CO449 Major Project Report

Ontology enrichment using social and news media

Submitted in partial fulfillment of
the requirements for the award of the degree of

Bachelor of Technology
in
Computer Science and Engineering

Submitted by

Roll No	Names of Student
14CO103	Ananda Rao H
14CO127	Neha Mohan
14CO142	Sheetal Shalini

Under the guidance of
Dr. P Santhi Thilagam

Department of Computer Science and Engineering
National Institute of Technology Karnataka
Surathkal, Karnataka, India – 575 025

Even Semester 2017-18

Declaration

We hereby declare that the Project Work Report entitled **Ontology enrichment using social and news media** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** for the award of the Degree of **Bachelor of Technology** in Computer Engineering from the Department of Computer Science and Engineering is a *bonafide report of the work carried out by us.* The material contained in this Project Work Report has not been submitted to any University or Institution for the award of any degree.

Roll No	Names of Student	Signature
14CO103	Ananda Rao H	
14CO127	Neha Mohan	
14CO142	Sheetal Shalini	

Department of Computer Science and Engineering

Place :
Date :

Department of Computer Science and Engineering

National Institute of Technology Karnataka

Certificate

This is to certify that this is a bonafide record of the project presented by the students whose names are given below during the Even Semester 2017-18 in partial fulfilment of the requirements of the degree of Bachelor of Technology in Computer Science and Engineering.

Roll No	Names of Student
14CO103	Ananda Rao H
14CO127	Neha Mohan
14CO142	Sheetal Shalini

Dr. P Santhi Thilagam
(Project Guide)

Dr. Annappa
(Chairman - DUGC)

Place :
Date :

Acknowledgement

We hereby gratefully acknowledge the support and inspiration provided by our esteemed guide Dr. P Santhi Thilagam, Department of Computer Science and Engineering, NITK Surathkal, which was instrumental in realizing this major project. We are forever indebted to her for giving us the wonderful opportunity of working under her tutorage and thereby helping us gain immense knowledge and experience. We would like to express our sincere gratitude to her for her thoughtful advice, encouragement, guidance, critics and valuable suggestions throughout the course of our project work. We would also like to thank her for her constant cooperation, support and for providing necessary facilities throughout the BTech programme.

We would like to take this opportunity to thank the teaching and nont-teaching staff in the Department of Computer Science and Engineering for their invaluable help and support.

We also express our gratitude towards our classmates who have been a constant source of motivation and who made our time spent at NITK truly memorable.

Ananda Rao H
Neha Mohan
Sheetal Shalini

Abstract

Ontology is the formal representation of entities by defining their properties or attributes and the interrelationship between the entities. Ontologies so developed can be used by law enforcement agencies to track and prevent crime activities. The online data, including social media sites and newspaper articles, can give away a lot of useful information about trends of crime activities in a particular place, personal information about suspects, etc. In this project, we propose a method to develop a crime ontology by capturing information comprising of text and images from online news articles. We then extend and enrich the ontology using relevant information from well-known social media sites like Facebook, Flickr and Twitter.

Contents

Chapter 1

Introduction

Social media is an ever-growing source of text and multimedia information. The retrieval of this information can be utilized in many applications like crime detection, user behaviour tracking, user income predictions and so on. Retrieval of images and text from social media has been identified as an important research issue from past few years owing to the presence of unstructured data.

A problem associated with social media information is authenticity of information i.e. the information about rumours which includes the false information also. However the news media is an authenticated source of information and the information retrieved from such a media can be used to check the authenticity of the information retrieved from social media. Past researches have shown that ontologies improve the task of information retrieval by providing more concepts semantically related to a given user query. Hence it is necessary to build general purpose/domain specific ontologies in order to improve the accuracy of information retrieval.

In this project, we have extracted textual and visual information from news and social media and depicted it as an ontology. We aim to summarize the information, to make it more accessible and useful. The multitude of data has to be presented in the most optimized way. This leads to the need for a good information extraction and summarizing mechanism. Information Extraction refers to extracting information from structured or semi-structured documents consisting of human language. Extracting key information from several documents over the web and generating a knowledge base with a defined structure requires algorithms that suit the morphology of the language in which the text is written in and comes with challenges like non-uniform structure of documents, availability of less than required information in documents extracted, mix of languages in a single document and several others.

Information extraction is aided by Natural Language Processing (NLP)

which has solved the problem of modelling human language processing with considerable success when taking into account the magnitude of the task. Information Extraction deals with tasks involving Information Retrieval and NLP. In this project, we have implemented Information Extraction and depicted the extracted information with a concept ontology.

Chapter 2

Motivation

The Semantic Web relies heavily on the formal ontologies that structure underlying data for the purpose of comprehensive and transportable machine understanding. Therefore, the success of the Semantic Web depends strongly on the proliferation of ontologies, which requires fast and easy engineering of ontologies and avoidance of a knowledge acquisition bottleneck. Conceptual structures that define an underlying ontology are germane to the idea of machine understandable data on the Semantic Web. Ontologies are (meta)data schema, providing a controlled vocabulary of concepts, each with an explicitly defined and machine processed semantics. By defining shared and common domain theories, ontologies help both people and machines to communicate concisely,supporting the exchange of semantics and not only syntax.

The term *ontology* has a long history in philosophy, in which it refers to the subject of existence. In the context of knowledge management, ontology is referred as the shared understanding of some domains, which is often conceived as a set of entities, relations, functions, axioms and instances. There are several reasons for developing context models based on ontology:

- **Knowledge Sharing**. The use of context ontology enables computational entities such as agents and services in pervasive computing environments to have a common set of concepts about context while interacting with one another.

- **Logic Inference**. Based on ontology, context-aware computing can exploit various existing logic reasoning mechanisms to deduce high-level, conceptual context from low-level, raw context, and to check and solve inconsistent context knowledge due to imperfect sensing.

- **Knowledge Reuse**. By reusing well-defined Web ontologies of different domains (e.g., temporal and spatial ontology), we can compose

large-scale context ontology without starting from scratch.

The essence of Semantic Web is a set of standards for exchanging machine-understandable information. Among these standards, Resource Description Framework (RDF) provides data model specifications and XML-based serialization syntax as well as Web Ontology Language (OWL) which enables the definition of domain ontologies and sharing of domain vocabularies. OWL is modeled through an object-oriented approach, and the structure of a domain is described in terms of classes and properties. From a formal point of view, OWL can be seen to be equivalent to description logic (DL), which allows OWL to exploit the considerable existing body of DL reasoning including class consistency and consumption, and other ontological reasoning. Domain ontologies are formal descriptions of the classes of concepts and the relationships among those concepts that describe an application area.

With the advent of the Internet, huge volumes of data (also called 'big data') are available online. Electronic newspapers are increasingly being read by users from anywhere, anytime. Newspapers are a source of (mostly) authentic and timely information. There is a large amount of information available in newspaper articles. For example, newspaper articles contain information about crimes, accidents, politics, cultural events and sports events. Even though valuable information is available in human-readable form in online newspapers and electronic archives, software systems that can extract relevant information and present these information are scarce and this has been of significant interest to researchers in the field of Information Extraction. Hence, this project aims to fulfill the need for information extraction and summarizing technologies by creating a concept ontology of the information gathered from online newspaper articles and social media.

Chapter 3

Literature Survey

A fair amount of research has been done in the area of information extraction. The Web Ontology Language (OWL) is a computational logic-based language such that knowledge expressed in OWL can be exploited by computer programs. It has been used to create the Crime Ontology. As mentioned in the research work on Semantic Ontology Alignment [1], the similarity between two OWL entities Sim_X defined between two nodes of category X of an OWL graph follows two principles:

- It depends on the considered category
- It takes into account all the features F of this category (e.g., properties).

The entity pair whose similarity is under assessment is known as the anchor pair of the comparison and all the pairs which contribute individually to calculate the total similarity are known as the contributors. Aggregation of all contributor similarities is obtained through a weighted sum, which helps controlling the contribution of each feature.

A wide range of string matching metrics, along with string pre-processing strategies such as removing stop words and considering synonyms on different types of ontologies have been evaluated [5]. A set of guidelines on when to use which metric and to show that optimal string similarity metrics can alone produce alignments that are competitive with the state of the art approaches in ontology alignment systems are presented. The review on text similarity techniques [6] discusses several approaches for finding similarities of words like lexical similarity, semantic similarity and so on. Knowledge based similarity [8] is a semantic based similarity that identifies the degree of similarity between words using information derived from semantic networks. The popular semantic networks are "Word Net" and Natural language Toolkit (NLTK) to measure the knowledge based similarity between words. Knowledge based similarity also provides similarity on the basis of word relatedness.

An object recognition methodology [7] that uses a new class of local image features has been introduced. The features are invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection. These features share similar properties with neurons in inferior temporal cortex that are used for object recognition in primate vision. Features are efficiently detected through a staged filtering approach that identifies stable points in scale space. Image keys are created that allow for local geometric deformations by representing blurred image gradients in multiple orientation planes and at multiple scales. The keys are used as input to a nearest neighbor indexing method that identifies candidate object matches. Final verification of each match is achieved by finding a low residual least squares solution for the unknown model parameters. SURF [3] is one such scale-invariant feature extraction method. In SURF, emphasis is given on scale and rotation invariant images by developing a detector and descriptor for them. These descriptors are deemed robust enough to deal with second order effects like skewness, anisotropic scaling and perspective effects.

Chapter 4

Proposed Solution

We have created a crime ontology comprising of text and images from online newspaper articles and social media. The extracted information from newspaper articles forms the base ontology. Relevant information from social media has been used to enrich the ontology. In order to create the enhanced ontology, the three phases have been implemented. Figure 4.1 shows the flow diagram of the proposed solution.

1. Creating Newspaper Ontology
2. Creating Social Media Ontology
3. Creating Combined Ontology

4.1 Creating Newspaper Ontology

Newspaper ontology is created by scraping text and image data from online news articles. The data has been pre-processed and tokenized to extract the tags which are to be represented in the ontology. The following steps have been implemented to get the Newspaper Ontology.

4.1.1 Data Scraping

Online news articles have been scraped to extract textual and visual information. A summary of each article is created, and this summary is then further processed. A data scraping tool called *Newspaper3k* has been used for this purpose. Newspaper3k is a python library for extracting curating

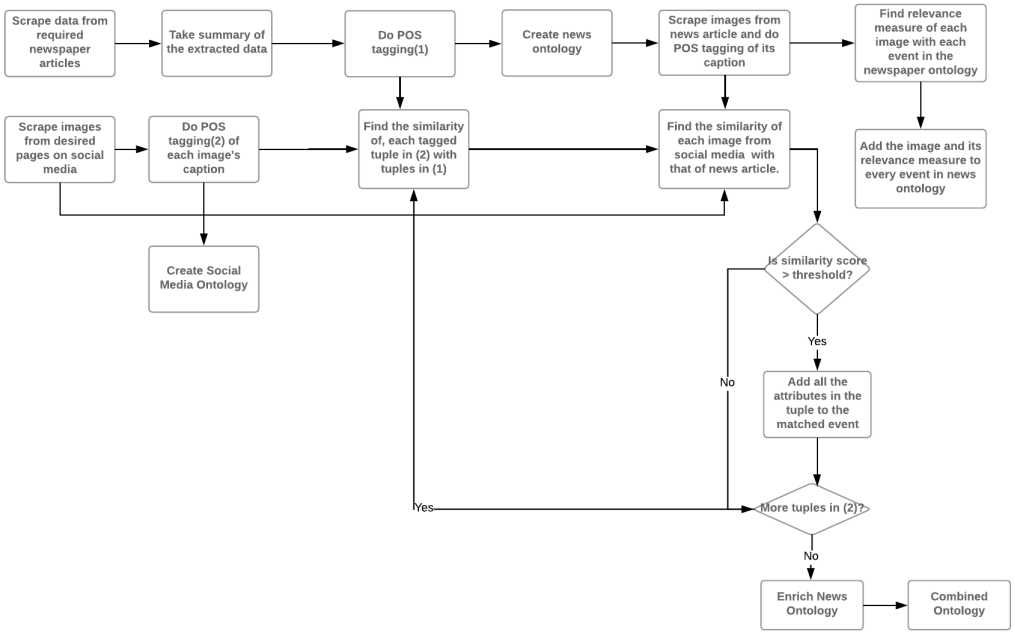

Figure 4.1: Flow Diagram of the system

articles. It supports languages like Arabic, Russian, Dutch, German, English, Spanish, French, Hebrew, Italian, Korean, Norwegian, Persian, Polish, Portuguese, Swedish, Hungarian, Finnish, Danish, Chinese, Indonesian, Vietnamese, Swahili, Turkish, Greek and Ukrainian. It has seamless language extraction and detection. If no language is specified, Newspaper3k will attempt to auto detect a language. Basic features of Newspaper3k are:

- Multi-threaded article download framework
- News url identification
- Text extraction from html
- Top image extraction from html
- All image extraction from html
- Keyword extraction from text
- Author extraction from text
- Google trending terms extraction

It also performs Automatic Summarization on a given article. Automatic summarization is a term which refers to the extraction of gist of a document with the help of a software. The basic idea is to create a subset of a set of data extracted, which include most informative sentences.

4.1.2 Event Tokenization

Each sentence in the summary extracted by Data Scraping is regarded as an event in the ontology. These events are incorporated into the ontology using a unique token which identifies them. The event is named as *event_articlenumber_sentencenumber*. Each sentence in the summary undergoes Part-Of-Speech (POS) tagging. For POS tagging, we use Natural Language Toolkit (NLTK). NLTK is a python framework, used for implementing Natural Language Processing in Python programs. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, and wrappers for industrial-strength NLP libraries.

4.1.3 Entity Extraction

For each *event* in the summary, 5 types of entities are extracted.

- **Image** : This includes the link of the top image of the article, linked with each sentence with a relevance measure depending on the similarity of the image caption with the sentence. The image link has been given a label of IMG and included as an entity.
- **LOC (Location)** : Any location/place mentioned in the sentence. It is straightforward that location of an event must be associated with it. Further, it helps to capture crime pattern in a particular place. It is extracted by Named Entity Recognition (NER).
- **PER (Person)** : Any person(s) in the predicate part of the sentence. It is extracted by Named Entity Recognition (NER).
- **ORG (Organization)** : Any Organisation(s) in the predicate part of the sentence. It is extracted by Named Entity Recognition (NER).
- **REL (Relation)** : Any verb(s) or action words in the predicate part of the sentence. It is extracted by POS tagging.

LOC, PER and ORG are extracted by Named Entity Recognition (NER). REL (Relations) are used to link 2 other entities, thereby forming a hierarchical ontology.

4.1.4 Relevance Measure Calculation

Relevance measure(RM) is a measure of the similarity of the image with an event. If there is no caption for any image in the article, the article's heading/title will be processed the same way the captions are processed. The similarity is calculated using synsets (sets of synonyms) in WordNet, which gives a score based on the semantic similarity of different words in the image caption and sentence.

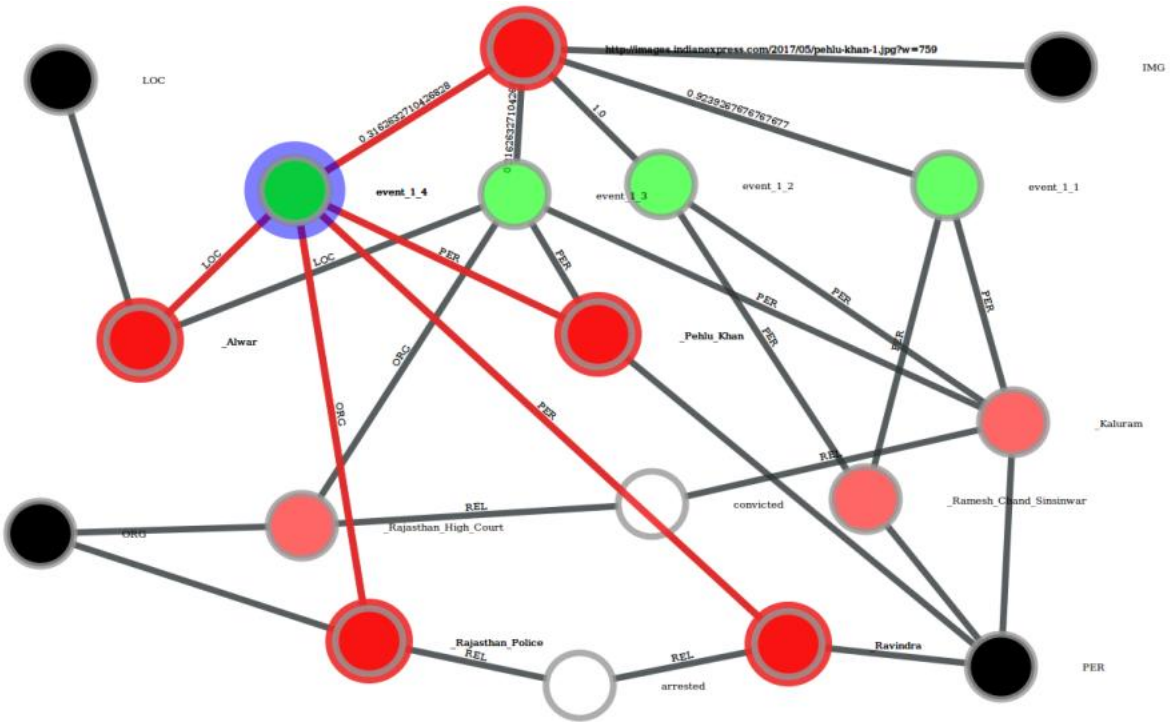

Figure 4.2: Newspaper Ontology

4.1.5 Ontology Construction

In the ontology as shown in Figure 4.2, entities and images are linked with each of the events of the article with the above calculated $RM(Image - Event)$. The green nodes represent the events, pink / orange nodes represent the entities, black nodes represent the concepts (entity type) and white nodes

represent the relations. All the entities are linked to their entity type. This is how the ontology is formed by scraping newspaper articles and extracting text and images.

4.2 Creating Social Media Ontology

Social Media ontology is created by scraping text and image data from Facebook, Flickr and Twitter. The data has been pre-processed and tokenized to extract the tags which are to be represented in the ontology. The following steps have been implemented to get the Social Media Ontology.

4.2.1 Choice of Social Media

The social media have been chosen based on factors like relevance, amount of information and popularity. The following have been considered.

- **Facebook** : Facebook has been chosen because it is a trustworthy source of information. The news and crime related pages on Facebook have relevant and ample data. This data is scraped using Facebook Graph API.
- **Flickr** : Images in Flickr are grouped based on labels. For instance, all the crime related images will belong to the groups labeled as crime, murder, news, blood and so on. This data is scraped using Flickr API.
- **Twitter** : Twitter has a good presence of politicians and news channels in India. The information from these pages are scraped using Tweepy API.

4.2.2 Social Media Information Extraction

For each page of social media that is scraped, for each image in it, if there is a caption for the image, then we process this caption and extract entities from it. Each caption now serves as an event in the Social Media Ontology. This event is linked with the extracted entities and images. No redundant events or entities are created. Before creating a new entity, the already existing ontology is checked and if this entity prevails, then that itself is used without creating another one.

4.2.3 Event Tokenization

Each caption of the image is regarded as an event in the ontology. These events are tokenized in the same way as that of Newspaper ontology events.

4.2.4 Entity Extraction

For each *event* in the ontology, entities are extracted in the same way as that of Newspaper ontology entities.

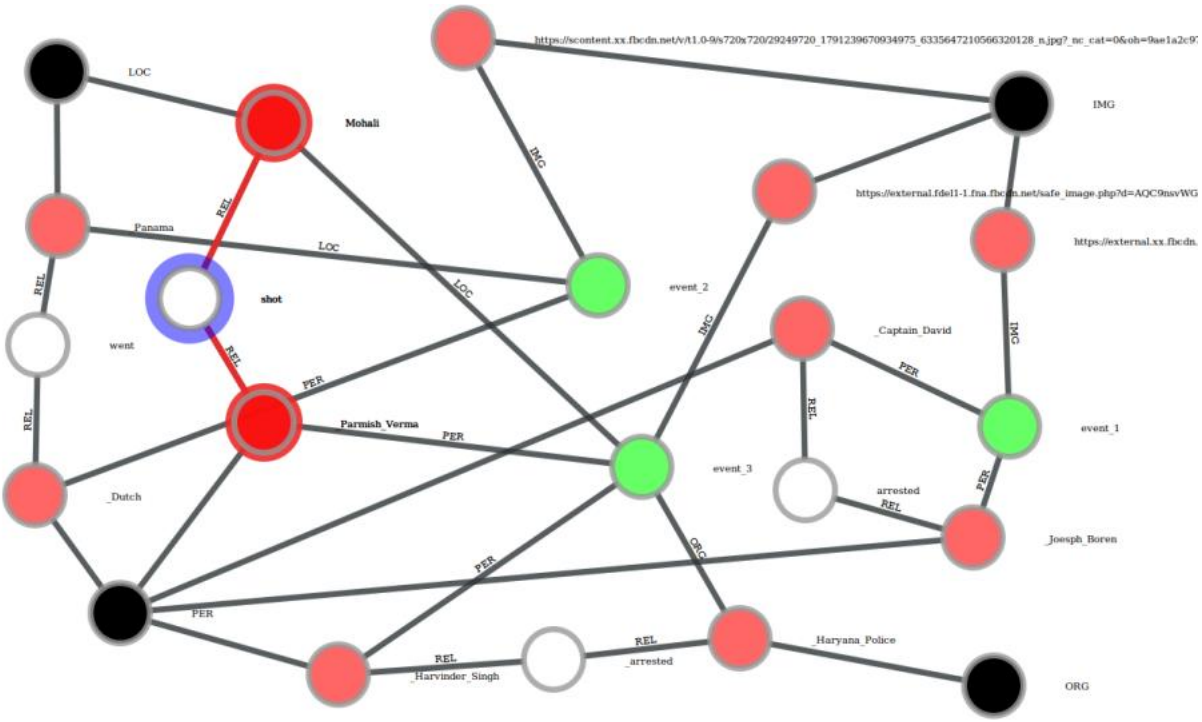

Figure 4.3: Social Media Ontology

4.2.5 Ontology Construction

In the ontology as shown in Figure 4.3, entities and images are linked with each of the events of the post. All the notations are the same as that of the Newspaper Ontology. This is how the ontology is formed by scraping social media pages and extracting text and images.

4.3 Creating Combined Ontology

The Combined Ontology consists of newspaper ontology along with social media ontology with events relevant to any of the events of newspaper ontology. We decide if a particular event of social media ontology is relevant or not using the methodology similar to that suggested in the research work on Ontology Alignment [1]. We compare each tuple obtained after performing POS tagging on scraped data from social media with those from newspaper articles to calculate similarity score. For this purpose, weights have been assigned to each entity type.

$$w_{relation} = 0.1$$

$$w_{person} = 0.25$$

$$w_{location} = 0.1$$

$$w_{organisation} = 0.25$$

$$w_{image} = 0.3$$

Let w_i represent the weight of the entity i and Sim_i be its similarity function. We calculate the total similarity score for a particular event of social media ontology as

$$Sim_t = \sum Sim_i * w_i$$

which sums up over all the entities linked to that event and where $\sum w_i = 1$. We compare Sim_t against a threshold value. If

$$Sim_t > threshold$$

the event must be added to the combined ontology. $Threshold = 0.5$ has been set using trial and error method.

4.3.1 Similarity Functions

The similarity functions for each of the attributes depend upon the type of the attribute. The attributes person, location and organisation needs to have similarity of the strings(since they are noun), while relation must be similar semantically(since it's a verb). The similarity of the images are calculated using feature matching. These are binary functions i.e., they return 1 if the attributes match and 0 otherwise.

4.3.2 String Similarity

We calculate the similarity of the strings using cosine similarity [2]. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0 degree is 1, and it is less than 1 for any other angle.
Calculating cosine similarity between two strings:

- Identify all distinct letters in both words.
- Identify the frequency of occurrences of these letters in both the words and treat them as vectors, say a and b.
- Apply cosine similarity function, i.e.,

$$Cos\theta = (a.b)/(\sqrt{a^2} * \sqrt{b^2})$$

If the value of $Cos\theta > 0.7$, we conclude that the strings are similar or the same, and hence the similarity function returns 1.

4.3.3 Semantic similarity

We calculate the semantic similarity of two attributes using WuPalmer Algorithm [4]. The Wu & Palmer algorithm calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer),using the formula,

$$similarityscore = (2 * Depth(lcs))/(depth(s1) + depth(s2))$$

This means that 0 < similarity score <= 1. The score can never be zero because the depth of the LCS is never zero (the depth of the root of a taxonomy is one). The score is one if the two input concepts are the same.
If similarity score >= 0.6, we the words match semantically.

4.3.4 Image Matching

This function calculates the similarity between two images using a supervised Machine Learning Algorithm[9]. First we find feature points in both the images using SURF(Speeded-Up Robust Features)[3]. SURF approximates Laplacian of Gaussian with Box Filter. It extracts some unique keypoints and descriptors from an image. A set of SURF keypoints and descriptors can be extracted from an image and then used later to detect the same image. SURF uses an intermediate image representation called Integral Image,

which is computed from the input image and is used to speed up the calculations in any rectangular area. It is formed by summing up the pixel values of the x and y co-ordinates from origin to the end of the image. This makes computation time invariant to change in size and is particularly useful while encountering large images. The SURF detector is based on the determinant of the Hessian matrix. The SURF descriptor describes how pixel intensities are distributed within a scale dependent neighbourhood of each interest point detected by Fast Hessian.
We then find the best matches between the features with FLANN (Fast Library for Approximate Nearest Neighbours). The FLANN library contains a collection of algorithms optimized for fast nearest neighbour search in large datasets and for high dimensional features. We use k-nearest neighbour algorithm to find k best matches with k=2. Finally we check if there exists at least 10 good matches using Lowe's ratio test [4], if yes, then the images match.

4.3.5 Ontology Construction

The general structure of the ontology remains same as newspaper ontology. The combined ontology contains all the instances of newspaper ontology, with new attributes added from matched instances of social media ontology, see Figure 4.4. The yellow nodes represent the matched attributes, blue nodes represent the attributes from social media ontology, red nodes represent the attributes from newspaper ontology, white nodes represent the relation attribute, and black nodes represent the concepts.

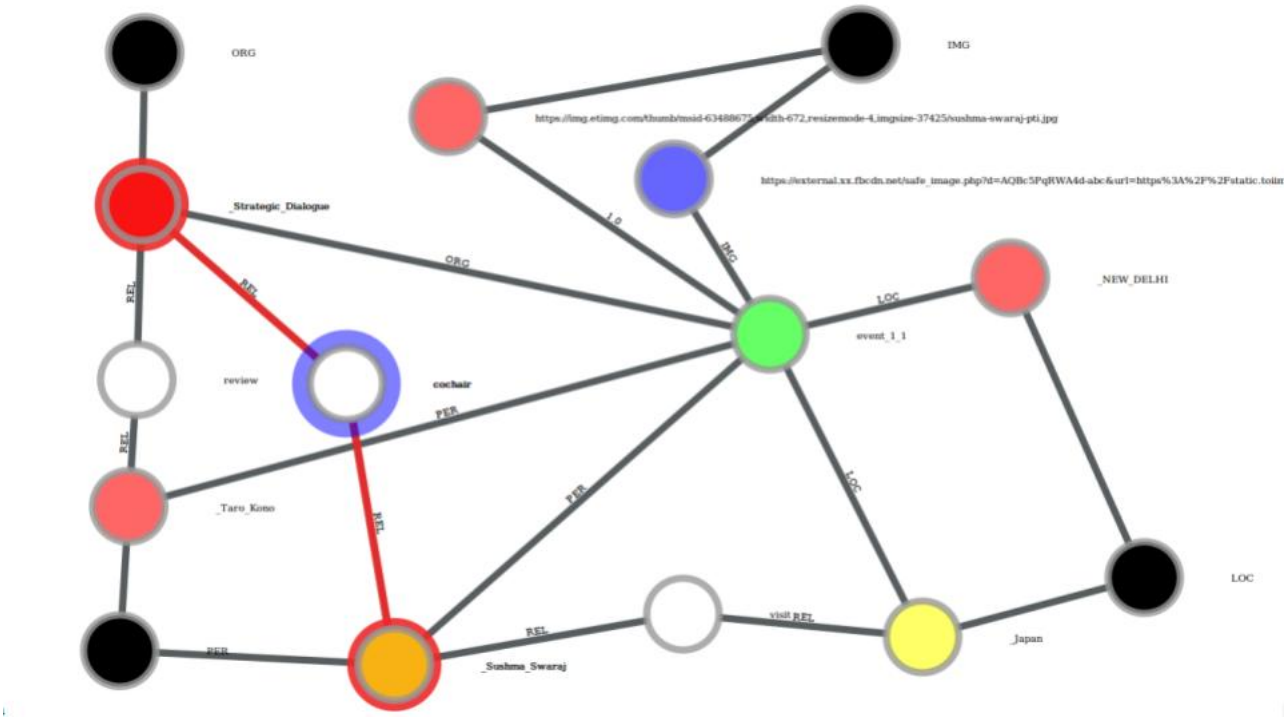

Figure 4.4: Combined Ontology

Chapter 5

Conclusion

A method to extract information from online news articles and social media pages and present them in an interactive form has been implemented in the project. We start by introducing various important concepts associated with information extraction and ontology generation. We then propose a method to meet our objective. The brief design of our approach is as follows.

- Collecting data to be parsed
- Extraction of entities and relationships
- Visualization of the information

Each of these major stages involve multiple steps, which we explain in the respective sections. This would be immensely useful for law enforcement and intelligence agencies to detect and prevent online radicalization, civil unrest and other antisocial activities.

Thus in this project we crawl through the multitude of information available on the Internet and propose a method to develop an ontology of entities, images and events that connect these entities, thus providing an aggregated, birds-eye-view of the information present across multiple sources.

References

[1] , Jérôme Euzenat, Petko Valtchev *Similarity-based ontology alignment in OWL-Lite*, Proc. 16th european conference on artificial intelligence (ECAI), Aug 2004, Valencia, Spain. IOS press, pp.333-337, 2004, Proc. 16th european conference on artificial intelligence (ECAI). <hal-00918127>

[2] Gang Qian, Shamik Sural, Yuelong Gu, Sakti Pramanik, *Similarity between euclidean and cosine angle distance for nearest neighbor queries*, Proceedings of ACM Symposium on Applied Computing, 2004.

[3] Bay H., Tuytelaars T., Van Gool L. (2006) *SURF: Speeded Up Robust Features* In: Leonardis A., Bischof H., Pinz A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3951. Springer, Berlin, Heidelberg.

[4] Z. Wu and M. Palmer., *Verb semantics and lexical selection.* In Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics, pp 133-138. 1994.

[5] Michelle Cheatham, Pascal Hitzler, *String Similarity Metrics for Ontology Alignment*, H. Alani et al. (Eds.),ISWC 2013.

[6] Nitesh Pradhan, Manasi Gyanchandani, Rajesh Wadhvani, *A Review on Text Similarity Technique used in IR and its Application.* International Journal of Computer Applications (0975 – 8887), June 2015.

[7] D.G. Lowe, *Object recognition from local scale-invariant features.* The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999.

[8] Budanitsky, Hirst, *Semantic distance in Word-Net: An experimental application-oriented evaluation of five measures.* In Proceedings of the NAACL Workshop on Word-Net and Other Lexical Resources, 2001.

[9] Upendra Singh,Sidhant Shekhar Singh, Manish Kumar Srivastava, *Object Detection and Localization Using SURF Supported By K-NN*, International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 2, Mar-Apr 2015.