Information Processing and Management 56 (2019) 102059

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers

Srinivasa K*, P. Santhi Thilagam

Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, India

ARTICLE INFO

Keywords:
Ontology
Natural Language Processing
Integration
Information Extraction
Knowledge Representation

ABSTRACT

In the current era of internet, information related to crime is scattered across many sources namely news media, social networks, blogs, and video repositories, etc. Crime reports published in online newspapers are often considered as reliable compared to crowdsourced data like social media and contain crime information not only in the form of unstructured text but also in the form of images. Given the volume and availability of crime-related information present in online newspapers, gathering and integrating crime entities from multiple modalities and representing them as a knowledge base in machine-readable form will be useful for any law enforcement agencies to analyze and prevent criminal activities. Extant research works to generate the crime knowledge base, does not address extraction of all non-redundant entities from text and image data present in multiple newspapers. Hence, this work proposes Crime Base, an entity relationship based system to extract and integrate crime related text and image data from online newspapers with a focus towards reducing duplicity and loss of information in the knowledge base. The proposed system uses a rule-based approach to extract the entities from text and image captions. The entities extracted from text data are correlated using contextual as-well-as semantic similarity measures and image entities are correlated using low-level and high-level image features. The proposed system also presents an integrated view of these entities and their relations in the form of a knowledge base using OWL. The system is tested for a collection of crime related articles from popular Indian online newspapers.

1. Introduction

With the emergence of the internet, vast volume of information related to news events is available digitally in the form of newspapers. A crime knowledge base with information enriched from various newspapers is of high interest to multiple end users (Arulanandam, Savarimuthu, & Purvis, 2014). Even though, Law Enforcement Agencies have information available with them that comes within their jurisdiction, prevention or tracking of the criminal activities is limited to some specific regions. With access to a shared knowledge base having information collected from various sources, helps them to prevent or track the criminal activities in other regions also. Integrated information from multiple news articles represented as a knowledge base, can be used for various applications like crime intelligence analysis, business intelligence analysis, disaster management, etc. Currently, crime monitoring and prevention is of great interest to most law enforcement agencies across the globe to keep the world safe. Although valuable, authentic and timely information in online newspapers can be extracted manually by reading through all the available newspapers, this is a difficult and error-prone job that needs a lot of human resources. Hence, a system that automatically extracts, integrates and

* Corresponding author.
E-mail addresses: srinivas.karur@gmail.com (S. K), santhi@nitk.ac.in (P.S. Thilagam).

https://doi.org/10.1016/j.ipm.2019.102059
Received 13 August 2018; Received in revised form 3 May 2019; Accepted 14 June 2019
Available online 21 June 2019

represents crime related text and image entities as a knowledge base is essential.

A primary challenge in developing a system to generate a crime knowledge base automatically from online newspapers is in enriching the knowledge base with text as well as image data without any loss and redundancy, unlike the existing works that consider only text information (Alruily, Ayesh, & Zedan, 2014; Arulanandam et al., 2014; Chau, Xu, & Chen, 2002; Dasgupta, Naskar, Saha, & Dey, 2017). The knowledge base so enriched, provides complete information about an entity in a single place without modality barriers. Identification of entities and their relations from multimodal data through Named Entity Recognition (NER) and semantically integrating the similar entities are still open problems in the domain of knowledge base construction. In NER entities like person, location, organization, etc. are extracted from unstructured texts by Natural Language Processing (NLP) tools like the Natural Language Tool Kit (NLTK). However, these tools generate untagged and erroneously tagged entities too. For example, for the phrase “Bangalore Police”, NLTK interprets “Bangalore” properly as LOCATION without assigning any tag to “Police”. Whereas the phrase “Lawrence Bishnoi Gang member” tagged by NLTK as either ORGANIZATION or PERSON. While integrating information from multiple newspapers, untagged entities will be ignored from adding to the knowledge base and lead to loss of information. Similarly, semantically similar entities will be added to the knowledge base more than once due to wrong tags assigned to them and cause redundancy. Thus, an entity and relation extraction mechanism that minimizes wrongly tagged and un-tagged entities are significant for a complete and non-redundant knowledge base.

Similar to the extraction of text data, extraction of image data is also necessary to keep the knowledge base complete. After extracting multimodal data from different newspapers, it is vital to correlate the entities from various newspapers semantically. Finally, it is also critical to populate the knowledge base with an enormous amount of integrated information using a knowledge representation model. Among many knowledge representation models like distributed, symbolic, probabilistic and rule-based, ontology is now widely used in many areas like artificial intelligence, biomedical informatics, semantic web, system engineering, forensic data analysis, and information architecture (Jalil, Ling, Noor, & Mohd., 2017). Knowledge base represented in the form of ontology can be used further by many of the data mining tasks such as clustering. Clustering entities based on ontological relationships like hypernym (Dhuria, Taneja, & Taneja, 2016), only part of the ontology can be accessed by the users from the whole knowledge base. Similarly, the knowledge base can be further used in question answering systems to retrieve information about a particular entity from multiple sources by means of SPARQL queries (Hazrina, Sharef, Ibrahim, Murad, & Noah, 2017) over a global schema described by the ontology (scar Ferrndez, Izquierdo, Ferrndez, & Vicedo, 2009).

The primary contributions of this paper are as follows:

- Effective named entity recognition using rule-based approach.
- Semantic integration of crime-related text and image data.
- Development of a crime knowledge base using ontology representation model.

The rest of the paper is organized as follows. The next section discusses the Background and Related work. Section 3 describes the problem along with research objectives. The proposed methodology explained in Section 4. Section 5 presents the results and analysis of experiments conducted. Conclusion and future works are explained in Section 6.

2. Background and related work

This section aims to provide an overview of the work in the domain of crime information extraction, integration, and ontology-based knowledge representation.

Information Extraction techniques concentrate on extracting some specific and pre-defined entities from text related documents (Derczynski et al., 2015). Data mining and crowdsourcing are the Standard techniques used by several researchers in the area of crime information extraction. Entity Extraction, Clustering Techniques, Association Rule, Classification Techniques and Social Network Analysis are the typical data mining techniques used for crime information extraction and analysis (Hossein, Huang, Silva, & Ghodsi, 0000). These techniques are used to identify most complex crime patterns along with identifying the crime related entities. Here, the discussion is limited to the entity extraction techniques as the extraction of entities is the basic requirement for integrating the information from multiple newspapers. Authors in Furtado et al. (2010) used crowdsourcing where they provide a platform for individual users to report crime-related information and is available to all other users to view and comment. The main problem in such an environment is difficulty in verifying the reliability of reported crimes. The solution provided by authors to address this issue is not applicable to data generated by online newspapers and applies to only the platform they created.

The effective identification of named entities using Named Entity Recognition (NER) techniques and/or tools (Goyal, Gupta, & Kumar, 2018) is the primary goal of entity extraction. The existing approaches to achieve NER are based on hand-crafted Lexicons and Rules, Statistical models and Machine learning algorithms (Chau et al., 2002). Lexical lookup based approaches maintain a list of hand-written lexicons which contain some known entities of interest. These systems check for lexicons in the given sentence that match any of the lexicons in the list to identify the entities in the given sentence. Rule-based approaches use hand-crafted rules to identify the entities. Systems based on Statistical methods make use of some training data along with statistical models like Conditional Random Fields (CRF) to identify specific patterns for entities in texts. Machine learning based systems use some machine learning algorithms like Hidden Markov Models (HMM), Maximum Entropy Markov Model (MEMM), Neural Networks and Decision Trees to identify and extract patterns from texts. Even though any of the available open NLP tools (URL) like NLTK can be used to achieve the task, the results obtained from tools alone can not be used in its entirety due to inaccuracies in generating the tags (Arulanandam et al., 2014). Hence tools like NLTK can be best utilized in combination with some NER approaches to improve the

accuracy in identifying the named entities. Statistical models and machine learning algorithms need a large set of training data to achieve accuracy. Whereas, developing the lexicons of all categories in the crime domain to cover the entities extracted from various sources is an impossible task. Hence the rule-based approach is chosen in combination with NLTK to achieve the extraction task. There has been much work in adopting NER for identifying entities belongs to crime domain. For example in Arulanandam et al. (2014), authors used NER to identify locations and adopted CRFs to assign each sentence in an article either as crime location sentence or not a crime location sentence. Many of the works like Ku, Iriberri, and Leroy (2008) used lexical lookup based approaches along with NLP Techniques like POS tagging for extracting crime-related data from various sources. But these works extract the entities without focusing on the impact of the extracted entities towards the integration of information from multiple sources.

Related to the integration of information, the capability of the system is restricted to find the similarity between the entities of similar types based on their contextual and semantic similarity. Many of the string similarity measures exist like character n-gram similarity (Kondrak, 2005), Leven-shtein Distance (Miller, Vandome, & McBrewster, 2009), Jaro-Winkler measure (Xiao, Wang, & Lin, 2008), Jaccard similarity (Chaudhuri, Ganti, & Kaushik, 2006), tf-idf based cosine similarity (Salton & Buckley, 1988) and Hidden Markov Model-based measure (Miller, Leek, & Schwartz, 1999). However, these metrics are limited to measure the similarity between strings syntactically but not semantically like synonyms using external knowledge sources like Wikipedia (Qu, Fang, Bai, & Jiang, 2018), which provides a better semantic correlation between the entities. To improve the quality of syntactic similarity measures, many of the machine learning-based methods like Tsuruoka, McNaught, Tsujii, and Ananiadou (2007) have been proposed. Even though these methods try to capture the semantic similarities between strings, they are limited to some pre-defined domains and need a huge set of training data for the effective capturing of semantics. Corpus-based methods like word embeddings are also introduced to find semantic similarity. The authors in Zhu and Iglesias (2018) have conducted experiments for checking the semantic similarity using the knowledge and corpus-based methods. Based on their observation, corpus-based methods like embeddings do not consider various meanings of words, and when words have some relations, the learned word vectors are not as accurate as knowledge-based methods. Moreover, for corpus-based methods when the training corpora changes, the similarity between the words are different due to change in word vectors. Whereas, in knowledge-based methods, similarity scores are different only when the corresponding similarity metric changes as ontologies are normally stable and fixed. However, the corpus-based methods are more suitable to find the semantic relatedness between the two phrases based on the word co-occurrences or surrounding words and their frequency of occurrences in two phrases (Zhu & Iglesias, 2018). Hence, in this work, we explore both corpus-based as-well-as knowledge-based methods to find similarity between semantically related entities.

The existing systems vary in the way the ontology is adopted in the extraction and the knowledge representation process. Some of the systems use domain-based ontology as an input in assisting the extraction work and produces the extracted entities and relations as output. For example, in Dasgupta et al. (2017), authors proposed a method for entity and relation extraction from crime news articles by performing some NLP tasks and extracting the knowledge from crime ontology. Some of the works produce the ontology as an output which can be used for further analysis. For example in Jalil et al. (2017), a prototype to analyze motorcycle theft is developed. Here, an ontology model is developed to represent the crime related concepts which are used to identify any cases related to the new theft case an investigation officer is looking for. Few other works take ontology as input in assisting the extraction process, and the extracted entities and relations are added to the knowledge base as an instance of ontology. For example in Gregory, McGrath, Bell, O'Hara, and Domico (2011), authors proposed a system that automatically populates a knowledge base by extracting entities and events from structured and unstructured sources with the help of ontology. In this work, the ontology is used as a knowledge representation model for the entities extracted from online newspapers.

A system called SOBA (Buitelaar, Cimiano, Frank, Hartung, & Racioppa, 2008) generates the knowledge base for the sports domain by extracting and integrating the sports-related entities from heterogeneous sources. Even though the system addressed the problem of removing the duplicates from the knowledge base, the loss of information while integrating is ignored. A system for open intelligence platform like CAPER (Aliprandi et al., 2014), considers multi-lingual texts belongs to 13 foreign languages along with English and also other modality data like audios and videos along with texts and images. However, the system ignored the effect of both duplicity and loss of information while integrating the data from multiple sources. The proposed work is limited to only English language due to unavailability of proper NLP tools in the context of Indian languages. Though audio and video modality data is considered by Aliprandi et al. (2014), the processing of audio and video is treated as the processing of text by using appropriate speech to text conversion tools and reducing videos into frames and audio. Hence, the proposed system considers only text and image data for processing. Some of the works like Kalemi, Yildirim-Yayilgan, Domnori, and Elezaj (2017) considers extraction of information from social media and Aliprandi et al. (2014) considers any web sources specified as URL along with social media. Due to privacy issues in accessing a large data set from social networks, this work considers only online newspapers with the primary concern as knowledge base generation.

From the above-related works, to the best of our knowledge, the system is unique to the domain of crime information extraction, integration and knowledge base generation without loss and duplicity from multiple online newspapers.

3. Problem description

This paper aimed at developing a system-"Crime Base", for creating a knowledge base of text and image entities and their relations extracted from text and image content of online newspapers in the English language. The proposed system focused at extracting and integrating entities and their relations with the desire to upkeep the knowledge base without duplicity and loss of information. The image captions are used to extract the image entities from images and both image captions as well as low-level

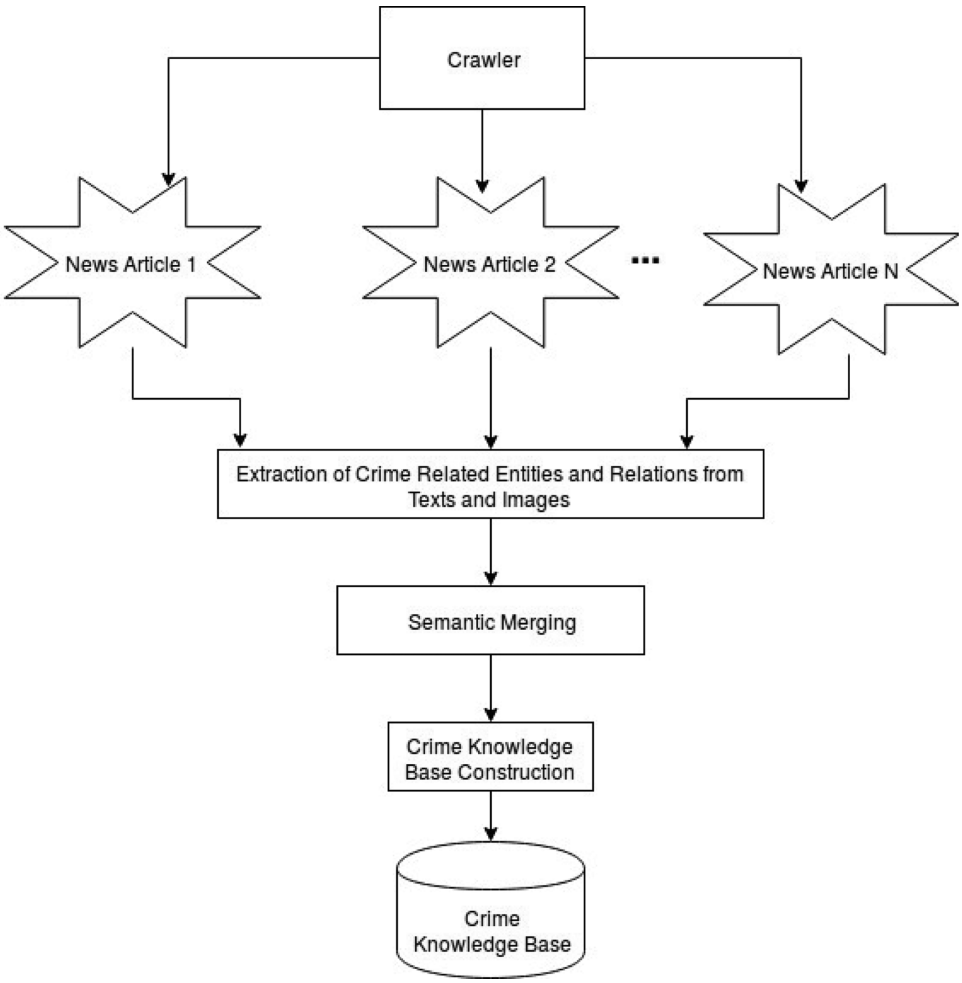

Fig. 1. Steps used in proposed system architecture.

features of images are considered for integrating the images from multiple news articles.

To address the problem described above, the following research objectives are set:

- To effectively recognize named entities using the rule-based approach to minimize the tagged or untagged entities obtained from NLTK.
- To semantically integrate crime related text and image entities.
- To develop a crime knowledge base using ontology representation model.

4. Methodology

The overall architecture of the proposed system is shown in Fig. 1. The proposed system consists of three main stages such as Information Extraction, Semantic integration and knowledge base generation. Each of the stages is discussed in detail in the following subsections.

4.1. Information extraction

The primary goal of any information extraction system is in assigning the named entities like PERSON, ORGANIZATION, LOCATION etc. to different segments of the text and extracting their relations (Popov et al., 2003). In this work some hand-crafted rules are applied over the POS and NER tags generated by NLTK to extract the entities and relations. The process involved in the generation of entities and their relations by the proposed approach is shown in Fig. 2 and is explained in the following steps:

(a) Data Acquisition

In this work the topic modeling and knowledge base aided domain-aware crawler is developed which gathers and stores crime

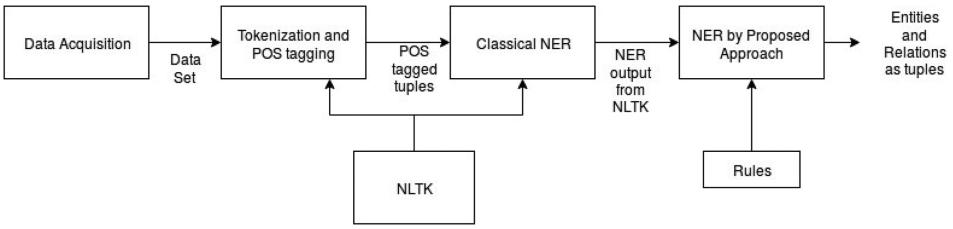

Fig. 2. Generation of entities and relations.

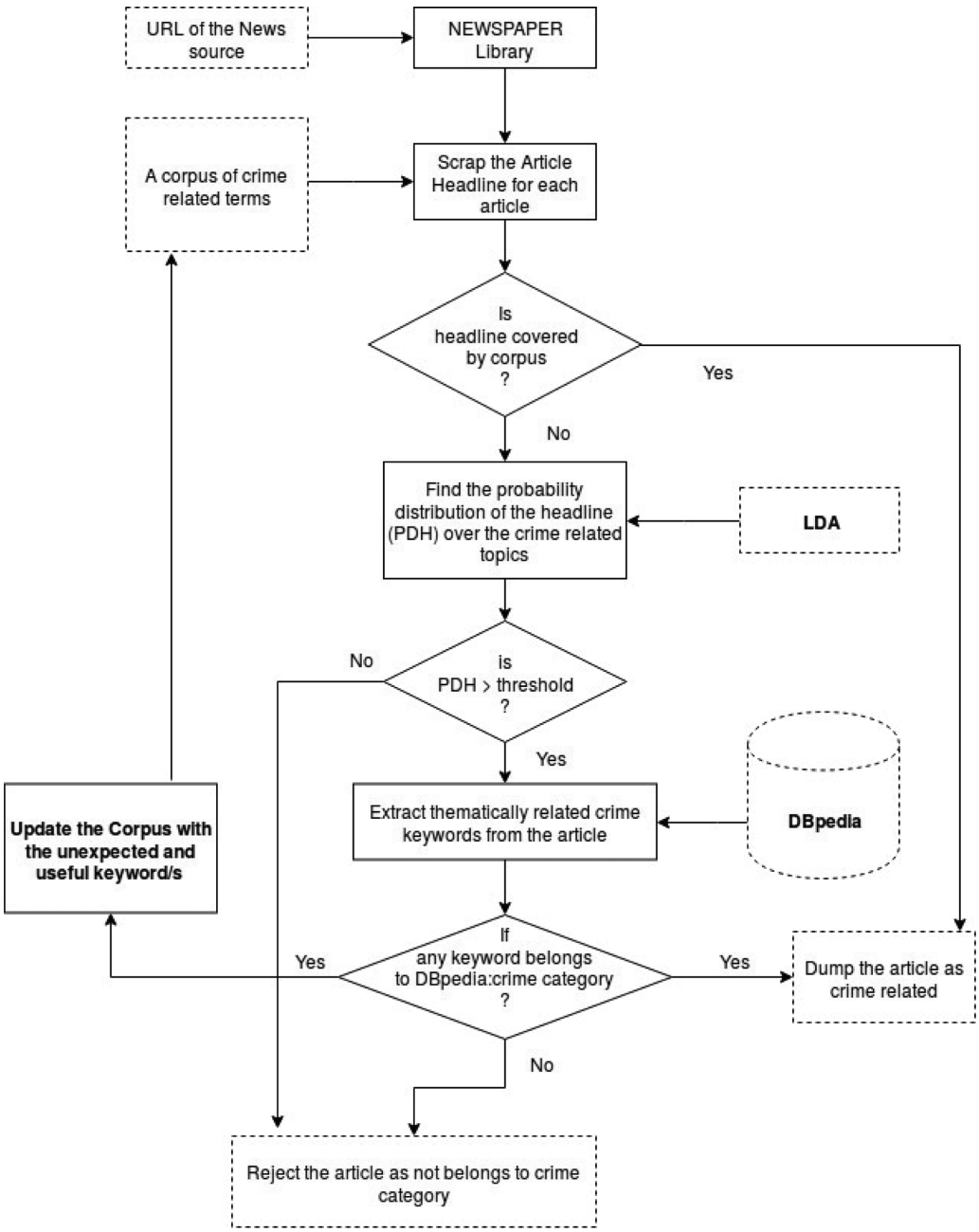

Fig. 3. Topic modeling and Knowledge base aided data acquisition.

related news articles from the online newspapers. The detailed process involved in gathering the crime related texts and images via crawling of URLs specified by the user is shown in Fig. 3 and is explained in the following steps:

- The main-URL from which articles are to be scrapped is specified. For example, "https://indianexpress.com/ " is the main-URL to scrap the articles from Indian Express newspaper.
- For each of the articles present in the main-URL, the domain-aware articles are selected using article's headline, obtained from the sub-URLs. A sample of a sub-URL related to an article from Indian express newspaper is shown below: "http://indianexpress.com/article/india/gurgaon-police-arrests-key-lawrence-bishnoi-gang-member-sampat-nehra-underworld-gangster-5207714/ "
- Selection of articles are initially bootstrapped by a bag of keywords, a corpus of 100 crime related terms constructed by collecting the terms manually from URL.
- For each news article, the terms present in the article's headline are compared with the keywords in the corpus. If the headline includes any of the keywords, the article is categorized as crime-related. Otherwise, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), a well-known topic modeling technique is applied to identify true positive or negative crime article based on the probability distribution score of the headline over a crime related topic. For the headline with a probability score less than the threshold, the article will be classified as true negative. Some times LDA will generate the false positives due to improper distribution of the headline over a crime topic. This is overcome by extracting the keywords from the article and finding their thematic relatedness towards crime domain using an external knowledge base, DBpedia. An article with a keyword belongs to DBpedia:crime category will be classified as true positive and corpus will be updated with the respective keyword. The proposed data gathering guided by DBpedia along with LDA also updates the corpus with more useful and unexpected keywords present in the articles.
- For each selected sub-URL, web pages are scraped by selecting only text related to the main article and top image of the article.
- Scraped text and image content is dumped to an output file for further processing.

(b) Extraction of Entities and Relationships

A new algorithm is proposed to extract information from texts based on the combination of NLP and rule-based technique. This improves the performance of named entities obtained from NLTK in terms of reduced number of erroneously tagged and untagged texts. The steps followed in the process of entity and relation extraction is explained below:

Tokenization and POS Tagging. The texts gathered for each article during data acquisition step are tokenized into sentences and words. Tokenized words of each sentence are tagged with POS tags using NLTK POS tagger. POS tags generated for two sample sentences crawled from Indian express newspaper dated *june* 7th, 2018 URL are shown below. The sentences are referred as test sentences in the further discussion. Information about meaning for each tag can be obtained from Loper and Bird (2002).

Test Sentence-1: In a major breakthrough, Gurgaon Police on Wednesday night arrested Sampat Nehra, a member of the Lawrence Bishnoi gang, from Hyderabad.

POS tags: In/IN,a/DT,major/JJ,breakthrough/NN,/,Gurgaon/NNP,police/NN,on/IN,Wednesday/NNP, night/NN,arrested/VBD,Sampat/NNP,Nehra/NNP,/,a/DT,member/NN,of/IN,the/DT,Lawrence/NNP,Bishnoi/NNP,gang/NN,/,from/IN,Hyderabad/NNP,./.

Test Sentence-2: According to police, 28-year-old Nehra was the gangs sharp-shooter, and entered the underworld through the route of student politics.

POS tags: According/VBG,to/TO,police/NN,/,28-year-old/JJ,Nehra/NNP,was/VBD,the/DT,gang/NN,'/NNP,s/VBD,sharp-shooter/NN,/,and/CC,entered/VBD,the/DT,underworld/NN,through/IN, the/DT,route/NN,of/IN, student/NN,politics/NNS,./.

Named Entity Recognition and Relationship Extraction. Based on the POS tags obtained from the previous step, meaningful entities and their relations are identified with the help of $\langle subject - verb - object \rangle$ rule. The overview of the proposed entity and relation extraction approach is given in Algorithm 1. This work identifies seven different types of named entities like PERSON, ORGANIZATION, LOCATION, DAY, SUBJECT, OBJECT and PERORG. Unlike many works that extract the relations after extracting the entities (Dasgupta et al., 2017), the proposed work follows open relation extraction (Zouaq, Gagnon, & Jean-Louis, 2017) which first extracts the relations to identify the related entities properly and also to assign tags to untagged terms to avoid loss of information while integrating data from multiple sources. For example, the NER tags assigned by NLTK for test sentence-2 is shown below:

According/VBG,to/TO,police/NN,/,28-year-old/JJ,(PERSON Nehra/NNP),was/VBD,the/DT,gang/NN,/NNP,s/VBD,sharp-shooter/NN,/,and/CC,entered/VBD,the/DT,underworld/NN,through/IN,the/DT,route/NN,of/IN,student/NN,politics/NNS

Here, only "Nehra" is recognized as an entity and tagged with PERSON. Other attributes like the age of the person is not recognized and hence will be missing in the knowledge base if not considered for tagging.

In order to address the issue, initially, the relations are extracted by applying the following rules and are tagged with REL:

- Any term with POS tag like 'VB', 'VBD', 'VBN', 'VBG', 'VBZ' and 'VM' is considered as a verb and is a primary candidate for relation. Similarly, terms "on", "from" and "at" are also selected as relations.

Input: P: Set of POS tagged tuples, R: Set of Rules
Result: T: Set of tuples that represent entities and their named entity tags

```
for each tuple t of P do
for each Relation term RE of t do
Divide the sentence at the RE if RE has non-empty left and right part
Subject ← Left(RE)
Object ← Right(RE)
end for
Apply the rules from R, assign named entities and generate the tuples in the form of [key- value] pair into T
end for
```

Algorithm 1. Algorithm for Entity and Relation extraction.

Table 1
Rules for NER tagging.

Rules	Description
(*"from"* | *"at"* | *"in"* (*Noun*) +) → *LOCATION*	One or more Noun terms preceded by "from" or "at" or "in" clause is labelled as LOC(LOCATION)
(*"on"* (*Noun*) +) → *DAY*	One or more Noun terms preceded by "on" clause is labelled as DAY
(((*Noun*)*(*Adjective*)*(*Noun*)*)*Verb) → *PERSON/ORGANIZATION/PERORG*	Any combination of Noun and adjective terms succeeded by verb are labelled with either (*PER*) PERSON or (*ORG*) ORGANIZATION based on the label provided by NLTK to any of the noun phrases. If both PERSON and ORGANIZATION terms are present, then the combination is labelled as PERORG. If none of the labels assigned by NLTK, it is treated as subject and is represented as SUB.
(*Verb*((*Noun*)*(*Adjective*)*(*Noun*)*)*) → *PERSON/ORGANIZATION/PERORG*	Similar to the previous pattern but the pattern is preceded by verb. Here, If none of the labels assigned by NLTK, it is treated as object and is represented as OBJ.

- Any verb followed by a term with POS tag as 'IN' are considered together as a single relation. For example "of" in "accused of" is taken along with verb "accused" to represent it as a relation.
- Two or more continuous verb terms are considered as a single relation.

Once the relations are extracted, terms from the English stop word list which are not identified as a relation and delimiters from POS tagged list are removed before proceeding to the next step. Then each sentence is divided based on relationships as a pivot in such a way that it has non-empty left and right part and are considered as 'subject' and 'object' respectively. The dividing process will continue until there is no relation found in each part. The rest of the work is in assigning the possible named entity tags to the words in subject and object part of the sentences which is done by applying the hand-crafted rules listed in Table 1.

NER tags assigned by the proposed method for the test sentences is shown in Fig. 4. The output is generated in the form of tuples that include key-value pairs to represent named entity tag and the corresponding name of the entity.

Similarly, the entities associated with the images are extracted by applying the proposed algorithm to the caption associated with them.

4.2. Knowledge base construction

A knowledge base of entities and relations are developed by mapping the named entity tuples generated by the previous step to equivalent components of ontology developed using OWL. The procedure to generate the knowledge base is shown in Algorithm 2. The significance of the work is in representing image data along with text data in the knowledge base which provides complete details of an entity in a single place. The general structure of the knowledge base constructed in this work is shown in Fig. 5.

Here the knowledge base is generated using Owlready which is a Python 3 library for working with OWL 2.0 ontologies (URL). It provides API calls to create ontologies, load existing ones as Python objects, modify them, and save them as OWL XML files. The tuples generated from the named entity recognition step are considered to include the entities into the knowledge base. Corresponding to each tuple, an event is created with an event id represented as *event_article number_sentence number* to identify the events. Edges are added between the event object and the entity object. These edges are labeled according to the nature of the entity like location, organization, etc. Relation entities may also be connected to other entities participated in the relation. In the ontology, image entities are linked with each of the events if the images are associated with any of the respective entities belongs to the events and are labeled with the respective Relevance Measure (RM). RM represents the relevance of the image with an event and is calculated as:

$$RM = \frac{(Number\ of\ entities\ of\ an\ Image\ matched\ with\ entities\ of\ an\ Event)}{(Total\ number\ of\ entities\ associate\ with\ the\ Image)}$$

For better visualization of the knowledge base in OWL format, knowledge base is visualized as an interactive graph with the help of a graph tool with all the entities displayed as nodes and relationships as edges between them. Knowledge graph so generated presents the information in a concise and interactive form. The user can zoom in/out and also drag the nodes to filter out only the

```
--------for Test sentence-1---------

['PER-major breakthrough Gurgaon police','REL-on','DAY- Wednesday night','REL
-arrested','PERORG-Sampat Nehra member Lawrence Bishnoi gang','REL-from','LOC
-Hyderabad']


--------for Test sentence-2---------

['PER-police 28 year old Nehra','REL-was','OBJ-gang sharp shooter','REL-enter
ed','OBJ-underworld route student politics']
```

Fig. 4. NER tagged tuples.

Input: T: List of NER tagged tuples
Result: Knowledge Base in OWL format

for each tuple in T do Event = New Event(Random Event ID) do
$REL\_FOUND \leftarrow 0; PREV\_ENTITY \leftarrow NULL; count \leftarrow 0$
for each item in tuple do Switch(item type)
if Relation then:
Event.action.append(item name);$REL\_FOUND \leftarrow 1; REL\_item \leftarrow item$
end if
if Location then:
Event.location.append(item name) //Similar for other entity types except Image
end if
if Image then:
for each item1 in NER tagged tuple obtained from Image caption do
for each item of the Event do
if ((item1 name) is part _ of or equal _ to (item name)) or vice versa
$count++$
end if
end for
if $count > 0$
$RM = count/|item1|$
Event.Image.append(item1.RM)
end if
end for
end if
end for
end for

Algorithm 2. Algorithm for Knowledge Base Construction.

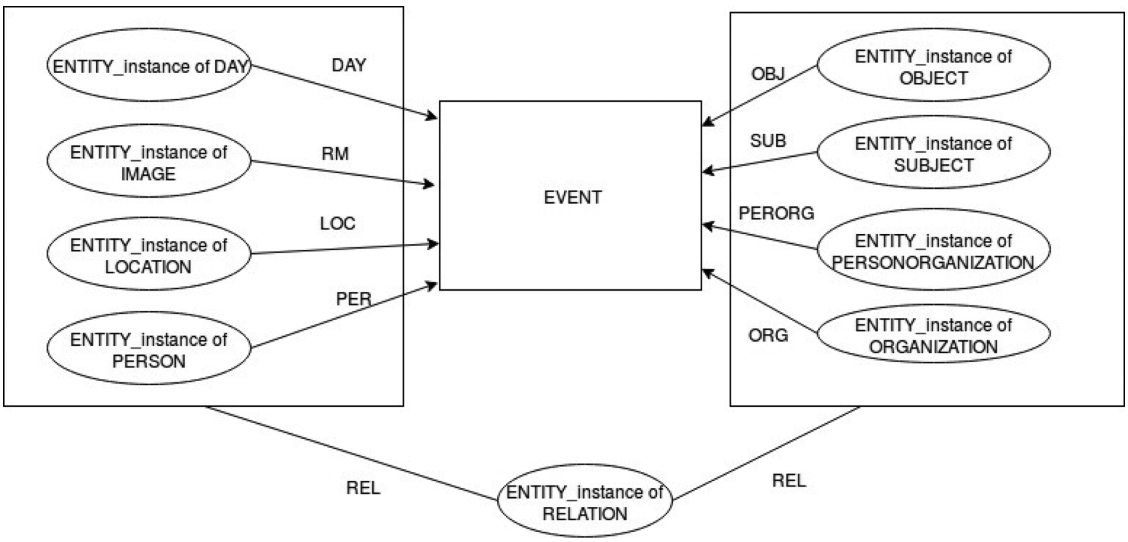

Fig. 5. General Knowledge Base Structure.

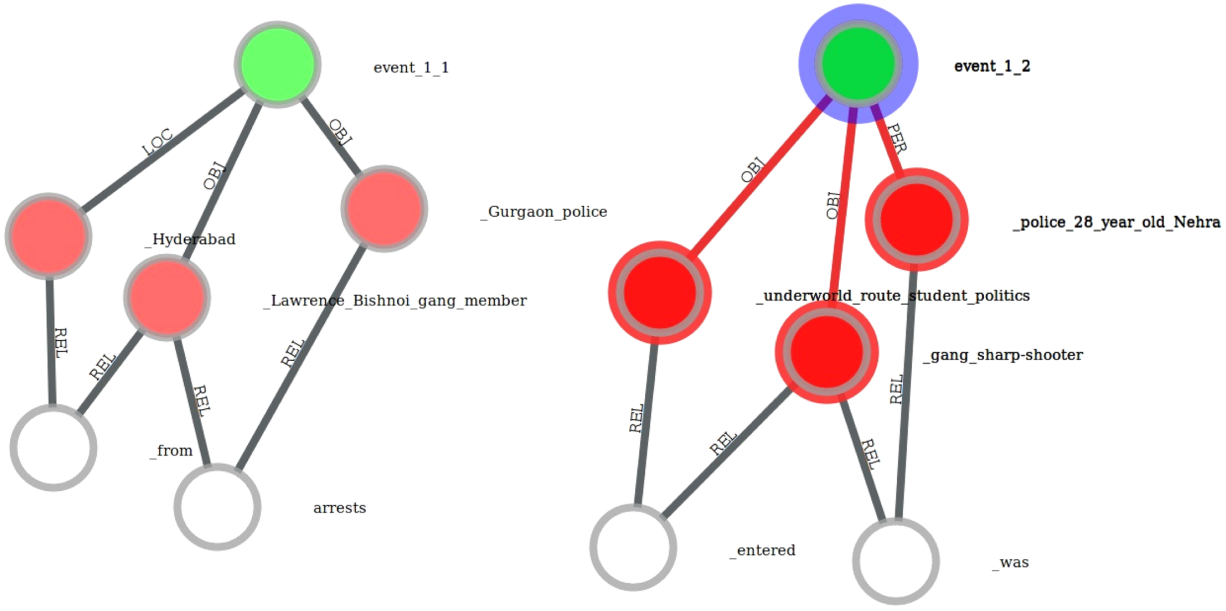

Fig. 6. Knowledge Graph without image entity.

entities he/she is interested in. Fig. 6 shows the interactive graph generated for the OWL file obtained for test sentences 1 and 2. Fig. 6 shows the graph for test sentences without an image entity being added to it. To illustrate the graph for the image entity, an image with the caption ”Bishnoi key gang member Sampat Nehra arrested by Gurgaon police”, is considered. Fig. 7 shows the updated graph with the image as an entity added to the graph in Fig. 6. Since the entities extracted from the caption of the image are matched with all the entities of test sentence-1 and a single entity of test sentence-2, the image is associated with event 1 and 2 with RM 1 and RM 0.3 respectively. Here, Green, Red and White nodes represent Event, Entity and Relation nodes respectively.

4.3. Semantic merging

The goal of semantic merging is to integrate the events and entities extracted from multiple sources to avoid the duplication and to enrich the knowledge base with the new information about an entity or event (Dragos, 2013). This is achieved by checking correlation or similarity between the entities of two events obtained from different newspapers so that, the final knowledge base is enriched with a single instance of the correlated events and entities.

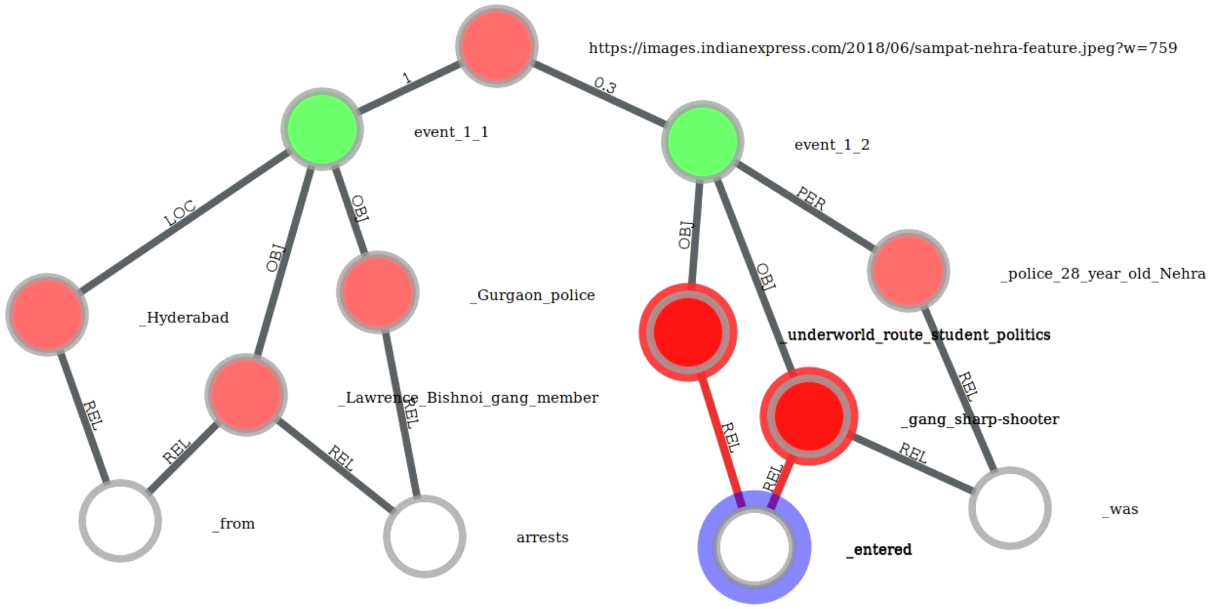

Fig. 7. Knowledge Graph with image entity.

Due to the extraction of image data along with texts, the similarity between the entities of similar types is considered at two phases i.e. similarity between text entities and image entities. Similarity measures followed by each phase and the calculation of the final similarity measure between two events are discussed in the following subsections.

4.3.1. Similarity between text entities

In this work, we explore the power of contextual as well as semantic similarity to find the similarity between the entities of two events. Contextual similarity measure helps to identify the semantic relatedness between the events based on the context in which the entities participating in the events are used. In general, events that are not semantically related can't be semantically similar. For instance, the phrases "death at the bank of a river" and "fraud at canara bank" are semantically non-related and hence are also dissimilar. Similarly, the phrases "ATM fraud at canara bank" and "robbery at SBI bank" are semantically related but are not similar. Whereas, the phrases "ATM fraud at canara bank" and "ATM robbery at canara bank" are semantically related as-well-as semantically similar. Hence it is necessary to consider the contextual meaning of the entities along with their synonymic relations to capture more semantics about the entities. And also, finding the contextual similarity in prior to semantic similarity restricts the identification of similarity between only semantically related events and hence avoids the semantic similarity check between two events that are semantically non-related.

Fig. 8 shows the proposed method for semantically merging the two events using contextual as-well-as semantic similarity measures and are discussed below:

Contextual Similarity Measure: In this work, we used Word2Vec, a predictive based word embedding model to represent the entities in a dense and low dimensional vector space due to its high performance in many applications (Zhu & Iglesias, 2018). The vector for each entity represents its description of how it occurs in context with other entities in an event. Hence, if two entities are used in a similar manner in two different events, the chances of obtaining similar vector representation for those entities are more. For example, if a relation "killed" is related by two-person entities in an event, its probability of related by the same or different person entities in other events is high. After mapping events into the vector space, their similarity is computed using standard cosine similarity measure (Salton & Buckley, 1988). Events whose contextual similarity score is less than a threshold are treated as independent. Otherwise, they are examined for their semantic similarity by using the following method.

Semantic Similarity Measure: The capability of Word2Vec models lies is handling large corpus and trains the word vectors efficiently. However due to the use of only the word sequences or co-occurrences of words for training the word vectors, Word2Vec model fails to handle words with synonymous and hierarchical relations precisely (Zhu & Iglesias, 2018). As a result, Wu & Palmer method, a knowledge base algorithm to achieve semantic similarity is adapted by using the information available from a well known ontological repository, WordNet. The semantic similarity measure finds similarity based on the depth of the two synsets of the words in WordNet taxonomy along with the depth of the Least Common Subsumer (LCS) (Wan & Angryk, 2007). A similarity score of 1 indicates highly correlated entities.

In case of events with multiple entities of same type, the similarity between them is calculated by finding the sum of the similarity between each of the entities. Let R_1 denotes the set of relations in the first event and R_2 denotes the set of relations in the second event, then the similarity of relation entities for the two events is calculated as:

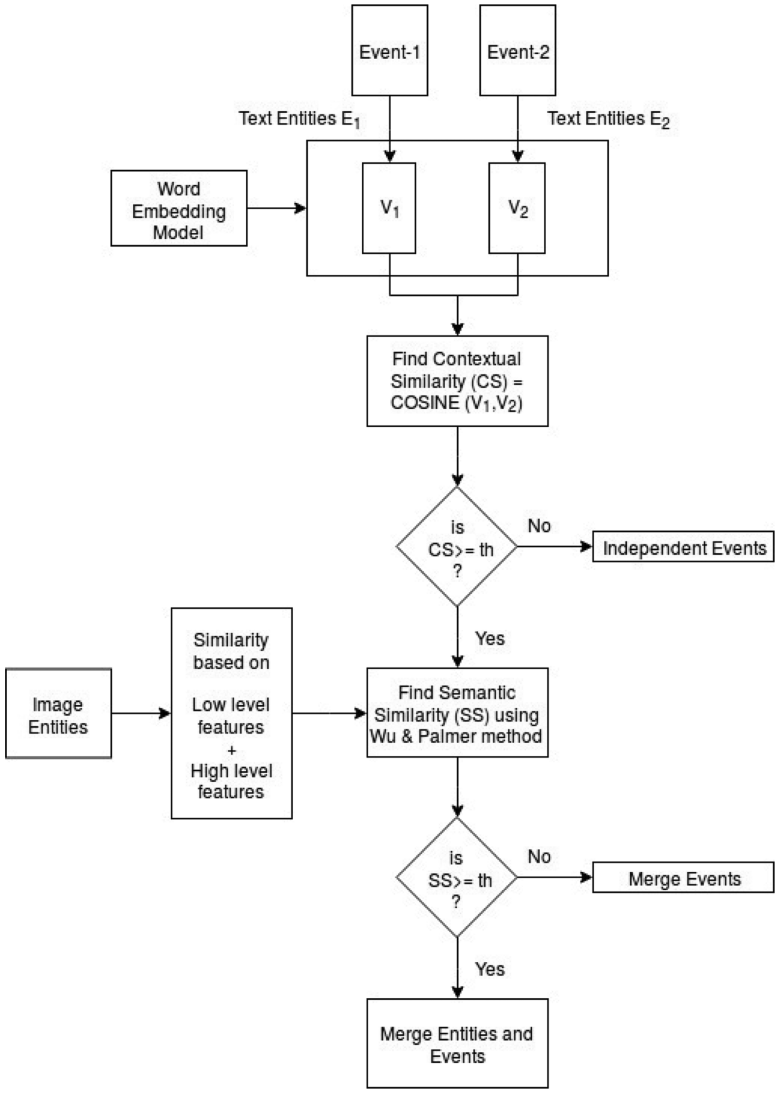

Fig. 8. Semantic Merging of events and entities.

$$sim(R_1, R_2) = \frac{\sum_{r_1 \in R_1} \sum_{r_2 \in R_2} wup\_sim(r_1, r_2)}{|R_1| * |R_2|}$$

4.3.2. Similarity between image entities

In this work, two image entities are compared for semantic similarity by matching low-level features as well as high-level captions. For the former case, the Structural Similarity Index(SSIM) method (Wang, Bovik, Sheikh, & Simoncelli, 2004) is used that perceives changes in the structural information of the image by comparing its local regions. Small subsamples of the whole image of two images are compared and a similarity score is obtained by calculating the structural similarity index on various windows of same sizes.

For finding the similarity of images using high-level captions, a vector-based method is proposed that interprets the images semantically rather than just interpreting images based on low-level features like color, shape, etc.

For the two images to be compared, a two dimensional vector V of captions and named entity chunks identified for the captions are formed. An entry $V[i, j]$ into the vector is equal to 1, if the ith caption includes the jth named entity, otherwise is equal to 0. Final similarity score based on the high level captions is calculated as:

$$caption\_score = \frac{\sum_{i \in caption} \sum_{j \in NamedEntity} V[i][j]}{Size(V)}$$

The final similarity score for two image entities i_1 and i_2 is the average of the similarity score obtained from low-level features and

Table 2
Weights given to entities for event similarity calculation.

Entity	LOC	REL	SUB	OBJ	ORG	PER	PERORG	DAY	IMG
Weight	1	1	0.75	0.75	1	1	0.5	0.25	0.25

high-level captions and is calculated as:

$$image\_similarity(i_1, i_2) = \frac{SSIM(i_1, i_2) + caption\_score(i_1, i_2)}{2}$$

If events contain multiple image entities, then the similarity between the image entities is calculated by the sum of the similarity between each of the multiple image entities. If I_1 denotes the set of images in the first event and I_2 denotes the set of images in the second event, then the similarity of image entities for the two events is calculated as:

$$image\_score(I_1, I_2) = \frac{\sum_{i1 \in I_1} \sum_{i_2 \in I_2} image\_similarity(i_1, i_2)}{|I_1| * |I_2|}$$

4.3.3. Similarity between two events

The similarity between two events say E_1 and E_2 is calculated by the weighted sum of the similarity between different entities from the two events as follows:

$$sim(E_1, E_2) = \frac{\sum_{i=1}^{9} w_i * sim(E_{1i}, E_{2i})}{|E_1| * |E_2|}$$

where w_i represents the weight of the ith entity of the event, E_{1i} represents the ith entity of the first event and E_{2i} represents the ith entity of the second event. Based on the repeated experiments the weights are assigned to each entity type that indicates the contribution of each in finding similarity among the entities and is given in Table 2.

Finally, two events are said to be semantically similar if the similarity score is greater than a threshold and in such case knowledge base is enriched with a single instance of the correlated events. Otherwise, two events are decided to be only semantically related but not similar and hence are merged into a single event without merging their respective entities.

5. Experimental analysis

The system considers three prominent newspapers *Indian Express, Times of India and Deccan Chronical* which have articles available online. The corpus includes the data collected from *Jan* 2018 to *Jun* 2018. The following sections describe the experimentation and evaluation results for data gathering, named entity recognition and semantic merging for knowledge base construction.

5.1. Evaluation of topic modeling and knowledge base aided data gathering

The proposed method for data gathering using topic modeling and DBpedia has achieved a significant improvement in identifying the true positive articles and also in filtering the true negative and false positive articles compared to the selection of articles based only on a corpus of crime-related terms. The experiment is conducted by training the LDA model using ABC News headlines dataset (URL). To illustrate the advantage of the proposed approach, a corpus of ten crime related terms are considered. Table 3 shows the result for six articles and their status of acceptance or rejection based on their relatedness to the crime domain. First three rows of the table are evident for true positive crime related articles. The first row is an example of a trivial case where the article is selected based on the presence of the keyword ”weapon” in both the headline and the corpus. Second and third rows show the selection of articles by applying LDA and using DBpedia as an external knowledge base. The probability score indicates the highest probability distribution score of an article towards a crime topic. The articles are selected as the probability score is more than the threshold and having the keywords whose entry is found in the DBpedia crime category. This helped to update the corpus with the useful and unexpected keywords like “gang”, “gangster” and “rape” which were not available in the corpus previously. The fifth row shows the true negative article whose rejection as crime article is trivial due to its probability score which is less than the threshold. The improper distribution of true negative articles towards a crime topic by the LDA is overcome by the use of DBpedia which is evident from the sixth row. Even though the article in the sixth row has the highest probability score, it is rejected due to non-availability of any keyword related to DBpedia crime category. The limitation of the proposed method is in filtering false negative articles which is also evident from the fourth row. Although the article in the fourth row is confirmed to be crime related by the probability score, due to the non-availability of any keyword related to DBpedia crime category, the article is categorized as a false negative. Such false negatives can be avoided by introducing more external knowledge bases along with DBpedia which will be considered in future.

5.2. Evaluation of semantic merging

In this work, experiments are conducted to find most similar events using both word embedding and knowledge base algorithm. The contextual similarity between the events is procured by utilizing a large collection of text as a training corpus to develop a word embedding model. To achieve this, a pre-trained Word2Vec model of news corpus provided by Google and Gensim, a python

Table 3
Data gathering results for 6 articles.

Initial corpus	Updation to the corpus	Sub-URL	Probability score	Keywords	DBpedia crime entry identified	Status
		https://www.indiatoday.in/crime/story/ j-k-2-held-for-weapon-snatching- 1475975-2019-03-12				✓
	gang, gangster	https://timesofindia.indiatimes.com/city/gurgaon/ member-of-haryanas-lawrence-bishnoi- gang-held-in-hyderabad/articleshow/64495307.cms	0.9956219	Gang,Murder Gangster	http//dbpedia.org/resource/Category:Crimes http://dbpedia.org/resource/Category:Organized_crime_members_by_role	✓
arrest, murder, kidnap, corrupt, theft, harassment, assault, blackmail, crime, weapon	rape	https://www.indiatoday.in/crime/story/ 10-year-jail-for-raping-minor-in-mumbai-1477005-2019-03-13	0.7532135	Rape	http://dbpedia.org/resource/ Category:Sex_crimes	✓
		https://timesofindia.indiatimes.com/city/mumbai/ 6-year-old-abducted-girl-found-dead-in-railway-toilet-in-navsari/articleshow/63462708.cms	0.9934596	Dead,Abduct	No DBpedia crime entry	✗
		https://www.deccanchronicle.com/lifestyle/health-and-wellbeing/300319/ exercise-can-help-in-containing-arthritis.html	0.5414266			✗
		http://odishasuntimes.com/ mahaprayan-ambulance-failure-in-odisha- body-carried-on-rickshaw-baby-delivered-in-auto/	0.9629939	Ambulance,Rickshaw, Delivered,Body, Baby,Auto,Odisha	No DBpedia crime entry	✗

Table 4
Events and their description.

Event- id	Event description
E_1	Gurgaon police arrests key Lawrence Bishnoi gang member from Hyderabad
E_2	death at the bank of a river
E_3	ATM fraud at canara bank
E_4	10 year jail for raping minor in Mumbai
E_5	Woman drug officer shot in Kharar office; assailant dies of own bullet
E_6	Haryana Police Arrests Key Lawrence Bishnoi Gang Member From Hyderabad
E_7	fraud at canara bank
E_8	robbery at SBI bank
E_9	Woman officer shot dead at office in Punjab's Kharar
E_{10}	50-Year-Old Telangana Man Shot Dead In Florida Departmental Store

Table 5
Similarity scores between the events.

Events	Contextual similarity score					Semantic similarity score				
	E_6	E_7	E_8	E_9	E_{10}	E_6	E_7	E_8	E_9	E_{10}
E_1	***0.97***	0.23	0.22	0.20	0.18	***0.99***	0.12	0.10	0.09	0.00
E_2	0.12	0.17	0.16	0.25	0.26	0.07	0.20	0.22	0.15	0.14
E_3	0.11	***0.98***	***0.68***	0.10	0.09	0.07	***1.00***	***0.70***	0.11	0.10
E_4	0.13	0.12	0.10	0.11	0.11	0.25	0.22	0.19	0.00	0.02
E_5	0.13	0.11	0.12	***0.78***	0.23	0.00	0.00	0.00	***0.96***	0.15

Table 6
Knowledge Base before and after semantic merging- using only semantic similarity measure.

Before semantic merging			After semantic merging		
Knowledge Base size (in KB)	Num of events	Num of entities	Knowledge Base size (in KB)	Num of events	Num of entities
49.8	6	29	49.8	6	29
137	22	105	91.8	22	90
283.3	56	126	166.9	56	125
410.5	85	221	269.7	82	206
1024	152	334	536.9	152	328

Table 7
Knowledge Base before and after semantic merging- using both contextual and semantic similarity measure.

Before Semantic Merging			After Semantic Merging		
Knowledge Base size (in KB)	Num of events	Num of entities	Knowledge Base size (in KB)	Num of events	Num of entities
49.8	6	29	40.4	4	29
137	22	105	88	16	90
283.3	56	126	160.1	51	95
410.5	85	221	261.7	78	198
1024	152	334	529.2	144	301

framework for modeling the vector space is adapted to load the pre-trained model.

Empirically it is found that the contextual meaning of the entities can be better captured using word embedding models due to vectors created by these models based only on the co-occurrences of entities in the events. However, these models do not provide a higher similarly score for semantically similar events due to the failure of measuring the synonymic relations. In contrast, the knowledge base algorithms provide a higher similarly score for semantically similar events provided they are contextually similar. To illustrate, Tables 4 and 5 show the description of ten events and the similarity score between them using contextual and semantic similarity measures respectively. From the Table 5 it can be observed that for most similar events like $E_1 - E_6$ and $E_3 - E_7$, the contextual and semantic similarity scores are high and almost equal. However, a higher semantic similarity score always does not guarantee that the events are semantically similar. For example- unlike E_5 and E_9, E_3 and E_8 are not semantically similar even though their semantic similarity score is greater than the contextual similarity score. Hence, using either of the similarity measures alone do not capture complete semantics about the events. Consequently, both the similarity measures are adapted and empirically fixed a threshold of 0.5 for checking semantic relatedness and 0.9 for checking semantic similarity.

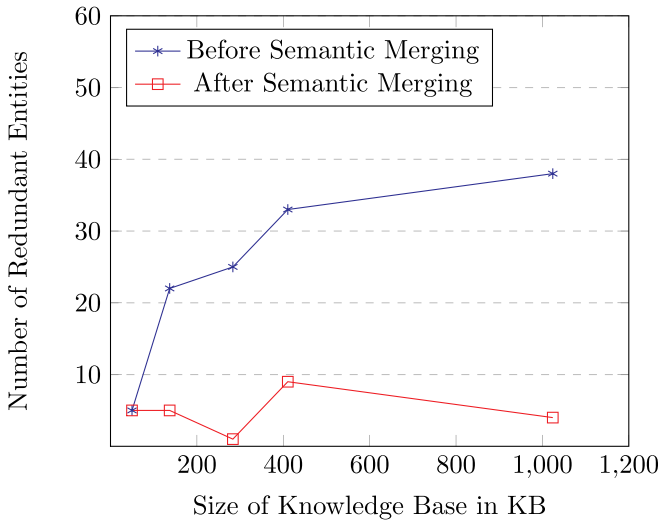

Fig. 9. Redundant entities before and after semantic merging.

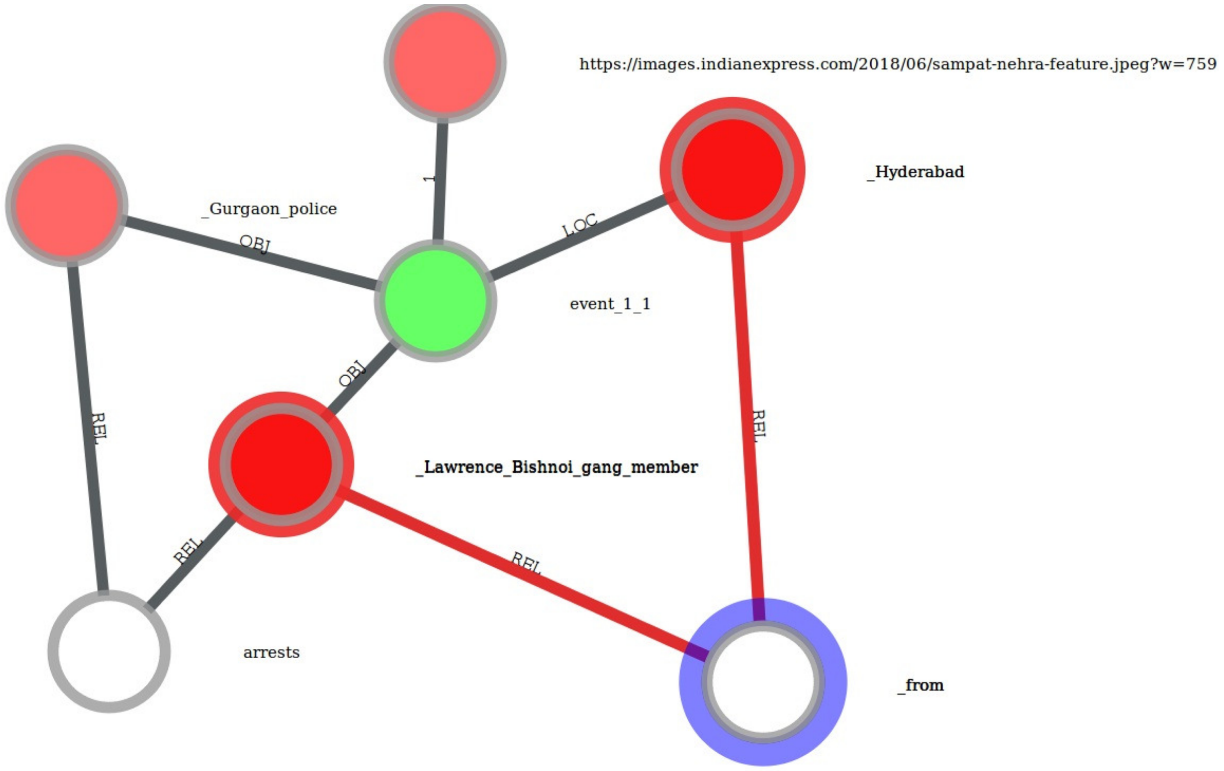

Fig. 10. Knowledge Graph for Event-1.

Here the experiment is also conducted over the knowledge bases of different sizes that consist of events and entities extracted from multiple news articles. Tables 6 and 7 shows the impact of semantic merging using only semantic similarity measure and using both semantic and contextual similarity measure respectively. From the tables, it is clear that the number of events is reduced significantly using contextual similarity in combination with semantic similarity measure. The only reduction in the number of events after semantic merging indicates contextually similar events which is evident from the first two tuples of Table 7.

Although the Table 7 is evident for the reduction in the size of the knowledge base after semantic merging, Fig. 9 shows the existence of redundancy even after semantic merging for varying sizes of the knowledge base. The presence of redundancy even after semantic merging is due to the wrong tags assigned to two or more semantically similar entities.

To better illustrate the integration of data obtained from multiple news articles, two semantically related and similar events

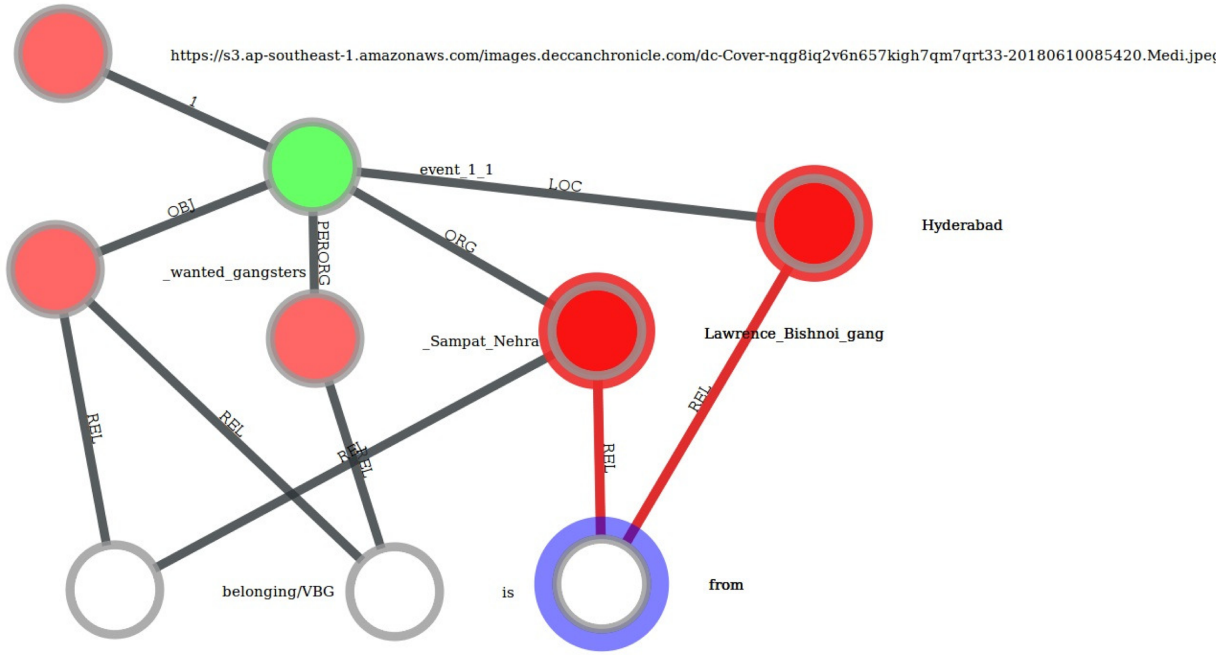

Fig. 11. Knowledge Graph for Event-2.

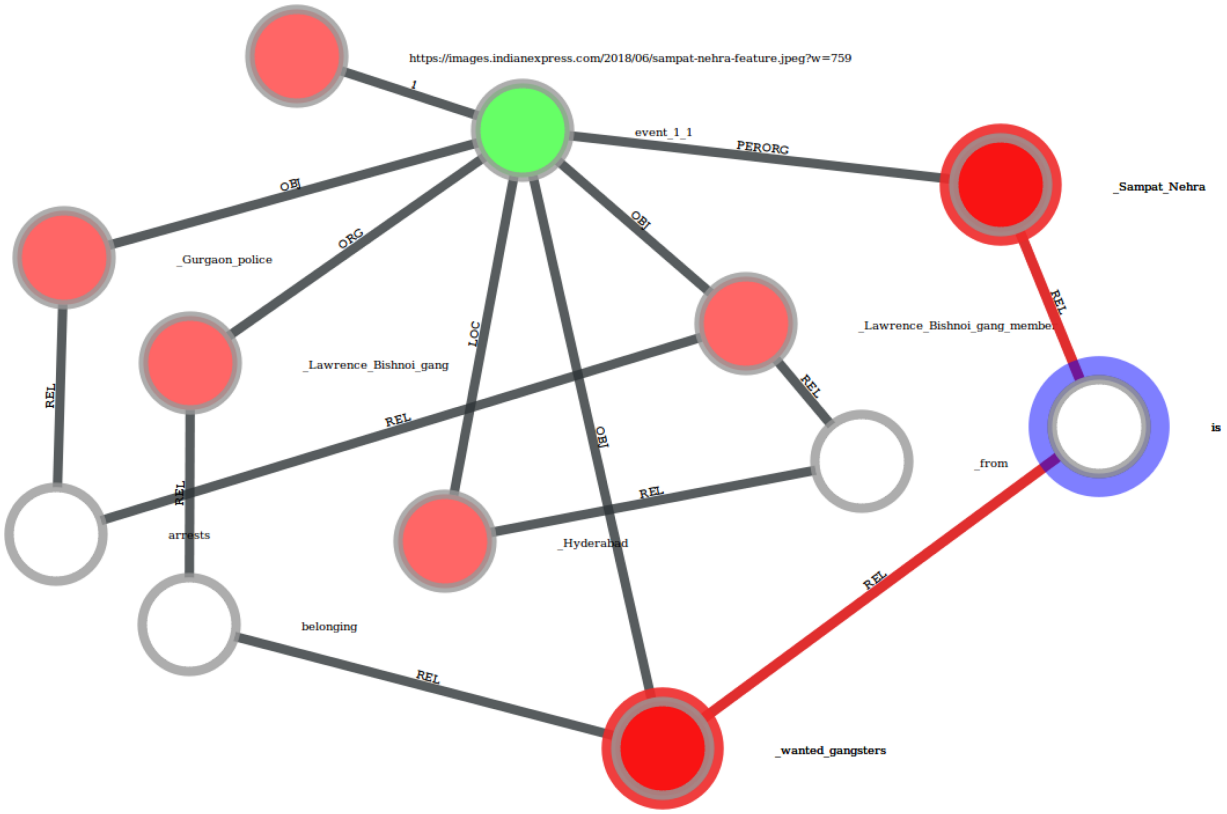

Fig. 12. Knowledge Graph after semantic merging.

captured from newspaper articles URL and URL are considered as follows:

Event-1: ['SUB- Gurgaon police','REL- arrests','OBJ- Lawrence Bishnoi gang member','REL- from', 'LOC- Hyderabad','IMG- https://images.indianexpress.com/2018/06/ sampat-nehra-feature.jpeg?w=759', 'RM-1']

Event-2: ['PERORG- Sampat Nehra','REL-is','SUB- wanted gangsters','REL-belonging','ORG- Lawrence Bishnoi gang','REL-from','LOC-Hyderabad','IMG- https://s3.ap-southeast-1.amazonaws.com/ images.decca nchronicle.com/ dc-Cover-nqg8iq2v6n657-kigh7qm7qrt33- 20180610085420.Medi.jpeg','RM-1']

The knowledge graph generated before and after applying the semantic merging for Event-1 and Event-2 is shown in Figs. 10–12 respectively. Two original event nodes are replaced with a single event node due to their semantic similarness. For entity types LOC and REL, a single instance of entities 'hyderabad' and 'from' are retained in the merged knowledge graph. Similarly, a single instance of the image is retained by checking for the similarity between the low-level features and the high-level captions. Even though 'Lawrence Bishnoi gang member' and 'Lawrence Bishnoi gang' are semantically similar, due to wrong identification of their entity types as OBJECT and ORGANIZATION in Event-1 and Event-2 respectively, they are retained as it is in the merged graph.

6. Conclusions and future works

This paper presents a methodology to create a knowledge base of crime entities extracted and integrated from text and image data of online newspapers in the English language without redundancy and loss of information. Rule-based approach and semantic as-well-as contextual similarity measures are used to identify untagged and wrongly tagged entities to keep the knowledge base complete and free from duplicates. However, wrong tags assigned to more than one semantically similar entities limits its capability to remove redundancy entirely and also the number of named entities recognized by the current system depends on the syntactic rule set. These limitations can be overcome by emphasizing more on semantics using external knowledge bases like DBpedia. Furthermore, completeness of knowledge base can be enhanced by enrichment of the knowledge base with crime entities from other Indian language newspapers. Also, keyword-based or image-based SPARQL query interfaces can be integrated with the current system to retrieve the information from the knowledge base effectively.

References

Aliprandi, C., Arraiza Irujo, J., Cuadros, M., Maier, S., Melero, F., & Raffaelli, M. (2014). Caper: Collaborative information, acquisition, processing, exploitation and reporting for the prevention of organised crime. In C. Stephanidis (Ed.). *HCI international 2014 - posters' extended abstracts* (pp. 147–152). Cham: Springer International Publishing.

Alruily, M., Ayesh, A., & Zedan, H. (2014). Crime profiling for the arabic language using computational linguistic techniques. *Information Processing & Management, 50*(2), 315–341. https://doi.org/10.1016/j.ipm.2013.09.001 http://www.sciencedirect.com/science/article/pii/S0306457313000988

Arulanandam, R., Savarimuthu, B. T. R., & Purvis, M. A. (2014). *Extracting crime information from online newspaper articles. Proceedings of the second australasian web conference - volume 155AWC '14*Darlinghurst, Australia, Australia: Australian Computer Society, Inc31–38 http://dl.acm.org/citation.cfm?id=2667702.2667706

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Buitelaar, P., Cimiano, P., Frank, A., Hartung, M., & Racioppa, S. (2008). Ontology-based information extraction and integration from heterogeneous data sources. *International Journal of Human-Computer Studies, 66*(11), 759–788. https://doi.org/10.1016/j.ijhcs.2008.07.007 http://www.sciencedirect.com/science/article/pii/S1071581908000906

Chau, M., Xu, J. J., & Chen, H. (2002). *Extracting meaningful entities from police narrative reports. Proceedings of the 2002 annual national conference on digital government researchdg.o '02*Digital Government Society of North America1–5 http://dl.acm.org/citation.cfm?id=1123098.1123138

Chaudhuri, S., Ganti, V., & Kaushik, R. (2006). *A primitive operator for similarity joins in data cleaning. Icde*. Institute of Electrical and Electronics Engineers, Inc https://www.microsoft.com/en-us/research/publication/a-primitive-operator-for-similarity-joins-in-/data-cleaning/

Dasgupta, T., Naskar, A., Saha, R., & Dey, L. (2017). *Crimeprofiler: Crime information extraction and visualization from news media. Wi*.

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., et al. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management, 51*(2), 32–49. https://doi.org/10.1016/j.ipm.2014.10.006 http://www.sciencedirect.com/science/article/pii/S0306457314001034

Dhuria, S., Taneja, H., & Taneja, K. (2016). *Nlp and ontology based clustering an integrated approach for optimal information extraction from social web. 2016 3rd international conference on computing for sustainable global development (indiacom)*1765–1770.

Dragos, V. (2013). Developing a core ontology to improve military intelligence analysis. *International Journal of Knowledge-based and Intelligent Engineering Systems, 17*(1), 29–36.

scar Ferrndez, Izquierdo, R., Ferrndez, S., & Vicedo, J. L. (2009). Addressing ontology-based question answering with collections of user queries. *Information Processing & Management, 45*(2), 175–188. https://doi.org/10.1016/j.ipm.2008.09.001 http://www.sciencedirect.com/science/article/pii/S0306457308000861

Furtado, V., Ayres, L., de Oliveira, M., Vasconcelos, E., Caminha, C., DOrleans, J., et al. (2010). Collective intelligence in law enforcement the wikicrimes system. *Information Sciences, 180*(1), 4–17. https://doi.org/10.1016/j.ins.2009.08.004 Special Issue on Collective Intelligence http://www.sciencedirect.com/science/article/pii/S0020025509003454

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review, 29*, 21–43. https://doi.org/10.1016/j.cosrev.2018.06.001 http://www.sciencedirect.com/science/article/pii/S1574013717302782

Gregory, M. L., McGrath, L., Bell, E. B., O'Hara, K., & Domico, K. (2011). *Domain independent knowledge base population from structured and unstructured data sources. Flairs conference*.

Hazrina, S., Sharef, N. M., Ibrahim, H., Murad, M. A. A., & Noah, S. A. M. (2017). Review on the advancements of disambiguation in semantic question answering system. *Information Processing & Management, 53*(1), 52–69. https://doi.org/10.1016/j.ipm.2016.06.006 http://www.sciencedirect.com/science/article/pii/S0306457316302102

Hossein, H., Huang, X., Silva, S. E., & Ghodsi, M.. A review of data mining applications in crime. Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(3), 139–154. 10.1002/sam.11312https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11312.

Jalil, M.. M. A., Ling, C. P., Noor, N. M. M., & Mohd., F. (2017). Knowledge representation model for crime analysis. *Procedia Computer Science, 116*, 484–491. https://doi.org/10.1016/j.procs.2017.10.067 Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017) http://www.sciencedirect.com/science/article/pii/S1877050917321178

Kalemi, E., Yildirim-Yayilgan, S., Domnori, E., & Elezaj, O. (2017). Smont: An ontology for crime solving through social media. *International Journal of Metadata, Semantics and Ontologies, 12*(2–3), 71–81. https://doi.org/10.1504/IJMSO.2017.090756 https://www.inderscienceonline.com/doi/abs/10.1504/IJMSO.2017.090756

Kondrak, G. (2005). *N-gram similarity and distance. Proceedings of the 12th international conference on string processing and information retrievalSPIRE'05*Berlin, Heidelberg: Springer-Verlag115–126. https://doi.org/10.1007/11575832_13.

Ku, C. H., Iriberri, A., & Leroy, G. (2008). *Natural language processing and e-government: Crime information extraction from heterogeneous data sources. Proceedings of the 2008 international conference on digital government researchdg.o '08*Digital Government Society of North America162–170 http://dl.acm.org/citation.cfm?id=1367832.1367862

Lamy J.-B., URL. Visited as on 31-05-2018. https://pythonhosted.org/Owlready/.

Loper, E., & Bird, S. (2002). *Nltk: The natural language toolkit. Proceedings of the ACL-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics - volume 1ETMTNLP '02*Stroudsburg, PA, USA: Association for Computational Linguistics63–70. https://doi.org/10.3115/1118108.1118117.

Miller, D. R. H., Leek, T., & Schwartz, R. M. (1999). *A hidden markov model information retrieval system. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrievalSIGIR '99*New York, NY, USA: ACM214–221. https://doi.org/10.1145/312624.312680.

Miller, F. P., Vandome, A. F., & McBrewster, J. (2009). *Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau?levenshtein distance, spell checker, hamming distance.* Alpha Press.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A., & Goranov, M. (2003). *Towards semantic web information extraction. Human language technologies workshop at the 2nd international semantic web conference (ISWC2003)20*.

Qu, R., Fang, Y., Bai, W., & Jiang, Y. (2018). Computing semantic similarity based on novel models of semantic representation using wikipedia. *Information Processing & Management, 54*(6), 1002–1021. https://doi.org/10.1016/j.ipm.2018.07.002 http://www.sciencedirect.com/science/article/pii/S0306457317309226

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0 http://www.sciencedirect.com/science/article/pii/0306457388900210

Tsuruoka, Y., McNaught, J., Tsujii, J., & Ananiadou, S. (2007). Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics, 23*(20), 2768–2774. https://doi.org/10.1093/bioinformatics/btm393.

URL. Visited as on 10-04-2019. https://www.kaggle.com/therohk/million-headlines.

URL. Visited as on 12-07-2018. http://en.wikipedia.org/wiki/Category:Crimes.

URL. Visited as on 31-05-2018. https://opensource.com/business/15/7/five-open-source-nlp-tools.

URL. Visited as on 31-05-2018. http://indianexpress.com/article/india/gurgaon-police-arrests-key-lawrence-bishnoi-gang-member-sampat-nehra-underworld-gangster-5207714/.

URL. Visited as on 31-05-2018. https://www.deccanchronicle.com/nation/current-affairs/100618/gangster-sampat-nehra-wanted-to-kill-salman-khan-say-police.html.

Wan, S., & Angryk, R. A. (2007). *Measuring semantic similarity using wordnet-based context vectors. 2007 IEEE international conference on systems, man and cybernetics*908–913. https://doi.org/10.1109/ICSMC.2007.4413585.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861.

Xiao, C., Wang, W., & Lin, X. (2008). Ed-join: An efficient algorithm for similarity joins with edit distance constraints. *Proceedings of the VLDB Endowment, 1*(1), 933–944. https://doi.org/10.14778/1453856.1453957.

Zhu, G., & Iglesias, C. A. (2018). Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications, 101*, 8–24. https://doi.org/10.1016/j.eswa.2018.02.011 http://www.sciencedirect.com/science/article/pii/S0957417418300897

Zouaq, A., Gagnon, M., & Jean-Louis, L. (2017). An assessment of open relation extraction systems for the semantic web. *Information Systems, 71*, 228–239. https://doi.org/10.1016/j.is.2017.08.008 http://www.sciencedirect.com/science/article/pii/S0306437916304999