

Distributed Computing Final Project

Research Topic: Reasoning with knowledge

G.Savitha,201501019

1. Scalable Distributed Semantic Network for knowledge management in cyber physical system

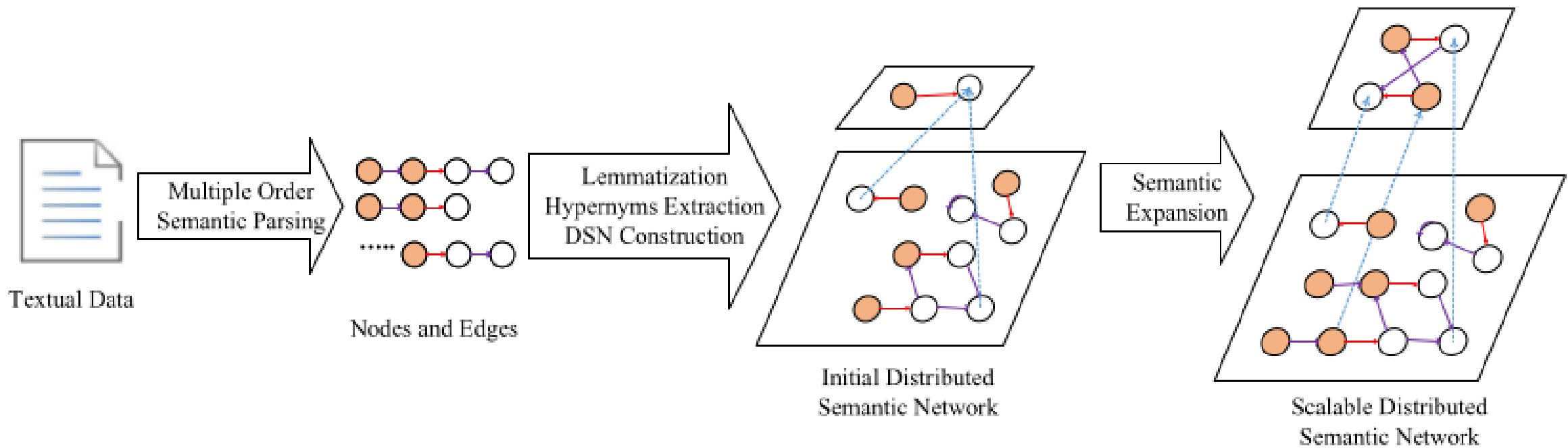
Authors: Shengli Song, Yishuai Lin, Bin Guo, Qiang Di, Rong Lv

Year: 2017

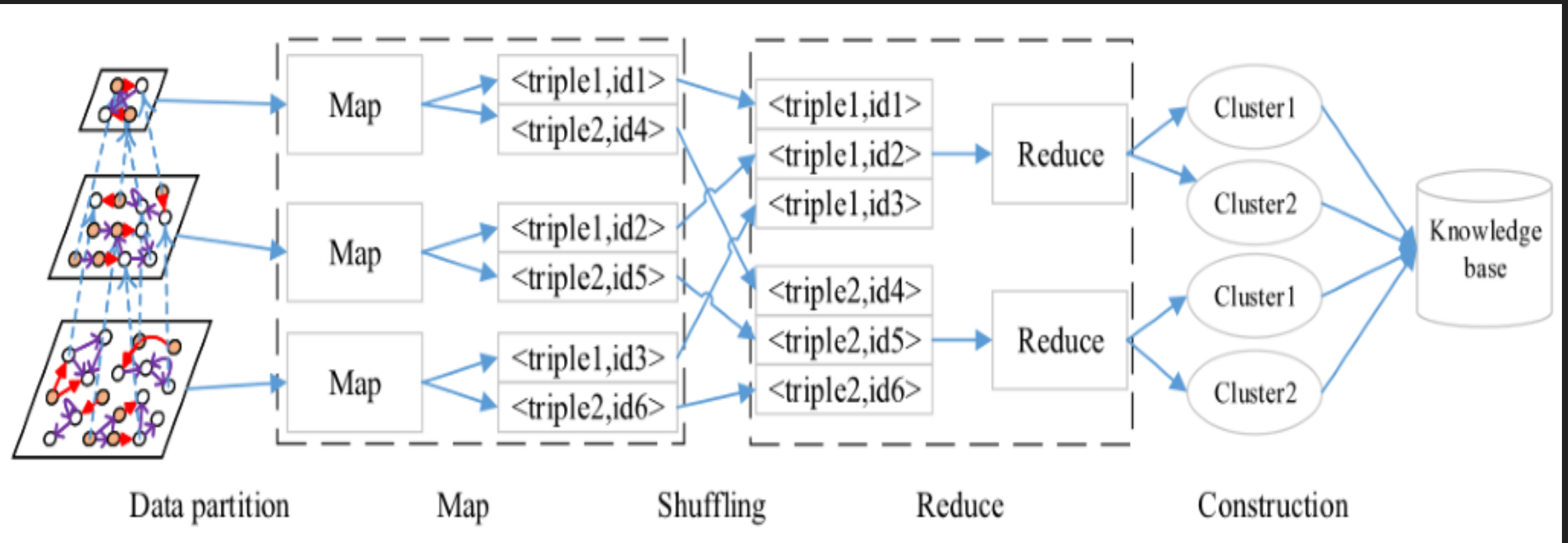
Journal: Journal of Parallel and Distributed Computing

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

Distributed Semantic Network



Proposed Architecture

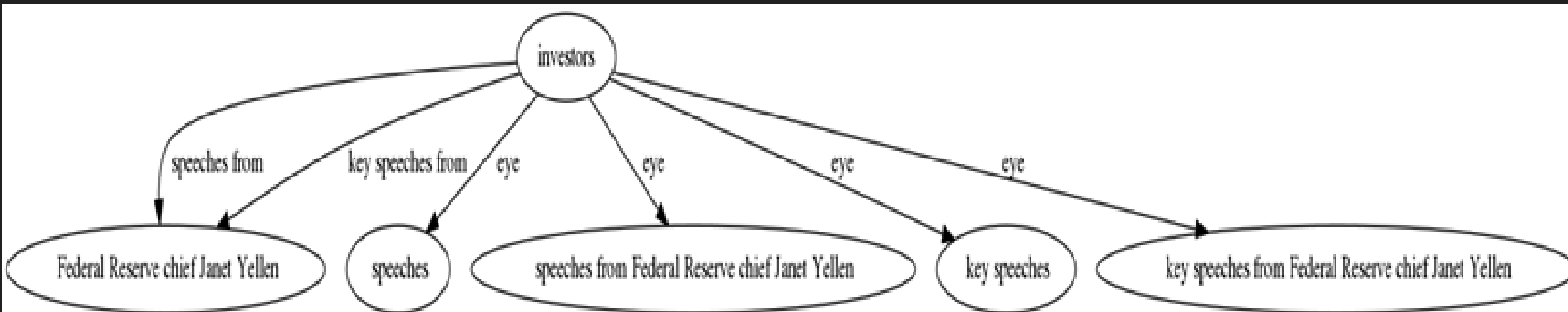


Implementation

1. Multiple Order Semantic Parsing

Extraction of NP-VP-NP triplets using Stanford Open Information Extraction

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



2. 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 is utilised.

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

For example: The word bank can have multiple meanings – river bank, bank as a financial institution, turn, belief etc.

3.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.

4.Markov Clustering

Markov Clustering 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

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

2. Resource discovery for distributed computing systems

Authors: Javad Zarrin , Rui L. Aguiar , João Paulo Barraca

This paper addresses the challenge of discovering resource in large-scale distributed computing environments .This paper is literature survey on current state of resource discovery protocols, mechanisms, and platforms for large-scale distributed environments, focusing on the design aspects.

Informed vs Uninformed

The primary distinction between informed and uninformed search methods is, uninformed search is a blind search – it has no prior knowledge about other nodes or resources in the network whereas informed search has each requester has some sort of heuristic or probabilistic measure of where the resource is located.

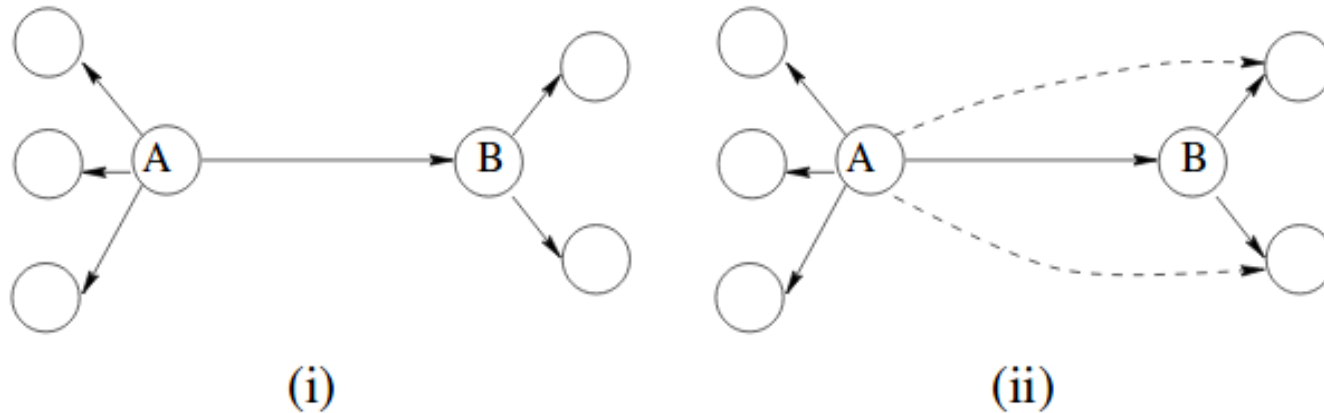
ALG-Flooding, Breadth First Search (BFS), and Depth First Search (DFS), are the most well-known systematic search methods. Moreover, there are several other systematic and random search methods such as Depth Limited Search, Iterative Deepening, Uniform Cost Search, Random Walk etcetera.

Synchronous vs. Asynchronous

Synchronous algorithms proceed synchronously, in parallel rounds, where each round is defined as the time required for each node in the network to communicate with one or more other nodes, learning about them and gathering information. Asynchronous methods are based on the asynchronous communication model, where each node can send a message in arbitrary size to any of its neighbors in a way that the outgoing message eventually arrives in the destination node after an unbounded finite time. ALG-Flooding, Swamping, Random Pointer Jump and Name-Dropper are well known synchronous resource discovery search algorithms.

Random Pointer Jump Algorithm

In each round $u \in V$ contacts a random neighbor $v \in V$. It then sends a message $M(u)$ to V which then merges u with v .



A randomly chooses neighbor B. A sends a message to B. B is merged with A. New edges are shown as dotted lines.

References

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.
4. J. Taheri, Y. Choon Lee, A.Y. Zomaya, H.J. Siegel, A bee colony based optimization approach for simultaneous job scheduling and data replication in grid environments, *Comput. Oper. Res.* 40 (6) (2013) 1564–1578.
5. M. Harchol-Balter, T. Leighton, D. Lewin, Resource discovery in distributed networks, in: *Proceedings of the Eighteenth Annual ACM Symposium on Principles of Distributed Computing, PODC '99*, ACM, New York, NY, USA, 1999, pp. 229–237.
6. Mor- Harchol Barter, Tom Leighton, Daniel Lewin, Resource Discovery in Distributed Networks
7. V.K. Garg, A. Aziz, An Efficient Deterministic Algorithm for the Resource Discovery Problem, Technical Report, Technical Report, ENS 527, The University of Texas, Austin TX 78712, 2000.