2004-37, Purdue University, West Lafayette, 2004.

[69] R. Srikant. *Fast Algorithms for Mining Association Rules and Sequential Patterns, The thesis of Computer Science.* University of Wisconsin, Madison, 1996.

[70] R. Sriram. Digital forensic research : Current state-of-the-art. *in CSI Transactions on ICT, the Journal of Computer Society of India(Springer) , Vol. 1, No. 1, pp. 91-114, Brisbane, Australia*, (2012).

[71] P. Stephenson. Modeling of post-incident root cause analysis. *in IJDE'03, International Journal of Digital Evidence, Vol. 2, No. 2*, (2003).

[72] J.R. Vacca. *Computer Forensics : Computer Crime Scene Investigation.* Second Edition, Charles River Media, USA, (2005).

[73] H.B. Veena, G.R. Prasanth, R.V. Abhilash, S.P. Deepa, K.R. Venugopal, and P. LM. A data mining approach for data generation and analysis for digital forensic application.

IACSIT'10, International Journal of Engineering and Technology, Vol. 2, No. 3, pp. 313-319, Singapore, (2010).

[74] W. Wang and T.E. Daniels. Network forensic analysis with evidence graphs. *in DFRWS'05, Proceedings of the 5th annual digital forensic research workshop, French Quarter, New Orleans*, (2005).

[75] Y. Wang, J. Cannady, and J. Rosenbluth. Foundations of computer forensics : A technology for the fight against computer crime. *Computer Law and Security, Vol. 21, No. 2, pp. 119-127*, (2005).

[76] S.Y. Willassen. Hypothesis-based investigation of digital timestamps. *in IFIP International Federation for Information Processing, Advances in Digital Forensics IV, Vol. 285, pp. 75-86, USA*, (2005).

[77] S.Y. Willassen. *Methods for Enhancement of Timestamp Evidence in Digital Investigations*. The PhD Doctoral thesis in Mathematics and Electrical Engineering, University of Science and Technology, Norwegian, (2008).

[78] S. Yamuna and N. Sudha Bhuvaneswari. Datamining techniques to analyze and predict crimes. *in IJES'12, The International Journal of Engineering And Science, Vol. 1, No. 2, pp. 243-247, India*, (2012).

[79] X. Yan. Study on the dynamic forensics method based on bp neural network. *in ICFCSA'11, International Conference Future Computer Science And Application, pp. 224-226, Hong Kong*, (2011).

[80] Y.Chung-Huang and L.Yen-Ting. Design and implementation of forensic systems for android devices based on cloud computing. *In Appl. Math. Inf. 2012 Sci An International Journal of Applied Mathematics and Information Sciences, Vol. 06, No. 1, pp. 243-247*, (2012).

[81] A. Zareen and S. Baig. Mobile phone forensics challenges, analysis and tools classification. *in SADFE'10, Proceedings of the 2010 Fifth IEEE International Workshop*

on Systematic Approaches to Digital Forensic Engineering, pp. 47-55, Oakland, CA, USA, (2010).