Digital Evidence Identification/Classification Study Using Causal Relationship Structure Information

Jong Jin Jung, Jong Bin Park, Ji Hyun Lee

***Abstract*—In this paper, we propose causal information necessary to effectively identify key clues and causes in order to analyze the causal relationship of the crime by extracting words and vocabulary corresponding to the criminal context from evidence files obtained for digital evidence analysis.**

I. Introduction

While traditional crimes are decreasing, cyber crimes are rapidly increasing due to the non-face-to-face economy and the spread of digital culture. In particular, cyber fraud/financial crimes, which require analysis of various digital evidence such as SNS and financial transactions, increased by 9.3% in 2022 compared to the previous year. Due to the spatial nature of digital communication channels, the rapid increase in crime-related information, such as short conversations between parties involved and slang, increases the need to effectively recognize and interpret it. Additionally, due to the unstructured nature of most crime data, the accuracy and scope of crime prediction and classification are determined by the accuracy and scope of the crime information.[1] In this study, we propose a causal information composition system to effectively identify key clues and causes for analyzing the causal relationship of crime by extracting words and vocabulary corresponding to the crime context from digital crime evidence in cyberspace.

II. Causal Relationship Information Analysis

Causal relationship configuration information is a predefined schema to know what key information to extract from digital crime evidence and what relationships exist between these clues. In this paper, this information is used to extract key causal information and use it for search. At the time of securing digital evidence through seizure and search from digital devices possessed by major criminal suspects, all possible records that may indicate potential criminal charges, such as documents and voice records, are secured. However, many of the evidence obtained are files unrelated to the investigation, or even in one file, there are many unnecessary parts that are not related to criminal charges, such as daily conversations.[2] Therefore, causal analysis should be conducted with only information related to the case, excluding information irrelevant to the investigation, to induce concentration of investigative capabilities. As digital media and means of digital communication increase, various types of evidence containing criminal circumstances exist in the form of audio and video. Analysts working at the investigation site talk about not only evidence that exists in the form of text such as documents and messenger conversations, but also a lot of suspected criminal content in recordings of voice conversations between suspects, captured web data, and CCTV footage, all of which are automatically analyzed. Therefore, there is a high demand for causal automatic connection analysis.

III. Pre-training Language Model Needed to Extract Causal Information

In order to identify/classify only those with high relevance to the characteristics of the case under investigation, digital evidence and investigation documents/data data are additionally trained on Pol-BERT_small, a pre-learning language model related to general crimes specialized in the public security domain. -tunning and specialized language model (Pol-Robust-BERT) is required.[3] Using these two models, it is possible to select digital evidence according to investigation characteristics after identifying the contents of individual evidence through key word/theme analysis using Document-to-Sequence. Evidence selected using the two BERT models consists of multimodal data such as messenger conversation capture screen voice conversation recordings and financial transaction record photos, which are converted into structured text for analysis. The structured text is extracted as triple (SPO) type information to be used in the knowledge graph by using NLP technology such as entity name/ relationship analysis based on deep learning.[4]

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2023-00225661, Development of causal reasoning and expression technology to enhance the proof of digital evidence)t

References

[1] Hee-Dou Kim, Heuiseok Lim, A Named Entity Recognition Model in Criminal Investigation Domain using Pretrained Language Model, Journal of The Korea Convergence Society, Vol. 13, No 2. Pp. 13-20, 2022

[2] K. R. Rahem & N. Omar. (2014). Drug-related crime information extraction and analysis. Proceedings of the 6th International Conference on Information Technology and Multimedia, pp.250-254. DOI : 10.1109/ICIMU.2014.7066639

[3] A. Alkaff & M. Mohd. (2013). Extraction of nationality from crime news.Journal of Theoretical and Applied Information Technology, 54, 304-312.

[4] S. Sathyadevan, M. S. Devan & S. S. Gangadharan (2014). Crime analysis and prediction using data mining. 2014 First International Conference on Networks & Soft Computing (ICNSC2014),406-412.DOI : 10.1109/CNSC.2014.6906719.

*Research supported by Korea MIST.

F. A. JongJin Jung is with the Korea Electronics Technology Institute, Korea (corresponding author to provide phone: +82-02-6388-6651; fax: +82-02-6388-6659; e-mail: mozzalt@keti.re.kr).

S. B. JongBin Park and T.C JiHyun Lee are with the Korea Electronics Technology Institute, Korea (e-mail: jpark@keti.re.kr, jihyunlee@keti.re.kr).