International Journal of Multidisciplinary and Scientific Emerging ResearcH (IJMSERH)

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 12, Issue 1, January 2024 ||

Efficient Mining of Criminal Networks from Unstructured Textual Documents

V.Vinodhini[1], M.Hemalatha[2]

Department of Computer Science, Karpagam University, Coimbatore, Tamil Nadu, India[1].

Department of Computer Science, Karpagam University, Coimbatore, Tamil Nadu, India[2].

ABSTRACT: Digital data unruffled for forensics analysis often contain expensive information about the suspects' social networks. However, most collected records are in the form of amorphous textual data, such as e-mails, chat messages, and text documents. An investigator often has to manually extract the useful information from the text and then enter the important pieces into a structured database for further investigation by using various criminal network analysis tools. Obviously, this information extraction process is monotonous and error-prone. Moreover, the quality of the analysis varies by the experience and expertise of the investigator. In this paper, we propose a systematic method to discover criminal networks from a collection of text documents obtained from a suspect's machine, extract useful information for investigation, and then visualize the suspect's criminal network. Furthermore, we present a hypothesis generation approach to identify potential indirect relationships among the members in the identified networks. We evaluate the usefulness and recital of the method on a real-life cybercriminal case and some other datasets.

KEYWORDS: Textual data, unstructured data, criminal network, cybercriminal.

I. INTRODUCTION

In many criminal cases, computer devices owned by the suspect, such as desktops, notebooks, and smart phones, are target objects for forensic seizure. These devices may not only contain important evidences relevant to the case under investigation, but they may also have important information about the social networks of the suspect, by which other criminals may be identified. Most collected digital evidence are often in the form of textual data, such as e-mails, chat logs, blogs, web pages, and text documents. Due to the unstructured nature of such textual data, investigators usually employ some off-the- shelf search tools to identify and extract useful information from the text, and then manually enter the useful pieces into a well-structured database for further investigation. Obviously, this manual process is tedious and error- prone; the completeness of a search and the quality of an analysis pretty much relies on the experience and expertise of the investigators. Important information may be missed if a criminal intends to hide it.

In this paper, we propose a data mining method to discover criminal communities and extract useful information for investigation from a collection of text documents obtained from a suspect's machine. The objective to help investigators efficiently identifies relevant information from a large volume of unstructured textual data. The method is especially useful in the early stage of an investigation when investigators may have little clued to begin with.

II. RELATED WORK

The computer devices owned by suspect are intention objects for forensic convulsion, these devices does not contain important evidences related to the case by which other criminals may be identified. Most collected data's are in the form of textual document, which are in the form of e-mails, chat logs, blogs, web pages, text documents .

Criminal network analysis has customary great attention from researchers. a unbeaten application of data mining techniques to extract criminal relations from a large volume of police department's incident summaries. they use the co-occurrence frequency to determine the weight of relationships between pairs of criminals. yang and ng (2007) present a method to extract criminal networks from web sites that provide blogging services by using a topic-specific exploration mechanism. In their come close to, they identify the actors in the network by using web crawlers (program that collects online documents and allusion links) that search for blog subscribers who participated in a discussion related to some criminal topics. After the network is constructed, they use some text classification techniques to analyze the content of the documents. Finally they propose a visualization of the network that allows for either a concept network view or a social network view. Our work is different from these works in three aspects.

First, our study focuses on unstructured textual data obtained from a suspect's hard drive, not from a well- structured police database. Second, our method can discover high-flying communities consisting of any size, i.e., not limited to

pairs of criminals. Third, while most of the previous works focus on identifying direct relationships, the methods presented in this paper can also identify indirect relationships.

A criminal network follows a social network archetype. Thus, the approaches used for social network analysis can be adopted in the case of criminal networks. Clustering is often used to perceive the crime pattern and speed up the course of action. Many studies have introduced various approaches to construct a social network from text documents. A framework to extract social networks from text document that are available on the web. A method to rank companies based on the social networks extracted from web pages. These approaches rely mainly on web mining techniques to search for the actors in the social networks from web documents.

Another direction of social network studies targets some specific type of text documents such as e-mails. propose a probabilistic approach that not only identifies communities in email messages but also extracts the relationship information using semantics to label the relationships. However, the method is applicable to only e-mails and the actors in the network are limited to the authors and recipients of the e-mails. Researchers in the field of knowledge discovery have proposed methods to scrutinize relationships between terms in text documents in a forensic context. A concept association graph-based approach to search for the best evidence trail across a set of documents that connects two given topics. In passed research they proposed the open and closed discovery algorithms to extract evidence paths between two topics that occur in the document set but not necessarily in the same document. The open discovery approach to search for keywords provided by the user and return documents containing other different but related topics. They further apply clustering techniques to rank the results and present the user with clusters of new information that are conceptually related to their initial query terms. Their open discovery approach searches for novel links between concepts from the web with the goal of improving the results of web queries. In contrast, this paper focuses on extracting information for investigation from text files.

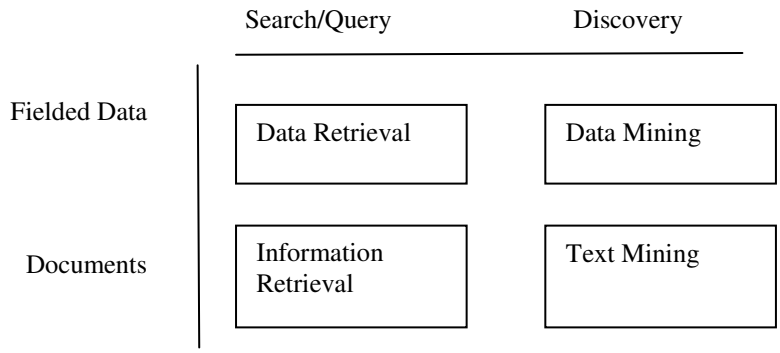

FIG I: Concept Association graph based approach

III. PROPOSED ALGORITHM

A. *Communities discovery from unstructured textual data*

Several social network analysis tools are available to support investigators in the analysis of criminal networks. However, these tools often assume that the input is a structured database. So, structured data is often not available in real-life investigations. Instead, the available input is usually a collection of unstructured textual data. Our first contribution is to provide an end-to-end solution to automatically discover, analyse, and visualize criminal communities from unstructured textual data.

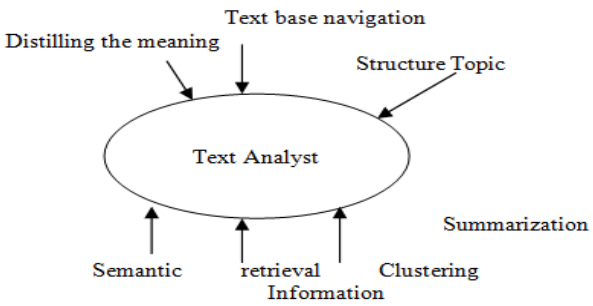

FIG II-Unstructured Textual Document

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 12, Issue 1, January 2024 ||

B. *Introduction of the notion of prominent communities.*

In the context of this paper, two or more persons form a community if their names appear together in at least one investigated document. A community is prominent if its associated names frequently appear together in some minimum number of documents, which is a user- specified threshold. We propose a method to discover all prominent communities and measure the closeness among the members in these communities. To measure the closeness between the communities clustering techniques is used which helps in identifying the centroid location and flanking distance appraise.

C. *Generation of indirect relationship hypotheses.*

The philosophy of well-known community and convenience among its members detain the direct relationships among the persons identified in the investigated documents. Our recent work presents a prelude study on direct relationships. In many cases, indirect relationships are also interesting since they may reveal buried relationships. For example, person a and person b are indirectly related if both of them have mentioned a meeting at hotel x in their written e-mails, even though they may not have any direct communications. We present a method to generate all indirect relationship hypotheses with a maximum, user- specified, depth. 4. Scalable computation.

The computations of prominent communities and closeness from the investigated text document set is non-trivial. A naive approach is to enumerate all 2juj combinations of communities and scan the document set to determine the prominent communities and the closeness, where juj is the number of distinct personal names identified in the input document set. Our proposed method achieves scalable totalling by efficiently pruning the non- prominent communities and examining the closeness of the ones that can potentially be prominent. The scalability of our method is supported by experimental fallout .by doing so we can increase the efficiency and reduce the error also can assist police work and enable investigators to distribute their time to other valuable errands.

IV. PSEUDO CODE

Step 1: Let D be a set of documents.
Step 2: Let U be a set of distinct names identified in D
Step 3: Let C4U be a prominent community and p ¸ (U ? C) be a person name that is not in C.
.Step 4: Let D denote the set of documents containing the enclosed argument where the enclosed argument can be a community, a personal name, or a text term.
Step 5: Let D(C) and D(p) be the sets of documents in D that contain C and p, respectively. An indirect relationship of depth d between C and p is defined by a sequence of terms [t1,., td] such that
D(C) X D (p) ¼
ðt1¸ DðCÞÞ^ðtd¸ DðpÞÞ
ðtr¸ Dðtr?1ÞÞ^ðtr¸ Dðtrþ1ÞÞ for 1 < r < d
D(tr ? 1) X D(tr þ 1) ¼
for 1 < r < d
Step 6: End.

Condition (1) requires that a prominent community C and a personal name p do not co-occur in any document. Condition (2) states that the first term t1must occur in at least one document containing C and the last term must occur in at least one document containing p. Condition (3) requires that the intermediate terms co-occur with the previous term in at least one document, and must co-occur with the next term 1ine at least one document. This requirement defines the chain of documents linking C and p. Condition (4) requires that the previous term and the next term do not co-occur in any document. The problem of indirect relationship hypothesis generation is formally defined as follows:

Let D be a set of text documents. Let U be the set of distinct personal names identified in D. Let G be the set of prominent communities discovered in D according to Definition 3.2. The problem of indirect relationship hypothesis generation is to identify all indirect relationships of maximum depth max_depth between any prominent community C ¸ G and any personal name p ¸ U in D, where max_depth is a user-specified positive integer threshold.

V. SIMULATION RESULTS

The proposed algorithm is implemented with MATLAB. The dataset File system contains 40 GB of files obtained from the first author's personal computer. As the minimum support threshold increases, the number of high-flying communities quickly decreases because the number of documents containing all members in a community decreases very quickly. Next, we weigh up the scalability of our proposed methods by measuring its runtime. The evaluation is con- ducted on a PC with Intel 3 GHz Core2 Duo with 3 GB of RAM, with respect to the size of the document set

which varies from 10 GB to 40 GB with min_sup ¼ 8. The program takes 1430 s to complete the entire process for 40 GB of data, excluding the time spent on reading the document files from the hard drive. As shown in the figure, the total run- time is dominated by prominent community discovery procedure. The runtime of the indirect relationship cohort and hallucination procedures is insignificant with respect to the total runtime.

VI. CONCLUSION AND FUTURE WORK

We have proposed an approach to discover and analyse criminal networks in a collection of investigated text documents. Previous studies on criminal network analysis mainly focus on analysing links between criminals in structured police data. As a result of extensive discussions with a digital forensics team of a law enforcement unit, we have introduced the notion of high-flying criminal communities and an efficient data mining method to viaduct the gap of extracting criminal networks information and unstructured textual data. Furthermore, our proposed methods can discover both direct and indirect relationships among the members in a criminal community. The developed software tool has been evaluated by an experienced crime investigator and future work can be concentrated on predicting crime network using Density based approach in order to condense missing values.

REFERENCES

1. Agrawal R, Imieli? nski T, Swami A.,Mining association rules between sets of items in large databases. ACM SIGMOD Record 1993;22(2):207–16.
2. Al-Zaidy R, Fung BCM, Youssef AM. Towards discovering criminal communities from textual data. In: Proc. of the 26th ACM SIGAPP symposium on applied computing (SAC); 2011. TaiChung, Taiwan.
3. Chen H, Chung W, Xu JJ, Wang G, Qin Y, Chau M. Crime data mining:a general framework and some examples. Computer 2004;37(4):50–6.
4. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling. In: Proc. Of the 43rd annual meeting on association for computational linguistics (ACL); 2005. p. 363–70.
5. Friedl JEF. Mastering regular expressions. 3rd ed. O'Reilly Media; 2006.Geobytes Inc. Geoworldmap, http://www.geobytes.com/; 2003.
6. Getoor L, Diehl CP. Link mining: a survey. ACM SIGKDD Explorations Newsletter 2005;7(2):3–12.
7. Hope T, Nishimura T, Takeda H. An integrated method for social network extraction. In: Proc. Of the 15th international conference on world wide web (WWW); 2006. p. 845–6.
8. Jin W, Srihari RK, Ho HH. A text mining model for hypothesis generation. In: Proc. Of the 19th IEEE international conference on tools with artificial intelligence ICTAI; 2007. p. 156–62.
9. Jin Y, Matsuo Y, Ishizuka M. Ranking companies on the web using social network mining. In: Ting IH, Wu HJ, editors. Web mining applications services. Studies in computational intelligence.

BIOGRAPHY

Dr. M. Hemalatha completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa women's University, Kodaikanal). She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science at Karpagam University, Coimbatore. Twelve years of experience in teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. **She received best researcher award in the year 2012** from Karpagam University. Her research areas include Data Mining, Image Processing, Computer Networks, Cloud Computing, Software Engineering, Bioinformatics and Neural Network. She is a reviewer in several National and International Journals.

Ms.V.Vinodhini,M.Sc.,M.Phil.,She is pursuing her Ph.D in Karpagam University ,Department of Computer Science under the guidence of Dr.M.Hemalatha and working as Assistant Professor in Dr.N.G.P Arts And Science College (Department of Information Technology).Her research areas include Data Mining, Image Processing.