SPECIAL SECTION ON ADVANCED DATA MINING METHODS FOR SOCIAL COMPUTING

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 26, 2019, accepted July 10, 2019, date of publication July 17, 2019, date of current version August 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929597

A Graph Based Clustering Approach for Relation Extraction From Crime Data

PRIYANKA DAS[1], (Student Member, IEEE), ASIT KUMAR DAS[1], JANMENJOY NAYAK[2], DANILO PELUSI[3], AND WEIPING DING[4], (Senior Member, IEEE)

[1]Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur 711103, India
[2]Department of Computer Science and Engineering, Sri Sivani College of Engineering, Srikakulam 532402, India
[3]Department of Communications Sciences, University of Teramo, 64100 Teramo, Italy
[4]School of Information Science and Technology, Nantong University, Nantong 226019, China

Corresponding author: Weiping Ding (ding.wp@ntu.edu.cn)

This work was supported by the Academic Research Funds of the University of Teramo, Italy.

ABSTRACT Application of natural language processing techniques based on crime data can prove to be beneficial in several processes of the criminal justice industry. The availability of massive crime reports helps law enforcement agencies when a criminal investigation is launched. While investigating a crime, questions like what type of crime, who committed the crime, what happened at which place, on what time, and what actions are taken, keep arising. Now, it is not feasible for the law enforcement agencies to get into the detail of these available massive crime reports and get the answers. To tackle these problems associated with criminal justice industry, the proposed work considers a textual corpus containing information of crime against women in India and extracts substantial relations between the named entities present in the corpus by a hierarchical graph-based clustering technique. For extracting the relations, different types of entity pairs have been chosen and similarities among them have been measured based on the intermediate context words. Depending on the similarity score, a weighted graph has been formed and a similarity threshold is set to partition the graph based on the edge weights. With the iterative application of the clustering algorithm, all the named entity pairs are grouped into clusters, each of which signifies different crime aspects. Each cluster is characterized using the most frequent context word present in it. The proposed relation extraction scheme helps in crime pattern analysis that can aid in various criminal investigation requirements. The results with optimal cluster validation indices depict the effectiveness of this method.

INDEX TERMS Crime analysis, named entity recognition, relation extraction, graph based clustering, cluster validation index.

I. INTRODUCTION

Textual information from forensic as well as criminal justice industries are increasing enormously and along with it, the data complexity has also increased. Manual annotation of these huge volumes of data is a very difficult task. Therefore, natural language processing techniques are mostly used for handling and processing these unstructured data by criminal investigators. Identifying named entities present in a text document helps in gaining knowledge about the persons involved in crime and discovering substantial relations among the identified named entities play an important part for taking proper actions in the criminal justice industry. To overcome the problems associated with relation extraction from crime data, some researchers have focused on supervised learning techniques that require a lot of human supervision from the criminalistic industries. But the supervision results in inducing biases for the learning process. Hence, considering the associated disadvantages of supervised learning techniques, researchers took up the challenge of using unsupervised approaches and clustering is one of the widely used methods in this context. The unsupervised approach deals with identifying named entities from large corpus and extracts the existing relational phrase from the entities. Not only it helps in achieving useful information about the entities, but also assists in further analysis of the text data for crime investigation. For example, in a sentence ‘‘**Srinath** has been accused of killing **Latika**’’, the relational tuple is considered as ⟨**X**, **a**, **Y**⟩, where, **a** represents the relation between the entities **X** (Srinath) and **Y** (Latika).

The associate editor coordinating the review of this manuscript and approving it for publication was Shirui Pan.

Domain-specific knowledge and application of several text mining techniques help in the improvement of relation determination task. The extracted relations mainly focus on several crime aspects that can help law enforcement agencies to predict future crime and take proper crime prevention strategies.

A. RELATED WORK

Relation detection task is known to have drawn attention since the 6th Message Understanding Conference (MUC-6) and Automatic Content Extraction (ACE) made further progress in this field. Prior to this relation detection, there has been detailed research on named entity recognition. Named entities are mostly recognized as the name of certain things. The entity recognition process mainly focuses on the presence of proper nouns in a corpus. In the year 1996, MUC-6 introduced seven basic named entities and those seven basic entities are classified as PER (person), ORG (organization), LOC (location), TIME, DATE, MONEY and GPE (geopolitical entity). But later it was observed that acknowledging more number of entities along with their subtypes present in heterogeneous datasets is quite beneficial for different applications of information extraction task. Hence, Sekine and Sudo [1] further extended the entities up to 150 types by considering the most probable subtypes for each basic entity. His work was able to describe that more effective relationship extraction can be performed by considering all feasible entities. In the past, newspapers [2] and heterogeneous data sources [3] have been explored for recognizing named entities. Brin [4] introduced 'DIPRE' (Dual Iterative Pattern Relation Expansion), where the vast World Wide Web was used for relation extraction using a semi-supervised approach called bootstrapping. The problems encountered in this method were solved by 'Snowball' [5] which used the elementary concepts related to 'DIPRE' and discovered novel methods for pattern extraction. Apart from this, some research works described in [6], [7] and [8] extensively applied semi-supervised approaches for relation discovery task. An unsupervised approach described in [9] used a named entity tagger for recognizing the entities present in 'The New York Times (1995)' newspaper corpus and the intervening context words of the entities have been hierarchically clustered for discovering the relations. But this approach performed the experiment with newspaper articles for one year, which failed to extract less frequent relations between the entity pairs and consecutively could not find paraphrases from them. Zhang *et al.* [10] showed better results than [9] by using a tree similarity-based clustering for relation detection from the same corpus of 'The NewYork Times (1995)' corpus. In Unsupervised Web Relation Extraction System (URES) [11], the target relations were considered as the input to the system and relations were extracted from the web pages in an independent manner. Few research works based on unsupervised approaches for relation discovery are described in [9]–[12]. Relations were discovered between noun categories in [13] by extensively analyzing several statistical approaches for suitable selection of the clusters. Particular patterns for the discovered relations were developed in [14]. This informed that feature generation technique results in better clustering of the entity pairs. Though most of the research works deal with candidate entities and identifying relations between them, Wang *et al.* [15] took all candidate relations into account and finally filtered and clustered the relations with a Conditional Random Field (CRF) model. Recently, Boujelben *et al.* [16] have applied linguistic models for determining relationships between Arabic Named Entities.

Basili *et al.* [17] introduced a system called 'REVEAL' that employed variants of support vector machine (SVM) for automatic relation extraction for crime investigation. Arulanandam *et al.* [18] extracted crime information from online newspapers. This work chose three different newspapers of three different countries and extracted the locations where theft related crimes occurred. They did it using entity recognition algorithms along with conditional random field. Again, Shabat and Omar [19] detected named entities present in a crime corpus and these identified named entities helped in the crime pattern analysis. Apart from named entity recognition, significant research works exist on extracting relevant relation among the entities. A project called "ASTREA" [20] developed a relation extraction system for crime investigation.

B. CONTRIBUTION

Nowadays, a lot of crime related information are available online. Though it is apparently easy to acquire knowledge from a single crime report at a glance, it takes a huge effort to deal with massive data for gaining perception of the crime pattern for a certain period of time. Many research works exist on extracting crime related information from crime reports, but there are very few of them analyze crime patterns using the concept of relation extraction [21]. To surpass the above mentioned problem, the proposed work demonstrates a graph based crime analysis scheme that emphasizes on determining relation between the entities present in a large corpus containing crime information of Indian states and union territories. The proposed work is considered to be a simple yet efficient contribution to the criminal justice industry. Initially, the method deals with extracting the crime data set from the electronic version of some classified newspapers. Several named entities like organizations, places, persons, etc have been recognized from the preprocessed data set by using an available named entity recognition module. The main objective of the proposed work is to discover the relation among the identified named entities by application of a top-down hierarchical graph based clustering technique. Different domain of entity pairs namely, PER-PER (person-person), PER-LOC (person-location) and ORG-PER (organization-person) are chosen for better visualization of the crime scene. For the relation discovery task, the entity pairs from each domain have been compared based on their intermediate context words and similarity has been measured among them.

Based on the similarity score, a complete weighted undirected graph has been generated where each node represents an entity pair and weight of an edge between two nodes is the similarity score between corresponding entity pairs. For the relation detection task, the primarily generated graph has been considered as a single partition [22]. The average value of all edge weights has been assigned as the threshold score. Two different subgraphs have been generated based on the threshold. The first subgraph contains the edges having equal and more weights than the threshold, whereas the second subgraph comprises the edges with weights below the threshold. The resultant two subgraphs may be the collection of several components which has been applied separately as input to the next iteration of the clustering algorithm. The threshold has been updated individually for each component and they are further partitioned into more compact subgraphs. At each level of iteration, a cluster validation index called Score Function (SF) [26] has been measured and the process continues only if the cluster quality improves. Finally, the graph clustering algorithm forms different groups of entity pairs. All the newly formed clusters have been labeled using the most frequent context words present in them. Context words for entity pairs belonging to PER-PER domain define the crime types like ‘rape’, ‘murder’, ‘abduction’, ‘molestation’ and many more, whereas the context words in PER-LOC domain describe the social status of the victim/offender like ‘teacher’, ‘mechanic’, ‘nurse’ etc. Likewise, the clusters from ORG-PER domain are characterized as the terms relating to the actions taken by the court or police against a criminal involved in crime. Some examples of those context words are ‘investigation’, ‘death sentence’, ‘penalty’ and many more. These relations obtained from the clusters help the criminal justice industry understand the crime pattern, the types of people involved in crime and actions taken against them. The clusters have been evaluated based on several external and internal cluster validation indices and finally, we have compared our proposed work with other existing relation detection methods.

Thus the contributions of the paper are concluded by the following steps:

1) The unstructured crime reports are collected and preprocessed by stopword removal, stemming and POS tagging. Then the named entities are recognized and paired as PER-PER, PER-LOC and ORG-PER domains.
2) For each domain of entity pairs, Word2Vec approach is applied to vectorize the context words present inside the entity pairs. Thus the structured data set of entity pairs is generated. Next a weighted undirected graph of entity pairs is constructed and the proposed hierarchical graph based clustering algorithm is applied to partition the entity pairs.
3) Each partition is labeled by the most influential context word. Then the labeled clusters are validated by various cluster validation indices to demonstrate its effectiveness. This simple but effective clustering technique is useful for both criminology and the criminal justice decision making.

The flowchart of the proposed work is given in FIGURE 1.

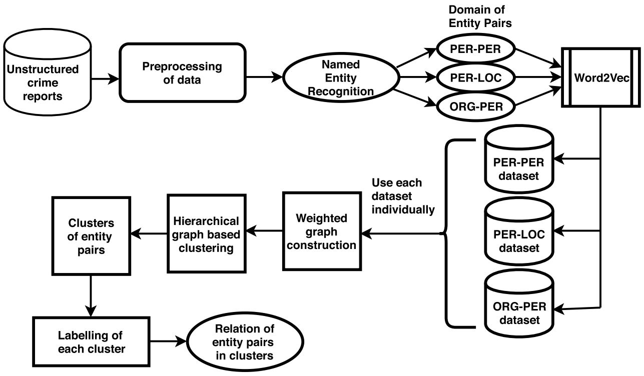

Unstructured crime reports
Preprocessing of data
Named Entity Recognition
Domain of Entity Pairs
PER-PER
PER-LOC
ORG-PER
Word2Vec
PER-PER dataset
PER-LOC dataset
ORG-PER dataset
Use each dataset individually
Weighted graph construction
Hierarchical graph based clustering
Clusters of entity pairs
Labelling of each cluster
Relation of entity pairs in clusters

FIGURE 1. Flowchart of the proposed methodology.

The remaining part of the article is organized as follows: Section II briefly discusses the background of the methods related to the proposed work. The proposed framework for graph based clustering is elaborately described in section III and section IV reflects the experimental results and the effectiveness of the method. Finally, section V draws the conclusion discussing the future scope of the paper.

II. PRELIMINARY CONCEPTS

This section demonstrates the preliminary concepts about the techniques used in the proposed methodology.

A. NAMED ENTITY RECOGNITION

The term ‘Named Entity’ evolved during the 6th Message Understanding Conference or MUC-6 consisting of the terms like ENAMEX (entity name expressions) and NUMEX (numerical expression). Named entities are mostly known as noun phrases or proper nouns that indicate person, organization, location, time, date, money and many more. The objective of any named entity recognition (NER) system is to find out all the named entities present in a text document. The procedure of named entity recognition comprises small steps given as follows:

1) Initially, the raw text document is being split into several sentences using the sentence segmenter.
2) Each word present in a sentence is represented as a token.
3) Parts-of-speech tagging of each token is done.
4) The noun phrases are identified as named entities.

These above mentioned steps involved the Natural Language Tool Kit (NLTK) module available in Python [23]. It is being performed easily by using NLTK’s in-built sentence segmenter, word tokenizer and parts-of-speech tagger. Next, chunking is done to segment and label multi-token sequences. FIGURE 2 shows the noun phrase chunking process for named entity recognition. For the sentence, “I saw the young girl”, the parts-of-speech tags are shown for each word. Here, PRP defines the pronoun, VBD refers to the past tense of verb,

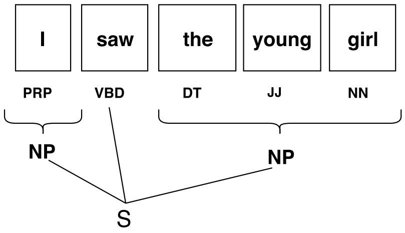


FIGURE 2. Noun phrase chunking for named entity recognition.

DT is the determiner, JJ is adjective and NN is the noun in singular form. Also, it is seen that 'I' as well as 'the young girl' are the noun phrases which are identified as NP.

Named entity recognition is often used as an essential step for relation extraction but it can also be used in other applications of information extraction.

B. RELATION EXTRACTION

Relation extraction explores distinct patterns between two entities that are present near one another in a text. Those explored patterns are used to form tuples that represent the relationship between two entities. To perform this task, two specified named entities are chosen as pairs and intermediate words between them are the context words that are known to represent the relation. For example, "**Srinath** has been accused of killing **Latika**", a tuple ⟨**X**, **a**, **Y**⟩ is considered where, **a** is the underlined intermediate context words and the tuple defining the relationship (is the accused murderer) between entities **X** (PER) and **Y** (PER). Here, the present work has focused on exploring the relationships between named entities identified from crime corpus using a graph based clustering technique.

III. GRAPH BASED CLUSTERING FOR RELATION EXTRACTION

The main objective of the proposed work is to determine the existing relationship between entities present in crime reports. The entity pairs having similar crime aspect are grouped together. This clustering scheme helps in criminal justice industry. Criminal investigators can consider data set for several years, apply this simple but efficient algorithm and gain insight on the criminal justice industry.

A. PREPROCESSING OF DATA

Once the data sets have been collected, those data have been preprocessed by removing all the stopwords present in them. NLTK has a predefined list of stopwords in it. After passing the sentences through NLTK, all the words that are present in the predefined stopwords list get removed from the texts of the datasets. We can append more words to the list on our own. The stopwords removal process is being followed by stemming that gives the root words discarding the suffixes. Then parts-of-speech or POS tagging is done that identifies the tags of each word and the noun phrases are considered for further processing. These noun phrases are acknowledged as named entities in the present work. All these mentioned preprocessing steps have been achieved by using the Natural Language Tool Kit (NLTK).

B. ACCUMULATION OF ENTITY PAIRS

The named entities present in the data are recognized by the method as mentioned in section II-A. The identified entities are paired as PER-PER (person-person), PER-LOC (person-location) and ORG-PER (organization-person) domains to facilitate our proposed crime analysis scheme. The intervening context words between the entity pair are known to represent the relationship between them. For example, a sentence, say "*Shamita* has been abused at work by *Rahul*". Here, both the italicized words define the entity PER (person) and the underlined words are the context words that define the relationship between *Shamita* and *Rahul*. Similarly say, "*Raman*, a software employee was stabbed to death at *Saltlake*" is a sentence in PER-LOC domain. Here, the italicized words are the entities like PER (person) and LOC (location) and underlined words are the context words defining the social status of *Raman*. Again, consider another example like, "*High Court* has declared imprisonment to *Anand*". The italicized words are entities like ORG (organization) and PER (person) and the intervening context words define the action taken by the *High Court* against *Anand*. Now, in each of the above examples, the context words reflect the relationship between the entities in their corresponding pair. There exist many such sentences in the crime reports that depict similar relationships. Therefore, the objective of the proposed work is to cluster the entity pairs based on these context words. These context words are also used to label the clusters. Thus, for each entity pair, relation of the first one has been determined with the latter. All the chosen entity pairs from different domains are accumulated separately and the presence of all stemmed intermediate words are considered as context of the pair of entities. For each entity pair, a context vector is created using the Word2Vec approach [24]. Word2Vec approach considers the intermediate context words from the entity pairs as the input and creates a potentially high dimensional vector space, where each unique context vector represents a p-dimensional feature vector that characterizes the relationship between the associated entity pair. Word vectors exhibiting contextual similarity stay in close proximity to each other in the vector space. The advantage of using Word2Vec approach other than frequency (Term Frequency-Inverse Document Frequency) based approaches is that Word2Vec method helps in semantic analysis of the corpus. Though both the Word2Vec and GloVe are models for generating word embeddings, but the proposed work is a relation extraction scheme which mainly emphasizes on the context words of the entity pairs for predicting the crime aspect, so we have used the Word2Vec model as it is a predictive model. Wor2Vec model learns the vectors in order to improve their predictive ability of the loss for predicting the target words from the context words provided the vector representations are given.

C. SIMILARITY MEASURE AMONG ENTITY PAIRS

Once all the entity pairs are represented by context vectors, the aim is to measure similarity among the entity pairs based on the context vector of the associated context words. For this purpose, *Cosine Similarity* has been measured between every pair of entity pairs, using (1). It basically compares the context of say, one PER-LOC pair with another PER-LOC pair and same in case of other entity pairs of different domains.

$$S_{xy} = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{||\overrightarrow{x}||||\overrightarrow{y}||} \tag{1}$$

where $\overrightarrow{x}$ and $\overrightarrow{y}$ are two context vectors of associated context words of two corresponding entity pairs.

Next, a complete weighted undirected graph termed as *Entity-pair Similarity Graph*, $G = (N, E, W)$ is formed, where N represents set of nodes (entity pairs), E is the set of edges (connection between entity pairs) connecting the nodes and W defines the set of weights (similarity between entity pairs) of the edges. For the present work, n_1 and n_2 are two nodes in N representing two context vectors, say a and b, whereas the edge between the nodes n_1 and n_2 has the weight equal to the similarity score between a and b, computed using (1). All the entity pairs have been considered as nodes. Higher the similarity among the entity pairs, more the weight assigned to their corresponding edges. Cosine similarity factor ranges in [0,1], where 1 denotes the entity pairs having the most similar context words and 0 defines the maximum dissimilarity.

D. CLUSTERING AND LABELLING OF ENTITY PAIRS

Once the *Entity-pair Similarity Graph*, $G = (N, E, W)$ has been constructed, the aim is to discover relations between the named entity pairs. The present work has used a graph based hierarchical clustering algorithm for extracting relations between named entities. Upon constructing the complete graph G, the average value W_{avg} of all edge weights present in the original complete graph has been calculated using (2) and considered as a threshold.

$$W_{avg} = \frac{\sum_{e \in E(G)} W(e)}{|E|} \tag{2}$$

where, $W(e)$ defines the weight of the edge $e \in E$.

Based on the threshold, the initial complete graph G has been partitioned into two subgraphs:

1) $G_1 \rightarrow$ subgraph with edges having weights at least equal to the threshold.
2) $G_2 \rightarrow$ subgraph with edges having weights below the threshold.

Obviously, G_1 and G_2 may be disconnected graphs and thus after the application of this clustering algorithm, a disconnected graph with many connected components has been obtained as the resultant graph. Thus at LEVEL 0, G is the singleton cluster for the data set. But at LEVEL 1, when G_1 and G_2 are constructed, all the components are the clusters for the data set. For further clustering of the data set, each individual component obtained at LEVEL 1 is treated as a new graph G' and partitioned similarly into G_1 and G_2. If the components obtained in G_1 and G_2 give better clusters than G' w.r.t some quality measures then G' is replaced by G_1 and G_2; Otherwise G' is passed to LEVEL 2. Thus the components at LEVEL 1 are either further partitioned into a new set of components or simply remained as the same component. This resultant set of components is the set of clusters at LEVEL 2. In this way, the clusters are generated hierarchically using top down approach starting from the singleton cluster G at LEVEL 0. The hierarchical top down approach is illustrated using an example in FIGURE 3. In the worst case a component may be partitioned in each level until all the subcomponents are individual edges. But in real life applications, many objects together form a cluster and so the method of partitioning a component into subcomponents terminates based on many different conditions. Few of them are describes as follows:

1) Edge weights of all the edges in the component are same.
2) Based on the user's choice, after a certain number of levels when desired number of clusters are achieved.
3) After partitioning a component into subcomponents, various cluster validation indices are measured based on new set of clusters. If the index values degrade then partitioning is not allowed and the previous component remains intact.

We have applied condition (3) to terminate the partitioning of a component. As there are many cluster validation indices [26], we may use any one or subset of indices in our task for measuring cluster quality. In this paper, a bounded validity index, called Score Function (SF) ia used to achieve the correct number of quality clusters. The main reason for using *SF* index is that it is applicable for a data set with single cluster too where the other indices require at least two clusters of the data set. But in the proposed hierarchical clustering algorithm, initially whole data set is a single cluster. So if the data set itself is the actual single cluster G then index value of G must be better than that of clusters obtained by G_1 and G_2. But without using SF index, we cannot measure quality of the cluster G. The other reason to use this index is that it is computationally less expensive than most of the other validation indices. It runs in $O(|N|)$ time where N is the set of objects in the data set. The computation of *SF* index value is discussed in the Experimental Results section of the paper.

Though the partitioning of a component is terminated computing SF index, but other components may be partitioned. Thus levels of the tree are increased. Also all the components of current level are examined before examining the components of the next level. To achieve it, a queue is implemented where all the components generated in a level are inserted together. So when a component is removed for further partitioning, either it is not partitioned at all or partitioned into subcomponents which are the components in the next level of the tree. As these subcomponents are inserted into queue, they will be examined after removing all components of current

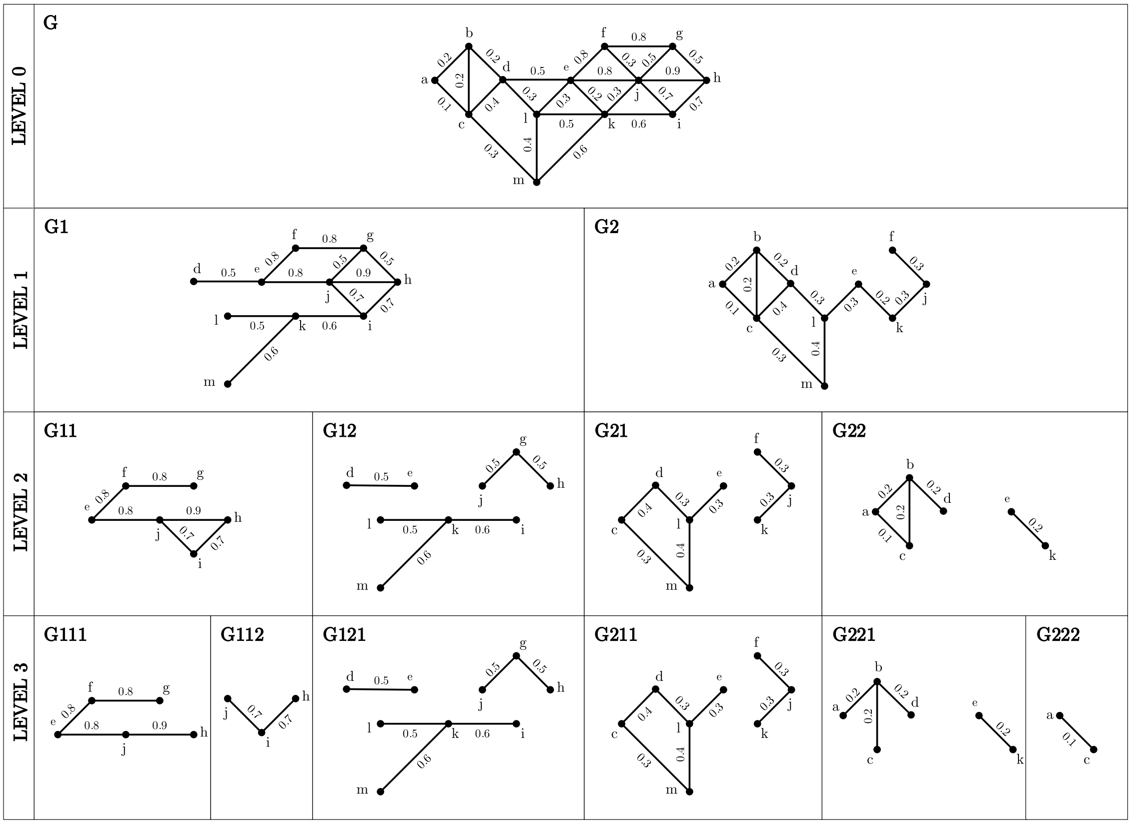


FIGURE 3. Illustration of the steps performed in the proposed methodology.

level from the queue. Thus the components are partitioned level wise starting from LEVEL 0. The methodology of entity pair clustering is described in Algorithm 1.

After convergence, the above mentioned graph based clustering algorithm creates few groups of entity pairs where each group contains the entity pairs which are similar in relation. We have calculated the frequency of each context word present in the entity pairs in each cluster and then the groups of named entity pairs have been tagged with the most frequent context word present in them. Here, the term tagging is similar to labeling or characterizing the clusters. The context word for entity pairs belonging to PER-PER (person-person) domain defines the crime types like 'rape', 'murder', 'abduction', 'molestation' etc., whereas the context word for PER-LOC (person-location) domain describes the social status of the victim/offender. Likewise, the clusters from ORG-PER (organization-person) domain are characterized by the terms relating to the actions taken by the court or police against a criminal involved in crime. Thus, all the clusters from the all three domains have been labeled by this process. This cluster labeling process helps in identifying the crime patterns extracted from criminological data and the criminal investigators are benefited by gaining insight on the persons involved in crime, the types of crime that are taking place for a certain period of time and how the organizations are acting against the criminals. This simple yet effective clustering technique can contribute to both criminology and criminal justice decision making.

E. TIME COMPLEXITY ANALYSIS

Initially, the graph contains $|N|$ number of nodes and $|E|$ number of edges. In first iteration the graph is partitioned into two subgraphs G_1 and G_2 both of which may be disconnected graphs. To construct G_1 and G_2, total time require is $O(|N| + |E|)$, since the graphs are represented by adjacency list. Let k_1 be the total number of components in G_1 and G_2. Simply by using breadth-first search or depth-first search, k_1 components can be computed in linear time, in terms of number of nodes and edges of the graph. So time needed to construct k_1 components is $O(|N||E|)$. Also the *SF* index is computed in $O(|N|)$ time. So each iteration of **repeat until** loop contributes to the running time of $O(|N|) + O(|E|) + O(|N||E|) + O(|N|) = O(|N||E|)$. The loop is continued until the queue is empty. Now the following cases are considered:

- If there is a single cluster for the whole data set, then the loop will execute only once. So the time complexity is $O(|N||E|)$. This is the best case scenario.
- If all the clusters are the components of single edge, then $|E|$ number of clusters are formed. So the time complexity is $O(|E|)O(|N||E|)$. This is the worst case scenario.

Algorithm 1 Graph Based Clusters of Entity Pairs

```
Input: $G = (N, E, W)$, where $N$ = set of nodes (i.e., set
       of entity pairs) in $G$, $E$ = set of edges and $W$ =
       set of weights of the edges in $E$.
Output: Set $C_E$ of clusters of entity pairs.
begin
    $C_E = G$ /* Initially the whole graph is a single
    cluster */;
    Compute cluster validation index $SF_{old}$ for cluster
    $C_E$ using (8–10);
    Insert $G$ into the queue $q$ /*each entry of $q$ contains
    one graph */;
    repeat
        Remove next graph $G'$ from $q$;
        Let $G' = (N', E', W')$ where $N'$ and $E'$ are the
        set of nodes and edges, respectively and $W'$ is
        the set of weights of edges in $G'$;
        Calculate average weight $W_{avg}$ in $G'$ using (2).
        $G_1 = \phi$, $G_2 = \phi$ /* $G'$ is partitioned into $G_1$ and
        $G_2$, both of which are initially empty */;
        for each edge $e \in E'(G')$ do
            if $W(e) \geq W_{avg}$ then
                $G_1 = G_1 \cup \{e\}$;
            end
            else
                $G_2 = G_2 \cup \{e\}$;
            end
        end
        $C_{temp} = \phi$ /* temporarily generated clusters
        from $G_1$ and $G_2$ */ ;
        for each component g of $G_1$ and $G_2$ do
            /* component is the connected subgraph of a
            graph */
            $C_{temp} = C_{temp} \cup \{g\}$;
            /* each component represents one cluster */;
        end
        $C_{new} = C_{temp} \cup C_E - \{G'\}$ /* new set of clusters
        */;
        Compute cluster validation index $SF_{new}$ for
        clusters $C_{new}$ using (8–10);
        if $SF_{new} > SF_{old}$ then
            $C_E = C_{new}$ /* old set of clusters are replaced
            by new set of clusters */;
            $SF_{old} = SF_{new}$;
            for each $g \in C_{temp}$ do
                insert $g$ into queue $q$;
            end
        end
    until q is empty;
    Return $C_E$;
end
```

- If the number of components is constant, say k then k number of clusters are formed. In this case, time complexity of the algorithm is $O(k|N||E|) = O(|N||E|)$.

IV. EXPERIMENTAL RESULTS

The proposed work has been implemented using Python 3.6 with its several modules like numpy 1.14, networkx 2.1, matplotlib 2.2.

A. DATA COLLECTION

Online version of Indian classified newspapers like ‘The Times of India’, ‘The Hindu’ and ‘The Indian Express’ have been chosen for collecting the newspaper reports on crime against women in Indian states and union territories. A Python based site crawler has been designed to look through the aforementioned newspaper websites and search for terms related to crime like ‘rape’, ‘abduction’, ‘molest’ and many more. Reports containing any of the tags have been extracted from the corresponding sources. The extracted data is based on different crimes committed against women in several states and union territories of India. The collected data set comprises a total of 200,150 crime reports for 29 states and 4 union territories of India for over a time period of 2004–2016. The obtained reports contain information on the locality, which includes names of the cities and districts. Once the data set has been collected, the basic preprocessing has been done and then we have considered 5,447 entity pairs from PER-PER domain, 5,341 entity pairs from PER-LOC domain and 6,214 number of entity pairs from ORG-PER domain.

After applying the proposed graph based clustering algorithm, several clusters of named entity pairs are formed. The algorithm recognizes the clusters of PER-PER domain in 14 *ms* ± $165\mu s$ per loop (mean ± standard deviation of 7 runs. 100 loop each) in a computer running Ubuntu GNU/Linux version 16.04 on an Intel(R) Core i3-5005U CPU @ 2.00 GHz processor. TABLE 1 shows the run time required for each domain of entity pairs by the proposed graph based clustering technique.

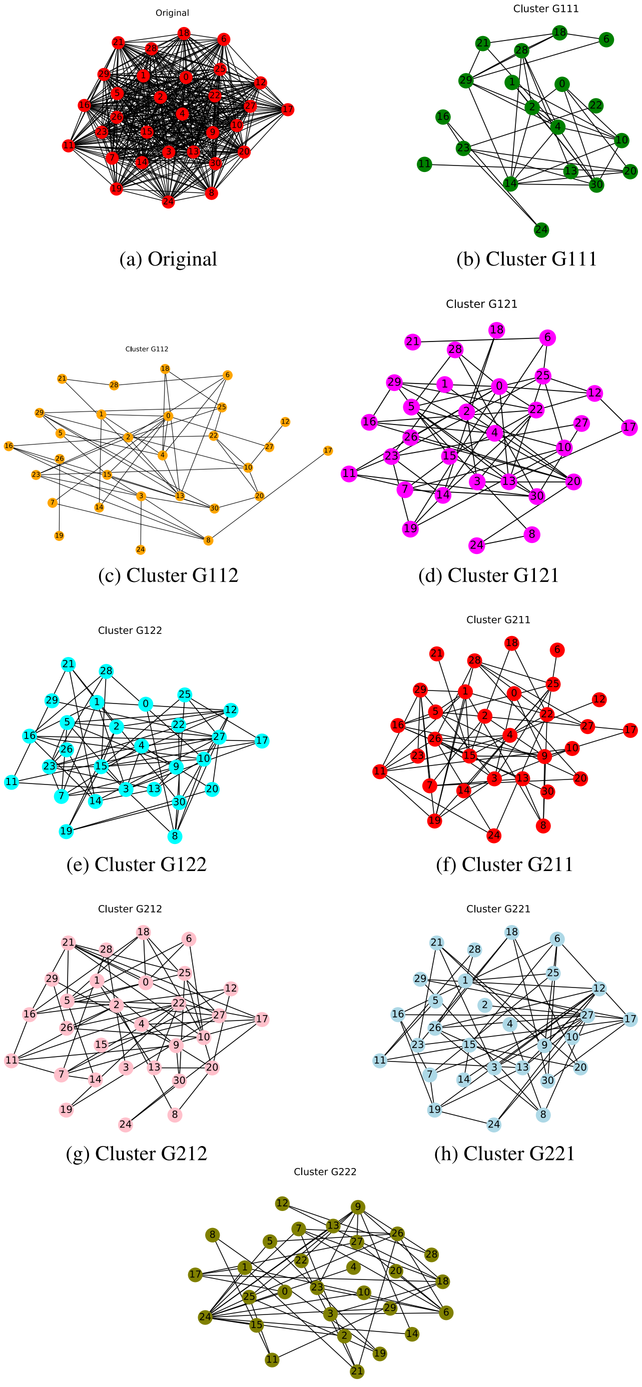
TABLE 1. Processing time of the proposed method for different dataset of entity pairs.

Domain of Dataset	Run Time
PER-PER	$14ms \pm 165\mu s$
PER-LOC	$7.07ms \pm 121\mu s$
ORG-PER	$7.07ms \pm 118\mu s$

TABLE 2. Number of clusters formed by the proposed clustering technique.

Domain	Clusters formed
PER-PER	12
PER-LOC	14
ORG-PER	10

TABLE 2 shows number of clusters formed by the proposed hierarchical graph based clustering algorithm. FIGURE 4 shows the original graph and the resulting subgraphs formed by the proposed clustering algorithm. This figure has been generated by considering 31 entity pairs from



(a) Original (b) Cluster G111

(c) Cluster G112 (d) Cluster G121

(e) Cluster G122 (f) Cluster G211

(g) Cluster G212 (h) Cluster G221

(i) Cluster G222

FIGURE 4. Subgraphs generated by the proposed graph based clustering for PER-LOC named entity pairs.

TABLE 3. Number of entity pairs for relation tagging/labeling in PER-PER domain.

Type	NE Pairs	Type	NE Pairs
Murder	225	Domestic Violence	234
Rape	376	Acid Attack	128
Molestation	349	Abuse	321
Kidnap	307	Human Trafficking	273
Sexual Harassment	265	Street Harassment	170
Dowry Death	137	Foeticide	113

TABLE 4. Number of entity pairs for relation tagging/labelling in PER-LOC domain.

Type	NE Pairs	Type	NE Pairs
Teacher	174	Bishop	162
Driver	246	Housemaid	351
Student	363	Techie	348
Labourer	269	Mechanic	169
Housewife	223	Party Leader	141
Doctor	119	Businessman	209
Relative	217	Teenager	341

TABLE 5. Number of entity pairs for relation tagging/labelling in ORG-PER domain.

Type	NE Pairs	Type	NE Pairs
Arrested	379	Investigation	381
Convicted	343	Order Probe	212
Penalty	265	Seize Property	219
Death Sentence	167	Order DNA test	139
Lifetime Imprisonment	216	Produce Chargesheet	183

PER-LOC (person-location) domain as total of 5,341 pairs are difficult to visualize by the figure.

Once the clustering is done, the next task is to assign a label to the clusters. For this purpose, the particular context word with maximum frequency is chosen as the label of the cluster. Thus the relation labeling has been done for all the clusters. TABLE 3–5 shows the result of relational labeling for the clustered entity pairs in each domain. It shows the distribution of the total number of entity pairs primarily considered for the present work. The relational labeling of the clusters caters various aspects of crime analysis. It not only focuses on the crime types but also emphasizes on the social status of the victims or actions taken by both the victims and governmental organizations for prevention of the crime. This analysis part holds the main significance for the proposed relation detection scheme.

B. EVALUATION RESULTS

For the evaluation purpose, the authors have assessed the generated clusters representing different crime aspects with respect to the ground truth clusters obtained by domain experts. External cluster evaluation techniques [25] like Purity (Pr), Precision (P), Recall (R), and F-measure (F) and Random Index (RI) have been computed using (3 - 7):

$$\text{Purity (Pr)} = \frac{1}{N}\sum_{i=1}^{c} \max_{j,1\leq j\leq c} |k_i \cap k'_i| \tag{3}$$

Here, N refers to the number of objects or entity pairs, c is the number of clusters, k_i and k'_i are clusters generated by the proposed graph based clustering algorithm and domain experts, respectively.

$$\text{Precision (P)} = \frac{T_p}{T_p + F_p} \tag{4}$$

$$\text{Recall (R)} = \frac{T_p}{T_p + F_n} \tag{5}$$

$$\text{F-measure (F)} = \frac{2PR}{P + R} \tag{6}$$

$$\text{Random Index (RI)} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{7}$$

where, the terms T_p, F_p, T_n and F_n refer to true positive, false positive, true negative and false negative, respectively.

Also, few internal cluster evaluation indices [26] like Score Function, Dunn, Davies-Bouldin, Silhouette, NIVA and Calinski-Harabasz [21] indices are computed, where Euclidean distance is used to measure the similarity between the objects.

The Score Function (SF) [26] uses "between class distance" called separability and "within class distance" called compactness of clusters.

The separability is given by (8),

$$\text{Sep} = \frac{1}{nc} \sum_{i=1}^{c} d(c_i, c_{all})^2.n_i \tag{8}$$

where n is the total number of objects, c is the number of clusters, c_i is the centroid of the i-th cluster, c_{all} be the centroid of all the data objects and n_i be the number of objects in i-th cluster. In (8), each distance is weighted by the cluster size n_i to limit the influence of outliers. It has the effect to reduce the sensitivity to noise. Here, n is used to avoid the sensitivity of separability to the total number of objects. Finally, c in the denominator is used to penalize the addition of new clusters. Thus Sep reduces as c increases. On the other hand, the Compaction is given by (9),

$$\text{Comp} = \frac{1}{c} \sum_{i=1}^{c} \sqrt{\frac{1}{n_i} \sum_{x \in X_i} d(x, c_i)^2} \tag{9}$$

where, n_i is the number of objects in cluster X_i and x be the object in X_i. Then the score function is defined by (10).

$$\text{SF} = 1 - \frac{1}{e^{e^{sep-comp}}} \tag{10}$$

The Score Function is in between 0 and 1, i.e., $0 < SF < 1$, which deals the one cluster case. Larger the SF index implies better the clusters are.

The Dunn Index (DN) determines clusters which are compact and well separated. Thus it minimizes the intra-cluster distance and maximizes the inter cluster distance. The Dunn Index (DN) for c clusters is defined by (11),

$$\text{DN} = \min_{1 \le a \le c} \{\min_{a \neq b} \{\frac{d(X_a, X_b)}{\max_{1 \le k \le c}(d(X_k))}\}\} \tag{11}$$

where, $d(X_a, X_b)$ is the inter cluster distance between the clusters X_a and X_b, $d(X_k)$ is the intra cluster distance of cluster X_k and c is the number of clusters. Higher value of Dunn index represents good clustering.

Similar to the Dunn Index, Davies-Bouldin Index (DB) determines clusters which are compact and well separated from each other. The DB index for a set of c clusters is defined by (12),

$$\text{DB} = \frac{1}{c} \sum_{a=1}^{c} \max_{a \neq b} \{\frac{d(X_a) + d(X_b)}{d(X_a, X_b)}\} \tag{12}$$

where, a and b are cluster labels, c is the number of clusters. $d(X_a)$ and $d(X_b)$ are intra cluster distance of clusters X_a and X_b respectively and inter cluster distance $d(X_a, X_b)$ between clusters X_a and X_b is measured as the distance between the cluster centroids. The minimum value of DB index denotes good clustering.

The Silhouette (SL) index of a set of c clusters is another very useful statistic to estimate the actual number of clusters in a data set. This index is computed for each sample point i in each of the c clusters and finally, average of all computed values is the SL index of the set of c clusters. The SL index of clusters is defined using (13),

$$\text{SL} = \frac{1}{n} \sum_{i=1}^{n} \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{13}$$

where, n is the number of objects, a_i is the average distance between i-th sample and all other samples in its own cluster and b_i is the distance of i-th sample to its nearest cluster. The maximum value of SL index provides the optimal set of clusters.

NIVA index has been computed as mentioned in [21], using (14),

$$\text{NI} = \frac{Comp}{Sep} \tag{14}$$

where *Comp* and *Sep* represent the compactness and separability of the set of clusters c. The minimum values of NIVA index represent good clustering.

Similarly, Calinski-Harabasz index is also calculated as discussed in [21], using (15).

$$\text{CH} = \frac{InterScat}{IntraScat} . \frac{n - c}{c - 1} \tag{15}$$

Here, the terms Interscat and Intrascat are intercluster scatter and intracluster scatter, respectively. n is the number of objects and c is the number of clusters. Higher values of CH index indicate optimal clustering.

TABLE 6. Results in (%) for external cluster validity indices.

Domain	P	R	F	Pr	RI
PER-PER	**79**	**82**	**80**	**81**	**76**
PER-LOC	76	81	78	77	75
ORG-PER	76	81	78	74	75

TABLE 6 provides the external cluster evaluation result for relational labeling of entity pairs. Focusing on the F-measure, it is observed that relational labeling in PER-PER domain has been done efficiently with the highest F-measure of 80 and almost similar result of 78 F-measure score has been achieved for both PER-LOC and ORG-PER domains. This result also provides the insight on how good the clusters are formed. Also, the highest Purity score has been obtained for PER-PER domain. The best scores corresponding to each domain and metric are marked in bold face.

TABLE 7. Results in (%) for internal cluster validity indices.

Domain	DN	DB	SL	NI	CH	SF
PER-PER	**81**	42	74	**54**	**76**	**82**
PER-LOC	75	62	**76**	61	69	79
ORG-PER	71	**40**	73	64	71	77

TABLE 7 describes the internal cluster evaluation result. It is known that lower values of Davies-Bouldin and NIVA indices, higher values of Dunn, Silhouette, Calinski-Harabasz and Score Function indices are given by optimal clustering. Therefore, from the results, it is observed that Dunn index provides the best result for PER-PER domain, whereas, Silhouette index provides good result almost in all cases. The lowest value of 40 of DB index in case of ORG-PER domain provides the best result. Smallest value of 54 in NIVA index and highest value of 76 in Calinski-Harabasz index provide the best results for PER-PER domain. Also, the SF index yields a value of 82 for PER-PER domain. The main reason behind obtaining the best scores for PER-PER domain is that the proposed work recognizes the crime types most efficiently.

C. COMPARATIVE STUDY

The present work has carried out the comparative study by three set of experiments. Initially, the proposed graph based algorithm has been compared with other graph based clustering algorithms that exist in the literature. Next, we have compared the present work with other existing relation extraction techniques. Finally, we have considered some other real data sets and applied our proposed methodology on those data sets.

1) COMPARISON WITH GRAPH BASED CLUSTERING ALGORITHMS

Four existing methods such as Infomap [27], Louvain [28], Girvan-Newman [29] and Fastgreedy [30] algorithms have been considered. All these methods are graph based and do not require any prior mentioning of the number of clusters. We have used the present crime data for these graph based clustering techniques and also used the external as well as internal cluster evaluation indices for measuring the effectiveness of the relational labeling of the clusters formed by these graph based methods. TABLE 8 represents the number of clusters formed by these comparative graph based techniques. The clusters have been labeled in the same way as the present scheme and TABLE 9 shows the result for evaluating the labeling of the clusters. It is observed that all the internal as well as external indices obtained by the proposed method as shown in TABLE 6 and TABLE 7 are better than that obtained by the comparative methods as shown in TABLE 9. The best scores for each method for the corresponding metrics are marked by bold face.

TABLE 8. Number of clusters formed by Infomap, Louvain, Fastgreedy and Girvan clustering algorithms.

Domain	Infomap	Louvain	Fastgreedy	Girvan
PER-PER	29	21	19	25
PER-LOC	21	26	18	13
ORG-PER	24	21	26	28

TABLE 9. Comparative result in (%) for external and internal cluster validity indices for Infomap, Louvain, Fastgreedy and Girvan algorithm.

Domain	Measures	Infomap	Louvain	Fastgreedy	Girvan
PER-PER	P	72	66	**74**	69
	R	74	67	**76**	72
	F	73	66	**75**	70
	Pr	**71**	68	67	70
	RI	67	71	69	**72**
	DN	**77**	74	75	71
	DB	51	49	**46**	51
	SL	72	**73**	**73**	69
	NI	59	61	**56**	59
	CH	73	67	70	**74**
	SF	**77**	74	71	73
PER-LOC	P	69	**71**	**71**	70
	R	70	70	**73**	**73**
	F	69	70	**72**	71
	Pr	71	63	67	**73**
	RI	66	**71**	69	63
	DN	**71**	69	69	65
	DB	37	33	**28**	29
	SL	69	**74**	**74**	69
	NI	52	**51**	70	61
	CH	**75**	70	68	71
	SF	**76**	74	75	73
ORG-PER	P	71	69	**74**	71
	R	71	71	**74**	**74**
	F	71	70	**74**	72
	Pr	**73**	63	69	70
	RI	60	67	**74**	71
	DN	**74**	72	71	64
	DB	43	41	31	**23**
	SL	69	**74**	**74**	69
	NI	**53**	58	61	63
	CH	**75**	68	66	70
	SF	71	73	**75**	71

TABLE 10. Comparative result in (%) for F-measure based on The New York Times (1995) corpus.

Domain	Proposed Work	Work in [10]	Work in [9]
PER-GPE	89	87	82
COM-COM	83	80	77

TABLE 11. Comparative result in (%) for F-measure based on present crime corpus.

Domain	Present Work	Work in [10]	Work in [9]
PER-PER	80	74	75
PER-LOC	78	72	74
ORG-PER	78	71	72

TABLE 12. Description of the datasets used for evaluating the present algorithm.

Dataset	Description
MLB	This is about Major League Baseball dataset for the year 2008. It has been collected from the site of Maximal Information-based Non parametric Exploration (MINE) application statistics [31] which helps in identifying important relationships in large datasets and also characterises them. MLB dataset contains individual offensive statistics from 2008 Major League Baseball season. It comprises the names of 337 baseball players with 134 different attributes related to the baseball prospectus.
WHO	The WHO dataset contains a set of social, economic, health, and political indicators using data from the World Health Organization and partner organizations. It holds the information about 202 countries with 358 attributes. This has also been collected from the MINE database.
Tweet data	This tweet data has been collected from a twitter API. It contains 1065 tweet ids and 63 attributes related to the ids.
Road Accident	This dataset has been collected from Open Government Data (OGD) platform, India. This data provides the information about road accidents occurred in Indian states during the year 2014.

TABLE 13. Number of clusters generated by the proposed and existing graph based clustering algorithms.

Dataset	Proposed	Infomap	Fastgreedy	Girvan	Louvain
MLB	9	10	5	6	6
WHO	10	8	5	7	7
Tweet	11	3	4	11	4
Road Accident	10	3	2	6	4

TABLE 14. Result in (%) of external cluster evaluation indices for Proposed work, Infomap and Fastgreedy approach on real data sets.

Dataset	Proposed					Infomap					Fastgreedy				
	P	R	F	Pr	RI	P	R	F	Pr	RI	P	R	F	Pr	RI
MLB	71	73	72	73	**75**	69	**73**	71	69	67	66	69	67	59	63
WHO	**72**	**75**	**73**	69	73	**71**	**73**	**72**	63	68	70	71	70	65	64
Tweet	69	73	71	69	67	**71**	72	71	63	**69**	**71**	**72**	**71**	**70**	63
Road Accident	69	71	70	**74**	73	61	64	62	**71**	62	62	62	62	67	**65**

TABLE 15. Result in (%) of external cluster evaluation indices for Girvan and Louvain method on real data sets.

Dataset	Girvan					Louvain				
	P	R	F	Pr	RI	P	R	F	Pr	RI
MLB	**70**	**73**	**71**	62	58	**69**	**72**	**70**	**66**	63
WHO	69	72	70	67	59	**69**	71	**70**	61	57
Tweet	68	**73**	70	63	61	61	65	63	60	62
Road Accident	67	71	69	**74**	**63**	62	65	63	61	**67**

The highest F measure of 80 as mentioned in TABLE 6 and highest SF score of 82 as shown in TABLE 7 for entity pairs from PER-PER domain implies that the proposed approach works best for identifying the crime types reflecting from the existing context words between the entities. It may also be noted that we have achieved better results for PER-PER domain than others as more instances of crime types were present in the dataset than other instances reflected by other domain of entity pairs.

TABLE 16. Results in (%) for internal cluster evaluation indices for the proposed graph based clustering algorithm on real data sets.

Dataset	Present Work					
	DN	DB	SL	NI	CH	SF
MLB	77	52	65	**46**	67	72
WHO	**81**	37	**74**	51	**71**	**79**
Tweet	79	**32**	67	54	68	68
Road Accident	71	38	64	59	65	74

As mentioned previously, the terminating condition of our experiment depends on the SF index. We have terminated the hierarchical clustering when we achieved the optimal results for the mentioned SF index. It is observed that all the other methods have provided better result (minimum values) than ours for DB index in case of PER-LOC domain. Here also, the best results have been achieved for PER-PER domain as it obtains both the higher and lower values for each indices.

2) COMPARISON WITH OTHER EXISTING RELATION DETECTION TECHNIQUES

Hasegawa *et al.* [9] considered newspaper reports 'The New York Times' for a single year (1995) and proposed a relation detection that ultimately resulted in a huge time

TABLE 17. Results in (%) for internal cluster evaluation indices for Infomap and Fastgreedy clustering algorithms on real data sets.

Dataset	Infomap						Fastgreedy					
	DN	DB	SL	NI	CH	SF	DN	DB	SL	NI	CH	SF
MLB	74	34	61	52	**66**	61	65	41	67	61	66	**72**
WHO	72	47	62	53	59	**65**	70	45	67	57	67	71
Tweet	74	**24**	54	**46**	52	62	**79**	**34**	**69**	**49**	58	69
Road Accident	**75**	27	**74**	63	61	64	73	38	66	59	**69**	71

TABLE 18. Results in (%) for internal cluster evaluation indices for the Girvan-Newman and Louvain clustering algorithms on real data sets.

Dataset	Girvan						Louvain					
	DN	DB	SL	NI	CH	SF	DN	DB	SL	NI	CH	SF
MLB	71	50	61	63	71	73	68	37	57	**54**	69	71
WHO	**74**	**24**	**71**	56	69	75	70	34	**73**	69	**74**	68
Tweet	67	**24**	**71**	61	**74**	**79**	**71**	**27**	63	65	69	71
Road Accident	69	45	67	**55**	61	74	**71**	39	71	62	71	**75**

complexity issue when tested on a larger dataset. They considered entity pairs of PER-GPE (person-geo-political entity) and COM-COM (company-company) domain for their work. Again, Zhang *et al.* [10] showed that their proposed tree similarity based clustering outperforms the result of [9]. For the comparison purpose, at first we have considered the same corpus of 'The New York Times (1995)', collected the PER-GPE (person-geo political entity) and COM-COM (company-company) domain of entity pairs and applied our proposed method. TABLE 10 shows the comparative result based on F-score. Next, we have applied the methods of [9] and [10] on our crime corpus. The result is shown in TABLE 11.

3) APPLICATION OF THE PROPOSED GRAPH BASED ALGORITHM ON DIFFERENT DATASETS

Finally, we have performed an experiment on real data sets in order to analyze the effectiveness of the proposed method on data sets other than crime data. To accomplish this purpose, four different real data sets has been used. The details of the data sets used in this comparison task is provided in TABLE 12. All the four data sets have been converted into similarity graphs and the rows of each data set contain some objects and the columns have been filled with their corresponding values of the attributes. Cosine similarity factor measures the similarity among the objects and a weighted graph has been constructed, where the objects are represented as nodes and the similarity score defines the weight of the edges. Then the average similarity score is taken to be the threshold and based on the threshold, the complete graph has been made sparse. Finally, the application of the present graph based clustering technique creates clusters of objects. TABLE 13 shows the number of clusters generated by the existing graph based algorithms.

TABLE 14 and 15 shows the comparison result tested on different datasets for the present and other existing graph based clustering algorithms. The result shows that though this part of the work has used comparatively smaller datasets than those extracted in crime data for present work, still the proposed algorithm provides good results than most of the existing graph based algorithms.

Generation of many clusters provide meaningful classification of phrases rather than one or two big clusters. In support of this comment, TABLE 16 to TABLE 18 shows the measure of cluster validation indices (11) – (18) for the datasets mentioned in TABLE 12. This comparative result signifies that not only in crime data set but the proposed method also works well for other data sets. The best scores corresponding to each metric obtained for the data sets are marked bold.

V. CONCLUSION

The present work demonstrates an unsupervised approach of extracting relations from newspapers based on criminological data. The proposed clustering technique identifies significant crime patterns that can help both in criminology and criminal justice industry. Three different aspects of crime performed against women in India are brought into light by this experimental research work. We have labeled the clusters according to the most frequent context word, but it may happen that some of the context words existing in the cluster do not reflect the same crime aspect as the label of the cluster. In that case, we can collect the context words defining the same meaning. This task is known as paraphrase extraction which is considered as a future work. The paraphrase extraction can significantly improve the relation labeling scheme. Apart from the chosen domain of entity pairs, other different domains can also be considered as future research work. This method can also be applied on general datasets. Improvisations in the methodology will further provide a vast description of crime related activities by exploring other aspects of crime pattern analysis and eventually it will help the law enforcement agencies to analyze crime at a faster pace.

REFERENCES

[1] C. N. Satoshi Sekine and K. Sudo, "Extended named entity hierarchy," in *Proc. 3rd Int. Conf. Lang. Resour. Eval. (LREC)*, Feb. 2002, pp. 1818–1824.

[2] G. R. S. Weir and N. K. Anagnostou, "Exploring newspapers: A case study in corpus analysis," in *Proc. ICTATLL Workshop*, Aug. 2007, pp. 1–9.

[3] C. H. Ku, A. Iriberri, and G. Leroy, "Natural language processing and e-government: Crime information extraction from heterogeneous data sources," in *Proc. 9th Int. Conf. Digit. Government Res.*, May 2008, pp. 162–170.

[4] S. Brin, "Extracting patterns and relations from the world wide Web," in *Proc. Int. Workshop World Wide Web Databases*, 1999, pp. 172–183.

[5] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proc. 5th ACM Conf. Digit. Libraries*, 2000, pp. 85–94.

[6] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka, Jr., and T. M. Mitchell, "Coupled semi-supervised learning for information extraction," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 101–110.

[7] D. S. Batista, B. Martins, and M. J. Silva, "Semi-supervised bootstrapping of relationship extractors with distributional semantics," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2015, pp. 499–504.

[8] C. Zhang, W. Xu, Z. Ma, S. Gao, Q. Li, and J. Guo, "Construction of semantic bootstrapping models for relation extraction," *Knowl.-Based Syst.*, vol. 83, pp. 128–137, Jul. 2015.

[9] T. Hasegawa, S. Sekine, and R. Grishman, "Discovering relations among named entities from large corpora," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics*, 2004, p. 415.

[10] M. Zhang, J. Su, D. Wang, G. Zhou, and C. L. Tan, "Discovering relations between named entities from a large raw corpus using tree similarity-based clustering," in *Proc. 2nd Int. Joint Conf. Natural Lang. Process. (IJCNLP)*, 2005, pp. 378–389.

[11] B. Rosenfeld and R. Feldman, "Ures: An unsupervised Web relation extraction system," in *Proc. 21st Int. Conf. Comput. Linguistics 44th Annu. Meeting Assoc. Comput. Linguistics*, 2006, pp. 667–674.

[12] Z. Syed and E. Viegas, "A hybrid approach to unsupervised relation discovery based on linguistic analysis and semantic typing," in *Proc. NAACL HLT 1st Int. Workshop Formalisms Methodol. Learn. Reading*, 2010, pp. 105–113.

[13] T. P. Mohamed, E. R. Hruschka, Jr., and T. M. Mitchell, "Discovering relations between noun categories," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2011, pp. 1447–1455.

[14] A. Akbik, L. Visengeriyeva, P. Herger, H. Hemsen, and A. Löser, *Unsupervised Discovery of Relations and Discriminative Extraction Patterns*. Mumbai, India: Citeseer, 2012.

[15] W. Wang, R. Besançon, O. Ferret, and B. Grau, "Filtering and clustering relations for unsupervised information extraction in open domain," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 1405–1414.

[16] I. Boujelben, S. Jamoussi, and A. Ben Hamadou, *RelANE: Discovering Relations between Arabic Named Entities*. Cham, Switzerland: Springer, 2014, pp. 233–239.

[17] R. Basili, C. Giannone, C. Del Vescovo, A. Moschitti, and P. Naggar, "Kernel-based relation extraction for crime investigation," in *Proc. AI*IA*. Citeseer, 2009, pp. 161–171.

[18] R. Arulanandam, B. T. R. Savarimuthu, and M. A. Purvis, "Extracting crime information from online newspaper articles," in *Proc. 2nd Australas. Web Conf. (AWC)*, vol. 155, Jan. 2014, pp. 31–38.

[19] H. A. Shabat and N. Omar, "Named entity recognition in crime news documents using classifiers combination," *Middle-East J. Sci. Res.*, vol. 23, no. 6, pp. 1215–1221, 2015.

[20] *Astrea, Information and Communication for Justice*, IRSIG-CNR, Rome, Italy, 2006.

[21] P. Das, A. K. Das, J. Nayak, and D. Pelusi, "A framework for crime data analysis using relationship among named entities," in *Neural Computing and Applications*. London, U.K.: Springer, 2019, pp. 1–19.

[22] H. Shachnai and M. Zehavi, "Parameterized algorithms for graph partitioning problems," *Theory Comput. Syst.*, vol. 61, no. 3, pp. 721–738, Oct. 2017.

[23] E. Loper and S. Bird, "NLTK: The natural language toolkit," in *Proc. ACL Workshop Effective Tools Methodol. Teach. Natural Lang. Process. Comput. Linguistics, Assoc. Comput. Linguistics*, vol. 1, 2002, pp. 63–70.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, pp. 1–12, Jan. 2013.

[25] P. Das, A. K. Das, and J. Nayak, "Feature selection generating directed rough-spanning tree for crime pattern analysis," in *Neural Computing and Applications*. London, U.K.: Springer, 2018, pp. 1–17.

[26] S. Saitta, B. Raphael, and I. F. C. Smith, "A comprehensive validity index for clustering," *Intell. Data Anal.*, vol. 12, no. 6, pp. 529–548, 2008.

[27] M. Rosvall. 2009. *Infomap*. [Online]. Available: http://www.mapequation.org/code.html

[28] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 10, pp. 1–12, Oct. 2008.

[29] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 99, no. 12, pp. 7821–7826, Apr. 2002.

[30] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 103, no. 23, pp. 8577–8582, 2006.

[31] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

PRIYANKA DAS received the B.Tech. degree in applied electronics and instrumentation engineering from the West Bengal University of Technology and the M.Tech. degree in electronics and communication engineering from Techno India University, Saltlake, India. She is currently pursuing the Ph.D. degree in engineering from the Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology Shibpur, Howrah, India. She has published research articles in peer-reviewed journals and international conference proceedings. Her current research interests include data mining, natural language processing, and text analytics.

ASIT KUMAR DAS received the B.Tech. and M.Tech. degrees in computer science and technology from Calcutta University and the Ph.D. degree in engineering from Bengal Engineering and Science University Shibpur, Howrah. He is currently a Professor with the Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah. He likes to teach the subjects like discrete structures, data structures and algorithms, database management systems, theory of computations, design and analysis of algorithms. He has published one research monograph, three edited books, many book chapters, and over 100 research articles in peer-reviewed journals and international conferences. His current research interests include data mining and pattern recognition, evolutionary computing, and text, audio and video processing.

Dr. Das has been associated with the international program committees and organizing committees of several regular international conferences, including the International Conference on Computational Intelligence in Data Mining (ICCIDM), Soft Computing in Data Analytics (SCDA), the International Conference on Emerging Technologies in Data mining and Information Security (IEMIS), and Computational Intelligence in Pattern Recognition (CIPR). He was a Guest Editor for special issues on "Nature Inspired Optimization and Its Application to Engineering" in Evolutionary Intelligence (Springer); "Hybrid Intelligent Techniques: Foundations, Applications and Challenges" in the *International Journal of Automation and Control* (Inderscience), and "Advances and Challenges of Soft Computing in Data Mining" in the *International Journal of Computational Systems Engineering* (Inderscience).

JANMENJOY NAYAK received the M.Tech. degree (Gold Medalist and Topper of the Batch) in computer science from Fakir Mohan University, Balasore, Odisha, India, and the M.Sc. degree (Gold Medalist and Topper of the Batch) in computer science from Ravenshaw University, Cuttack, Odisha, India. He is currently an Associate Professor with the Sri Sivani College of Engineering, Srikakulam, Andhra Pradesh, India. He was a recipient of the INSPIRE Research Fellowship from the Department of Science and Technology, Government of India (both as JRF and SRF level) for doing his Doctoral Research in the Department of CSE, Veer Surendra Sai University of Technology, Burla. He has published more than 70 research papers in various reputed peer-reviewed international conferences, referred journals, and book chapters. His research interests include data mining, nature-inspired algorithms, and soft computing. He was a recipient of Young Faculty in Engineering Award from the Venus International Foundation, Chennai, in 2017, and the Young Researcher in Computer Science Engineering Award from the ITSR Foundation, India. He has been serving as an active member of reviewer committee of various reputed peer-reviewed journals, such as the IEEE Transactions on Emerging Topics in Computational Intelligence, the *IET Intelligent Transport Systems*, and many more reputed Elsevier and Springer journals.

DANILO PELUSI received the degree in physics from the University of Bologna and the Ph.D. degree in computational astrophysics. He is currently an Associate Professor with the Faculty of Communication Sciences, University of Teramo at Coste Sant'Agostino Campus, Teramo, Italy. His research is on coding theory and artificial intelligence. Moreover, he is interested in signal processing, patterns recognition, fuzzy logic, neural networks, and genetic algorithms. He has developed research activity on control systems' optimization and database design at the Astronomic Observatory Collurania "V. Cerulli" of Teramo. He received the Elsevier Reviewer Recognition Award. He is an Associate Editor of the IEEE Transactions on Emerging Topics in Computational Intelligence and a Reviewer of the "U.K. Modelling and Simulation Society (UKSim)" and of several international journals and conferences. He is a Keynote Speaker at several conferences and a Guest Editor of Inderscience and Springer. He is an Editorial Board Member of international journals and a Technical Program Committee Member of international conferences.

WEIPING DING (M'16–SM'19) received the Ph.D. degree in computation application from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2013. He was a Visiting Scholar with the University of Lethbridge (UL), AB, Canada, in 2011. From 2014 to 2015, he was a Postdoctoral Researcher with the Brain Research Center, National Chiao Tung University (NCTU), Hsinchu, Taiwan. In 2016, he was a Visiting Scholar with the National University of Singapore (NUS), Singapore. From 2017 to 2018, he was a Visiting Professor with the University of Technology Sydney (UTS), Ultimo, NSW, Australia. To data, he holds 13 approved invention patents in total over 20 issued patents. He has published over 60 papers in flagship journals and conference proceedings, including the IEEE Transactions on Fuzzy Systems, the IEEE Transactions on Neural Network and Learning Systems, the IEEE Transactions on Cybernetics, the IEEE Transactions on Industrial Informatics, the IEEE Transactions on Biomedical Engineering, the IEEE Transactions on Systems, Man, and Cybernetics: Systems, and the IEEE Transactions on Emerging Topics in Computational Intelligence. His main research interests include deep learning, data mining, evolutionary computing, granular computing, machine learning, and big data analytics.

Dr. Ding is a Senior Member of the IEEE-CIS, the ACM, the IAENG and CCF. He is a member of the Technical Committee on Soft Computing of the IEEE SMCS, the Technical Committee on Granular Computing of the IEEE SMCS, and the Technical Committee on Data Mining and Big Data Analytics Technical Committee of the IEEE CIS. He is also a member of the IEEE CIS Task Force on Adaptive and Evolving Fuzzy Systems.