# AN INFORMATION EXTRACTION FRAMEWORK FOR DIGITAL FORENSIC INVESTIGATIONS

Min Yang, Kam-Pui Chow

Chapter 4

# AN INFORMATION EXTRACTION FRAMEWORK FOR DIGITAL FORENSIC INVESTIGATIONS

Min Yang and Kam-Pui Chow

**Abstract** The pervasiveness of information technology has led to an explosion of evidence. Attempting to discover valuable information from massive collections of documents is challenging. This chapter proposes a two-phase information extraction framework for digital forensic investigations. In the first phase, a named entity recognition approach is applied to the collected documents to extract names, locations and organizations; the named entities are displayed using a visualization system to assist investigators in finding coherent evidence rapidly and accurately. In the second phase, association rule mining is performed to identify the relations existing between the extracted named entities, which are then displayed. Examples include person-affiliation relations and organization-location relations. The effectiveness of the framework is demonstrated using the well-known Enron email dataset.

**Keywords:** Information extraction, named entity recognition, relation extraction

## 1. Introduction

Investigations involving text documents and data rely on expression-based or keyword-based searches [6]. Current digital forensics tools implement search methods that rely on accurate keywords. For example, given the name of a suspect, a search tool would return all the exact and similar occurrences in the data [24]. However, in the vast majority of cases, complete keyword information is not available; this makes it necessary to uncover and use additional information to improve searches.

Information summarization and event extraction from data are extremely useful in forensic investigations. In general, the more information that can be extracted, the more accurate the search. However,

digital forensic examiners are often unable to comprehensively review all the keyword matches in a corpus [5]. Text summarization can reduce the time spent on reviewing search hits while ensuring that all the instances of interest are located.

Information extraction systems can extract abstract knowledge from a text corpus or extract concrete data from a set of documents [30]. Information extraction has two phases. The first phase, involving the detection and classification of proper names, is called named entity recognition [31]. Since digital forensics focuses on the who, what, where, when and why of a case, it is essential to recognize named entities such as persons, organizations and locations. For instance, in the sentence "Jeff Skilling (born November 25, 1953) is the former CEO of the Enron Corporation headquartered in Houston, Texas," "Jeff Skilling," "Enron Corporation" and "Houston, Texas" correspond to person, organization and location entities, respectively.

The second phase of information extraction is relation extraction. In a forensic investigation, it is useful to discover relations that are hidden in named entities extracted from large data sets. For example, by analyzing newspaper text, it is possible to discern that an organization is located in a particular city or that a person is affiliated with a specific organization [44]. The relations discovered can then be represented in two forms, as association rules or as sets of frequent items.

This chapter describes a novel information extraction framework that helps find valuable evidence in text documents. The framework has two phases. In the first phase, a named entity recognition approach is applied to the raw forensic data to extract the named entities (i.e., persons, organizations and locations). To assist investigators in finding coherent evidence rapidly and accurately, informative named entities are visualized and highlighted using cloud tags and text clouds. The extracted entities provide a useful overview of the data when a forensic investigator does not know what to look for or has a large number of hits to review. The second phase uses a modified version of the Apriori algorithm [34] to identify the relations among the extracted named entities and visually display the extracted relations. Examples of relations are person-affiliation and organization-location. Experiments are conducted on the Enron email corpus to demonstrate that the information extraction framework is very effective at helping find relevant information.

## 2. Related Work

The pervasiveness of information technology has led to an explosion of evidence. In order to handle vast amounts of evidence in a limited time,

a number of text analysis approaches have been proposed for forensic investigations [12, 14, 16, 21, 25, 32, 35]. Information extraction, which seeks to extract useful information from unstructured text, has become an active research area in digital forensics. The first step in information extraction is the detection and classification of proper names, which is often referred to as named entity recognition [31]. Named entity recognition is a well-studied problem and has many applications, including focused searches over massive collections of textual data [15], social network analysis [19] and information summarization [36].

Named entity recognition has been applied to news articles [28] and scientific articles [7]. These articles are written for fairly broad audiences and the authors generally take great care when preparing them. However, less time and effort are spent on preparing informal documents; as a result, they contain many grammatical and spelling errors. In addition, informal text generally contains many abbreviations. Techniques have been proposed to improve named entity recognition performance for informal text such as email [29] and web postings [19].

Named entity recognition techniques have also been applied in forensic investigations [4, 9, 21, 22, 24, 33, 41]. Louis et al. [24] have used dynamic Bayesian networks to identify named entities. Kuperus et al. [22] have presented a probabilistic named entity recognition model based on multiple candidate labels that exploits user feedback to obtain increased recall with little loss in precision. However, both techniques do not attempt to discover the relations existing between named entities in forensic investigations.

The second step in information extraction is relation extraction. Relation extraction can be characterized as a classification problem: if two entities are potentially related, it is necessary to determine if they are indeed related [3]. Supervised methods [44, 45], semi-supervised methods [10, 38] and unsupervised methods [43] have been applied to solve this binary classification problem. In the case of structured data, association rules can be used to simplify relation extraction [26, 30]. This work uses the Apriori algorithm [34], a classical association rule approach, to discover the relations among named entities. Al-Zaidy et al. [2] have proposed a similar approach. However, they focus on identifying indirect relationships between a person and a prominent community. An indirect relationship starts from a person and forms a chain connecting to a prominent community; each internal node in the chain is a person that links two documents. In contract, the work described in this chapter attempts to discover named entities and their semantic relations.

Colombe and Stephens [11] discuss visualization techniques for assisting forensic investigations. The Tilebars method [17], for instance, is

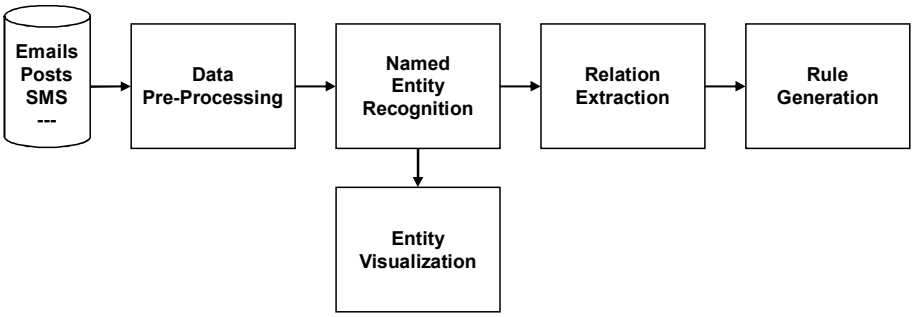



*Figure 1.* Information extraction process.

designed to be a navigational aid for analysts. Whittaker et al. [42] visualize term distributions in histograms to support information retrieval from speech archives. Schwartz and Liebrock [37] apply visualization to digital forensic string searches. On the other hand, the information extraction framework presented in this chapter uses tag clouds [18] to visualize named entities and also leverages visualizations of the relations between named entities, providing forensic investigators with an intuitively appealing means to comprehend and analyze large bodies of textual evidence.

## 3. Information Extraction Framework

This section describes the information extraction framework for finding evidence. As shown in Figure 1, the framework has two main phases. In the first phase, a named entity recognition approach is applied to the raw forensic data to extract the named entities (i.e., persons, organizations and locations). In the second phase, a data mining tool is used to identify the relations existing between the extracted named entities. To assist investigators in finding coherent evidence faster and more intuitively, relevant informative named entities are highlighted with cloud tags and the relations among the named entities are visually displayed.

### 3.1 Named Entity Recognition

Given a sentence, the named entity recognition approach segments words that are parts of named entities and then classifies each entity by its type (person, organization, location, etc.). Existing named entity recognition approaches are divided into two categories: (i) rule-based approaches [31]; and (ii) machine learning approaches [8, 39]. A rule-based approach relies on linguistic knowledge, in particular, grammar

rules, while a machine learning approach relies on a labeled corpus [1]. A rule-based approach achieves better results in many domains because the rules can be adapted very precisely and are, therefore, able to detect complex entities. However, in the case of an unrestricted domain, it would be expensive in terms of cost and time to derive rules. On the other hand, a machine learning approach is more flexible and is able to identify uncommon patterns of named entities that are not specified in the regular expressions of rule-based systems [1]. The proposed named entity recognition framework exploits the complementary performance associated with the rule-based and machine learning approaches by taking the union of the results provided by the two approaches to improve the recall.

The machine learning approach used in the proposed framework engages the conditional random fields model [23] to identify named entities, primarily because it provides excellent performance for many information extraction tasks [39]. A document is expressed as a vector $X = (x_1, \ldots, x_i, \ldots, x_n)$ where $n$ is the number of features and $x_i$ is the relative frequency of feature $i$ in the document. After all the documents have been represented in this way, the conditional random fields model can be applied to extract the named entities. Lafferty et al. [23] define the probability of a label sequence $y$ given observation sequence $x$ to be the normalized product of potential functions of the form:

$$\exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum \mu_k s_k(y_i, x, i)) \tag{1}$$

where $t_j(y_{i-1}, y_i, x, i)$ is a transition feature function of the entire observation sequence and the labels at positions $i$ and $i-1$ in the label sequence; $s_k(y_{i,}, x, i)$ is a state feature function of the label at position $i$ and the observation sequence; and $\lambda_j$ and $\mu_k$ are parameters that are estimated from training data.

When defining feature functions, a set of real-valued features $b(x, i)$ of the observation are constructed to express some characteristic of the empirical distribution of the training data that should also hold in the model distribution. An example of such a feature is:

$$b(x,i) = \begin{cases} 1 & \text{if the observation at position } i \text{ is the word "September"} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Each feature function takes on a value of one for real-valued observation features $b(x, i)$ if the current state (in the case of a state function)

or previous and current states (in the case of a transition function) take on particular values.

In the remainder of this chapter, the notation is simplified by writing:

$$F_j(y, x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i) \tag{3}$$

where each $f_j(y_{i-1}, y_i, x, i)$ is either a state function $s(y_{i-1}, y_i, x, i)$ or a transition function $t(y_{i-1}, y_i, x, i)$. This allows the probability of a label sequence $y$ given an observation sequence $x$ to be written as:

$$p(y|x, \lambda) = \frac{1}{Z(x)} \exp(\sum \lambda_j F_j(y, x)) \tag{4}$$

where $Z(x)$ is a normalization factor. In this work, the conditional random fields model was executed using publicly-available Python code [13]. Default settings were used for all the parameters.

The rule-based approach largely follows the guidelines listed in [27]. However, several changes have been introduced to make the extracted named entities more practical and to improve the recall of the named entity recognition system. For example, when analyzing email, a check was made to see if a token in the header was equal to some token $w$.

## 3.2 Relation Extraction

Forensic investigators usually browse forensic data using queries such as "Select all the people who have a relation with the suspect" or "Select all the people who are related to a specific organization." This can be implemented using relation extraction, which identifies the significant relations existing between entities.

This work incorporates a modification of the Apriori algorithm [34] to perform relation extraction. The Apriori algorithm is a classic algorithm that finds relations in data. For example, the rule, Diapers $\rightarrow$ Beer, suggests that a strong relation exists between the sale of diapers and the sale of beer because many customers who purchase diapers also purchase beer [40].

An indication of how frequently an item or a relation occurs is measured by the support function $supp(s)$. The support of a rule $A \rightarrow B$ is the percentage of documents in the corpus that contain $A \cup B$. The *supp* measure is important because a rule that has very low support may occur simply by chance.

The confidence measure *conf*, on the other hand, measures the reliability of the inference made by the rule. For a rule $A \rightarrow B$, the higher

*Table 1.* Illustrative example.

| | **Coffee** | $\overline{\textbf{Coffee}}$ | **Total** |
|---|---|---|---|
| **Tea** | 15 | 5 | 20 |
| $\overline{\textbf{Tea}}$ | 75 | 5 | 80 |
| **Total** | 90 | 10 | 100 |

the confidence, the more likely it is that a document containing $A$ also contains $B$.

Given a set of documents containing extracted named entities, it is necessary to find all the rules with $supp \geq minsupp$ and $conf \geq minconf$ where $minsupp$ and $minconf$ are two constant thresholds. Interested readers are referred to [34] for details about the Apriori algorithm.

In the original Apriori algorithm, confidence is the only measure used to verify rules. However, although the confidence of a rule may be high, the rule could still be misleading. For instance, consider the situation shown in Table 1 where there are 100 transactions that contain either tea or coffee. Furthermore, assume that the *conf* threshold is 70%. From Table 1, for the rule Tea $\rightarrow$ Coffee, *conf* = P(Coffee|Tea) = 0.75, indicating that the rule is valid. Nevertheless, a person who does not purchase tea is more likely to purchase coffee since P(Coffee|$\overline{\text{Tea}}$) = 0.9375. To address this problem, the *lift* score = P(Y|X)/P(Y) is used to evaluate rules of the form $X \rightarrow Y$. Only rules whose *lift* scores are larger than one are considered.

## 4. Experiments and Analysis

This section describes the experiments performed on a real-world forensic dataset and presents an analysis of the results.

### 4.1 Dataset Description

Because no authoritative real-world forensic datasets are available to evaluate the performance of a named entity recognition system in forensic investigations, a portion of the texts from the Enron corpus was manually labeled. The raw Enron corpus made public by the U.S. Federal Energy Regulatory Commission contains 619,446 messages belonging to 158 users. Klimt and Yang [20] subsequently cleaned up the corpus by removing certain folders associated with each user because they appeared to be computer-generated. The cleaned Enron corpus used in this work has 200,399 messages belonging to 158 users with an average of 757 messages per user.

*Table 2.* Named entity recognition results.

| Named Entities | Precision | Recall | F1 |
|---|---|---|---|
| Persons | 0.84 | 0.92 | 0.88 |
| Organizations | 0.88 | 0.95 | 0.91 |
| Locations | 0.78 | 0.87 | 0.82 |

## 4.2 Data Pre-Processing

The texts were tokenized using the NLTK natural language toolkit. Next, non-alphabet characters, numbers, pronouns, words with two characters or less, punctuation and stop words (common words appearing frequently) were removed from the text. Finally, the WordNet stemmer was applied to reduce the size of the vocabulary and address data sparseness.

## 4.3 Experimental Results

This section discusses the experimental results.

**Named Entity Recognition.** The experiment to evaluate the effectiveness of the named entity recognition system used 200 randomly-chosen messages from the Enron corpus containing more than 100 words as test data. Three natural language processing researchers were then asked to manually label each entity tag (i.e., person, organization, location, etc.).

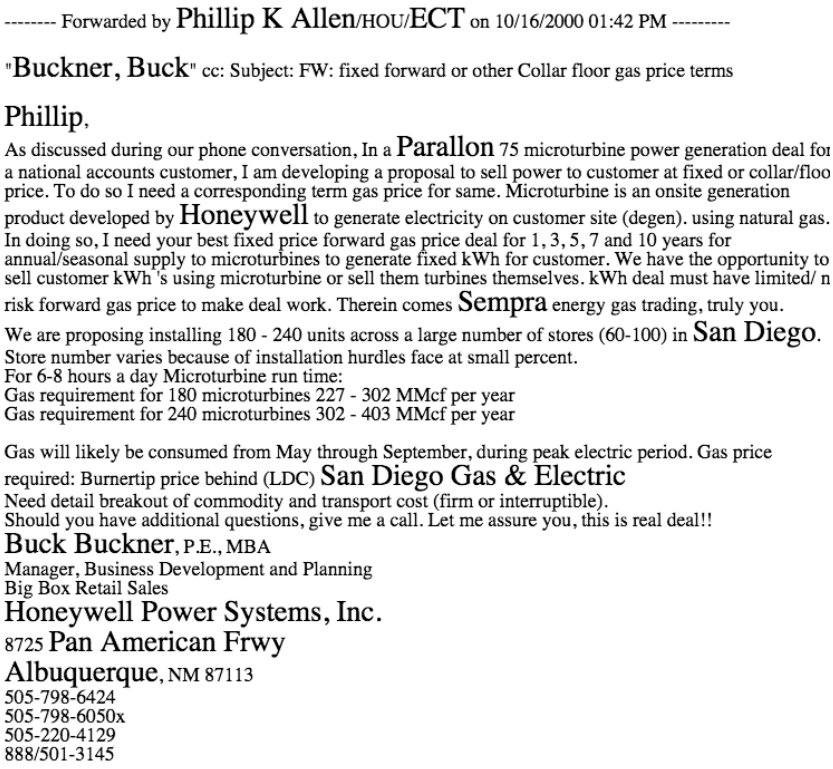
Recall, precision and $F_1$ measures were used to evaluate the performance of the named entity recognition system:

$$precision = \frac{\#\text{correct identified entities}}{\#\text{total entities found}} \tag{5}$$

$$recall = \frac{\#\text{correct identified entities}}{\#\text{total correct entities}} \tag{6}$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

Table 2 summarizes the results. Since higher values for the three measures indicate more accurate extraction, the results indicate that the named entity recognition system is effective at identifying persons, organizations and locations. In particular, the system has a relatively high recall for persons and organizations.



-------- Forwarded by Phillip K Allen/HOU/ECT on 10/16/2000 01:42 PM ---------

"Buckner, Buck" cc: Subject: FW: fixed forward or other Collar floor gas price terms

Phillip,

As discussed during our phone conversation, In a Parallon 75 microturbine power generation deal for a national accounts customer, I am developing a proposal to sell power to customer at fixed or collar/floo price. To do so I need a corresponding term gas price for same. Microturbine is an onsite generation product developed by Honeywell to generate electricity on customer site (degen). using natural gas. In doing so, I need your best fixed price forward gas price deal for 1, 3, 5, 7 and 10 years for annual/seasonal supply to microturbines to generate fixed kWh for customer. We have the opportunity to sell customer kWh 's using microturbine or sell them turbines themselves. kWh deal must have limited/ n risk forward gas price to make deal work. Therein comes Sempra energy gas trading, truly you.

We are proposing installing 180 - 240 units across a large number of stores (60-100) in San Diego. Store number varies because of installation hurdles face at small percent.
For 6-8 hours a day Microturbine run time:
Gas requirement for 180 microturbines 227 - 302 MMcf per year
Gas requirement for 240 microturbines 302 - 403 MMcf per year

Gas will likely be consumed from May through September, during peak electric period. Gas price required: Burnertip price behind (LDC) San Diego Gas & Electric
Need detail breakout of commodity and transport cost (firm or interruptible).
Should you have additional questions, give me a call. Let me assure you, this is real deal!!

Buck Buckner, P.E., MBA
Manager, Business Development and Planning
Big Box Retail Sales
Honeywell Power Systems, Inc.
8725 Pan American Frwy
Albuquerque, NM 87113
505-798-6424
505-798-6050x
505-220-4129
888/501-3145

*Figure 2.* Highlighted named entities in the analyzed text.

**Named Entity Visualization.** Forensic investigators are often overwhelmed by the number of keyword matches when dealing with large datasets. To assist investigators in rapidly finding coherent evidence, a text cloud visualization of word importance is used to discriminate the extracted named entities from the original text. Figure 2 shows a text cloud display of a random sample of text from the Enron corpus. The named entities are represented using larger fonts than the other entities.

Suppose that an investigator has no prior knowledge of the data and, thus, does not what to look for. In such a scenario, tag clouds [18] may be used to quickly visualize all the extracted named entities with respect to the name category. Figure 3 shows three tag clouds for persons, organizations and locations, respectively. The importance of each tag is expressed using its font size. For example, "John" and "Jeff" are the top person names in Enron Corp. Note that the tag cloud representation can enable a forensic investigator to quickly draw conclusions from a massive volume of data.

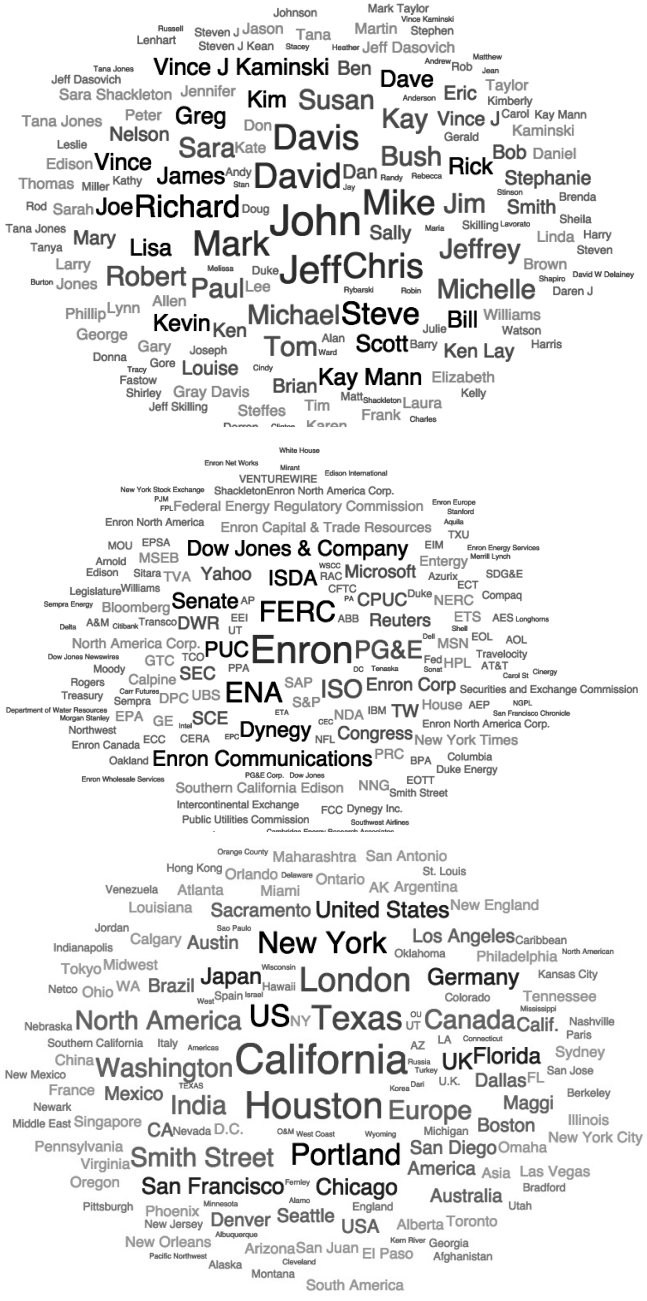



*Figure 3.* Identified named entities (persons, organizations and locations).

*Table 3.* Example entity relationships.

| **Entity Relations** | **Relation Description** |
|---|---|
| Enron North America Corp → Smith Street, Houston | Enron is located at Smith Street, Houston, Texas |
| Ken Lay → Enron | Ken Lay is the CEO of Enron |
| California → Gray Davis | Gray Davis served as the Governor of California |
| Enron Capital and Trade Resources Corp → ECT | ECT is short for Enron Capital and Trade Resources Corp |
| Frank L. Davis, Tana Jones → ECT | Frank L. Davis and Tana Jones are ECT traders |
| California Public Utilities Commission → California | California Public Utilities Commission is located in California |
| Mike Swerzbin → Joe Stepenovitch | Joe Stepenovitch is the VP of energy marketing and trading, the boss of Mike Swerzbin |
| Enron Wholesale Services → Enron | Enron Wholesale Services is Enron's largest business unit |
| Tanya Tamarchenko → Vince J. Kaminski | Vince J. Kaminski is the director of research, the boss of Tanya Tamarchenko |
| BUSH → Houston | BUSH International airport is located in Houston |

**Relation Extraction.** In a forensic investigation, it is also important to discover interesting relations that are hidden in named entities. For example, a useful query might be: Select all the persons who have a relation with the suspect. The Apriori algorithm was modified to discover relations existing between name entities (i.e., persons, organizations and locations). The threshold *minsupp* was set to 150. Table 3 shows some of the relations with relatively high support. Forensic investigators can use this technique to identify a suspect using logically-related queries. Note that traditional exact matching techniques often fail to provide useful results.

In addition to discovering direct relations, the information extraction framework can also construct implicit social networks from email activities. For example, although Jeffrey K. Skilling and Matthew Lenhart had no direct email exchanges, a relation between them exists as a result of the rules:"ECT → Jeffrey K. Skilling" and "Matthew Lenhart → ECT." In particular, Jeffrey K. Skilling is the CEO of ECT while Matthew Lenhart is an ECT trader, which manifests the relation that Skilling is the boss of Lenhart.

## 5. Conclusions

The information extraction framework presented in this chapter is specifically designed to enhance forensic investigations. It applies a named entity recognition approach on raw forensic data to extract named entities. Following this, association rule mining (i.e., the Apriori algorithm) is applied to identify the relations existing between the extracted named entities. Relevant and informative named entities are visualized using tag clouds, and new relations existing between the named entities can also be discovered. Experiments using the well-known Enron email corpus demonstrate the effectiveness of the framework.

Future research will extend the named entity recognition system to identify addresses, vehicles, narcotics and personal characteristics. It will also attempt to develop social networks of criminals and suspects based on the extracted entities. Additionally, it will apply user interest profiling to reveal indirect relations between individuals and identify individuals who have specific interests over time.

## References

[1] S. Abdallah, K. Shaalan and M. Shoaib, Integrating rule-based system with classification for Arabic named entity recognition, *Proceedings of the Thirteenth International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 1, pp. 311–322, 2012.

[2] R. Al-Zaidy, B. Fung, A. Youssef and F. Fortin, Mining criminal networks from unstructured text documents, *Digital Investigation*, vol. 8(3-4), pp. 147–160, 2012.

[3] N. Bach and S. Badaskar, A Review of Relation Extraction, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2007.

[4] N. Beebe, Digital forensic research: The good, the bad and the unaddressed, in *Advances in Digital Forensics V*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 17–36, 2009.

[5] N. Beebe and J. Clark, Dealing with terabyte data sets in digital investigations, in *Advances in Digital Forensics*, M. Pollitt and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 3–16, 2005.

[6] N. Beebe and L. Liu, Clustering digital forensic string search output, *Digital Investigation*, vol. 11(4), pp. 314–322, 2014.

[7] R. Bunescu and R. Mooney, Collective information extraction with relational Markov networks, *Proceedings of the Forty-Second Annual Meeting of the Association for Computational Linguistics*, article no. 438, 2004.

[8] C. Cardie, Empirical methods in information extraction, *AI Magazine*, vol. 18(4), pp. 65–79, 1997.

[9] M. Chau, J. Xu and H. Chen, Extracting meaningful entities from police narrative reports, *Proceedings of the Annual National Conference on Digital Government Research*, 2002.

[10] J. Chen, D. Ji, C. Tan and Z. Niu, Relation extraction using label propagation based semi-supervised learning, *Proceedings of the Twenty-First International Conference on Computational Linguistics and the Forty-Fourth Annual Meeting of the Association for Computational Linguistics*, pp. 129–136, 2006.

[11] J. Colombe and G. Stephens, Statistical profiling and visualization for detection of malicious insider attacks on computer networks, *Proceedings of the ACM Workshop on Visualization and Data Mining for Computer Security*, pp. 138–142, 2004.

[12] A. de Waal, J. Venter and E. Barnard, Applying topic modeling to forensic data, in *Advances in Digital Forensics IV*, I Ray and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 115–126, 2008.

[13] J. Finkel, T. Grenager and C. Manning, Incorporating non-local information into information extraction systems by Gibbs sampling, *Proceedings of the Forty-Third Annual Meeting of the Association for Computational Linguistics*, pp. 363–370, 2005.

[14] S. Garfinkel, Digital forensics research: The next ten years, *Digital Investigation* vol. 7(S), pp. S64–S73, 2010.

[15] J. Guo, G. Xu, X. Cheng and H. Li, Named entity recognition in query, *Proceedings of the Thirty-Second International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–274, 2009.

[16] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer and D. Benredjem, Towards an integrated e-mail forensic analysis framework, *Digital Investigation*, vol. 5(3-4), pp. 124–137, 2009.

[17] M. Hearst, Tilebars: Visualization of term distribution information in full text information access, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 59–66, 1995.

[18] M. Hearst and D. Rosner, Tag clouds: Data analysis tool or social signaller? *Proceedings of the Forty-First Annual Hawaii International Conference on System Sciences*, 2008.

[19] J. Jung, Online named entity recognition method for microtexts in social networking services: A case study of Twitter, *Expert Systems with Applications*, vol. 39(9), pp. 8066–8070, 2012.

[20] B. Klimt and Y. Yang, Introducing the Enron corpus, *Proceedings of the Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, 2004.

[21] C. Ku, A. Iriberri and G. Leroy, Natural language processing and e-government: Crime information extraction from heterogeneous data sources, *Proceedings of the Ninth Annual International Conference on Digital Government Research*, pp. 162–170, 2008.

[22] J. Kuperus, C. Veenman and M. van Keulen, Increasing NER recall with minimal precision loss, *Proceedings of the European Intelligence and Security Informatics Conference*, pp. 106–111, 2013.

[23] J. Lafferty, A. McCallum and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, 2001.

[24] A. Louis, A. De Waal and C. Venter, Named entity recognition in a South African context, *Proceedings of the Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries*, pp. 170–179, 2006.

[25] A. Louis and A. Engelbrecht, Unsupervised discovery of relations for analysis of textual data, *Digital Investigation*, vol. 7(3-4), pp. 154–171, 2011.

[26] A. Maedche and S. Staab, Ontology learning for the semantic web, *IEEE Intelligent Systems*, vol. 16(2), pp. 72–79, 2001.

[27] D. Maynard, V. Tablan, C. Ursu, H. Cunningham and Y. Wilks, Named entity recognition from diverse text types, *Proceedings of the Conference on Recent Advances in Natural Language Processing*, pp. 257–274, 2001.

[28] A. McCallum and W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, *Proceedings of the Seventh Conference on Natural Language Learning*, vol. 4, pp. 188–191, 2003.

[29] E. Minkov, R. Wang and W. Cohen, Extracting personal names from email: Applying named entity recognition to informal text, *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 443–450, 2005.

[30] R. Mooney and R. Bunescu, Mining knowledge from text using information extraction, *ACM SIGKDD Explorations Newsletter*, vol. 7(1), pp. 3–10, 2005.

[31] D. Nadeau and S. Sekine, A survey of named entity recognition and classification, *Lingvisticae Investigationes*, vol. 30(1), pp. 3–26, 2007.

[32] J. Okolica, G. Peterson and R. Mills, Using PLSI-U to detect insider threats from email traffic, in *Advances in Digital Forensics II*, M. Olivier and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 91–103, 2006.

[33] D. Newman, C. Chemudugunta, P. Smyth and M. Steyvers, Analyzing entities and topics in news articles using statistical topic models, *Proceedings of the Fourth IEEE International Conference on Intelligence and Security Informatics*, pp. 93–104, 2006.

[34] S. Orlando, P. Palmerini and R. Perego, Enhancing the Apriori algorithm for frequent set counting, *Proceedings of the Third International Conference on Data Warehousing and Knowledge Discovery*, pp. 71–82, 2001.

[35] M. Pollitt and A. Whitledge, Exploring big haystacks, in *Advances in Digital Forensics II*, M. Olivier and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 67–76, 2006.

[36] L. Rau, P. Jacobs and U. Zernik, Information extraction and text summarization using linguistic knowledge acquisition, *Information Processing and Management*, vol. 25(4), pp. 419–428, 1989.

[37] M. Schwartz and L. Liebrock, A term distribution visualization approach to digital forensic string search, *Proceedings of the Fifth International Workshop on Visualization for Computer Security*, pp. 36–43, 2008.

[38] A. Sun, R. Grishman and S. Sekine, Semi-supervised relation extraction with large-scale word clustering, *Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 521–529, 2011.

[39] C. Sutton and A. McCallum, An introduction to conditional random fields for relational learning, in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar (Eds.), MIT Press, Cambridge, Massachusetts, pp. 93–128, 2007.

[40] P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison-Wesley Longman, Boston, Massachusetts, 2005.

[41] J. Venter, A. de Waal and C. Willers, Specializing CRISP-DM for evidence mining, in *Advances in Digital Forensics III*, P. Craiger and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 303–315, 2007.

[42] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira and A. Singhal, SCAN: Designing and evaluating user interfaces to support retrieval from speech archives, *Proceedings of the Twenty-Second International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 26–33, 1999.

[43] Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang and M. Ishizuka, Unsupervised relation extraction by mining Wikipedia texts using information from the web, *Proceedings of the Joint Conference of the Forty-Seventh Annual Meeting of the Association for Computational Linguistics and the Fourth International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, vol. 2, pp. 1021–1029, 2009.

[44] D. Zelenko, C. Aone and A. Richardella, Kernel methods for relation extraction, *Journal of Machine Learning Research*, vol. 3, pp. 1083–1106, 2003.

[45] S. Zhao and R. Grishman, Extracting relations with integrated information using kernel methods, *Proceedings of the Forty-Third Annual Meeting of the Association for Computational Linguistics*, pp. 419–426, 2005.