Generating Regular Expressions from Semantically and Pragmatically Analysed Text as Forensic Search Support in Evidence Mining

Abiodun Abdullahi Solanke.[0000−0002−6082−893X]

CIRSFID
Alma Mater Studiorum - Universita di Bologna
Via Galliera, 3 - 40121, Bologna, Italy
abiodun.solanke@unibo.it

Abstract. Extracting evidence from text in a typical digital forensic investigation operation could involve searching for patterns that are somewhat complex and generally difficult to conceive, which is why in most current digital forensic tools, string or text pattern searching is handled by Regular Expressions (Regex). However, timeliness, accuracy and optimization of regex to extract meaningful patterns could be a complex and time-consuming task. This research work focuses on evidence mining by semantically and pragmatically annotating unstructured textual entities with strong evidential substance. Then, subsequently generate Regex sequences of the annotated text through carefully structured algorithms. The generated Regex could then be used in forensic analysis tools in furtherance of the reconstruction of events for investigative purposes.

Keywords: Regular Expressions · Pragmatic Analysis · Semantic Analysis · Ontology · Genetic Programming · Dynamic Programming.

1 Introduction

The domain of Digital Forensics (DF) is gaining popularity largely because of the rapid advancements in technology and the growing complexities of cybercrime. Electronic devices are becoming smarter, data generated are getting larger, so also the susceptibility and intention to commit crime. Therefore, forensic investigators and law enforcements are now facing, more than ever before, the growing challenges of searching for concrete digital evidence that is admissible in the court. Swift conclusion of cybercrime trial can be viewed as a function of time, accuracy and relevance. These functional parameters constitute a major challenge during a digital forensic operation.

Digital forensics involve the use of scientific methods to identify, preserve, collect, validate, analyse, document and present evidence from digital sources to proof complicity, intent, guilt; or otherwise, in a criminal case. For every electronic crime committed, there is a footprint left, which could be traced and extracted for investigative purposes.

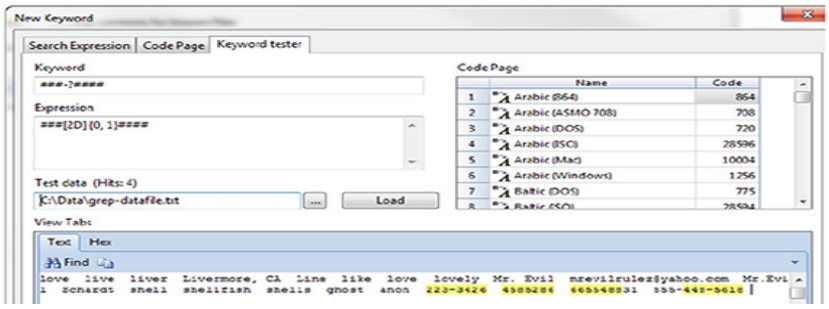

Fig. 1. Extract from EnCase Forensics showing different matched patterns of phone numbers (Source: https://static1.squarespace.com/static/Chapter+9+Sample.pdf)

Extracting this proof may involve analysing a large volume of unstructured textual data with hidden facts which might seem meaningless ab initio. Also, in a bid to extract instances of an event from the physical and logical levels of a digital medium, matching string pattern is always a key process. In current DF tools - such as EnCase, searching for and extracting correlating or similar string patterns from data stream is mostly addressed by regular expressions, because, evidential entities, such as phone numers or credit card numbers, in most cases, follow a syntactical pattern. Fig. 1 is an extract from EnCase Forensics, a popular DF tool, showing matches for different phone numbers with a regex.

A regular expression, regex or regexp (sometimes called a rational expression) is a sequence of characters that define a search pattern. [1]. Regex as a concept has grown in robustness, complexity and popularity over the years and its applications are used over a wide range of computing domains. Most especially, regex is one of the main structure of search engines and searching algorithms.

Writing regex is a task that requires great expertise, even experts often find it difficult to come up with regex sequences that can match all instances of a string. Therefore, programming machines to surrogate the work of an expert to dynamically construct complex regex will be more challenging and will require thorough research. StackOverFlow has as at the time of this compilation, 215,355 question tags on regular expressions [2], which makes it one of the most popular domain on that platform. This shows how complex the writing of a regex can be and the amount of skills required to accomplish a major task.

Several domain researches have been conducted on optimization of regular expression for advanced searches, in fact, there are free online tools to help generate regex either by singling out text segments that is to be matched from

a textual corpus [3] or by entering input strings which are parsed as individual tokens, and then used to generate the corresponding regex and the associated code [4].

With pragmatics analysis in Natural Language Processing, we can study, not only the linguistic knowledge in a text, but also how the transmission of meaning depends on the context of utterances, pre-existing knowledge and intent in language. In fact, research work in this domain is constantly growing and almost countless. There have been quite appreciable number of works on the combination of these domains, i.e., on natural language and generation of regular expression as available in [5,6].

However, as much as there have been several works in these domains, there is no previous study that has directly addressed the problem of automatically generating regex from forensically valuable annotated entities in natural language for the purpose of evidence mining. Generally, this project proposes to create a tool that synthesises selected novel approaches towards the overall optimization of automatic generation of regular expression in order to support accurate and relevant evidence extraction.Also, pragmatic and semantic-based approaches will be used to analyse and extract evidential entities from Natural Language text by annotating strings or patterns that holds greater evidential values.

The goal of this project is to automatically and dynamically generate robust regex patterns from the annotated evidential entities. Subsequently, these regex patterns will be used in DF tools to extract further instances of the evidence. To achieve this goal, two algorithmic approaches have been identified: the first one relies on dynamic programming with Finite Automata (FA) theory approach; while the second relies on genetic programming, i.e., an approach that is currently gaining popularity in computing domain which is biologically inspired and domain-independent. Although one of these two approaches will be used, the choice of these approaches and details will be explained further in subsequent sections of this work description.

2 Problem Statement

Mining evidence in a digital forensics investigation is quite a complex task in all facets of the operation. The fact that the outcome of the investigation presented in court could be easily disproved, makes it extremely important for investigators to find all instances that could be used for or against in an electronic crime or civil case. Admissibility of presented evidence in a criminal case is hinged on relevance, correctness and most often, the ability to show that the scientific methods used were novel and provable.

The features available on DF tools allow the extraction or retrieval of information. This extracted information may contain a lot of non-relevant data, and interpreting it to show potential or subtantial evidence may largely depend on the experience of the investigator. In order to optimize the extraction of relevant evidence, different domain independent approaches have been deployed to aid better sensitivity. Several semantic ontologies for different DF operations have

been either proposed or developed, but the observable lapses are that majority of the works were based on ontological models: for evidence management purposes; to classify threats; or to determine the relevance of retrieved documents. Creating ontologies that can infer knowledge from mulitple sources to annotate new and exisiting entities will help to create new knowledge-base that can enhance evidence extraction.

Many works have adopted semantic analysis approach to analyse natural language text. This may have succeeded in many instances, but, semantic analysis can only examine the relationship between words and determine the conventional meaning inherent in sentences. Also, conventional natural language preprocessing on text alone does not seem to be sufficient to address the complexity of computational linguistics, especially in digital evidence mining. The problem of Polysemy (words with multiple meanings for the same lexical form) and Synonymy (terms that describes a word which has other words with exactly, close related or subtitutable meaning) makes it even harder to find hidden facts necessary to establish events.Tthere could be extended inferable meanings in natural language as communicated amongst actors. This work intends to explore the provisions of pragmatic analysis to evidence mining.

One can be a forensic expert without necessarily able to write codes. Which means we can have a DF expert who may not know how to write a simple regular expressions. Extracting patterns with regex is efficient, makes the work easier, and it is largely provable provided the right sequence of characters are used. Major literatures and tools on automatic generation of regex have been deployed for other purposes except in the domain of digital forensics. The current state of the art work only generate regex based on users' manual annotations of the strings they want to generate regex for, and the result is presented in a different format not understandable by an ordinary investigator. Specifically generating regex that is compatible with major DF tool regex engine will optimize searching pattern for evidence extraction. The complexity of the algorithms to be used in this project is believed to be sufficient enough to optimize this generation process. Details of these algorithms are in the methodology section.

One of the observed efficiency gap is when investigators spend much effort coming up with a regex that someone else has created previously. Previous studies have not adequately investigated the impact of optimized (reduced) timing on pattern generation to the overall operational time of evidence extraction. This work proposes to bridge the efficiency gap and optimize the general operational process, which is why it intends to synthesize multiple domains, through separate-and-conquer strategy, towards a common goal.

3 Research Questions

1. How can a dynamic algorithm, support and optimize the efficiency of evidence extraction from texts with automatic generation of multi-patterned regular expression sequences?
2. To what extent could the re-usability of previously generated regex pattern optimize the overall operational time in a digital evidence mining operation?

4 Aim and Objectives

4.1 Aim

This research work proposes to develop a tool that helps investigators to make more efficient use of the regex patterns during digital evidence extraction from text. The main aim is to drastically reduce the entire time required to extract evidence in furtherance of the establishment of events in a digital crime case.

4.2 Objectives are:

1. to develop a model that dynamically auto-generate multiple regex patterns from forensically valuable annotated entities; and,
2. allow for optimal re-use of previously structured patterns

5 Literature Review

Discussed in this section are further proposals that broadens the general scope of the works that have been conducted earlier, in order to put in perspective this work, with respect to the existing ones. In data mining, several studies have shown plausible improvements on the application of data mining techniques to digital forensics. [7] used data mining techniques to discover common trends in certain criminal text by creating clusters based on document similarities that revealed either authorship or type of crime. Their result showed that most of text written or spoken by certain individual had higher similarity indices and, in some cases, the result showed the type of crime. [8] and [9] surveyed data mining techniques for crime detection. Their works analysed techniques for entity extraction, clustering, association rule mining, sequential pattern mining, deviation detection, classification, locate group of latent facts (clustering) and discover patterns in data that may lead to useful predictions (forecasting). Tsochataridou et al. [10] used K-means algorithm on the publicly available Enron email dataset as a vehicle to analyse how clustering techniques can effectively identify malicious collaborative activities. The result however showed that the approach was time-consuming and may be inaccurate.

Kahvedzic and Kechadi [11] introduced an evidence management methodology to encode semantic information. The work used terms from DIALOG, a digital forensic ontology developed to encompass knowledge associated with the digital investigations to model metadata, file content and event evidence in an application independent and semantic manner. However, the work only annotated the meaning of evidence and helped to tag and structure extracted evidence. Amato et al., [12] in two distinct but related works, presented a method, based on semantic web technologies, that helped investigators to correlate and present information acquired from data, with the aim of getting valuable reconstruction of events. In [13], the system was modelled to find series of relevant and peculiar terms in order to detect the set of concepts that allowed the resource identification. Their approach used information retrieval text pre-processing techniques on

textual data and later determined the semantic relevance of the retrieved documents with Tf-Idf (Term Frequency-Inverse document frequency). [14] proposed a research and theory-based set of over 20 potential linguistic risk indicators that may discriminate credible from non-credible threats within online threat message corpora. The result of the annotations were classified as non-threat, false threat and credible threats with computational linguistics. Close to the idea of this work in the area of semantic text analysis is the work of Raskin et. al. [15] which demonstrated how existing ideas, methods, and resources of ontological semantics can be applied to detect deception in natural language text (and, eventually, in data and other media as well). As the author claimed, the work was the first attempt to linguistically explore cyber forensics through semantic identification of clues, and, to establish events – rather than reconstruct events.

Elkhatib [16] pragmatically analysed Obama's speech in the wake of Arab spring that led to the stepping down of Former Egyptian President, Hosni Mubarak. The author adopted 'The theory of Grice H.P (Conversational) Implicature' as a model to analyse the nature and power of pragmatics to explain the general principle of co-operation. l [17] investigated the vague expressions in news articles on the security situation in Iraq. The cited work showed that this vagueness was logically calculated to obscure the context; as too much clarity in the news article would cause the text to lose its spontaneity and acceptability by readers. Abdulameer [18] showed that the possible choice of some selected deictics is dependent on the context in which they are used.

Also, Bartoli et al. [19] investigated whether a machine may automatically construct regular expressions for task of realistic complexity, trying to reduce the human efforts in building regex. An experiment was conducted to compare the automatic solution constructed by a tool based on genetic programming and those of human experts. The result showed the that the automatic and the skilled users' generation were indeed the same in quality and timing. In a similar but distinct works, Bartoli et al. [20] proposed and presented the design and implementation of a system capable of addressing extraction task, from semi-structured and unstructured text, based on evolutionary procedure to generate regex from examples, adopting a seperate-and-conquer strategy. Their tool can discover whether the extraction may be solved by a single pattern, or rather a set of multiple patterns. Part of the approach proposed in their works include active learning for minimizing user annotation effort: the user annotates only one desired extraction and the system constructs and selects the most optimized candidate solution with genetic programming.

Several other related works are [6,21,22,23] where alternative approaches were explored. They commonly implemented the automatic generation of regular expression with dynamic programming through finite automata theory; neural models and; Bayesian inferencing through stochastic process recognition model.

For pattern searching in EnCase, GREP (Globally search for Regular Expression and Print it) is used. GREP originates from UNIX operating systems, but its implementation in EnCase, just like the regex language itself, is not fully standardized [24]. Different flavours (programming languages) implement

regex matching, in most cases, differently. The reason it is necessary to generate specific regex patterns that are compatible with major DF tools.

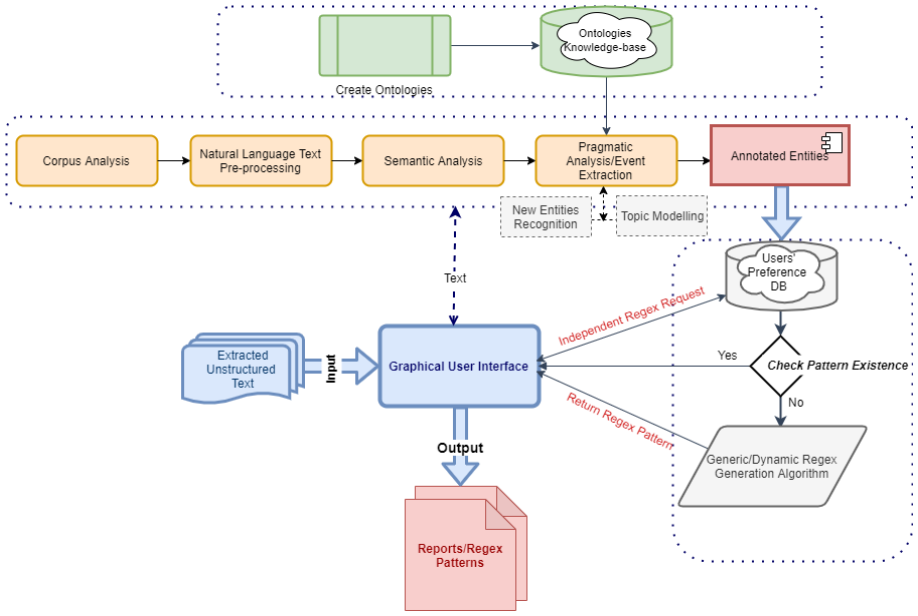

Fig. 2. Research Work Diagramatic Sketch

6 Proposed Research Methodology

This work is proposed to support digital evidence mining. In no way this work will interrupt or pre-empt the general procedural provisions of DF operations, rather, it proposes to efficiently support and optimize the operations. This work shall only be concerned with textual data (especially unstructured text) extracted from devices with the approved standard procedures maintained. Fig. 2 shows the diagramtic process flowchart of the proposed project.

First, some ontologies and specific DF domain knowledge-bases will be created. These ontologies will be stored in a knowledge-base so other application layer, specifically, the paragmatic and the event extraction module, can infer new knowledge from it.

Textual extract from devices, prepared in the form of documents, will be inserted through the GUI and computational linguistics pre-processing will be applied on the text with the sole aim of extracting and semantically annotating concepts and entities with circumstantial and corroborating evidential values, such as: credit cards; passwords; dates; phone numbers; email addresses; encryption patterns; organization names; places; URLs; twitter hashtags; ambiguous

terms and so on, that are relevant to digital forensic knowledge base. Relationship between known entities will be identified while new ones will be inferred.

Then, in the pragmatic analysis phase, understanding that there could be more to what was ordinarilly written or communicated in a text, the structural organization of the thought-process of the writers/authors of the text will be analysed. Also in this phase, investigators subjective input will be excluded, i.e., interpretations of the inferences and additional inputs can be done after reports have been generated. Dataset will be trained to identify evidence-based terms by labelling some parts of the dataset on pre-defined sets of domain-specific evidence classes. This will be used for topic modelling or crime description.

On the other part of the research work, which is the main goal, the extracted entities from the previous phase shall then undergo an algorithmic process that automatically constructs and generates regex patterns that can match any instance of these entities when used on DF tools. This auto-generation process is where this work intends to apply either the genetic or dynamic programming concept. The former is inspired by evolutionary natural genetic process to select fittest programs for reproduction (crossover) and mutation according to a predefined fitness measure, while the latter uses finite automaton to recursively self-evolve within a predefined sequence of operations. Regex sequence for a single string could be constructed in different forms – from simple to complex ones. The general idea behind using one of these techniques is not only to craft the most complex and robust regex, but also to asymptotically analyse the computational complexity of the process. The approach with the most efficient and optimized complexity will be used (research on this is still ongoing).

Three (3) to five (5) different patterns from the fittest candidates constructed will be presented to the user with a selection option against each pattern. Usage will be ranked by counting users' selection on any of the patterns. These will guide future algorithmic process behaviour (the algorithm will consider first, users' highest ranked previous selection before presenting other eligible candidates). In strict consideration of the domain in which this research work is focused (digital forensics), and as earlier discussed, regex engine comes in different flavours, i.e.; programming language compilers interpret regex pattern, in some cases, differently – this work will structure the algorithms to generate only regexes that are compatible with major DF tools. The generated regexes with other extracted evidential inferences will be reported as the output.

Validation of the results of this work shall be done in different phases.The first phase will involve comparing experts' skills at constructing regex against what the system automatically generates, and, the asymptotic complexity of the generating algorithms as compared to previous state of the art works. The second phase will aim to evaluate the correctness and accuracy of the prediction of potential evidential entities as extracted, annotated and classified. The third phase will evaluate the entire operational time of investigation when the system is used against human time when same is done manually.

7 Proposed Contribution to Knowledge

The contribution of this research work will be significant in all the phases of the entire work. The first contribution will be the creation of a new approach to textual evidence mining by adopting pragmatic techniques to infer contextual meaning while adding new ontologies to the knowledge base. The efficient support towards easing the burden and complex task of constructing regex patterns that is compatible with DF tools' regex engine will be the second contribution. Lastly, the general optimization of the operational time of digital forensic investigation will be greatly reduced, especially with the framework this research work proposes to present for textual evidence mining.

References

1. Regular Expressions Wikipedia, `https://en.wikipedia.org/wiki/Regularexpression`. Last accessed 1 Oct 2019
2. Question Tagged [Regex]. Stackoverflow, `https://stackoverflow.com/questions/tagged/regex`. Last accessed 3 Oct 2019
3. Bartoli, A., Lorenzo, A., Medvet, E., Tarlao, F.: Inference of regular expressions for text extraction from examples. IEEE Transactions on Knowledge and Data Engineering, **28**(5), pp. 1217–1230 (2016). IEEE Press (2016). doi: 10.1109/TKDE.2016.2515587
4. Text 2 Regular Expression Hompage, `http://www.text2re.com`. Last accessed 25 Sep 2019
5. Zhong, Z., et al.: SemRegex: A Semantics-Based Approach for Generating Regular Expressions from Natural Language Specifications. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1608 – 1618. Association of Computing Linguistics (2018). doi: 10.18653/v1/D18-1189
6. Kushman, N., Barzilay, R.: Using Semantic Unification to Generate Regular Expressions from Natural Language. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology, pp. 826–836. Association of Computing Linguistics (2013)
7. Vatter, C., Mozgovov, M.: Data Mining in Forensics: a Text Mining Approach to Profiling Criminals (2015) `http://web-ext.u-aizu.ac.jp/~mozgovoy/homepage/papers/vm15.pdf`. Last accessed 5 Aug 2019
8. Qayyum, S., Shareef, H.: A survey of Data Mining Techniques for Crime Detection. University of Sindh Journal of Information and Communication Technology (USJICT) **2**(1), pp. 1-6 (2018)
9. Nirkhi, S., Dharaskar, R., Thakre, V.: Data Mining: A Prospective Approach for Digital Forensics. Internaternational Journal of Data Mining and Knowledge Management Process (IJDKP) **2**(6) (2012)
10. Tsochataridou, C., Arampatzis, A., Katos, V.: Improving Digital Forensics through Data Mining. In: The Fourth International Conference on Advances in Information Mining and Management. IARIA (2014).
11. Kahvedzic, D., Kechadi, T.: Semantic Modelling of Digital Forensic Evidence. In: 2010 Proceedings of the International Conference on Digital Forensics and Cyber Crime. pp. 149–156 (2010). doi: 10.1007/978-3-642-19513-6-13

12. Amato, F., Cozzolino, G., Mazzeo, A., Mazzocca, N.: Correlation of Digital Evidence in Forensics Investigation through Semantic Technologies. In: 2017 Proceedings of the 31st International Conference on Advanced Information Networking and Application (WAINA). IEEE Press (2017) doi: 10.1109/WAINA.2017.4
13. Amato, F., Cozzolino, G., Mazzeo, A., Moscato, An Application of Semantic Techniques for Forensic Analysis. In: 2018 Proceedings of the 32nd International Conference on Advanced Information Networking and Application (WAINA). IEEE Press (2018) doi: 10.1109/WAINA.2018.00115
14. Spitzberg, B.H., Gawron, J.M.: Towards Online Linguistic Surveillance of Threatening Message. Journal of Digital Forensics, Security and Law **11**(2), Art. 7 (2016). doi: 10.15394/jdfsl.2016.1418
15. Raskin, V., Hempelmann, C.F., Triezenberg, K.E.: Semantic Forensics: An Application of Ontological Semantics to Information Assurance. In: 2004 Proceedings of the 2nd Workshop on Text Meaning and Interpretation. pp. 105–112. Association for Computational Linguistics (2004)
16. ElKhatib, A.S.: A Pragmatic Analysis of President Obama's Speech After Mubarak's Stepping Down: A Conversational Implicature Analysis. Scribd, `https://www.scribd.com/doc/127587786/A-Pragmatic-Analysis-of-President-Obama-s-Speech-After-Mubarak-s-Stepping-Down`. Last Accessed Oct 5 2019
17. Khalil, H.: A Pragmatic Analysis of Vague Langugae in the News Articles on the Iraqi Security Crisis. Theory and Practice in Language Studies **7**(5), pp. 327-335 (2017) doi: 10.17507/tpls.0705.01
18. Abdulameer, T.A.: A Pragmatic Analysis of Deixis in a Religious Text. International Journal of English Linguistics **9**(2) (2019) doi: 10.5539/ijel.v9n2p292
19. Bartoli, A., Lorenzo, A., Medvet, E., Tarlao, F.: Active Learning of Regular Expressions for Entity Extraction. IEEE Transactions and Cybernetics **48**(3), pp. 1067–1080 (2018). IEEE Press (2018) doi: 10.1109/TCYB.2017.2680466
20. Bartoli, A., Lorenzo, A., Medvet, E., Tarlao, F.: Learning Text Patterns Using Separate-and-Conquer Genetic Programming. EuroGP (2015) doi: 10.1007/978-3-319-16501-1-2
21. Locascio, N., Narasimhna, K., DeLeon, E., Kushman, N., Barzilay, R.: Neural Generation of Regular Expressions from Natural Language with Minimal Domain Knowledge. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1918 – 1923. Association for Computational Linguistics (2016) doi: 10.18653/v1/D16-1197
22. Luo, B., et al.: Marrying up Regular Expressions with Neural Networks: A Case Study for Spoken Language Understanding (2018) `https://www.arxiv.org/abs/1805.05588`. Last accessed 1 Oct 2019
23. Ouyang, L.: Bayesian Inference of Regular Expressions from Huma-Generated Example Strings `https://www.arxiv.org/abs/1805.08427`. Last accessed 3 Oct 2019
24. Widup, Suzanne: Computer Forensics and Digital Investigation with EnCase Forensic v7. McGraw-Hill, New York (2014)