Crime Ontology Enrichment using social and news media

Introduction

The final Crime Ontology has been formed based on the following 3 ontologies :

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

Newspaper Ontology

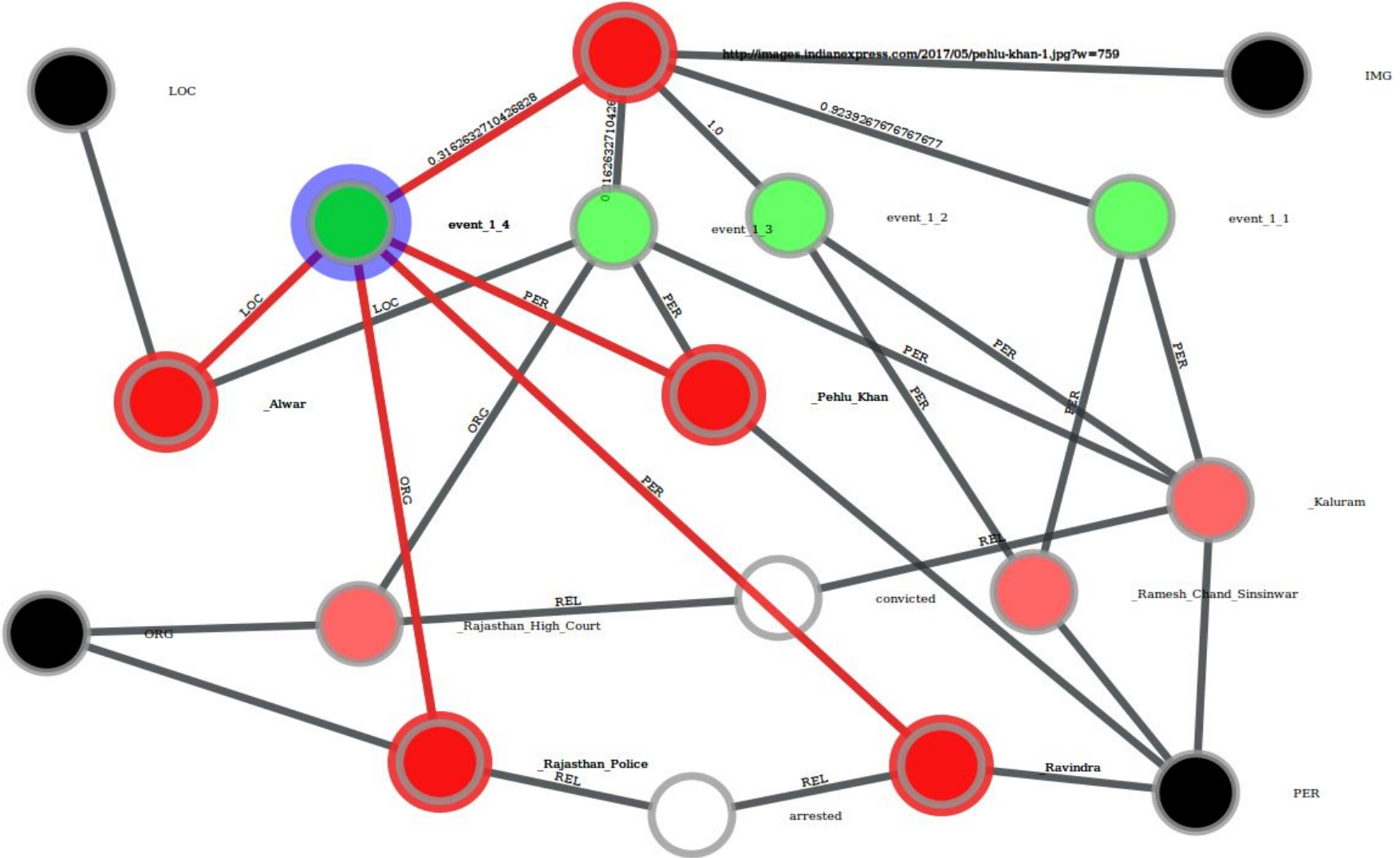
- News articles were scraped from some news websites.
- These articles were parsed to obtain entities from it.
- The images in the article was connected to the event with a particular relevance measure.
- Relevance measure was calculated using synsets.

Scrape Newspaper Text & Images

- Crawl online news articles. (Times of India, Indian Express)
- Extract summary of each article using newspaper API
- Each sentence of the summary of an article corresponds to an event in the ontology.
- The events can be named as Event_articleNumber_sentenceNumber.
- For each event, we will have 6 entities.
    a. Action
    b. Actor(s)
    c. Location
    d. Time
    e. PER
    f. ORG

Relevance Measure

- In this way, the ontology now has text information about 1 article.
- Suppose the article has an image associated with it.
- We process the caption of the image, and extract entities from it.
- We will then compare these entities with the entities of each event and get a relevance measure for each (Image -- Event) Link in the ontology.
- Relevance measure(RM) is a measure of the similarity of the image with that event/tag.
- RM is calculated using synsets from Wordnet.
- If there is no caption for any image in the article, the article's heading/title can be processed the same way the captions are processed.


http://images.indianexpress.com/2017/05/pehlu-khan-1.jpg?w=759
IMG
LOC
0.3162632710426828
0.3162632710426
1.0
0.9239267676767677
event_1_4
event_1_3
event_1_2
event_1_1
LOC
LOC
PER
PER
PER
PER
PER
PER
PER
_Alwar
_Pehlu_Khan
_Kaluram
ORG
ORG
PER
REL
convicted
_Ramesh_Chand_Sinsinwar
ORG
REL
_Rajasthan_High_Court
_Rajasthan_Police
REL
REL
Ravindra
arrested
PER

Social Media Ontology

- Images and their respective captions were extracted from Facebook, Twitter and Flickr.
- The captions were processed to extract the entities from it.
- Each caption was then regarded as an event in the Social Media Ontology and its entities were represented by the entities extracted from it along with the image link.

Facebook Data Scraping

- Pages scraped : 'TimesofIndia', 'TOIIndianews', 'ManayunkTrueCrimeBookClub', 'crimefeed', 'crimetoday.tv'.
- Scraped based on Relevance of crime posts.
- Tool : Facebook Graph API, NLTK (preprocessing the caption)

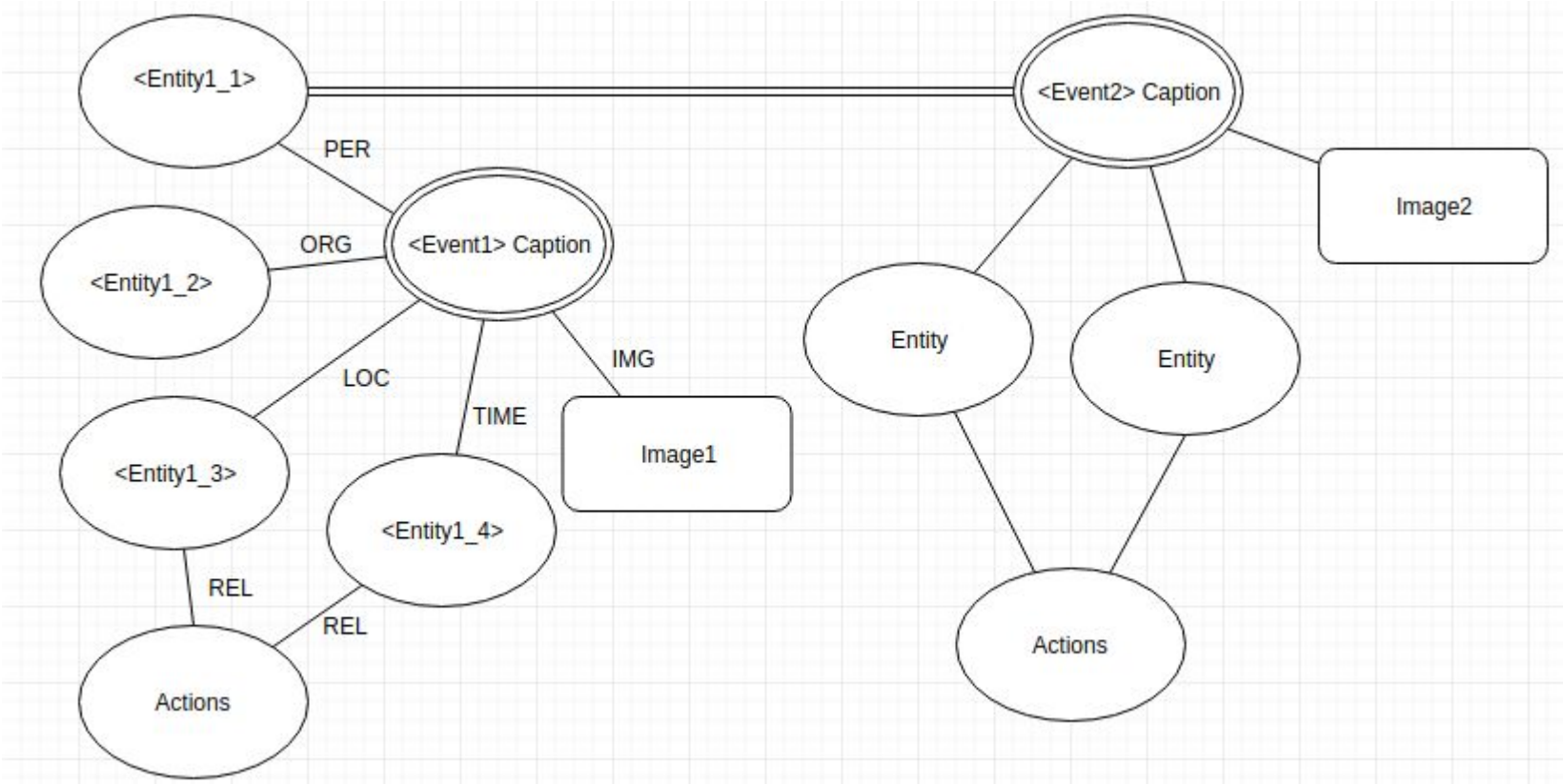
Twitter Data Scraping

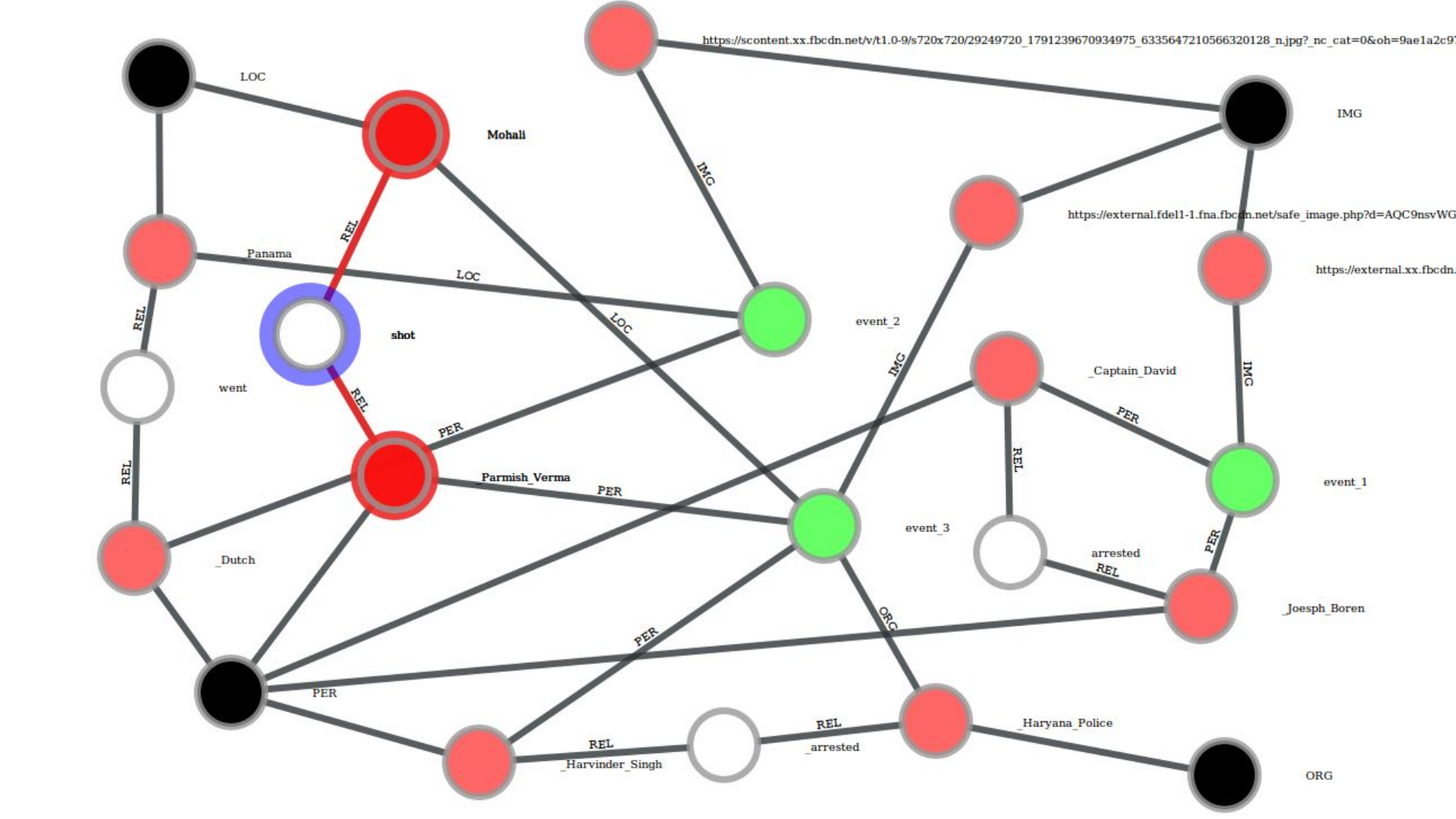
- Pages scraped : 'timesofindia', 'WeirdCrimeFacts', 'CyberCrimeNEWS', 'DNewsCrimeTeam', 'MadisonCrime', 'IowaMurderNews'.
- Scraped based on relevance of crime tweets.
- Tools : Tweepy API, NLTK (preprocessing the caption)

Flickr Data Scraping

- Image groups scraped : "Dainik Vijay News" : '2339229@N21', "Crime Scene Photographers" : '49606819@N00'
- Scraped based on relevance of crime images.
- Tools : flickrapi, NLTK (preprocessing the caption)

Structure of Social Media Ontology




https://scontent.xx.fbcdn.net/v/t1.0-9/s720x720/29249720_1791239670934975_6335647210566320128_n.jpg?_nc_cat=0&oh=9ae1a2c9
IMG
LOC
Mohali
IMG
https://external.fdel1-1.fna.fbcdn.net/safe_image.php?d=AQC9nsvWG
https://external.xx.fbcdn.
Panama
REL
LOC
REL
LOC
shot
event_2
IMG
_Captain_David
IMG
went
REL
PER
REL
PER
Parmish_Verma
PER
event_1
REL
event_3
PER
_Dutch
arrested
REL
_Joesph_Boren
ORG
PER
PER
REL
_arrested
REL
_Haryana_Police
_Harvinder_Singh
ORG

Combined Ontology

- Newspaper ontology along with social media ontology with events relevant to any of the events of newspaper ontology.
- We compare each tuple of data scraped from social media and compare it with those from newspaper to calculate similarity score.
- The total similarity score,

$$\text{Total Similarity score} = \Sigma\ (Sim_i * Weight_i)$$

where i represents an attribute and $\Sigma Weight_i = 1$

- If Total Similarity Score is greater than the specified threshold, then we add the event to our combined ontology

Similarity Functions

- Sim_i is a binary function which checks if the attributes match. The type of similarity function we use depends on the attributes.
- The attributes, person, location and organisation needs to have similarity of the strings, since they are noun.
- The attribute, action is a verb. Hence we find the semantic similarity.
- The similarity of the images are calculated using feature matching.

String Similarity

- We calculate the similarity of the strings using cosine similarity.
- Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.
- cosθ= (a.b)/(√(a^2) √(b^2))
- If the value of cosθ>0.7, we conclude that the strings are similar.

Semantic Similarity

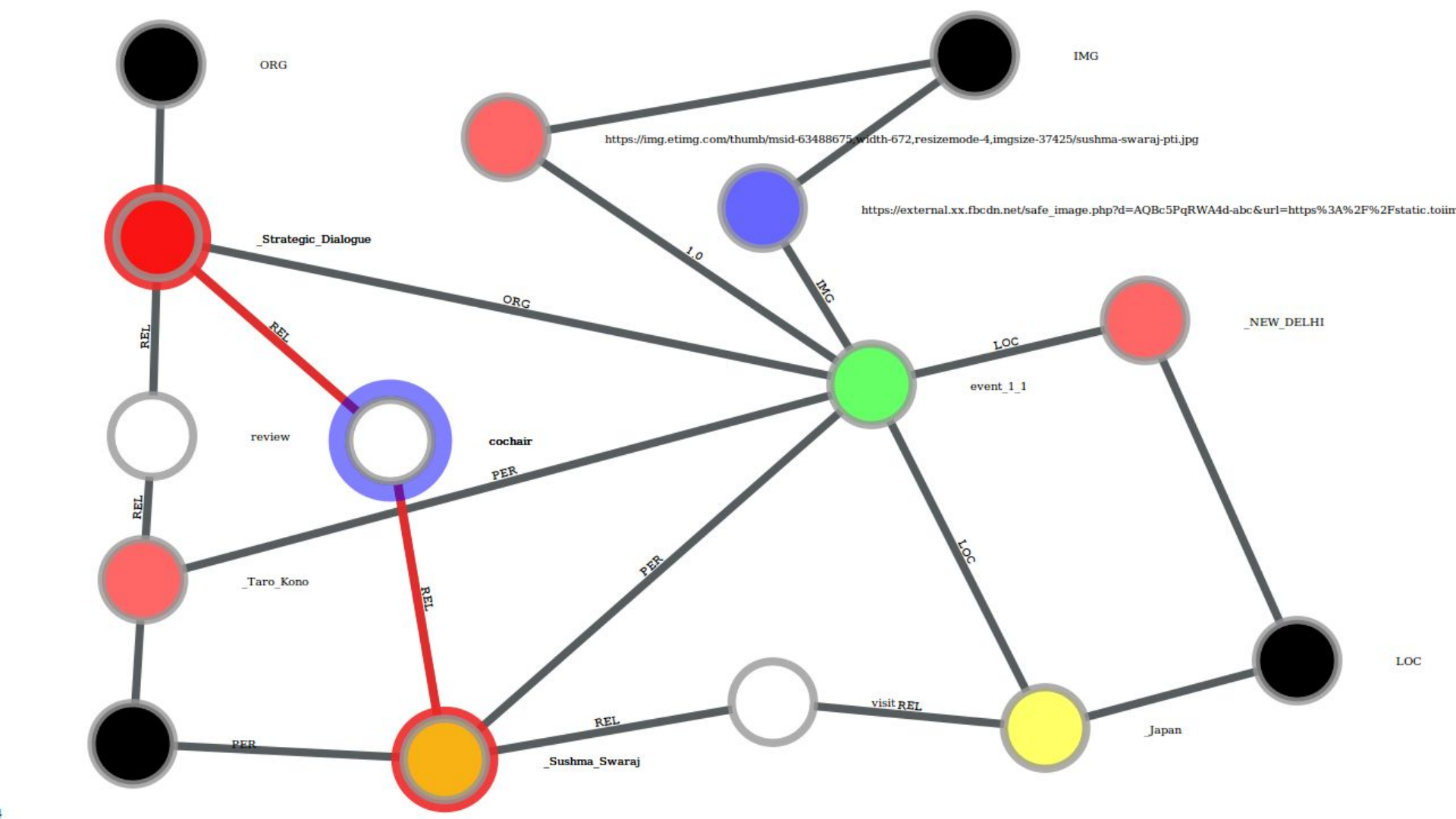
- We calculate the semantic similarity of two attributes using WuPalmer Algorithm.
- calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer), using the formula,

$$\text{similarity score} = (2 * \text{Depth}(lcs)) / (\text{depth}(s1) + \text{depth}(s2))$$

- The score is one if the two input concepts are the same.

Image Similarity using feature matching

1. We find feature points in both the images using SURF(Speeded-Up Robust Features).
2. We then find the best matches between the features with FLANN (Fast Library for Approximate Nearest Neighbours).
3. We use k-nearest neighbour algorithm to find k best matches with k=2.
4. Finally we check if there exists at least 10 good matches using Lowe's ratio test, if yes, then the images match.


ORG
IMG
https://img.etimg.com/thumb/msid-63488675,width-672,resizemode-4,imgsize-37425/sushma-swaraj-pti.jpg
https://external.xx.fbcdn.net/safe_image.php?d=AQBc5PqRWA4d-abc&url=https%3A%2F%2Fstatic.toiim
_Strategic_Dialogue
1.0
IMG
ORG
REL
REL
_NEW_DELHI
LOC
event_1_1
review
cochair
PER
REL
_Taro_Kono
REL
PER
LOC
LOC
visit
REL
REL
PER
_Sushma_Swaraj
_Japan