ISSN 2250-3153

An Ontological Approach for Digital Evidence Search

Venkata Krishna Kota

Central Research Laboratory
Bharat Electronics Limited, Bangalore, India
venkatakrishnav@bel.co.in

***Abstract*—** Usage of emails for the fraudulent activities is accelerating with higher pace. There is a thirst need for the tools to analyze large collections emails forensically. Traditional Information Retrieval tools can retrieve documents those are relevant to the given query. But directly answering the questions specific to the forensics domain will make the job forensic examiners easy. In this paper a system is presented to answer questions specific to email forensics. Ontology is designed with the basic concepts of email forensics domain. Information relevant to the case under investigation is retrieved using Information Retrieval techniques. Ontology is dynamically populated with the retrieved information. Knowledge which is of interest to the forensic investigators is inferred by firing the domain specific rules with the help of inference engine. Some domain specific questions have been answered with the help of inferred knowledge. The proposed system is a prototype and it can stand as a base to develop bigger systems.

***Index Terms*—**Digital Evidence Search; Ontology; Information Retrieval

I. INTRODUCTION

Email based crimes are becoming a major threat to the national as well as organizational security. Forensic Analysis of these emails can prevent, investigate or prove a crime committed. There is a thirst need for the tools to analyze large collections emails forensically. Information Retrieval (IR) is being used as the primary weapon for Digital Forensics to retrieve the relevant information from large corpus with the aim of gathering digital evidence. Lot of research is emerged in this field. New ranking and visualizing algorithms are developed to improve the performance of digital evidence search.

Keyword based IR systems retrieve the documents those are relevant to the user's information need. IR systems developed for Digital Forensics generally aims at high recall. Due to the huge size of digital data and high recall requirements of Digital Forensics, the IR system retrieves huge number of results. Even with the help of efficient ranking algorithms and visualization techniques, forensic examiner needs to spend a lot of time in analyzing each search hit.

In this paper, a system is presented to analyze the emails forensically and to answer some queries related email forensics. Ontological approach is chosen in this experiment to analyze the emails forensically. Ontology can be described as "Specification of Conceptualization". Ontologies have been widely used in many domains to formally represent the knowledge of that domain, to provide automated reasoning, to infer new Knowledge and to answer domain specific queries. Ontologies in forensics domain have been proposed by the researchers for intrusion detection & prevention, spam email detection & prevention, query expansion for IR to achieve high recall and so on.

There is a lot of research and guidelines exist to assist ontology design. But ontology needs to get populated with the relevant information before it can be used for reasoning or question answering. In digital forensics, the information that we need to analyze will change dynamically depending on the nature of case which is under investigation. Even if the ontology is well designed and populated with the information that is relevant to the case under investigation, it may not be useful while dealing other cases, because the populated information is irrelevant to other cases. So forensics ontology needs to get populated dynamically with the case relevant information.

While several methodologies for designing Ontologies and automating ontology learning have been proposed, ontology population has not received much attention so far [8]. Case relevant information will not be readily available all the time. One needs to extract it from the large collections of data. In this paper, a simple ontology for Email Forensics is designed. The ontology is dynamically populated with case relevant information which is extracted using IR. Some of the questions related to email forensics are answered using this ontology.

II. RELATED WORK

Usage of emails for the fraudulent activities is accelerating with higher pace. Email based crimes are becoming a major threat to national as well as organization security. Forensic Analysis of these emails can prevent, investigate or prove a crime committed. These issues have triggered us to focus on Email Forensics. Enron data set is a large collection of emails [15]. It has been used for testing the effectiveness of techniques used for counter terrorism and fraud detection by many researchers [10][11]. It has been chosen as the data set in this experiment.

IR techniques have been widely used to gather digital evidence from large data collections. Lot of researchers has developed many algorithms to improve the performance of digital evidence search. In [2], author used WordNet ontology for query expansion to achieve the high recall requirement of digital evidence search. In [1], [3] authors proposed a new ranking methodology to improve the performance of digital evidence search.

Traditional keyword based IR systems will retrieve documents those are relevant to the user's information need and present them to the user [9]. It will leave the responsibility of analyzing those documents to the user. User need to analyze them in order to get required knowledge. But a well defined and populated ontology can straight away answer the domain specific questions.

Ontologies have been widely used in many domains to formally represent the knowledge of that domain, to perform automated reasoning and to answer domain specific queries. Ontologies have also been developed to support digital forensics. In [4], Ontology is used for developing automated digital forensic tools. In [5], authors proposed ontology for network forensic analysis. This ontology represents both network forensics domain knowledge and problem solving domain knowledge. In [7] authors presented some challenges in conceptualizing ontology and specific techniques useful for ontology construction. They have expressed their efforts to construct ontology for a terrorism analysis. [6] Describes a framework to forensically analyze large volumes of data using ontology and machine learning.

While several methodologies for designing Ontologies and automating ontology learning have been proposed, ontology population has not received much attention so far [8]. In [9], authors presented an approach to semi-automatically instantiate Ontologies, with the support of human expert. A methodology to dynamically populate the ontology with the help of IR is presented n this paper.

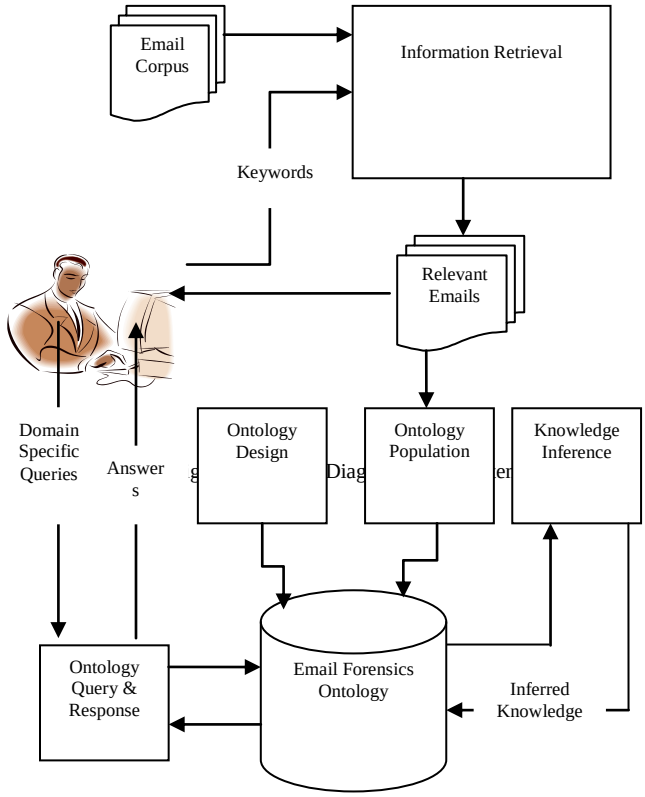
III. DESIGN

In this experiment ontology with dynamic instantiation is aimed for forensic analysis of emails. The block diagram of the system is given in Fig.1. Ontology Design Unit is meant to design ontology to assist emails forensic analysis by specifying the domain concepts, properties of concepts and relationships among the concepts. In Digital Forensics, the information that needs to be analyzed varies from case to case. The system presents a way to dynamically populate the ontology. Information Retrieval Unit retrieves the emails those are relevant to the user query (keywords) from the large email corpus.

These results will be presented to the user as well as used to populate the ontology. Ontology Population Unit extracts the relevant information from the results of Information Retrieval Unit and populates the ontology with it. Knowledge Inference Unit infers new knowledge by firing inference rules with the help of an inference engine and updates the ontology with the inferred knowledge. Ontology Query and Response Unit answers the domain specific queries of the user.

IV. IMPLEMENTATION

Implementation details of the system are explained in this section. As the first step, a simple Email Forensics Ontology is designed as described below.



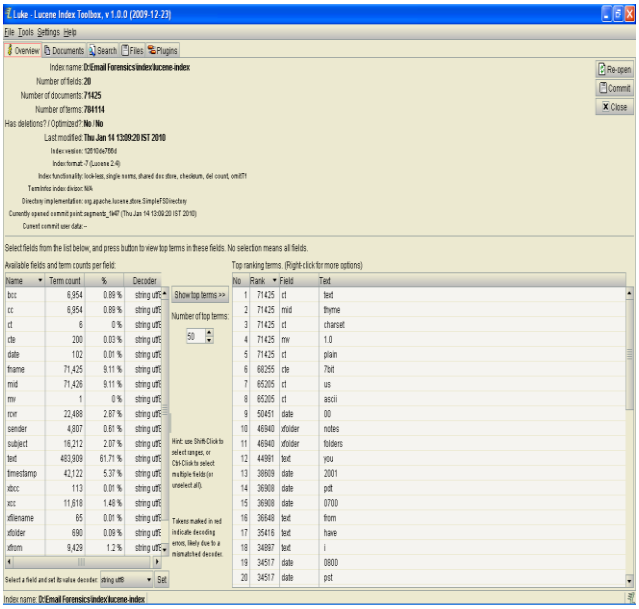


Figure 2. Luke's View of the Index.

A. *Ontology Design*

Ontology is described as the "Specification of Conceptualization". Some of the concepts of email forensics domain, properties of those concepts and relationships among the concepts are conceptualized initially. These details are formally represented in the form of ontology. OWL (Web Ontology Language) is chosen as the ontology representation language

[16]. It has rich representation for the semantics and it has support for many tools. Protégé tool is used to develop the ontology [15]. Protégé is a widely used open source ontology editor. Once the Ontology is designed, we need to extract the case relevant information to populate the ontology.

B. Information Retrieval

Enron email collection is chosen as the dataset for the experiment. Relevant information is retrieved from the data set for the given query using Information Retrieval techniques. Relevant information is retrieved using LUCENE [14] which is an open source search library.

Indexing refers to processing the original data into a highly efficient cross-reference lookup in order to facilitate rapid searching [12]. Every email in the corpus is parsed, analyzed, tokenized and indexed. Index is tested using Luke tool. The Luke's view of the index is given in the Fig. 2.

Searching is the process of looking up words in an index to find documents where they appear. User's query keywords are analyzed, refined, mapped against the index and relevant emails are retrieved [12].

Retrieved emails are ranked based on their relevance to the given query and presented to the user as well as used to populate the ontology. The retrieved emails for the phrase query "terrorist attack" are given in Fig. 3.

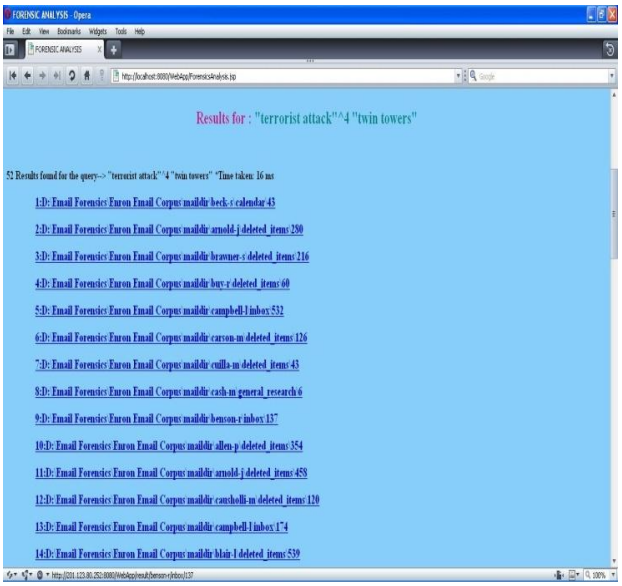


Figure 3. Retrieved Relevant Emails for the Query.

Once the relevant emails are available, we have to populate the ontology with them.

C. Ontology Population

From the retrieved emails (result of Information Retrieval Unit) details like sender, receiver, date and time of email sent/received, email is deleted or not, via path and other details are extracted. Using protégé's "DataMaster" plug-in, ontology is populated with the extracted information [17]. The Populated Ontology is shown in Fig.4.

From the results of Information Retrieval Unit, we can dynamically acquire other interesting details like sender IP, sender location using tools like EmailTrackerPro, SmartWhoIs [11]. With the help of email service provider, we can get the bio data information like when the email account is created, historical information like statistics of that email account and a lot other details. Considering these details to further enhance the ontology is well suggested.

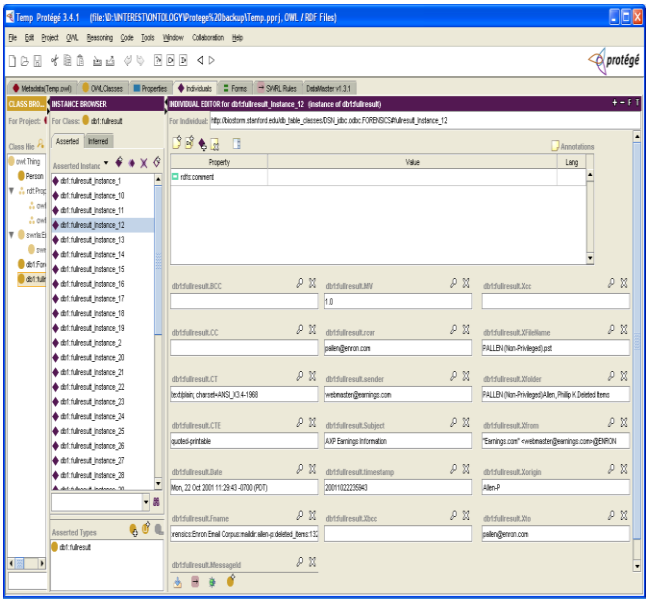


Figure 4. Populated Ontology.

D. Knowledge Inference

Inference rules relevant to the email forensics are developed in SWRL (Semantic Web Rule Language). It is widely used language to represent rules [17]. JESS inference engine is used to infer new knowledge. It maps the rules against the ontology, fires the rules accordingly and infers new knowledge [17]. The Ontology is updated with the inferred knowledge. This inferred knowledge may also cause firing other rules. Thus ontology will be updated in an iterative fashion. Updating sometimes causes the ontology to become inconsistent. Consistency checking is performed after every updation with the help of pellet reasoner [17].

E. Ontology Query & Response

Ontology queries are developed in SPARQL query language. SPARQL equivalents of the following domain specific queries are submitted to the ontology.

a. Are there any email transactions (directly or through via paths) between given 2 email ids or not?
b. Is Email id A, reachable to Email id B?

Ontology has answered the above questions. By enhancing the ontology as suggested earlier, the ontology can answer the queries like

c. List the emails originated from particular location?
d. Are there any deviations in the email account history during the period of crime incident?

Some SWRL rules and SPARQL query result are shown in Fig. 5.

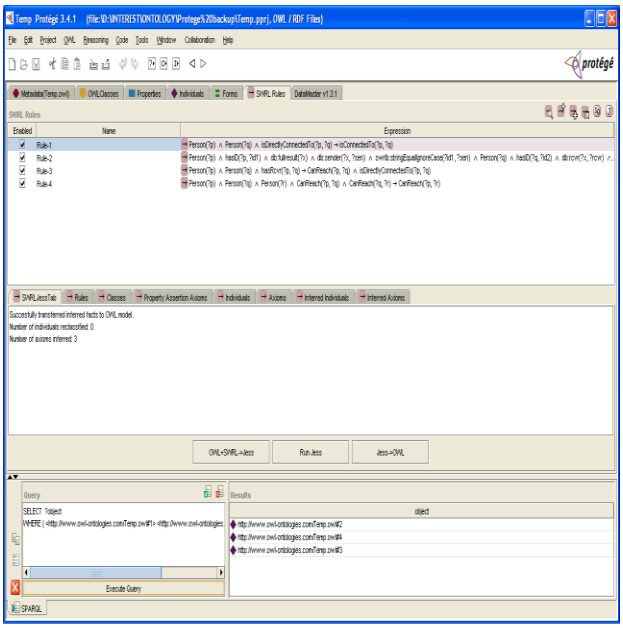


Figure 5. Sample SWRL Rules and SPARQL Result.

V. CONCLUSION

Ontology for Email Forensics is designed. Ontology is populated dynamically with the relevant information which is extracted from large corpus using IR techniques. New Knowledge is inferred by reasoning the ontology. Some domain specific questions related to email forensics have been successfully answered using the ontology. Some techniques have been proposed to further enhance the Email Forensics Ontology.

REFERENCES

[1] Jooyoung Lee, "Proposal for Efficient Searching and Presentation in Digital Forensics", The Third International Conference on Availability, Reliability and Security, 2008, pp. 1377-1381, doi:10.1109/ARES.2008.192.

[2] Report, Australian Phan Thien Son, "Ontology-Driven Text Mining for Digital Forensics", COMP6703 Project National University, 2007.

[3] Hong-Rong Yang, Ming Xu and Ning Zheng, "An Improved Method for Ranking of Search Results Based on User Interest", IFIP International Conference on Network and Parallel Computing, 2008, pp. 132-138, doi:10.1109/NPC.2008.6

[4] Hoss A.M, Carver D.L, "Weaving ontologies to support digital forensic analysis", IEEE International Confernce on Intelligence and Security Informatics, 2009, pp. 203-205, doi: 10.1109/ISI.2009.5137303

[5] Saad S and Traore I, "Method ontology for intelligent network forensics analysis", Eighth Annual International Conference on Privacy Security and Trust (PST), 2010, pp. 7 - 14, doi: 10.1109/PST.2010.5593235

[6] Jingshan Huang, Yasinsac A and Hayes P.J., "Knowledge Sharing and Reuse in Digital Forensics",Fifth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE), 2010, pp. 73 - 78, doi: 10.1109/SADFE.2010.18

[7] Mannes A and Golbeck J, " Ontology Building: A Terrorism Specialist's Perspective", IEEE Aerospace Conference, 2007, pp. 1 - 5, doi: 10.1109/AERO.2007.352794

[8] Ruiz-Martinez J.M, Miarro-Gimenez J.A, Guillen-Carceles L, Castellanos-Nieves D, Valencia-Garcia R, Garcia-Sanchez F, Fernandez-Breis J.T and Martinez-Bejar R, "Populating Ontologies in the eTourism Domain", International Conference on Web Intelligence and Intelligent Agent Technology, 2008, pp. 316 - 319, doi: 10.1109/WIIAT.2008.278

[9] Doherty L, Kumar V and Winne P, "Assisted Ontology Instantiation: a LearningKit perspective", Seventh IEEE International Conference on Advanced Learning Technologies, 2007. ICALT 2007, pp. 265 - 267, doi:: 10.1109/ICALT.2007.75

[10] Jitesh Shetty and Jafar Adibi, "The Enron Email Dataset Database Schema and Brief Statistical Report".

[11] Natarajan Meghanathan, Sumanth Reddy Allam and Loretta A. Moore, "Tools and Techniques for Network Forensics", International Journal of Network Security & Its Applications (IJNSA), Vol .1, No.1,April 2009

[12] Erik Hatcher, Otis Gospodnetic and Michael McCandless, "Lucene in Action, Second Edition", Manning Publications, 2009.

[13] Lucene search library, available at http://lucene.apache.org/nutch

[14] Enron email dataset available at http://www-2.cs.cmu.edu/~enron/

[15] Protégé Ontology Editing Tool, available at http://protege.stanford.edu/

[16] OWL guide, available at, http:// www.w3.org/TR/owl-guide/

[17] Protégé Wikipedia, available at, http://protegewiki.stanford.edu/wiki/Main_Page

AUTHOR

Venkata Krishna Kota received his B.Tech degree in Computer Science and Information Technology from Jawaharlal Nehru Technoligical University in 2005 and M.E degree in Computer Science from Anna University in 2008. He is working as Member (Research Staff) at Central Research Laboratory (CRL), Bharat Electronics Limited (BEL), Bangalore. His research interests are Information Retrieval and Complex Event Processing.