Criminal Event Detection and Classification in Web Documents Using ANN Classifier

J. Sheela and A. Vadivel
Department of Computer Application, National Institute of Technology, Tiruchirappalli, Tamil Nadu, India-620015
Email: {jsheela, vadi}@nitt.edu

***Abstract*—Text mining can be described as the process of extracting particular information from within unstructured data, thereby facilitating access to potentially valuable information for use in a wide variety of fields. In this paper, we selected the crime domain to explore the hidden important information using text mining techniques. Developing an effective and intelligent system for extracting the important and hidden information from crime reports and Social Networking websites would be useful for police investigators, for accelerating the investigative process which helps in crime prediction by conducting further analysis. The proposed approach deals with automatic construction of crime related thesaurus. The proposed information extraction approach relies on computational linguistic techniques. The domain related terms and its related sentences are selected to identify patterns of interest. Further, syntactic analysis is done based on POS Tagging. Sentences Classification and clustering is done based on the sentence patterns using ANN.**

***Index Terms*—event detection, pattern classification, detection and recognition, criminal activities, web patterns, NLP, knowledge mining**

I. Introduction

Information Extraction (IE) is the process for extracting structured information from unstructured text and this process is considered as one of the major challenges. The focus of information retrieval system revolves around providing users with information that are interesting to him/her. Up until the year 1998, the information extraction systems have scrutinized and analyzed by the Message Understanding Conferences (MUC). The successors to MUC, the Automatic Content Extraction program (ACE) has helped the development of extraction technology that facilitates in (automatic) processing of source language data. The ACM carried out the classification, selection and filtering and performed these on the language content of the source data or in other words, the actual meaning that the data convey [1]. There are three types of classification of information that the ACE program aims to define, such as entities, relations and events. The 'when', 'where', 'why' and authenticity of the events is also looking in the form of polarity, tense, generality, modality, etc. Event information is added as metadata to the text document, once the interest is identified. The definition of a process in ACE takes place through the identification of events from a single sentence. Event extraction is commonly regarded as one of the research points in IE that poses some of the toughest challenges in the overall process. Predicting time, place and other facts and actions that correspond to an incident are described using natural language in Event Extraction. Natural Language Processing (NLP) fields such as automatic summarization [2], questioning and answering [3] as well as information retrieval can be ascertained by using event extraction. A subtask of information extraction is temporal IE and it extracts time expressions and temporal relations from text containing natural language. NLP makes use of this processing temporal information in natural language. In temporal question and answering systems, temporal information processing is essential. This is an example how NLP makes use of this processing. The system has to temporally anchor the event in case of a 'when' question. If the question hinges on the time taken or 'how long', the duration of the event need to be measured. Some queries might want to find out the 'why' aspect of an event, and hence, the system should be able to provide the reason behind it. There exist thousands of electronic collections that contain information of high quality [4].

Search engines process the web content for creating indexes. Starting from a collection of unstructured documents, the indexer extracts a large amount of information that may include processing data, such as a list of documents, It contain terms and other details, the number of all the occurrences of each term within every document. An index maintains all this information and it is generally represented using an inverted file. The index consists of an array of the posting lists and contains the term as well as the identifiers of documents containing the term. Thus, the significance of the term in building the index is reduced and more prominence is given to the context of the document. Extra information is provided due to the context being considered and the eventual result is a marked improvement in the relevancy of search results. The relations extracted from Universal Networking Language (UNL) graphs helps to point out the context of a document.

An event in two news stories is defined as a specific happening at a certain time, in a specific place and involves two or more number of participants. Different news articles look at the same occurrence through

Manuscript received August 25, 2015; revised November 30, 2015.

doi: 10.18178/ijsps.4.5.382-388

different point of views. Gathering information about the same or similar events from different news corpora poses a tricky and interesting challenge [5]. In automatic summarization, events are ranked from documents. The Page Rank algorithm is the go-to tool for deciding the importance of events. The existing approaches have a couple of things working against them. Primarily, it is borderline impossible to extract elements from each and every event. Secondarily, there is variance in the associative strength of events and event relations being depicted might suffer from a lack of accuracy. Based on the above discussion, it is observed that detecting "event instance" at sentence level from Web documents is a challenging task. Most of the above stated methods have failed in understanding the semantics of "event instance".

In this paper, we propose scheme for detecting crime related sentence from web page. Pattern identification and classification is considered as an event detection problem at a sentence level in a document. Sentences are classified based on POS tag of event trigger term, immediate co-occurrence term and non-immediate co-occurrence term. Event instances present in crime related reports are processed at sentence level by understanding the semantic information and syntactic structure of sentence. Based on the event extracted, crime event detection corpus is constructed for complementing investigating agencies. Investigating agent can extract high intensity sentences from crime report. This paper is organized as follows. The related work is presented in Section II. The proposed method is presented in Section III, Section IV presents the experiment results and we conclude the paper in Section V.

II. Related Works

Various methods have been proposed for extracting events in the literature. Some results of crime mining have been presented using data mining techniques. Li *et al.* (2000) have proposed a Twitter-based Event Detection and Analysis System (TEDAS), which made it easier to spot new events to analyze the spatial/temporal patterns of events and to contrast the benefits of events. In order to help the users to understand the essence of data and concentrate on specific subsets, large, heterogeneous information resources are briefed and summarized. Questions like "what?", "where?" and "when?" are obvious entry points for narrative documents. The Message Understanding Conferences (MUC) has been developed for automatically identifying and analyzing military messages present in textual information. The main objective of the ACE program is to classify the popular events from news articles into various classes and subclasses. Previous ACE research and TDT research have focused on event detection at the term/phrase level and the document level [6]. However, no prior work was done at sentence level for event detection. Allan *et al.* (2000) argues that New Event Detection (NED) approach could always display low achievement. The Systems employing tracking technology for NED exhibits a characteristic and the occurrence is followed by tracking system for discovering a new event. Since tracking and filtering is not convincing, a backup approach to NED was required to improve the performance [7]. Smith *et al.* (2002) have proposed a method for detecting event based on co-occurrence term like dates and place in the document collection [8]. The relative significance of several events is determined by statistical measuring. McCracken *et al.* (2006) have combined statistics and knowledge based method for extracting events. It mainly focuses on the summary report genre. It focuses on developing a scheme that allows the utilization of statistical techniques without new training data [9]. Zhao *et al.* (2006) have proposed an integrated Web Event Detection (iWED) algorithm to extract events from web documents by integrating author–centric data and visitor centric data. Web document related data is organized as a multi graph each vertex signifies a web page and each edge signify the relationship between the connected web documents in terms of structure, semantics and usage pattern in web documents. Natural language applications such as Question Answering (QA) and Summarization depend on the appropriate organization of sentences that describe events [10]. Zhao *et al.* (2006) have proposed a work study and use log data of web search engines to detect events. To record the semantic and evolutionary relationships between queries and pages recorded and this sequence is denoted as a vector- based graph. The vector-based graph is transformed into its dual graph, where each node is a query-page pair that is used to represent real world events. Upon clustering the dual graph of this vector based graph, based on a two-phase graph cut algorithm, particular events are detected. The first phase clusters the query-page pairs semantically, such that each cluster contains pairs corresponding to a specific topic. In the second phase, each cluster is further clustered based on similar evolution pattern such that each cluster represents a specific event under a specific topic [11].

Naughton *et al.* (2008) have developed a methodology for event detection at sentence level. Support Vector Machine (SVM) classifier and a Language Modeling (LM) approach have been used to check the performance of the system. Earlier studies in identifying novelty have considered only the problem of finding novel material from the given document on a particular topic [12]. TREC 2002 is used to find the relevant sentences from the documents and find the novel sentences from the collection of relevant sentences. The results prove that finding the relevant sentences from the documents is the vital part and that the presence of non-relevant sentences degrades the performance of novelty measures. Long *et al.* (2010) have proposed a novel context based event indexing and event ranking model for detecting event in news articles [13]. The context event clusters made from the Universal Networking Language (UNL) graphs use the modified scoring system for segmenting events, which precedes event cluster. Three models have been developed based on the obtained context clusters, such as identification of main and sub events, Event Indexing and Event Ranking. The main events and associated sub-events are identified based on the properties reflected by the UNL graphs for the modified scoring. A hash map data structure is used to store the temporal details (place: where; person: who is involved; time: when) attained from the context cluster. This information is the basis for generating three indices to obtain all events from context clusters (Time index, person index, and place index). A new scoring scheme for event ranking gives weightage based on the priority level of the events, which include the occurrence of the event in the title of the document, event frequency, and inverse document frequency of the events.

Sun *et al.* (2011) have proposed a query-guided event detection method to detect event from two parallel document streams (news and blog). Changes in the query keywords and news/blog content reflect the evolution of an event. Two-stage real-time event detection frameworks consisting of event fragment detection and event detection is considered [14]. It integrates queries, news articles and blog posts through the notion of query profile. In [7], the event detection method is divided into four classification problems, that is, trigger identification, argument identification, and attribute assignment and event co-reference resolution. The lexical, Word Net and dependency tree features are used and argument identification is treated as a pairwise classification problem. A distinct classifier has been used for training each attribute and finally, event co-reference is treated as a binary classification task. Earlier studies have concentrated on event detection at either the term/phrase level Automatic Content Extraction (ACE research) or the document level. However, many Information Retrieval (IR), QA and Summarization applications favor a sentence level granularity. For example, QA systems that process complex questions such as "How many people were killed in Baghdad in March?" often depend on event detection systems to identify the sentences that contain all relevant event instances before formulating an answer. In addition, text summarization approaches rely on sentence extraction techniques that identify key event-related information for inclusion in the end summary. Most of the previous research work has focused on either the term/phrase level or the document level detection schemes and in this paper we propose a sentence level event extraction schema. Our consideration is also based on semantic similarities between words and sentences. Thus, it is imperative that a proper approach is required for specifying the events by considering specific events. The method also should categorize the events based the features (who, whom, place, time) of the sentence. This paper, addresses these issues by formulating rules that finds sentences based on semantic similarities between terms and sentence. The methodology detects the sentences in a collection of web document that describes one or more instances of a specified event type. The meaning of the sentence and hyponyms, Hypernyms is considered to improve the efficiency of event detection. Each sentence in a web document is classified into the periodic sentence and Non-period sentence based on the occurrence of conjunction in a sentence. The intensity of sentence is identified by the occurrence of conjunction, based on the POS tag of the event trigger term, immediate co-occurrence term and Non-immediate co-occurrence term. The performance of classification can be validated by Artificial Neural Network (ANN) Tool.

III. Proposed Work

In this paper, we propose a methodology that detects "event instance" by considering sentences in a web page. The detected event patterns are classified into sixteen groups based on the intensity value of a sentence pattern and the hierarchical classes are named as Periodic and Non Periodic classes and the hierarchical model is presented in Fig. 1. The first level is based on the presence of conjunction in sentences and named as Periodic and Non periodic sentences. The second level is based on the POS tag of Event Trigger (ET) terms in the sentences. The ET term may appear as noun/verb/adjective and further termed as Static and Dynamic sentences. The third level is based on the POS tag of co-occurrence terms (ET ±1) next to the ET terms. The co-occurrence (ET ±1) terms may appear as adjective/adverb and is named as Qualified and Non-Qualified sentences. The fourth level is based on the presence of cardinal number terms in a sentence and named as spatial sentence and no spatial sentence. Finally, the fifth level denotes the sentence cluster in which the sentence patterns are divided into sixteen clusters. The classes are realized in the form of rules. The periodic sentence is classified into eight classes with five levels and the hierarchical decision tree model is presented in Fig. 2 and the rules related to these sentences are also presented.

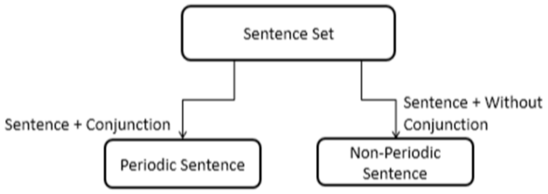


Figure 1. First level hierarchical structures for a sentence

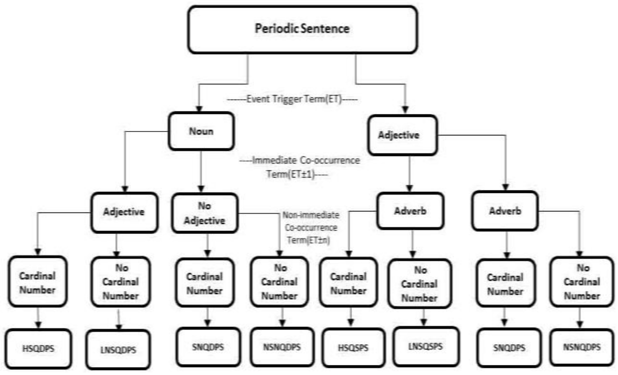


Figure 2. Hierarchical structure for periodic sentence set

For Periodic sentences the rules are as follows:

'*s*' is a sentence in document. Each sentence in a document contains more number of terms 't' and only certain terms represent the Event. The term that represents the Event '*et*' that event is called Event trigger. Here POS tagger (NN/VB/JJ) of crime related term is considered as Event trigger term such as kill, attack, damaged, etc. The *ct(I)* is a cooccurence term of Event trigger term and it may be adjective (JJ) or Adverb (RB). The *ct(I±n)* is a non-immediate cooccurence term of Event trigger Term. CD is considered as non-immediate cooccurrence term. In POS tagger the Number is denotes as cardinal Number.

Rule 1:
*if(*s=CC*)&&(et=NN)&&ct(I)=JJ&& ct(I±n)=CD*
*then C=*H*SQSPS*
else if(s=CC) && (et=NN)&&ct(I)=JJ&&ct(I±n)≠CD
*then C=L*NSQSPS
if(s=CC) && (et=NN) &&ct(I)≠ JJ&&ct(I±n)=CD
*then C=*SNQSPS
else if(s=CC) &&(et=NN)&&ct(I)≠ JJ&&ct(I±n)≠CD
then C= NSNQSPS
Rule 2:
if(s=CC) && (et=JJ/VB)&&ct(I)=RB&&ct(I±n)=CD
*then C=H*SQDPS
else if(s=CC)&&(et=JJ/VB)&&ct(I)=RB&&ct(I±n)≠CD
*then C=LN*SQDPS
if(s=CC)&&(et=JJ/VB)&&ct(I)≠ RB&&ct(I±n)=CD
then C= SNQDPS
else if(s=CC)&&(et=JJ/VB)&&ct(I)≠ RB&&ct(I±n)≠CD
then C= NSNQDPS

For Non-Periodic sentences the rules are as follows:

Rule1:
if(s≠ CC)&&(et=NN)&&ct(I)=JJ&&ct(I±n)=CD
*then C=H*SQSNPS
else if(s≠ CC) && (et=NN)&&ct(I)=JJ&&ct(I±n)≠CD
*then C=L*NSQSNPS
if(s≠ CC)&&(et=NN)&&ct(I)≠ JJ&&ct(I±n)=CD
then C= SNQSNPS
elseif(≠ CC) && (et=NN)&&ct(I)≠ JJ&&ct(I±n)≠CD
then C= NSNQSNPS
Rule2:
if(s≠ CC)&&(et=JJ/VB)&&ct(I)=RB&&ct(I±n)=CD
*then C=H*SQDNPS
*else if(s≠ CC)&&(et=JJ/VB)&&ct(I)=RB&&ct(I±n)≠ CD then C=LN*SQDNPS
i*f(s≠ CC)&&(et=JJ/VB)&&ct(I)≠ RB&&ct(I±n)=CD*
then C=SNQDNPS
*else if(s≠ CC)&&(et=JJ/VB)&&ct(I)≠ RB&&ct(I±n)≠ CD then C=*NSNQANPS

where *s*-sentence, *et*-Event triggered term, *ct(I)*-co-occurrence term, *VB*-verb, *NN*-noun, *JJ*-adjective, *RB*-adverb, *CD*-Cardinal Number and *C*-class. Periodic sentences are classified into eight classes with four levels and the hierarchical model is presented in Fig. 2.

A. Periodic Sentence Classification

The First level categorization is made based on the appearance of interested keywords. In the sentence, it may appear as Noun/Verb/Adjective. If the interested keyword is a Noun, then the sentence is considered as a static sentence, as it gives only the stable information such as where/what/who about a motion (Name of the person/place/activity). If the interested keyword is an Adjective/Verb, then the sentence is considered as a dynamic sentence, as it carries detailed information about the event. The part of speech is determined by how a word is used in a sentence. The same word may be a noun, verb, adjective, preposition, or conjunction and is based how it is used. In this paper, we used the Stanford NLP tool to generate POS tag for each term in sentence. For example, the sentence "*Gangsters fire the house*" and "*purchases a fire alarm*" are tagged by the POS Tagger as Gangsters_NNS fire_VBP the_DT house_NN._. and John_NNP purchases_NNS a_DT fire_NN alarm_NN._.. and the interested keyword is "*fire*". Event trigger Term in first sentence is fire_VBP and for second sentence is fire_NN. The POS tag of Term is determined by how the word used in a sentence. If the interested keyword (Event trigger Term) is a Noun, then the sentence is considered as a static sentence. If the interested keyword is an Adjective/Verb, then the sentence is considered as a dynamic sentence. First sentence is belonging to dynamic sentence and second sentence is belonging to static sentence**.** The second categorization is done based on the association of the interested keyword with its co-occurrence terms. The final categorization is based on occurrences of cardinal numbers in a sentence, as it gives spatial information about when an event/activity happened.

In the periodic sentences, for the first level, if the interested keyword is Noun and its association is with the Adjective co-occurrence term and a cardinal number, then the sentences belonging to such a class is classified as High Spatial Qualified Static Periodic Sentence (HSQSPS). In the periodic sentences, if the interested keyword is a Noun, its association is with Adjective co-occurrence term and it is not associated with a Cardinal Number, then the sentences belonging to such class are classified as Low Non-Spatial Qualified Static Periodic Sentence (LNSQSPS). In the periodic sentences, if the interested keyword is a Noun, its association is with a cardinal number and it is not in association with the Adjective co-occurrence term, then the sentences belonging to such class are classified, as Spatial Non-Qualified Static Periodic Sentence (SNQSPS). In the periodic sentences, if the interested keyword is Noun and it is not in association with the Adjective co-occurrence term and a cardinal number, then the sentences belonging to such class are classified, as Non-Spatial Non-Qualified Static Periodic Sentence (NSNQSPS). In the Periodic sentences, if the interested keyword is an Adjective/Verb and its association is with the Adverb co-occurrence term and a Cardinal Number, then the sentences belonging to such class are classified as High Spatial Qualified Dynamic Periodic Sentence (HSQDPS). In the periodic sentences, if the interested keyword is an Adjective/Verb its association is with Adverb co-occurrence term and it is not associated with a Cardinal Number, then the sentences belonging to such class are classified as Low Non-Spatial Qualified Dynamic Periodic Sentence (LNSQDPS). In the periodic sentences, if the interested keyword is Adjective/Verb, its association is with a Cardinal Number and it is not in association with Adverb co-occurrence term, then the sentences belonging to such class are classified, as Spatial Non-Qualified Dynamic Periodic Sentence (SNQDPS). In the periodic sentences, if the interested keyword is Adjective/Verb and it is not in association with Adjective co-occurrence term and a Cardinal Number, then the sentences belong to such class are classified, as Non-Spatial Non-Qualified Dynamic Periodic Sentence (NSNQDPS). The Non-Periodic sentences are classified into eight classes with four levels and the hierarchical structure is presented in Fig. 3.

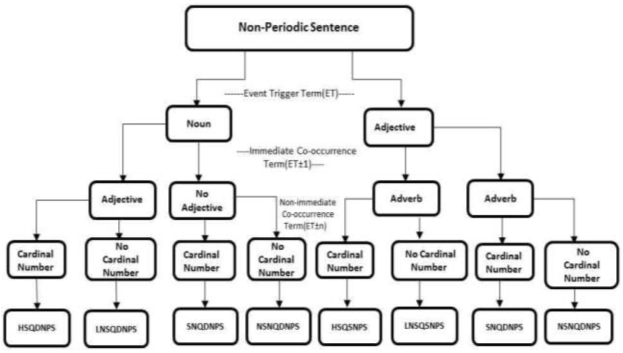
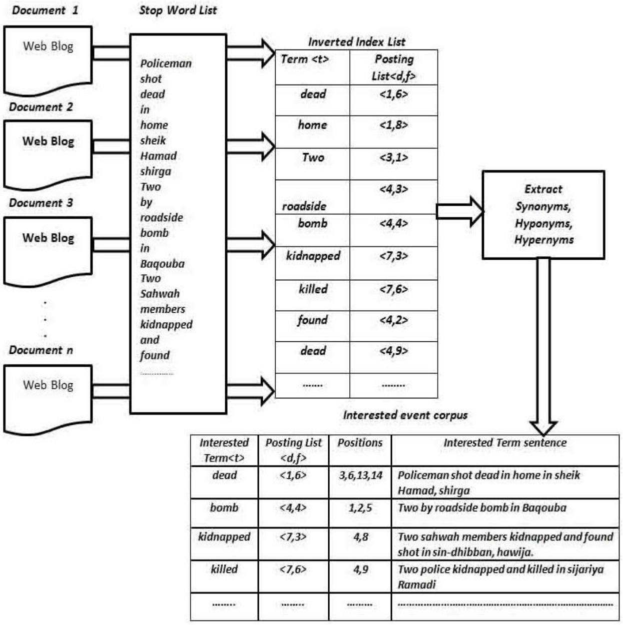


Figure 3. Hierarchical structure for non-periodic sentence set

B. Non-Periodic Sentence Classification

In the Non-periodic sentences, for the first level if the interested keyword is Noun and its association is with Adjective co-occurrence term and a cardinal number, then the sentences belonging to such class are classified as High Spatial Qualified Static Non-Periodic Sentence (HSQSNPS). In the Non-periodic sentences, if the interested keyword is Noun, its association is with Adjective co-occurrence term and it is not associated with a cardinal number, then the sentences belonging to such class are classified as Low Non-Spatial Qualified Static Non-Periodic Sentence (LNSQSNPS). In the Non-periodic sentences, if the interested keyword is Noun, its association is with a cardinal number and it is not in association with Adjective co-occurrence term, then the sentences belonging to such class are classified, as Spatial Non-Qualified Static Non-Periodic Sentence (SNQSNPS). In the Non-periodic sentences, if the interested keyword is Noun and it is not in association with Adjective co-occurrence term and a cardinal number, then the sentences belonging to such class are classified, as Non-Spatial Non-Qualified Static Non-Periodic Sentence (NSNQSNPS). In the Non-Periodic sentences, if the interested keyword is Adjective/Verb and its association is with Adverb co-occurrence term and a cardinal number, then the sentences belonging to such class are classified as High Spatial Qualified Dynamic Non-Periodic Sentence (HSQDNPS). In the Non-periodic sentences, if the interested keyword is an Adjective/Verb its association is with Adverb co-occurrence term and it is not associated with a cardinal number, then the Sentences belonging to such class are classified as Low Non-Spatial Qualified Dynamic Non-Periodic Sentence (LNSQDNPS). In the Non -periodic sentences, if the interested keyword is an Adjective/Verb its association is with a cardinal number and the association is with the Adverb co-occurrence term, then the sentences belonging to such class are classified, as Spatial Non-Qualified Dynamic Non-Periodic Sentence (SNQDNPS). In the Non-periodic sentences, if the interested Keyword is an Adjective/Verb and it is not in association with the Adjective co-occurrence term and a cardinal number, then the sentences belonging to such class are classified, as Non-Spatial Non-Qualified Dynamic Non-Periodic Sentence (NSNQDNPS).

C. Construction of Inverted Index Table



Term <t>	Posting List<d,f>
dead	<1,6>
home	<1,8>
Two	<3,1>
roadside	<4,3>
bomb	<4,4>
kidnapped	<7,3>
killed	<7,6>
found	<4,2>
dead	<4,9>
....	....

Interested Term<t>	Posting List <d,f>	Positions	Interested Term sentence
dead	<1,6>	3,6,13,14	Policeman shot dead in home in sheik Hamad, shirga
bomb	<4,4>	1,2,5	Two by roadside bomb in Baqouba
kidnapped	<7,3>	4,8	Two sahwah members kidnapped and found shot in sin-dhibban, hawija.
killed	<7,6>	4,9	Two police kidnapped and killed in sijariya Ramadi
....	....	....	....

Figure 4. Inverted index to interested event

In our approach, the Event trigger Term (ET) is identified by constructing an events corpus from inverted index. A huge number of documents from web pages with criminal related events are crawled. The sentences are tokenized, stemmed and stop words are removed. During preprocessing step, the stop words are removed from the web documents and however, the terms are conserved without stemming to avoid the term ambiguity. The terms are stored in the inverted index and for each term <*t*> in the inverted index, there is a posting list that has sentence *id* and frequency of occurrence <*s*, *f*>. *S* is a set of sentences and *T* is a set of terms present in *S*. The information found from the inverted index such as terms Frequency of occurrences and the related document *id*'s are used for event detection. The inverted index consists of interested as well as non-interested terms. We select few interested terms manually. Further, synonyms, hyponyms and hypernyms for those manual selected terms are found. This helps to find rest of the terms present in the inverted index. Since, our proposed approach aims to build criminal activities based document, the interested terms are criminal related terms such as crime, terrorism, kill etc. The domain interested terms, their posting list, position of occurrence and the sentences including interested terms are considered. The interested term sentence is so considered, since, it gives the information of the interested event i.e. criminal event. Here, position of occurrence of the term is used to find the interested term sentence and then sentence is called interested sentence. The sequence of the procedure is depicted in Fig. 4. Let *D* be a collection of documents retrieved from WWW and *T* be a set of terms present in *D*. The occurrence of a term in a document may be treated as a labeling approach denoted as follows:

$$l : T \times D \rightarrow \{True, False\} \qquad (1)$$

From (1), it is assumed that a term $t \in T$ presents in a document $d \in D$, if $l : (t, d) = True$. In document retrieval an application, the posting list is extracted from the inverted index.

The posting list is in the form of <t_i, d_i, f_i> where f_i is a frequency of occurrence assigned to term t_i in document d_i. Since, a term can be physically appearing in many documents, given a query term q_t, such that $q_t \subseteq Q_T$, q_t can be defined as the relationship of $< t,\ d,\ f >$ as follows. This relation is represented in (2).

$$C^D(q_t) = \\ \{< t, d, f > | d \in D, t \in T, f \in F and\ \forall q_t \subseteq Q_T, (q_t = t), (t, d) = True\} \qquad (2)$$

IV. Experimental Results

For experiments, we have used ANN Tool and it is motivated by human learning process. Web pages with crime related terms considered for evaluation and multilayer perception algorithm using 10 fold cross validation testing mode applied. The web pages collected from TV and Crime magazine are considered and the 500 interested keywords are identified. A total of 47,884 sentences are identified, where the interested keywords are present. Among these, 8325 are periodic and 39559 re non-periodic sentences. From the 47,884 interested sentences, 9 sentences fall into class A i.e., 0.001% of the sentences match for the High Spatial Qualified Static Periodic Sentences pattern (HSQSPS). Similarly, 2,202 sentences match for class B (Low Non-Spatial Qualified Static Periodic Sentence (LNSQSPS), 92 sentences match for class C (Spatial Non-Qualified Static Periodic Sentence (SNQSPS), 3003 sentences match for class D (Non-Spatial Non-Qualified Static Periodic Sentence (NSNQSPS), 1 sentences match for class E (High Spatial Qualified Dynamic Periodic Sentence (HSQDPS), 300 sentences match for class F (Low Non-Spatial Qualified Dynamic Periodic Sentence (LNSQDPS), 75 sentences match for class G (Spatial Non-Qualified Dynamic Periodic Sentence (SNQDPS), 2225 sentences match for class H (Spatial Non-Qualified Dynamic Periodic Sentence (SNQDPS) and so on. In Table I, the classification accuracy presented. The performance evaluation of sentence grouping using ANN is shown through a graphical representation in below Fig. 5. Here 'x' axis is denotes as Name of the class and 'y' axis is denotes as Frequency occurrence of sentence. Initially, the Periodic classes and Non-Periodic classes are manually identified for creating respective clusters. The process of annotation is carried out based on the verb POS nature of a sentence using the NLP tool. Further classification based on ET terms is also carried out for event type 'Die, Kill'… The classification based on the ET terms gives static and dynamic sub-classes of Periodic and Non-Periodic classes.

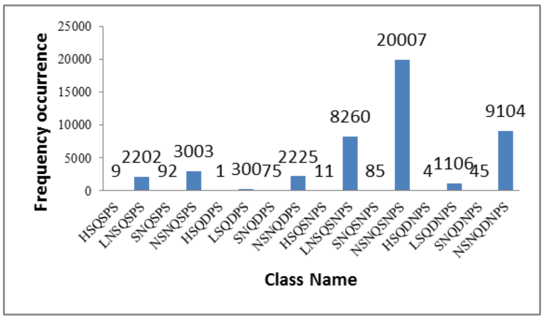


Figure 5. Sentence distribution

TABLE I. Classification Accuracy

ANN		
Periodic class patterns	Patterns Count (%)	Classification accuracy (%)
HSQSPS	0.11%	34.62
LNSQSPS	26.45%	96.92
SNQSPS	1.11%	82.14
NSNQSPS	36.07%	97.41
HSQDPS	0.01%	50.00
LSQDPS	3.60%	96.77
SNQDPS	0.90%	78.95
NSNQDPS	26.73%	91.75
Miss classified	5.02%	
Non-Periodic class patterns	Patterns Count (%)	Classification accuracy (%)
HSQSNPS	0.03%	44.00%
LNSQSNPS	20.88%	97.64%
SNQSNPS	0.21%	80.19%
NSNQSNPS	50.58%	99.50%
HSQDNPS	0.01%	66.67%
LSQDNPS	2.80%	98.22%
SNQDNPS	0.11%	60.00%
NSNQDNPS	23.01%	94.30%
Miss classified	2.37%	

Further classification is based on the immediate co-occurrence term $\text{ct}_{(\text{I})}$ of ET in the sentence which gives Qualified and Non-Qualified patterns of Periodic and Non-Periodic class. The final level classification is based on the Non-immediate co-occurrence term such as cardinal number $\text{ct}_{(\text{NI})}$ of ET in the sentence which gives spatial and Non-spatial patterns of Periodic and Non-Periodic class. All the classification output, based on Manual annotation, is re-evaluated by an ANN for obtaining eight event mention patterns for each class. It is observed from Table I and that the manual classification accuracy is very low. As a result, the ANN is used for improving the accuracy and is mentioned another column. The ANN model is used as a classifier with Multilayer perceptron algorithm used for training the data. The back propagation optimization technique is used for training the Multilayer perceptron structure. Four term features such as *et*, $ct_{(I)}$ and $ct_{(NI)}$ and the cardinal number are given as the input and eight rules are obtained for the classifying the patterns. Similarly, sentences that belong to Periodic class are classified and eight patterns are obtained and finally, sixteen event mention patterns are obtained for both Periodic and Non-Periodic sentence classes. We have used classification accuracy as a performance measure and it is defined as the ratio of sentences correctly classified by the ANN classifier to class type *n* to the human annotated sentences for the class type *n*. The ANN generated for sentence pattern classification and the results are shown in Table I. The misclassified sentences are represented as outlets for *Periodic* class and Non-*Periodic* class. This is due to the presence of multiple event types and event instances in the sentence, which conflicts and misleads the classifier during classification. It is noticed that the difference between classifier and manual annotation is less, i.e., in the range of 1-3% for each pattern. Also, the classification accuracy of ANN and human annotations match above 97% for a sample data set.

V. Conclusion

We have proposed an artificial neural network based novel method for event detection using event instance in in a sentence. It classifies criminal related sentence into hierarchical groups through rules. Four levels of further classification is based on conjunctions, POS tag of event trigger term, immediate co-occurrence term and Non-immediate co-occurrence term onto the obtained high intensity sentence and POS tagging tool is used to find patterns of interested sentences with intensity and other related sort out techniques. Criminal corpus was built to create domain-specific thesaurus. The performance of the system is evaluated using the ANN classification tool. The classification accuracy is encouraging compared to some of the other similar methods.

References

[1] H. Cunningham, " formation extraction, automatic," in *Encyclopedia of Language and Linguistics*, Elsevier, 2005, pp. 665-677.

[2] W. J. Li, M. L. Wu, and Q. Lu, "Extractive summarization using inter- and intra-event relevance," in *Proc. 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 369-376.

[3] D. Ahn, "The stages of event extraction," in *Proc. Workshop on Annotations and Reasoning about Time and Events*, Sydney, Australia, 2006, pp. 1-8.

[4] K. F. Wong, and Y. Xia, "An overview of temporal information extraction," *International Journal of Computer Processing of Oriental Languages*, vol. 18, no. 2, pp. 37-152, 2005.

[5] P. Gupta and A. K. Sharma, "Context based indexing in search engines using ontology," *International Journal of Computer Applications*, vol. 1, no. 14, 2010.

[6] R. Li, K. H. Lei, R. Khadiwala, and K. C. Chang, "Tedas: A twitter-based event detection and analysis system," in *Proc. IEEE 28th International Conference on Data Engineering*, April 2012, pp. 1273-1276.

[7] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level," in *Proc. 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2003, pp. 314-321.

[8] D. A. Smith, "Detecting and browsing events in unstructured text," in *Proc. 25th Annual IGIR Conference on Research and Development in Information Retrieval*, 2002.

[9] N. McCracken, N. E. Ozgencil, and S. Symonenko, "Combining techniques for event extraction in summary reports," in *Proc. AAAI Workshop Event Extraction and Synthesis*, 2006, pp. 7-11.

[10] Q. Zhao, "A survey of mining semi-structured and dynamic web data," in *Proc. ACM PODS*, Nanyang Technological University, Singapore, 2006.

[11] Q. Zhao, *et al.*, "Event detection from evolution of click-through data," in *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[12] M. Naughton, N. Stokes, and J. Carthy, "Investigating statistical techniques for sentence-level event classification," in *Proc. 22nd International Conference on Computational Linguistics*, 2008.

[13] R. Long, *et al.*, "Towards effective event detection, tracking and summarization on microblog data," in *Web-Age Information Management*, Springer Berlin Heidelberg, 2011, pp. 652-663.

[14] A. Sun and M. Hu, "Query-Guided event detection from news and blog streams," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, pp. 834-839, 2011.

J. Sheela (ME, BE) is a PhD student in Department of Computer Applications, at National Institute of Technology (NIT), Trichy. Prior to beginning the PhD program, Sheela worked as an Assistant Professor in the Hindusthan College of Engineering and Technology. She got her Masters of Engineering from Anna University Coimbatore, and Tamilnadu, India. She pursued Bachelor of Engineering from VLB Janaki Ammal College of Engineering and Technology, Coimbatore, India.

A. Vadivel is Associate Professor in the Department of Computer Applications, National Institute of Technology (NIT), Trichy. He has got his Masters of Science from National Institute of Technology (NIT), Tiruchirappalli, and Tamilnadu, India. He persued Masters of Technology and Philosophy in Doctorate from Indian Institute of Technology (IIT), Kharagpur, India. He has 12 years of technical experience as Network Engineer & Instrumentation Engineer at IIT-Kharagpur and Ten years of teaching experience in Bharathidhasan and National Institute of Technology (NIT), Trichy. He has published more than 80 papers in international journals and conferences and delivered invited Lectures in India and Abroad. His research areas are content-based image and video retrieval, multimedia information retrieval from distributed environment, medical image analysis, object tracking in motion video, data mining (association rule mining and frequent pattern analysis), and cognitive science. He was conferred with Young Scientist Award by Department of Science and Technology, Govt. of India in 2007, Indo-US Research Fellow Award by Indo-US Science and Technology Forum in 2008 and Obama– Singh Knowledge Initiative Award in 2013. He is carrying out many research and consultancy projects sponsored by Ministry of Human Resource Development (MHRD), Department of Science & Technology (DST) and United States & India Education Foundation (USIEF).