OPEN ACCESS 23

ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 6, Issue. 1, June 2020. pp. 23-30
doi: 10.21108/IJOICT.2020.61.481

Keyword Indexing And Searching Tool (KIST): A Tool to Assist the Forensics Analysis of WhatsApp Chat

Syafiqah Hanisah Shahrol Nizam [1], Nurul Hidayah Ab Rahman [1], Niken Dwi Wahyu Cahyani [2*]

[1]*Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia*
Batu Pahat, Johor, Malaysia

[2]*School of Informatics*
Telkom University, Bandung, Indonesia.

[*] nikencahyani@telkomuniversity.ac.id

Abstract

Digital forensics is a field that concerned with finding and presenting evidence sourced from digital devices, such as computers and mobile phones. Most of the forensic analysis software is proprietary, and eventually, specialized analysis software is developed in both the private and public sectors. This paper presents an alternative of forensic analysis tools for digital forensics, which specifically to analyze evidence through keyword indexing and searching. Keyword Indexing and Searching Tool (KIST) is proposed to analyze evidence of interest from WhatsApp chat text files using keyword searching techniques and based on incident types. The tool was developed by adopting the Prototyping model as its methodology. KIST includes modules such as add, edit, remove, display the indexed files, and to add WhatsApp chat text files. Subsequently, the tool is tested using functionality testing and user testing. Functionality testing shows all key functions are working as intended, while users testing indicates the majority of respondents are agree that the tool is able to index and search keyword and display forensic analysis results.

Keywords: Forensic analysis, keyword indexing, keyword searching, WhatsApp

I. INTRODUCTION

Digital forensics is a field of analyzing and interpreting electronic data to be used as digital evidence in a court of law (Reith, Carr, & Gunsch, 2002). The increasing use of computing devices posed digital forensics as a significant field to gather digital evidence that may be used to prosecute or to defend a suspect in a court. For instance, a personal computer that belongs to the suspect or the victim may indicate to the criminal activity, in which a forensic analysis on a personal computer is conducted to provide evidence that can be presented in the court proceedings. In Malaysia, for example, the government of Malaysia has amended the Malaysian Evidence Act 1950 in 1993 to include a section on electronic devices (Amandeep, 2012).

The digital evidences are collected from various sources such as Internet-based, stand-alone computers or devices, and mobile devices to be used in legal proceedings (Reith et al., 2002). The emerging of computing infrastructure (e.g. mobile computing, smartphones) resulting in various data formats and computing resources, thus posed challenges to digital forensics practices (Ab Rahman et al., 2017). Furthermore, the

Received on 23 Mar 2020. Accepted on 19 May 2020

increasing use of smartphones and many Instant Messaging (IM) apps (e.g. WhatsApp) led to a new landscape of cybercrimes platforms such as spreading fake news, spam messages, and cyberbullies. A study conducted by the Malaysian Communications Multimedia Commission (MCMC) showed that 84% of Malaysians received fake news via WhatsApp, followed by Facebook (8%), and remaining got from blogs (NSTP Team, 2018). With the growth popularity of WhatsApp app, it is undeniable that spreading fake news might involve a big impact and on a large scale. As an example, 30 people were arrested in India after a 2000 individual mob became rage by WhatsApp rumors of a child kidnapping and lynched by a man in the year 2018 (Ghosh, 2018).

As a result, the evidence of interest would be acquired from the suspects' WhatsApp chat history in a forensically sound manner. The evidence artifacts are subsequently being analyzed to find clues for a cybercrime. However, forensic analysis of the IM apps can be a quite complex task if the number of forensic data is increasing by thousands of gigabytes to be handled with limited storage, low CPU performance, time-consuming, and lack of keyword indexing (Carrier, 2003). These issues could affect directly or indirectly investigation costs, for instance, time-consuming when searching for keywords with limited storage to handle a large amount of data.

Keyword searching is an example of a technique that is used during a forensic analysis to browse for a word or a combination of words from a keyword index (i.e., a list of data, such as databases) that is provided for the indexer to load some words while performing keyword searching (Beebe & Dietrich, 2007) within digital evidence. We argued that keyword searching and indexing technique need to be integrated into analyzing evidence artifacts from WhatsApp chats' history to look for evidence of interest based on a particular incident type (e.g., harassment, trafficking, and fraud). Keywords associated with the incident types can be loaded into the keyword index to be used by investigators. For example, "banana split" and "candy man" are related keywords in the drug type of crime (DFIR Training, 2019).

In this study, therefore, a tool for keyword indexing and searching the WhatsApp chat text files based on incident types is proposed. The objectives are in three-fold:

1) to design a tool for keyword indexing and searching based on incident types for analyzing WhatsApp chat text files,
2) to develop the keyword indexing and searching tool using Microsoft Visual Studio, and
3) to test the functionalities of keyword indexing and searching tool.

The remaining of this paper is organized as follows: Section II describes a review of the study background. Section III presents the project methodology in which the Prototyping model is adopted to develop the proposed tool. Section IV presents results and discussion, and finally, Section V concludes the work and highlights directions for future works.

II. BACKGROUND OF THE STUDY: A REVIEW

A. Digital Forensics Investigative Process

The investigative process is structured to encourage a complete, accurate investigation, ensure proper evidence handling, and scale down the probability of errors created by preconceived theories and other possible pitfalls. The investigators and examiners work together in a systematic and determined way to deliver accurate and reliable evidence in the courtroom (Ademu & Imafidon, 2013). In the first of Digital Forensics Research Workshop (DFRW), digital forensics investigation is outlined into six phases (Palmer, 2001), as follows:

1) Identification - identify digital evidence from indicators and determines its type.
2) Preservation - the digital evidence found at the crime scene is preserved for further analysis.
3) Collection - record the physical scene and duplicate digital evidence using appropriate procedures.
4) Examination - identifying the potential evidence in order to do an in-depth systematic search with the related evidence to the suspected crime.
5) Analysis - forensics investigator uses the appropriate tool to analyze the digital evidence.
6) Presentation - the last phase that summarizes and explains the conclusion of the investigation process.

This study involves the analysis phase in which forensic investigators use the proposed tool to undertake forensic analysis in WhatsApp text file.

B. Keyword Indexing and Searching

Keyword indexing refers to the process of making articles able to be found in databases, while keyword searching is the process of looking for words in any form of documents. It is one of the techniques used in the forensic investigation to identify specific digital evidence from suspects' computers or digital storage that involves search parameters such as Boolean search, fuzzy search, and regular expression search (Mishra, 2007).

According to Carlson (2006), the Boolean search is a type of search using operators or modifiers like AND, NOT and OR that allows the investigator to combine keywords and produce more relevant results; fuzzy search can be used to return the appropriate words that are very close to the misspelled or partial keywords, and a regular expression is a method used for pattern matching with a special text string. Besides, a regular expression also enables the investigator to find any matching keywords in the form of phone numbers, IP addresses, URLs, and can be used to search for an email address.

TABLE 1
COMPARISON STUDY OF EXISTING DIGITAL FORENSIC KEYWORDS INDEXING AND SEARCHING

Forensic tools / **Features**	**OsForensics**	**Autopsy**	**Belkasoft Evidence Center**	**Kist-Keyword Indexing And Searching Tool**
Platform	Windows, Linux	Windows, Linux	Windows	Windows
Advanced search features	Keyword search, regular expression, hex search	Keyword search, Regular expression	Keyword search, regular expression, hex search	Keyword search, regular expression
Focus search areas	Search for files, emails, filenames, unallocated space and Internet history on disk image.	Search for files, unallocated space, filenames, files within archive files and Email search on disk image.	Search for unallocated space, files, filenames, Internet history, files within archive files, email search and file slack on disk image	Search keywords by specific incident types on chat history in communication database image.

C. Comparison of tools

This section reviews the existing forensic analysis tools that involve the keyword search features and compare them with our proposed tool (i.e., Keyword Indexing and Searching Tool (KIST)). OSForensics, Autopsy, and Belkasoft Evidence Centre are the three open-source forensic tools that have been compared with KIST as a benchmark (see Table 1). These open-source tools were selected due to their openness and availability to the researchers. It should be noted that the comparison is undertaken in terms of its functionality and does not involve any data.

OSForensics from PassMark Software is a built-in Timeline Viewer for e-discovery functions. Lahaie et al. (2012) observed that OSForensics used the advanced hashing algorithm with SHA-1, MD5, CRC32, and SHA-256 hashes to allow an investigator to calculate the hash file and compare it to validate a file has not been modified. OSForensics can perform full-text searches within files by its filename, size and date created, and support search within email archives. An index file is created to be able to search whether by files, images, or emails. This tool allows us to index various types of file formats such as DOC, PDF, PPT, XLS, RTF, WPD, SWF, JPG, GIF, PNG, MP3, ZIP, and more. The Search Index feature allows the analysis process of searching the contents of many files, make it easier to locate texts, and less time to wait for analysis results.

Autopsy provides an open-source digital forensics platform and graphical interface used by The Sleuth Kit and developed by Brian Carrier. The results of the forensic search are displayed in the graphical user interface, which makes it easier for investigators to point out important data. Autopsy's keyword searching features are strong, where the investigator can search the input for a keyword search as plain strings or a keyword search with regular expression. The Autopsy keyword search module supports regular expression mode, which permits the investigator to search for the phone number, IP addresses, URLs, and Email addresses (Basis Technology, 2015).

The Belkasoft Evidence Center developed by a global company in digital forensics technology, Belkasoft. Data evidence needs to be extracted into BEC software, and it will filter the number of items to review. All texts, dates, times, metadata, and everything in the data evidence will automatically be indexed. The BEC software offers several types of search options such as search by a word or phrase, words from files, regular expression, and predefined search.

Based on Table 1, we argued that the Belkasoft Evidence Center (BEC) is the most optimum as it offers many search areas advanced search features. However, none of the tools come out with a target search area where it is specifically according to incident types. Thus, we proposed KIST as a tool with the implementation of keywords searching that is based on incident types.

III. DESIGN AND DEVELOPMENT OF KIST

We adopt the Prototyping model to develop the KIST as the model allows users to interact with a prototype, which gives the users the experience of how the actual system works (Dennis, Wixom, & Roth, 2013). The KIST development started with a planning phase that collected the input and output requirements of the system and began to plan a prototype.

Analysis phase involves comparative studies of the existing systems in the focused area of keywords search for digital forensic analysis tools. Some similar systems were selected to be compared in terms of their functions, which can be improved and implemented in the proposed tool. The available keyword datasets were reviewed from the National Institute of Standards and Technology (NIST) and the Digital Forensics and

Incident Response (DFIR) Training websites. Five types of crime were selected from the DFIR website that includes Rape, Fraud, Hate, Trafficking, and Harassing.

In the design phase, the Unified Modelling Language (UML), such as the use case diagram, sequence diagram, and activity diagram was illustrated to help during the software design process. An example of a use case diagram is presented in Figure 1.

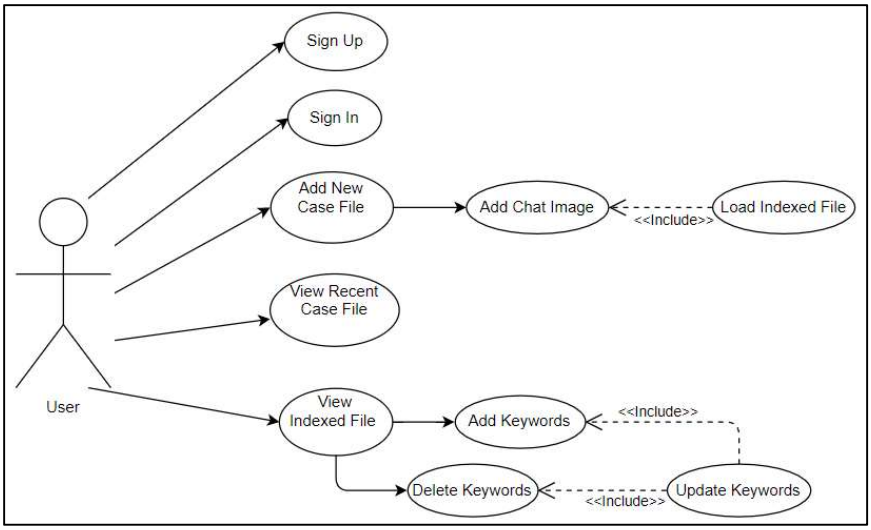

Fig. 1. Example of KIST use case diagram

There are three main components in a use case diagram, which are functional requirements, actors, and relationships. From Figure 1, the functional requirements represent the action of the system, and the actor is a forensic investigator who manages the system.

The implementation phase involves tool development using the Microsoft Visual Studio 2017 in a C# programming language. The database for this tool is developed using MySQL to store the case information and the keywords index. The datasets for keyword were loaded based on the selected types of crime. We created dummy WhatsApp chat text files that involve three sets of chat history. Overall, KIST includes five main modules as presented in Table 2.

TABLE 2
MAIN MODULES OF KIST

Module		Description
Sign up	i.	User registration
	ii.	Add user details
Indexed file	iii.	Add keywords
	iv.	Delete keywords
	v.	Update new keywords from existing keywords
	vi.	Display keywords
New case	i.	Load indexed file
	ii.	Import evidence file (WhatsApp chat text file)
	iii.	Search keywords
	iv.	Display matched keywords in evidence file
Recent case	i.	Display the existing case
	ii.	Edit existing case
	iii.	Generate report

IV. RESULTS AND DISCUSSION

This section highlights the outcomes of KIST and its advantages. One of the main modules of KIST – New Case File (see Figure 2) was programmed by uploading WhatsApp chat text file, and users load the keywords list based on the crime category. The matching keywords are highlighted, and users can view from the displayed text file. Another key feature is the keyword count that summarizes the number of associated keywords found in the text file. Both features of keyword match and keyword count are significantly assisting investigators in analyzing the frequency of the keyword used.

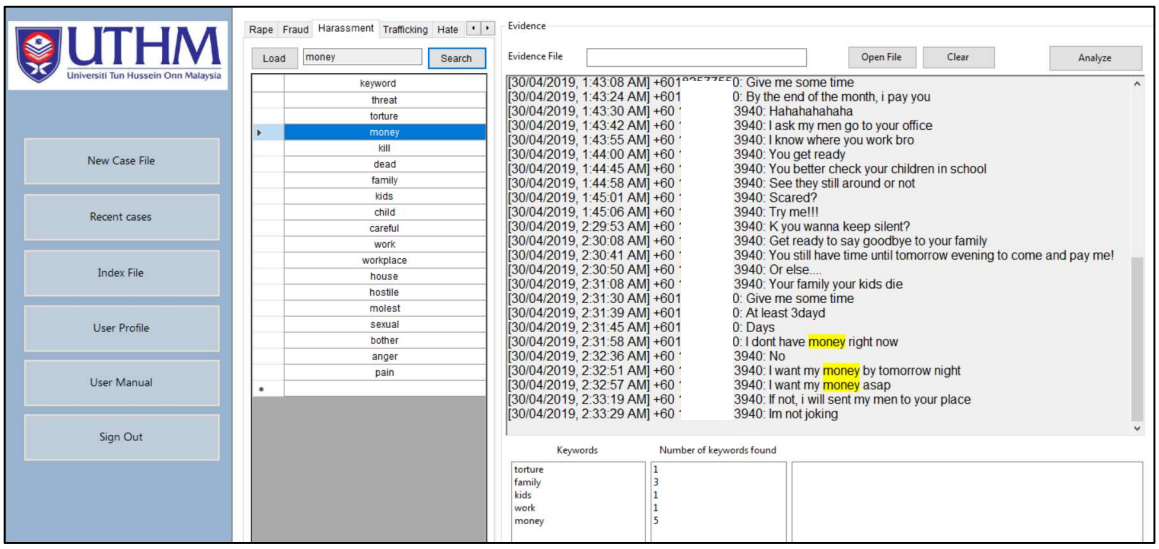

Fig. 2. Screenshot of Matching Keywords Interface

Two types of testing were undertaken to KIST that are: 1) Application Functionality Testing and 2) User Acceptance Testing. Testing of application functionality involved testing on user interfaces, database, security elements, and all modules KIST to ensure that the tool can be executed properly without any error. Subsequently, all modules are running as expected and successful (see Table 3).

User testing includes 25 respondents (i.e., students and staff) of the Faculty of Computer Science and Information Technology (FSKTM) from the Universiti Tun Hussein Onn Malaysia. The group is selected due to their (at least) minimum of 3 years of experience in digital forensics and information security as well as programming practices. Respondents are needed to evaluate KIST if the tool could search and index keywords, as well as display the analysis results. Figure 3 shows that 64% of the respondents are both strongly agree and agree that KIST can index keyword easily, 72% positively agree that users could search keywords easily, and 40% are satisfied with the display of matching keywords analysis.

Other suggestions, such as most of the respondents, noted that this tool should improve the user interface to provide a professional look. A few of the respondents suggest using WhatsApp chat database from the local phone database to collect more detailed information (e.g., sender, receiver, location, images, and video). The KIST program is publicly available in GitHub repository [1].

[1] https://github.com/nurulh/KIST-Keyword-Indexing-Tool

TABLE 3
RESULTS OF FUNCTIONALITY TESTING

MODULE	EXPECTED RESULT	ACTUAL RESULT	STATUS	REMARKS
Sign Up	Users can add their personal information such as full name, email, contact number, address and reset password.	Users manage to add their personal information such as full name, email, contact number, address and reset password.	Successful	-
New Case	Users can add WhatsApp chat ad evidence into the system, load keywords list and searching keywords.	Users manage to add WhatsApp chat as evidence into the system, load keywords list and searching keywords.	Successful.	Only can add WhatsApp chat with (.txt) format
Recent Case	Users can view case history and edit the existing case for further investigation.	Users manage to view the case history.	Successful.	-
Index File	Users can add keyword, update new keyword, delete keyword and load list of keywords.	Users manage to add keyword, update new keyword, delete keyword and load list of keywords.	Successful.	The application can do searching for similar keywords on evidence only after the keywords inserted.
User Profile	Users can view and update their personal information.	Users manage to view and update their personal information.	Successful.	-

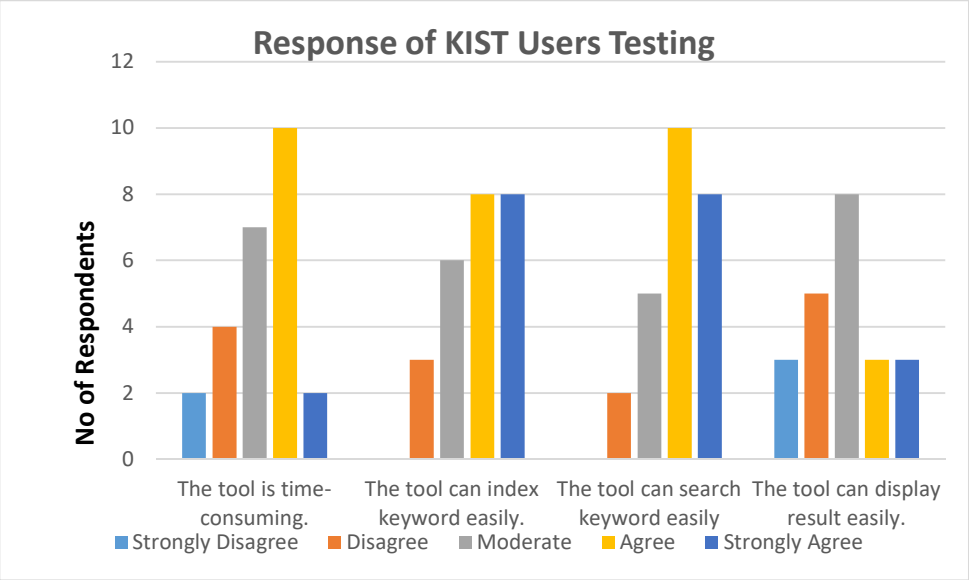

Fig. 3. Response of KIST Users Testing

V. Conclusion

This study aims to design, develop, and test the Keyword Indexing and Searching Tool that is based on incident types. As a result, all system functionalities are executing well, and therefore, this study successfully achieved the objectives.

However, limitations of KIST include the tool only can upload evidence file in the text (.txt) format and lack of graph visualization to enhance user analysis experience. Therefore, future works of KIST would involve:

i. Allow users to add WhatsApp evidence that is extracted from the database, so the forensic investigators know the details of the receiver and sender of a chat message.
ii. The report of the result analysis should be in .pdf format so the forensic investigator can easily go through the recent cases.
iii. Integrate information visualization elements to present analysis findings.

ACKNOWLEDGMENT

The authors express appreciation to the Ministry of Higher Education (MOHE) and Universiti Tun Hussein Onn Malaysia (UTHM). This research is supported by FRGS grant (Vot 1640). Thanks to the anonymous reviewers for valuable comments.

REFERENCES

Ab Rahman, N. H., Cahyani, N. D. W., & Choo, K. K. R. (2017). Cloud incident handling and forensic-by-design: cloud storage as a case study. *Concurrency Computation* , *29*(14), 1–16. https://doi.org/10.1002/cpe.3868

Ademu, I. O., & Imafidon, D. C. O. (2013). The Importance and Need for Digital Forensic Investigative Framework. In *International Conference on Artificial Intelligence (ICAI'13)*. Las Vegas,Nevada, USA. Retrieved from http://world-comp.org/proc2013/icai/ICAI_Contents__Vol_II.pdf

Amandeep, K. R. & K. (2012). Digital Forensics. *International Journal of Computer Applications*, *50*(5), 5–9. https://doi.org/10.5120/7765-0844

Basis Technology (2015). Autopsy User Documentation: Keyword Search Module - The Sleuth Kit. Retrieved October 18, 2018, from http://sleuthkit.org/autopsy/docs/user-docs/3.1/keyword_search.html

Beebe, N., & Dietrich, G. (2007). Chapter 12 A New Process Model For Text String Searching, *242*, 179–191.

Carlson, P. (2006). Apache Lucene - Query Parser Syntax. The Apache Software Foundation. Retrieved from http://lucene.apache.org/core/old_versioned_docs/versions/3_5_0/queryparsersyntax.html

Carrier, B. (2003). Defining digital forensic examination and analysis tools using abstraction layers. *International Journal of Digital Evidence*, *1*(4), 1–12. https://doi.org/10.1017/CBO9781107415324.004

DFIR Training. (2019). Drugs lists. Retrieved December 15, 2019, from https://www.dfir.training/keyword-lists/drug-lists

Ghosh, S. (2018). WhatsApp is dramatically cutting message forwarding after viral fake news led to lynchings. Retrieved November 12, 2019, from https://www.businessinsider.my/whatsapp-cut-message-forwarding-stop-viral-fake-news-2018-7/

Lahaie, C., Porto, K., & Leberfinger, D. (2012). *OSForensics Comparison*. Retrieved from http://www.champlain.edu/Documents/LCDI/archive/OSForensics-Comparison-ReportPDF.pdf

Mishra, S. (2007). *Keyword Indexing and Searching for Large Forensics Targets using Distributed Computing*. University of New Orleans Theses and Dissertations. https://doi.org/510

NSTP Team. (2018). WhatsApp, Facebook main sources of fake news for Malaysians. Retrieved December 15, 2019, from https://www.nst.com.my/news/nation/2018/03/349523/whatsapp-facebook-main-sources-fake-news-malaysians

Palmer, G. (2001). A Road Map to Digital Forensic Research. In *The Digital Forensic Research Conference DFRWS 2001 USA Utica, NY (Aug 7th - 8th)* (p. 32). Retrieved from http://www.dfrws.org/2001/dfrws-rm-final.pdf

Reith, M., Carr, C., & Gunsch, G. (2002). An Examination Of Digital Forensic Models. *International Journal of Digital Evidence*, *1*(3), 1–12. https://doi.org/10.1109/SADFE.2009/