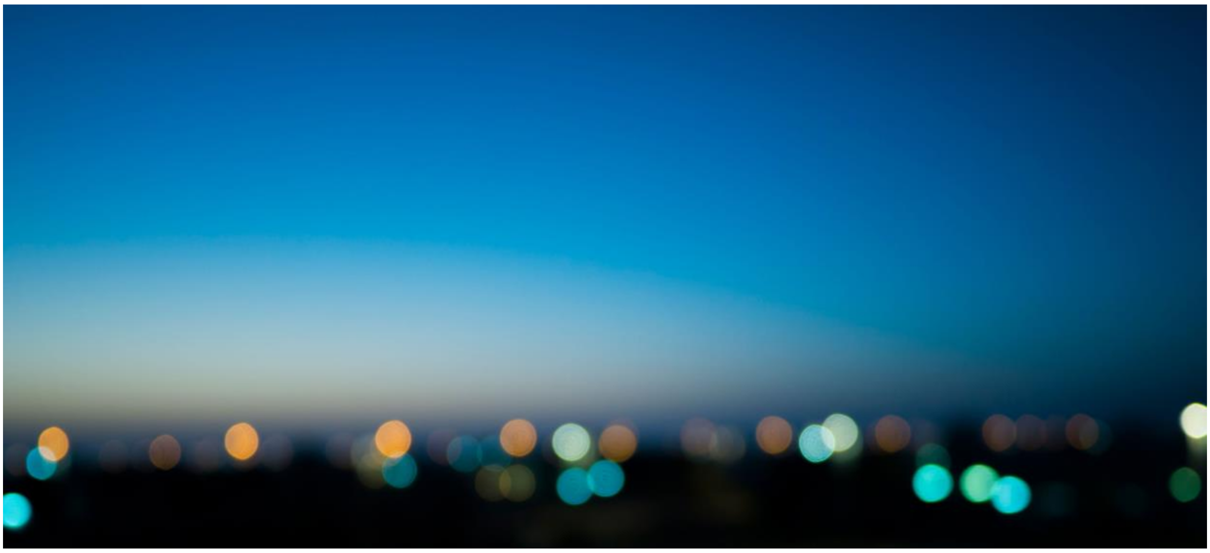


Distributed Computing Project Report

G. Savitha, 20150109 Instructor name: Dr. Rajendra Prasath

Research Topic: Reasoning with Knowledge

Paper 1: Scalable Distributed Semantic Network for knowledge management

in cyber physical system

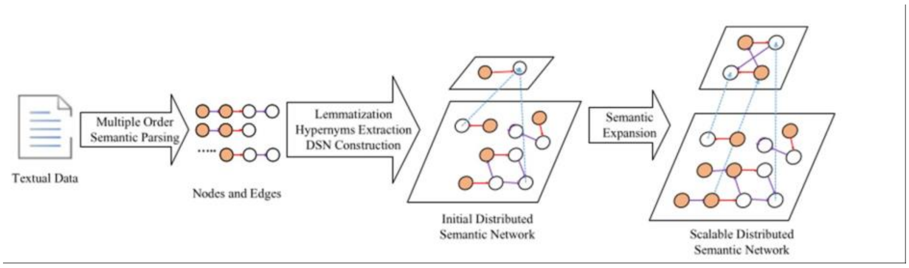
Authors: Shengli Song, Yishuai Lin, Bin Guo,Qiang Di , Rong Lv

Year: 2017

Journal: Journal of Parallel and Distributed Computing

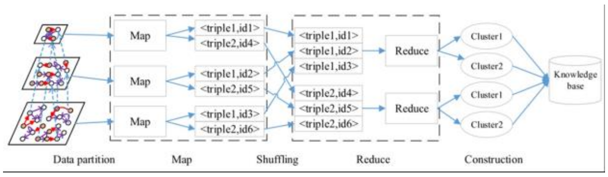
The accelerated development of emerging technologies in the internet and the digital physical frameworks yield enormous mass of information sources. Data sources can contain complementary or semantically equivalent information stored under different

Formats. This can influence semantics and meaning. This paper proposes a new scalable model, named **Distributed Semantic Network (DSN)**, for heterogeneous data representation and can extract more semantic information from different data sources.


Textual Data
Multiple Order Semantic Parsing
Nodes and Edges
Lemmatization Hypernyms Extraction DSN Construction
Initial Distributed Semantic Network
Semantic Expansion
Scalable Distributed Semantic Network

1 - Fig 1: Construction of Distributed Semantic Network

The paper describes the construction of multilevel distributed semantic network. Once network is constructed, each layer of the network is given as input to the Hadoop mapper, which extracts triplets, and assigns it a unique key. The reducer shuffles and removes and the duplicates. After obtaining the result from reducer, clustering is performed and the semantic similarity of the generated triplets from the existing knowledge base is calculated. If the triplet is is significantly different from triplets in knowledge base, the knowledge base is updated otherwise the triplet is discarded.

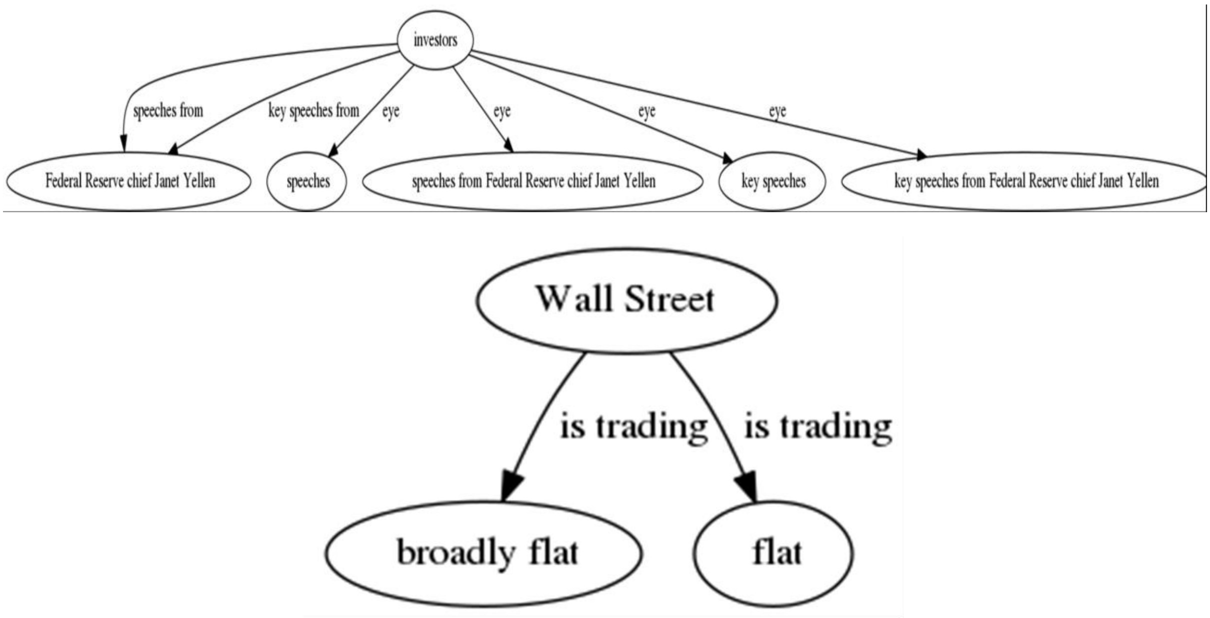


Implementation:

1. Multiple Order Semantic Parsing

Extraction of NP-VP-NP triplets using Stanford Open Information Extraction [1]

Example Sentence: Wall Street is trading broadly flat as investors eye some key speeches from Federal Reserve chief Janet Yellen and other central bankers.


investors
speeches from
key speeches from
eye
eye
eye
eye
Federal Reserve chief Janet Yellen
speeches
speeches from Federal Reserve chief Janet Yellen
key speeches
key speeches from Federal Reserve chief Janet Yellen
Wall Street
is trading
is trading
broadly flat
flat

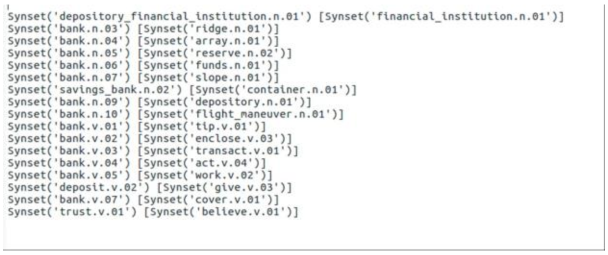
2 - Fig 2: Graph denoting triplets genereated through Multiple Order Parsing.

1. Find Hypernyms to construct second layer of the graph.

Finding hypernym of a given word, is well known research problem falling under word disambiguation. The scope of the paper does not address this. For this project provided if hypernyms exist, the first hypernym is taken.

To find hypernyms, WordNet[2] is utilised.

For a given words it's sysnsets are generated. A particular synset disambiguates the term. For the first synset generated, hypernym is found and stored.



```
Synset('depository_financial_institution.n.01') [Synset('financial_institution.n.01')]
Synset('bank.n.03') [Synset('ridge.n.01')]
Synset('bank.n.04') [Synset('array.n.01')]
Synset('bank.n.05') [Synset('reserve.n.02')]
Synset('bank.n.06') [Synset('funds.n.01')]
Synset('bank.n.07') [Synset('slope.n.01')]
Synset('savings_bank.n.02') [Synset('container.n.01')]
Synset('bank.n.09') [Synset('depository.n.01')]
Synset('bank.n.10') [Synset('flight_maneuver.n.01')]
Synset('bank.v.01') [Synset('tip.v.01')]
Synset('bank.v.02') [Synset('enclose.v.03')]
Synset('bank.v.03') [Synset('transact.v.01')]
Synset('bank.v.04') [Synset('act.v.04')]
Synset('bank.v.05') [Synset('work.v.02')]
Synset('deposit.v.02') [Synset('give.v.03')]
Synset('bank.v.07') [Synset('cover.v.01')]
Synset('trust.v.01') [Synset('believe.v.01')]
```

3 - Fig 3: Possible synsets and hypernyms of the word "bank"

1. Indexing the network

The DSN is indexed on Solr.

The paper doesn't address how the graph or the adjacency list is stored. An adjacency list is given as input to the mapper. This adjacency list can grow exponentially, I feel it's impractical to store as plain text files. Indexing aids in faster retrieval.

1. Markov Clustering

Markov Clustering[3] is an unsupervised cluster algorithm for graphs based on simulation of stochastic flow in graphs is applied to the network. It simulates flow within a graph and promotes flow in a highly connected region and demotes otherwise, thus revealing natural groups within the graph

Markov Clustering is performed on the graph with inflation factor-1.4

Possible Extension:

1. Figure out how to perform Hadoop map reduce from solr indexed data.

Observations:

1. Open IE extraction provides redundant triplets. Deduplication method has to be devised.
2. The paper addresses the fact that hadoop map reduce creates overhead. But since the paper hasn't used indexed data performance needs to be compared.

The lowest time reported by the paper is 23 secons for 200 statements in NYT dataset.

[1] Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. Leveraging Linguistic Structure For Open Domain Information Extraction. In Proceedings of the Association of Computational Linguistics (ACL), 2015.

[2] Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.

[3] Stijn van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000.