Scalable Analytics Framework for Context Driven Forensics Linguistic Investigation on Unstructured Data

Mots clés :
- Directeur de thèse : salima BENBERNOU
- Co directeur de thèse : salima BENBERNOU
- Co-encadrant(s) :
- Unité de recherche : Laboratoire d'Informatique PAris DEscartes
- École doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Sciences et technologies de l'information et de la communication

Résumé du projet de recherche (Langue 1)

Unstructured data is information that either does not have a predefined data model or is not organised in a predefined manner. Today, 80% of the data in most of the organizations is unstructured. Organizations need to understand the types of unstructured data they are accumulating and the best ways to process and store this data for complex analysis to gain advantage. Without effective data management strategies and guidance, organizations run the risks of not capitalizing on unstructured data.

Typically, unstructured information is text-heavy that represent subjective opinions however, it can be non-textual, and human- or machine-generated. Unstructured information contain critical data such as name, dates, locations, entities, actions, behavior, and facts. These data are very often suggestive, sensitive, and influential ingredients of mission-critical tasks. Digital forensics investigation is an example of crucial task which infer evidences related to an illicit event or actions. Traditionally, digital forensics investigation is carried out through examination of digital devices such as compter, cell phone, etc. to extract knowledge or evidence. However, forensics analysis of data – which could be highly effective in extracting intelligence or evidence – has been overlooked entirely until today. This type of analysis involve large volume of unstructured text stored within the repository of organizations and also data from the Web and social medias which do not have any predefined structure as well.

Résumé du projet de recherche (Langue 2)

Unstructured data is information that either does not have a predefined data model or is not organised in a predefined manner. Today, 80% of the data in most of the organizations is unstructured. Organizations need to understand the types of unstructured data they are accumulating and the best ways to process and store this data for complex analysis to gain advantage. Without effective data management strategies and guidance, organizations run the risks of not capitalizing on unstructured data.

Typically, unstructured information is text-heavy that represent subjective opinions however, it can be non-textual, and human- or machine-generated. Unstructured information contain critical data such as name, dates, locations, entities, actions, behavior, and facts. These data are very often suggestive, sensitive, and influential ingredients of mission-critical tasks. Digital forensics investigation is an example of crucial task which infer evidences related to an illicit event or actions. Traditionally, digital forensics investigation is carried out through examination of digital devices such as compter, cell phone, etc. to extract knowledge or evidence. However, forensics analysis of data – which could be highly effective in extracting intelligence or evidence – has been overlooked entirely until today. This type of analysis involve large volume of unstructured text stored within the repository of organizations and also data from the Web and social medias which do not have any predefined structure as well.

Informations complémentaires (Langue 2)

There are different types of challenges involved in this research. Some of the challenges can be tackled by the recent advancement of technologies. For instance, the scalability today can be tackled by using technologies that facilitate massive-scale data management in scalable architecture such as Hadoop Distributed File system. The major challenge lies within the investigation of unstructured text. According to our literature review, there are many natural language processing solutions. The most recent one is called BERT from Google. However, all the existing approaches including BERT are pre-trained algorithms which essentially are static approaches. The cognition of targeted terms cannot be achieved unless otherwise the algorithms are trained adequately. To sum up, the existing solutions are missing ability to cognize criminal and legal terms that are essential for forensics study of the text and apparently they cannot facilitate computational intelligence in terms of measuring the sensitivity of the context that would drive criminal investigation with high accuracy, transparency of evidence and evident legal components.