Evidence Analysis to Basis of Clustering
Approach Based on Mobile Forensic Investigation

Nabila Bermad
Faculty of Exact Sciences
University of Abderrahmane Mira
Bejaia, Algeria
bilabrmd@gmail.com

Mohand Tahar Kechadi
Department of Computer Science and Informatics
University College Dublin
Dublin, Ireland
tahar.kechadi@ucd.ie

Abstract—**Efficiency of mobile investigation process (Smartphone Forensics) is associated with its evidence analysis phase. This phase rests on collection and location of all evidence and their temporal, functional and relational combinations. High volume of these sets evidence, its complexity and size of relations between the different data types may complicate the evidence analysis phase and crime reconstruction. In this paper, we propose a temporal and functional analysis method based on Data mining (unsupervised classification). We introduce a new technique of clustering ascending based on dynamic causality and events reconstruction (SMS and Calls) in time, in this case, we can help an investigator to identify anomalies and information on crime and to provide a global vision of all events through all collected evidences.**

Keywords—Digital Forensic Investigation; Mobile Forensic Investigation; Evidences Analysis; Smartphone; Data Mining; Clustering; Dynamic Causal; Temporal and Functional Analysis.

I. Introduction

Digital Forensic Investigation is one process that uses the scientific methods of information analysis stored to determine the sequence of events leading to a particular incident. This process allows to collect, to examine and to analyze all evidences from numeric support to produce them in the context of legal action.

The purpose of digital forensics is to expose and to present truth, which frequently leads to answers of different questions (Fig.1) relating to digital crime [1].

- When: It refers the crime time interval.
- What: It concerns the activities executed on computer system.
- Who: It concerns the person responsible for the crime.
- Where: It refers the location of evidence.
- How: It treats the manner whose the activities have been achieved.
- Why: It searches the crime motivation.

Digital Forensic Investigation can be classified in three key domains, such as, computer forensics, network forensics and mobile forensics. In this paper, we put the important on this last. The mobile forensic investigation collects all digital evidences from mobile phones in legal conditions using the valid methods.

Fig. 1. Questions of investigating forensics

Digital forensics is the science of obtaining, preserving, analyzing and documenting digital evidence from electronic devices, such as tablet PC, server, digital camera, PDA, fax machine, computers, smart phone and various memory storage devices. According to Casey's definition [2], the "analysis" refers the process of organizing and structuring previously recovered, low-level data into higher-level evidence and narratives of behavior.

The process of investigating forensics (Fig.2) according to NIST model is composed of [4]:

1) Collection: In this phase, it is to identify, to acquire, to store, to transport and to preserve the evidence from various sources of trust in the crime scene (eg. smartphones).

2) Examination: This phase consists in systematic research deepened for evaluation and location of adequate elements from large volume of evidence data while preserving its integrity.

3) Analysis: It is the main and heart of investigating process, the analysis designates process of organization and structuring of exam results while using methods legal and justifiable techniques to derive useful knowledge that address questions to solve the investigation case. It exists three types of analysis [5]:

a) Temporal Analysis: It answers on question when? It is the organization of recovered evidence in time to provide a narrative sequence of events, to identify the anomalies on crime and to lead the other data sources.

b) Functional Analysis: It answers on questions how? And why? This involves determining which entities would be capable for different events related to incident.

c) Relational Analysis: It answers on questions who? What? And where? To show links between entities in crime.

4) Report: It is a phase in which conclusions of analysis phase are documented and presented to authority in the form of investigation report.

The rest of this article is organized as follows. In section 2, we present the motivation and discussions. In section 3, we present a detailed description of our approach. In section 4, we present the performance analysis of our proposed algorithm. Finally, we conclude this work in section 5.

II. MOTIVATION AND DISCUSSIONS

Today, the smartphones become an indispensable tool in all aspects of life. The smartphones are essentially some pocket-size computers, contain a great deal of data, can include, the historic of calls, the lists of contacts, the historic of navigation (internet research), accounts of social networks, calendar/notes, connection (mobile network), WI-FI (wireless fidelity), Bluetooth, maps (location favorites), message text/emails, medias (photos, videos, audios), software (software of document treatment), software of VoIP (Voice Protocol Internet), etc., therefore, the ability to use or to serve the some smartphones for malicious purposes is becoming more evident.

The high volume of smartphone data from the crime scene, as well as, the complexity and the size of the relationships between this type of evidential data have made the evidence analysis phase (step of reconstruction of evidence chains), one of the tasks the more consumer of time and the most complex during an investigation. This is why, data mining is a discipline that could benefit digital forensic. It is a process that extracts a new knowledge through types of models that are: cluster/concept description, characterization, discrimination, association, classification and prediction.

Several methods of data mining are applied to legal numeric investigation, either in stage of evidence collections or the one of evidence analysis, including the supervised classification for the extraction of entity [6, 7] or for the evidences localization [8, 9, 10, 11], the analysis of association for the generation of user profile [12], the analysis by clustering for the research of evidence [13, 14] and their visualization [1]. The clustering is discussed in this paper. The clustering is the process that regroups the set of objects (abstract physical) in similar clusters so that data of same cluster have some similar features, and those belonging to distinct groups are dissimilar [3].

III. TEMPORAL AND FUNCTIONAL ANALYSIS

The proposed approach uses the technique of clustering and the notion of dynamic causality in time to construct a generic and scalable time line (chronology clusters of SMS event and phone calls), in order to provide a comprehensive view of all events related to incident being investigated and to answer the questions when? And how?

The idea to use the unsupervised classification is motivated by fact that several events can be joined by relation causes and effect defined by Lamport [19].

We use the following concepts for a development of our unsupervised classification algorithm [15]:

1) Digital Event: It is one event that changes state of one or several numeric objects.

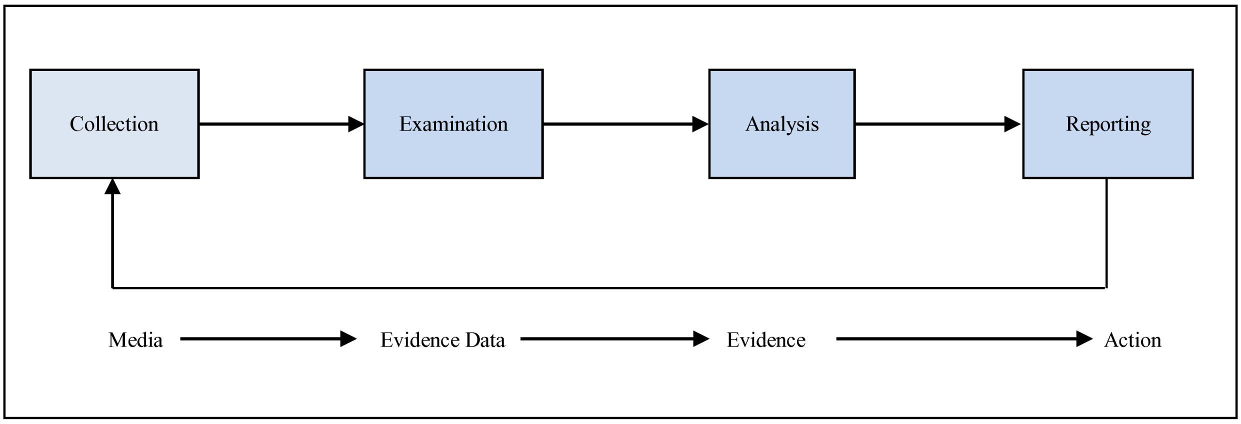

Fig.2. Process of investigating forensics

2) Event Reconstruction: It is one process of events identification occurring in scene of crime using the characteristics of evidence.

3) Evidence Object: An object is evidence of an event if event has changed the state of object. There are two types of evidence items:

a) Cause: An object plays the role of cause if its characteristics have been used in event.

b) Effect: An object plays the role of effect if his state was changed by another object.

4) Events Chains: An event chain is a sequence of events (e_0, e_1, e_2, ..., e_k) such as the effect of event e_i is a cause of event e_{i+1} for $i = 1, 2, ..., k$. A chain of events can be represented by a Direct Acyclic Graph (DAG) as they are illustrated in the Fig.3.

The architecture of our evidence analysis strategy is shown in Fig. 4. After the phase of evidences collection, we exploit the base of proof, the first step is to generate the classification context (clustering context) based on our algorithm that calculates similarity. We define the similarity measures and dissimilarity for group of events (SMS, Calls). For each cluster we are trying to build a chain of events through a hierarchical fusion process in order to build a timeline.

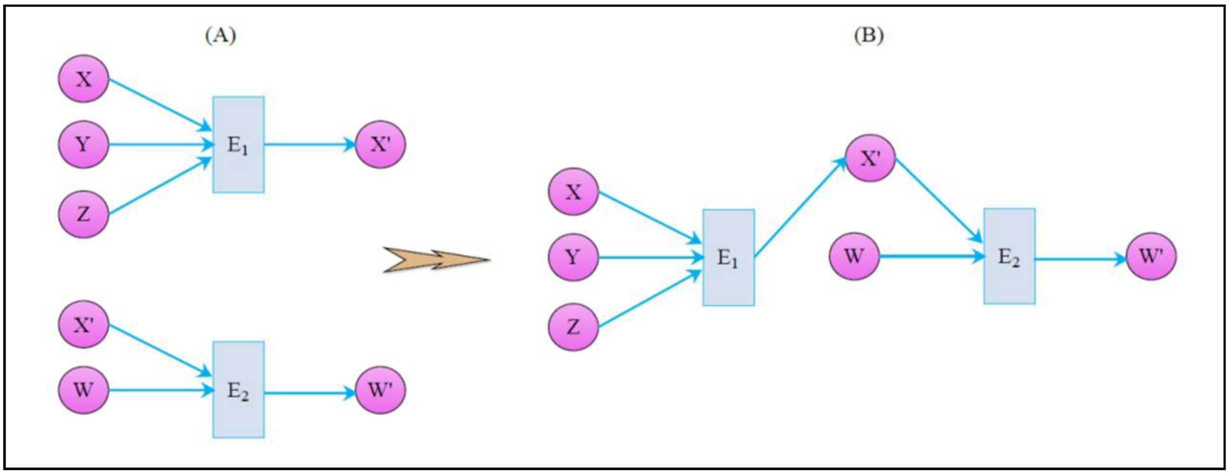

Fig.3. Graphical representation of (A) individual events, (B) chain of events after sequencing [15]

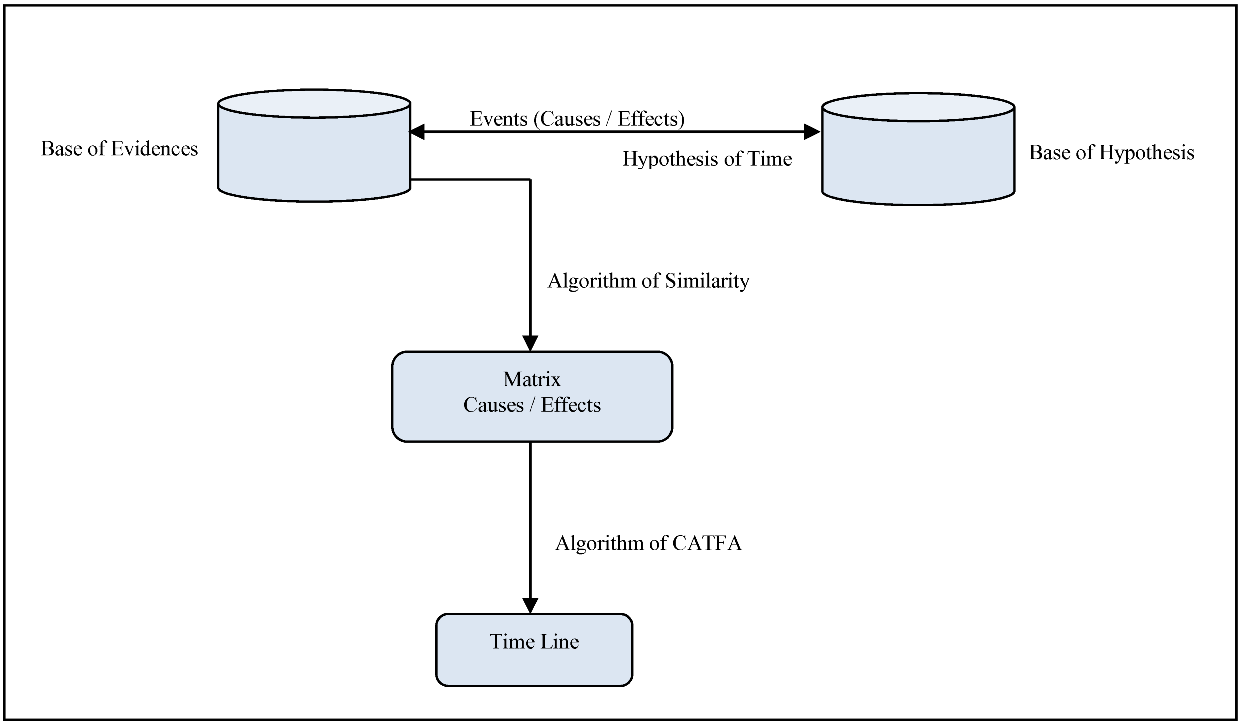

Fig.4. Temporal and functional analysis strategy of evidences

A. Causes/Effects Matrix (Context of Classification)

All events stored in the evidences base are classified to causes and effects in the base of hypothesis. An event on a smartphone is either a communication between a sender and a receiver (e.g. Call) or a communication between a sender and a group of receivers (e.g. SMS). An event e_i is seen as a line in the matrix of similarity composed of cells that correspond to representative attributes PhoneNumberSrc and PhoneNumberDes such as the PhoneNumberSrc is the cause of e_i and PhoneNumberDes is the effect of same event. The general term e_{ij} of this matrix is represented by natural values attached to source phone number and destination phone number of each event. The initial value of e_{ij} is 1 (e_{11}=1), the modification of this value depends on the arrival time of events and the causality principle.

Our algorithm of similarity (Program Matrix of Similarity) allows the filling of similarity matrix according to causality principle and time:

1) Causality (the relation between cause and effect): Can be formally expressed as a mathematical relation between events. We are picking he notion of causality presented in [16] to give a definition of dynamic causality according to events.

2) Time: We assume that each event has a time that associates it and that these moments in time can be controlled using the relations <and =.

The goal of our algorithm is the scheduling of different events according to time and to link between them apply the causality principle. The events e_i $\{1 \leq i \leq n\}$ represent the inputs of algorithm and similarity matrix represents its outputs. The algorithm is proposed in Pascal language.

```
Program Matrix of Similarity
{We will calculate the similarity matrix of events}
Var
N: Integer; {Number of event}
i, j, k: Integer;
matrixSimilarity: array[1..2,1..N] of Integer;
Begin
  j:=1;{The initial value that is to say for the event}
  matrixSimilarity[1,1] := j;
  j := j + 1 ;
  matrixSimilarity[1, 2]:=j ;
  j := j + 1;
  Repeat
    i := 2 ;
    Repeat
      k:=1;
      {Control the ordering of event in time}
      If (MomentEventSrc(e_i) = MomentEventSrc(e_k )) And
      (PhoneNumberSrc(e_i) = PhoneNumberSrc (e_k )) Then
        matrixSimilarity [i,1] := matrixSimilarity [k,1];
        k:=k+1
      Else
        If (MomentEventSrc (e_i) = MomentEventDes (e_k) +
        AjustementTime) And (PhoneNumberSrc( (e_i) =
        PhoneNumberDes(e_k )) Then
          matrixSimilarity [i,1] := matrixSimilarity [k,2]
        Else
          Begin
            matrixSimilarity [i,1] := j;
             j:= j+1;
            If (PhoneNumberDes(e_i) = PhoneNumberDes(e_j )
            Then
              Begin
                matrixSimilarity [i,2] := matrixSimilarity [k,2];
                i := i+1;
              End
            Else
              Begin
                matrixSimilarity [i,2] := j;
                j := j+1;
                i := i+1;
              End;
          End;
    Until (k = i-1);
  Until (i >N);
End.
```

The principle of filling the similarity matrix according to context of extraction and application of algorithm is represented in the TABLE I:

TABLE I. CAUSES/EFFECTS MATRIX

	Cause	effect
e_1	1	2
e_2	1	3
e_3	1	4
e_4	5	4
e_8	4	6
e_5	2	7
e_6	3	7
e_7	3	8
e_{11}	6	9
e_{10}	8	9
e_9	7	9

This table illustrates an example of the similarity matrix of eleven events that are ordered in time, it represents an input to next clustering algorithm (CATFA) and its values are natural. The events e_i and e_j are causally dependent if effect value of e_i equal cause value of e_j, via these values, we can cluster the events that are causally dependent, for example, e_1 (Effect=2) with e_5 (Cause=2) to build the cluster $\{e_1, e_5\}$, and e_5 (Effect =7) with e_9 (Cause=7) to build the cluster $\{e_5, e_9\}$.

B. Similarity and dissimilarity measure

1) *Similarity measure:* EN is a matrix Events-Numbers such as $E = \{ e_i , 1 \leq i \leq N\}$ represent the lines and numbers K = PhoneNumberSrc (Cause), PhoneNumberDes (Effect) are the columns. We assume that each event e_i is represented by its two corresponding values (Cause and Effect) in the similarity matrix: $e_i = [e_{i1}, e_{i2}]$. This model description can compare two events. For example, we can consider the events e_1 and e_2 as highly similar if their vectors $[e_{11},e_{12}]$ and $[e_{21},e_{22}]$ present an equality of values between the PhoneNumberDes (Effect) for (e_1) and PhoneNumberSrc (Cause) for (e_2).

2) *Similarity and dissimilarity between events:* We define the relation of similarity and dissimilarity between two events by two functions $\sigma_{Sim(ei,ej)}$ and $\sigma_{Dissim(ei,ej)}$, which respectively measurethe similarity and dissimilarity between two events e_i and e_j according to attributes cause and effectr espectively:

$$\sigma_{Sim(ei,ej)} = \begin{cases} 1 & \textbf{If } \; e_{i2} = e_{j1} \; (i \leftarrow 1 \text{ to } N, j \leftarrow i+1) \\ 0 & \textbf{Otherwise} \end{cases} \quad (1)$$

$$\sigma_{Dissim(ei,ej)} = \begin{cases} 0 & \textbf{If } \; e_{i2} = e_{j1} \; (i \leftarrow 1 \text{ to } N, j \leftarrow i+1) \\ 1 & \textbf{Otherwise } (e_{i2} \neq e_{j1}) \end{cases} \quad (2)$$

The first function defines the similarity between e_i and e_j according to attributes PhoneNumberSrc(cause), PhoneNumberDes(effect), two events e_i, e_j are considered similar if there have a phone number common between these two events(effect=cause in that order). The second function defines the dissimilarity between two events e_i and e_i. The two events e_i and e_j are considered dissimilar if there is not a phoneNumber common between two events, so there is a time sequencing between e_i , e_j ($e_i \rightarrow e_j$).

3) *Similarity and dissimilarity between a set of events:* As we have defined it for two events, we introduce two functions that take in account degree of similarity and dissimilarity between two sets of events.

E_a and E_b are two subsets of the set of events E. to give account level of similarity (respectively, dissimilarity) between sets of events, we use the function Sim (E_a,E_b) (respectively, Dissim (E_a,E_b) that determines the number of similarity (respectively, dissimilarity) between two sets of events E_a and E_b ($E_a \neq E_b$) as follows:

$$\text{Sim}(E_a,E_b) = \begin{cases} 1 & \textbf{If } \sum_{e_i \in E_a,\ e_j \in E_b} sim(e_i,e_j) = \\ & \quad (\text{Card }((E_a \cup E_b)\text{-}1) \times \\ & \quad \text{Card }((E_a \cup E_b))/2)) \\ 0 & \textbf{Otherwise} \end{cases} \quad (3)$$

$$\text{Dissim}(E_a,E_b) = \begin{cases} 1 & \textbf{If } \sum_{e_i \in E_a,\ e_j \in E_b} sim(e_i,e_j) \neq \\ & \quad (\text{Card }((E_a \cup E_b)\text{-}1) \times \\ & \quad \text{Card }((E_a \cup E_b)/2) \\ 0 & \textbf{Otherwise} \end{cases} \quad (4)$$

C. Clustering CATFA (Clustering Algorithm for Temporal and Functional Analysis) for the Generation of Timeline:

We propose a strategy of events reconstruction while exploiting the unsupervised classification. The problem is to group events with their causes and effects in order to obtain clusters of events ordered in time. These clusters are represented as directed acyclic graphs, which is why we use an ascending hierarchical clustering method. Moreover, we do not know a priori the number of cluster to be obtained its depends on the evidences base

The unsupervised classification algorithm that we propose is based on the principle of the Chameleon algorithm proposed by Karypis et al. [17] Chameleon is a hierarchical aggregation algorithm based on a dynamic model. We choose this algorithm because it allows not only to classify events in a way that interests us easily to integrate our measure of similarity to dynamically build the largest sequence of clusters ordered in time but also to integrate the notion of Event graph obtained during reconstruction of classification process during the sub-graph merging step (each cluster is a sub-graph)

1) *Hypotheses:* The clustering process is formalized as a search graph process ordered events. The algorithm of clustering (Program CATFA) is based on the following hypotheses:

- Each event can have a multiple causes and a multiple effects.
- We suppose that causality is preserved in time,i.e: MomentEventSrcc(e_i) < MomentEventSrcc(e_j) and MomentEventDesc(e_i) < MomentEventDesc(e_j) does not imply that $e_i \rightarrow e_j$, because, the different events can happen at different moments are related by →.

2) Notation:

- NotExist(cluster(i,j)): To verify the existence of cluster(i,j).
- EN: The similarity matrix (Event-Phones Numbers).
- EventFirstCluster(j,t):This method is used to return the first event of cluster j of size t.
- EventLastCluster(i,t): This method is used to return the last event of cluster i of size t.

```
Program CATFA
Var
  N: Integer; {It is the number of event}
  i, j, k, ClusterStage_i: Integer;
  ok, ok1:Boolean;
Begin
  k:=0;
  For  i=1 to N Do
    {To place each event in its own cluster};
    Begin
      GenerateCluster() ;
      k :=k+1;
      Cluster(k) :=e_i;
    End;
    ClusterStage_i := 1;
    i := ClusterStage_i; {To join the clusters} ;
    Cluster(i) ; {Method of clustering };
    Repeat
      ok := true;
      ok1:= true;
      j:=  i+1 ;
      While (ok) Do
        If ( (EN[EventLastCluster(i,t),2] =
          EN[EventFirstCluster(j,t),1]) and  (NotExiste
          (cluster(i,j))) Then
            Begin
              MergeCluster(i,j) ;
              ok1:= false ;
              k:= k+1 ;
              Cluster(j) ;
              j++ ;
            End
        Else
          If (ok1) Then
            j++
          Else
            Begin
              Cluster (j);
              j++;
              If  (j > N ) Then
                Begin
                  ok:=  false;
                  i++;
                  ClusterStage_i :=  N+1 ;
                  i :=  ClusterStage_i;
                  If (inside(interval Analysis)) Then
                    N:= N+K;
                End;
            End;
    Until (i > N);
End.
```

This algorithm has the some following interesting properties:

- Complexity of calculation is relatively low because, it is polynomial O (N^2) following the number of events.
- Results of calculation are precise because, every event is attached in its class adequate (the rate of mistakes is weak).

The algorithm is efficient (scalable) with a large number of events because it is parallel. It allows building several clusters of events at the same time, and it also permits the classification of other events causally dependent on the same event or several events dependent of this one, in addition, it allows memorizing the merged clusters at every stage to make the sweep on classes that are similar to fusion step.

In Fig.5, we illustrate the first steps of unsupervised classification of Causes/Effects Matrix (TABLE.I) to build clusters according to principle of our algorithm (Clustering CATFA). At the beginning of clustering, all events are clusters such as $\{\{e_1\}, \{e_2\}, \{e_3\}, \{e_4\}, \{e_5\}, \{e_6\}, \{e_7\}, \{e_8\}, \{e_9\}, \{e_{10}\}, \{e_{11}\}\}$, thereafter, the events $\{\{e_1\}, \{e_5\}\}$ can construct this cluster $\{e_1, e_5\}$ and $\{\{e_3\}, \{e_8\}, \{e_6\}\}$ can construct this cluster $\{e_3, e_8, e_6\}$ according to causality principle. These clusters $\{\{e_1, e_5\}, \{e_3, e_8, e_6\}\}$ are built at the same time according to principle of parallelism.

IV. Simulation results

In this section, we present the parameters and assumptions related to the environment of simulation, and then we present and discuss the obtained results

A. Parameters and assumptions:

We have implemented our simulations using Java environment. We have simulated a system of evidence analysis which analyzes the proofs of 1000 SMS and Call events of a crime, once the evidence was acquired from smart phones through the acquisition tool, we stored the SMS and calls as records in a database, and we identify the correlations and the associations in evidences (SMS, Calls). We constructed the events ordered in time to generate the timeline.

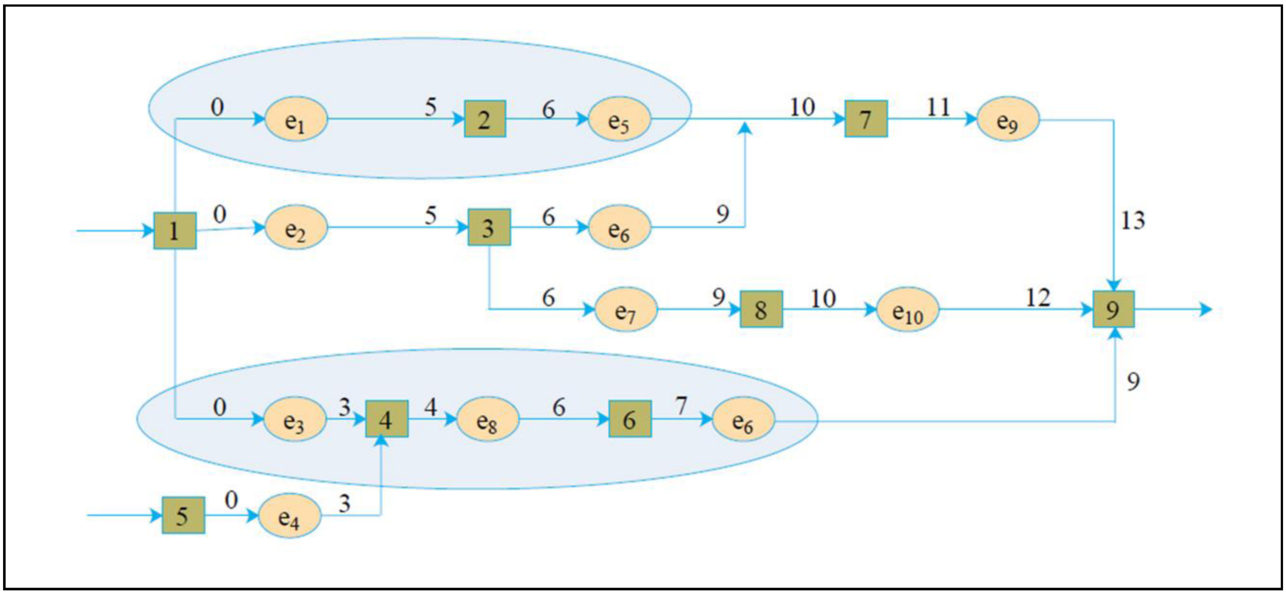

Fig.5. Principle of events classification

B. Results of comparison:

In Fig. 6, we illustrate the results of the comparison in terms of cluster generation time for a number of events. We measure the evolution in time of cluster generation of our proposed algorithm to construct the time line according to number of SMS and Call events of evidence base. We compared the performances of our algorithm with ROCK [18] and CHAMELEON [17]. Results show a better scalability of our approach.

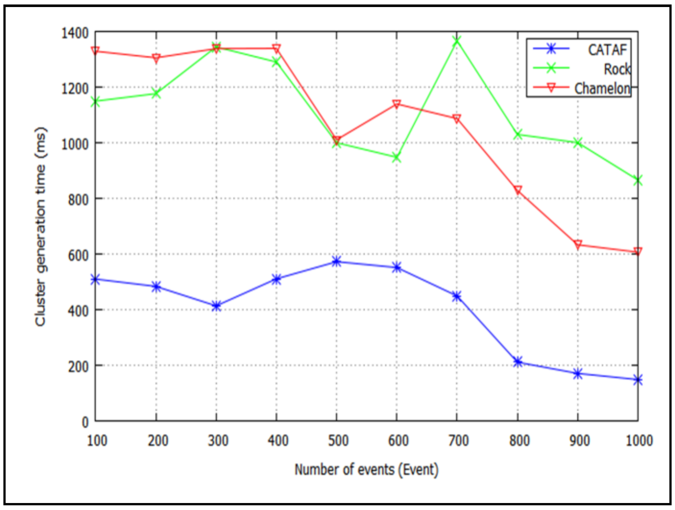

Fig. 6. Time of clusters generation versus number of events

We noticed that our approach (CATFA) offers high performance in comparison with the algorithm of ROCK and CHAMELEON, for example, the algorithm of ROCK has needs 1000 MS to generate the clusters corresponding to 800 events and algorithm of CHAMELEON has needs 800 MS to generate the clusters corresponding to 800, whereas, our algorithm has needs 200 MS to generate the clusters corresponding to 800 events. Therefore, we have gained time at moment of cluster generation. This gain is due to the parallelism principle used to construct all clusters with causally dependent events in the same iteration and to merge events causally dependent at the time of dendrogram construction. Whereas, the algorithms of ROCK and CHAMELEONS don't exercise any parallelism, they based on a simple sweep of evidence base to construct their dendrogram (construction and fusion of clusters), the growth of events number in evidence base misleads to reduction in time of these approaches. For our approach it is reasonable because of parallelism, in two algorithms, this reduction is on the other hand uncertain and induced to total sweep at every iteration to construct a number of clusters. This number is varied according to similarity between the generated classes.

V. Conclusion and Perspectives

The paper presents, a technique to automatically exploit the information of cellular phones with goal of helping crime investigation. We proposed in this article an approach based on data mining for temporal and functional analysis of proofs in the setting of legal investigation smartphone. Our strategy exploits the results of unsupervised classification to generate a timeline that is the set of clusters of event (SMS and Calls) ordinates in time, our algorithm is based on notion of dynamic causality, as well as, one of time, to spread our approach to treat different types data exit of smartphones as pictures, video, email account…etc.

REFERENCES

[1] B. Fei, "Data Vizualisation In Digital Forensics," Magister Scientia in computerscience. Department of Computer Science, University of Pretoria, South Africa , 2007.

[2] A. Morum, S.F. Caus, 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.

[3] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques, Management Systems (The Morgan Kaufmann Series in Data Management Systems)," Second Edition, Morgan Kaufmann, 2006.

[4] K. Kent , S. Chevalier, T. Grance, and H. Dang, "Guide to Integrating Forensic Techniques into Incident Response,"NIST SP 800-86,National Institut of Standards and Technology, Gaithersburg, MD, United States, 2006.

[5] B.D. Carrier, and E.H. Spafford, "Anevent-based Digital Forensic Investigation Framework," Center for Education and Research in Information Assurance and Security Proceedings of the fourth Digital Forensic Research Workshop, USA, Baltimore, 2004.

[6] O.D. Vel, A. Anderson, M. Corney, and G. Mohay, "E-mail Authorship Attribution for Computer Forensics," First Edition, Kluwer Academic, Boston, Dordrecht, London, 2002.

[7] C. Malcolm, O.D. Vel, A. Alison, and M. George, "Gender-Preferential Text Mining of E-mail Discourse," Proceedings of the 18th Annual Computer Security Applications Conference, pp. 282-289, San Diego, California, 2002.

[8] R. Brown , B. Pharm, and O.D. Vel, "Design of a Digital Forensics Image mining System," Proceedings of the International Workshop on Intelligent Information Hiding and Multimedia Signal Processing,Vol. 3683, pp. 395-404, 2005.

[9] Z. Liu, D. Lin, and F. Guo, "A Method for Locating Digital Evidences With outLier Detection Using Support Vector Machine," International Journal of Network Security, Vol. 6, N°. 3, pp. 301-308, 2008.

[10] Q.Lu, "Data Mining Method Based on Computer Forensics-based ID3 Algorithm," Published in The 2nd IEEE International Conference on Information Management and Engineering (ICIME), pp. 340-343, 2010.

[11] X. Yan, "Study on the Dynamic Forensics Method Based on BP Neural Network," International Conference on Future Computer Science And Application, pp. 224-226, 2011.

[12] T. Abraham, and O.D. Vel, "Investigative profiling with Computer Forensic log Data and Association Rules," Proceedings of the IEEE International Conferenceon Data Mining, pp. 11-18, 2002.

[13] S. Decherchi, S. Tacconi, J. Redi, F. Sangiacomo, A. Leoncini, and R. Zunino, "Text Clustering for Digital Forensics Analysis," Computation al Intelligencein Security for Information Systems, Vol. 63, pp. 29-36, 2009.

[14] H.B. Veena, G.R. Prasanth, R.V. Abhilash, S.P. Deepa, K.R. Venugopal and L.M. Patnaik, "A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application," International Journal of Engineering and Technology, Vol. 2, N°. 3, pp. 313-319, 2010.

[15] B.D. Carrier and E.H. Spafford, "Defining Event Reconstruction of Digital Crime Scenes,"Journal of Forensic Sciences, Vol. 49,N°. 6, 2004.

[16] 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.

[17] G. Karypis , E.H. Han , and V. Kumar, "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling," IEEE Computer, Technical Report n°99-007, USA, 1999.

[18] S. Guha, R. Rastogi and K. Shim, "Rock:A Robust Clustering Algorithm for Categorial Attributes," Vol. 25, No. 5, pp. 345-366, Published by Elsevier Science Ltd, 2000.

[19] L. Lamport, "Time, clocks, and the ordering of events in a distributed system," Communication of the ACM,volume 21, Number 7, Pages 558-565, New York, USA, 1978.