Information to Inference –A Process Flow using Knowledge Graph

Mercy Dol[1]
School of Computing Sciences
Hindustan Institute of Technology and Science
Chennai, India
mercydony@gmail.com
Angelina Geetha[2]
School of Computing Sciences
Hindustan Institute of Technology and Science
Chennai, India
angelinageetha@gmail.com

Article Info
Page Number: 188-197
Publication Issue:
Vol 71 No. 3s2 (2022)

Abstract— Every day a massive amount of data is created due to the constant use of the internet. During the COVID-19 pandemic, people are constantly learning to obtain knowledge at a faster pace. The knowledge graph can represent huge volumes of data and inference can be gained in lesser time. In our work, we have proposed a technique that initially cleans the scholarly data and then retrieves the entities and explores the relations between entities. Finally, we can split the knowledge graph using different methods to gain knowledge about the nodes and find the reasoning in their relation. The experimental results visually indicate that we can draw inference from the real triplets of the customized knowledge graph.

Keywords— Knowledge graph, CORD-19, inference, reasoning, knowledge

Article History
Article Received: 28 April 2022
Revised: 15 May 2022
Accepted: 20 June 2022
Publication: 21 July 2022

I. INTRODUCTION

The rise in enormous data in today's world combined with people's curiosity to learn and gain knowledge quickly has led to the evolution of the knowledge graph. A huge amount of vital structured data has not attracted humans greatly rather than a knowledge graph. Knowledge graphs can represent large-scale data in the form of a semantic network of entities and relations. As humans learn 80% of what they visualize hence a knowledge graph is a crucial form of knowledge representation. Google in 2012 has coined the knowledge graph as Linked data.

Features of knowledge graph

- A diverse relational network, comprising entities and relations represented as nodes and edges.
- Provides a precise source for information retrieval and suggestions by displaying connected entities of the real world.
- Helps people and machines to gain knowledge efficiently and solve difficult tasks.

- Displays data in a neat and structured manner that intelligently shows the relationship between objects.
- Single directional weighted graph with incoming and outgoing edges depending on the entities
- Derives and combines information with the help of intuition to create new ideas
- Knowledge representation and reasoning showcases the human’s ways of solving problems

Nowadays, Academicians and Industry professionals acquire knowledge through research done across the world. Knowledge acquisition can be done by reading and understanding a large number of research articles and finally deriving expertise. This process is time-consuming and tedious when done manually. Knowledge graphs can solve this challenging task quickly and efficiently.

Many research papers are being published daily on coronavirus disease (COVID-19). This pandemic has increased the demand for tools and libraries that help researchers to dive through a large scientific dataset to extract related information that envisions relations between data and derive knowledge.

In this work, we have taken the COVID-19 Open Research dataset (CORD-19) to derive the inference about the issues faced in the COVID-19 pandemic. CORD-19 is a collection of titles and abstracts of 500,000 scholarly articles. We constructed a knowledge graph with this dataset after pre-processing and sampled it into smaller graphs based on the triplet (source-relation-target) to acquire inference and reasoning for the challenges discussed in the research papers.

II. Literature Review

Yunrong Yang et al proposed a novel model [1] to extract proof from COVID-19 scholarly documents. COVID-19 public health evidence knowledge graph (CPHE-KG) was used for decision-making in public health. A knowledge graph was created from each abstract and graphs were joined together to form the final graph from the CORD-19 dataset. The entities of the graph were labeled automatically using the spaCy library from natural language processing. Acquiring evidence from lengthy documents was challenging.

Zhixue Jiang et al introduced a system [2] that acquired medical terms from websites and build knowledge graphs by placing disease at the center. Data acquisition followed by Information extraction to construct a knowledge graph. Named Entity Recognition (NER) was used to create nodes in graphs. The nodes were related using edges thus meaning was derived. The faster building, retrieval speed, and reasoning of knowledge graphs were still a limitation.

Huaxuan Zhao et al explained the methodology of building a standard knowledge graph [3]. Knowledge extraction included finding, comprehending, and framing rules. In the field of Artificial Intelligence, knowledge graphs acted as a storehouse where data from the internet were arranged and understood similar to human reasoning ability.

Rui Zhang [4] et al discovered existing and novel drugs that could be reused in COVID-19. The automatic semantic relation extraction was processed using the SemRep natural language processing (NLP) tool in the form of (subject, predicate, object) triplets. The triplet association strength was found by G2 score, which was high when the noticed and predicted frequency of triples were different. The prediction function on the triples identified the irrelevant triples based on the score generated by the open and closed discovery pattern in the knowledge graph.

Min Zhang et al invented a model [5] that depicted the human comprehending process in the QA system. It filtered the irrelevant content depending on the question using score. Then understood the relation between words in the question and context. The essential sentences were extracted by building a graph and ranked based on their importance.

Colby Wise et al built a COVID-19 knowledge graph (CKG)[6] from COVID-19 related research papers (CORD-19) using triplets (head, relation, tail). It is an extraction system based on the similarity between documents combined with topological order derived from CKG. String lemmatization was reduced using SciSpacy.

Our proposed work is able to provide better reasoning and draw improved inferences from the large dataset.

III. Materials and Methods

The architecture diagram of the proposed method is shown in figure 1.

A. Data Collection

CORD-19 dataset is the acronym of COVID-19 Open Research Dataset, which consists of frequently updated research articles regarding coronavirus built by the Allen Institute of AI. In our work, we have considered the dataset with 52K rows and columns including the title and abstract of scholarly articles. From this dataset, we can extract 46K entities and relations to build a knowledge graph.

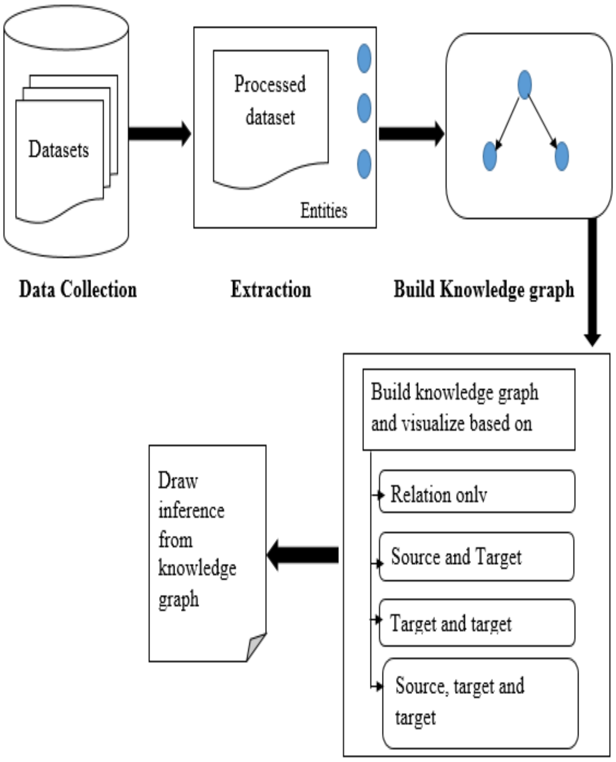


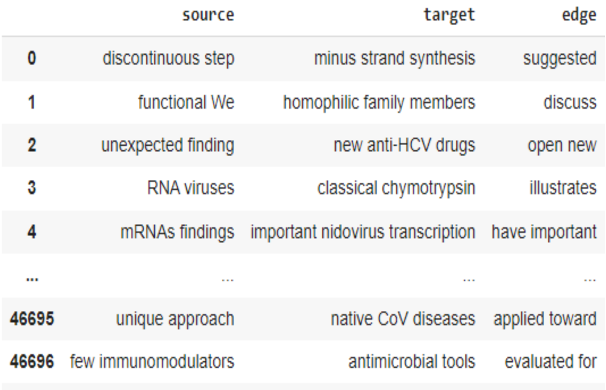
Figure 1. Architecture Diagram of Information to Inference

B. Extration

The dataset is initially cleaned by eliminating rows with null values and incomplete information and columns not needed for our processing. The abstracts of the research papers are considered for processing and sentences are split. Entities are extracted from the sentences by identifying the subject and object using dependency tags of spaCy library and considered as nodes. The nodes are not independent [7] but related hence relations are extracted between them in the form of edges with the help of dependency parsing and rule-based matching of spaCy.

C. Building Knowledge Graph

The nodes (subject-object pairs) and edges (relations) are extracted from the dataset in the form of triplets (entity-relation-entity) using the data frame. Then a uni-directed knowledge graph is created from the triplets using the networkx library. The triplets are shown in figure 2, are obtained from extraction and can be stored in a file for further processing.



	source	target	edge
0	discontinuous step	minus strand synthesis	suggested
1	functional We	homophilic family members	discuss
2	unexpected finding	new anti-HCV drugs	open new
3	RNA viruses	classical chymotrypsin	illustrates
4	mRNAs findings	important nidovirus transcription	have important
...	...	...	...
46695	unique approach	native CoV diseases	applied toward
46696	few immunomodulators	antimicrobial tools	evaluated for

Figure 2. Triplets obtained from CORD-19 dataset

D. Deriving inference from the knowledge graph

It is difficult to visualize the knowledge graph with the entire dataset hence we can create small knowledge graphs to identify the true triplets. Sampling is the need of the hour to efficiently access the data and identify the real truth [8]. The nodes with many neighbors and triplets are considered for inference.

Knowledge graphs are plotted based on the following for better visualization and understanding

- Relation only
- Source and target
- Target and target
- Source, target, and target

IV. Results And Discussion

A. Relation only

In figure 3, the knowledge graph created with the relation =' disease', which depicts all entities related to the disease that include inflammation, viral infection, novel coronavirus.

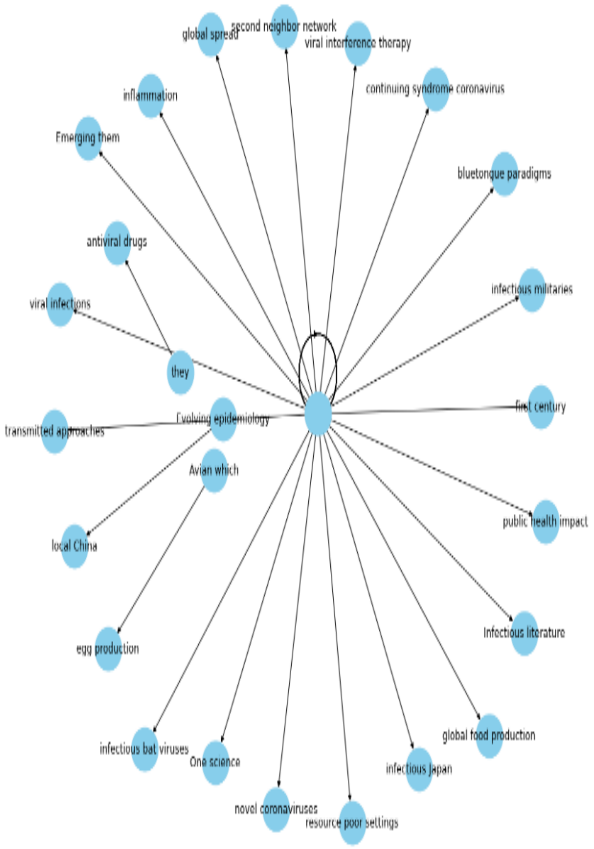


Figure 3: Knowledge graph with relation=‘disease’

B. Source and Target

To derive in-depth knowledge, we plotted a graph depending on source and target (source=’importance’ and target=’control’) depicted in figure 4. This graph shows incoming and outgoing links which depend on the entity. The importance of the edge [9] is defined by the number of links between two entities. The number of nodes closer to a node depicts its importance according to degree centrality [10].

The knowledge derived from this graph are

- The nodes to be considered for inference are control and importance as the nodes in its neighborhood are more compared to other nodes.
- The incoming links to control entity is 7 and outgoing links=0, while outgoing links of importance node is 7 and incoming links=0, which indicates that control is an object and importance is a subject in most sentences.
- The single edge between control and importance shows they are weekly related.
- The triplets give us the crucial features to control any disease
  - prevention strategies-importance-control
  - viruses-importance-control
  - initial identification-importance-control

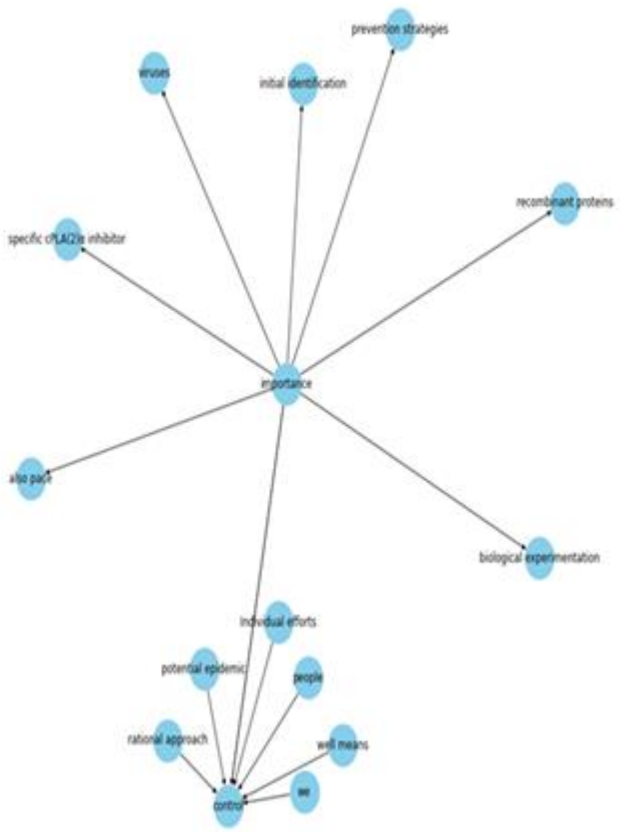


Figure 4: Knowledge graph with source=’importance’ and target=’control’

C. Source,target and target

To draw more inference, figure 5 displays the graph using source, target, and target combination (source=’virus’ target=’disease’ and target=’control’).

The inference derived from this graph are:

- the entity-relation-entity (A--→B--→C) draws the reason for A entity emerging to C entity with reason B
  - virus-emerging pathogen-disease
  - virus-respiratory tract-successful virus propagation
  - virus-near future-control

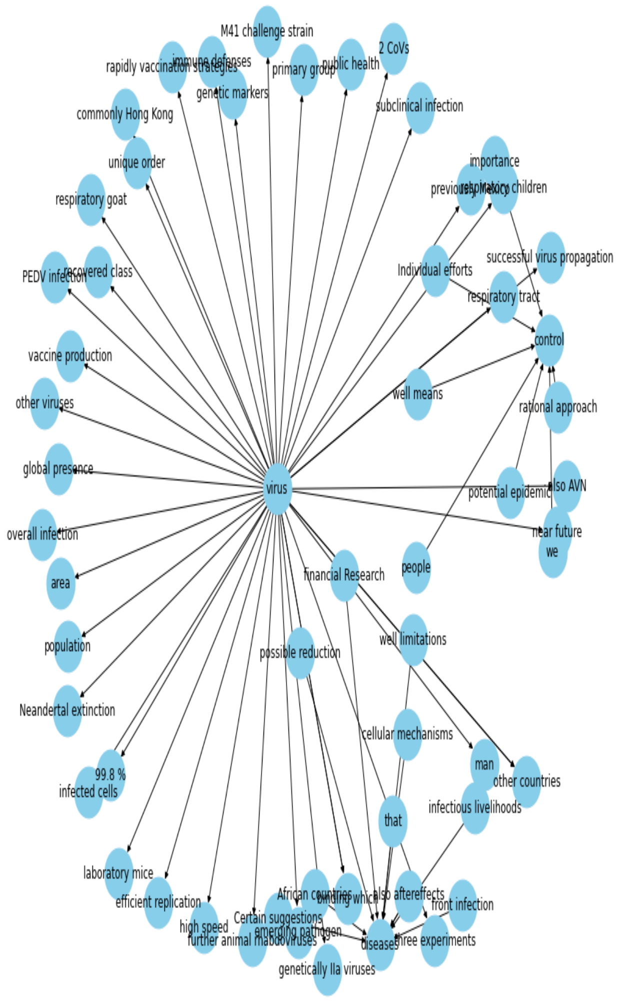


Figure 5: Knowledge graph with source='virus' target='disease' and target='control'

We can understand that when the graph is plotted with more input entities, we get more inference and thus triplet gives more knowledge from Table 1.

TABLE I. COMPARING INPUT INFORMATION AND OUTPUT INFERENCE

Sl. No	**Combination of entities used for plotting graph**	**Derived from knowledge graph**		
		Entities	**Relations**	**Inference**
1	Relation ('disease')	21	21	17
2	Source('response') target('immunity')	24	17	9

3	Relation('prevent')	45	23	17
4	Source('importance') target('control')	15	14	12
5	Source('virus') target('disease') and target('control')	59	52	48

TABLE II. COMPARING OUR PROPOSED METHODS WITH OTHER LITERATURES

Sl. No	**Literature**	**Methods used**	**Benefits**
1	Common Sense-Based Reasoning Using External Knowledge for Question Answering [11]	Paths with two or less hops	Unwanted edges removed
2	Question-Answering system based on the Knowledge Graph of Traditional Chinese Medicine	Dictionary matching method	Find the meaning of the word
3	Cascade embedding model for knowledge graph inference and retrieval	Graph embedding models	Anticipate the missing data
4	Knowledge Graph Completeness: A Systematic Literature Review [12]	Schema fullness	Completeness of graph

5	Information to Inference-A Process Flow using Knowledge graphs	Knowledge graph built based different combination of relation, source, target	Inference and reasoning ability

V. Conclusion

We have proposed a methodology for breaking down the knowledge graphs into different smaller networks based on the various combinations of triplet and finally to arrive at the conclusion on certain topics related to COVID-19 scholarly articles. Our work was able to derive strong evidence quickly by building a knowledge graph automatically. The results show how reasoning ability has been improved when compared with other related processes. The experiments also help us to find the true triplet and draw inferences from it.

This technique can be experimented with other datasets to draw knowledge. In the future, we would be using this technique to derive inference from unstructured data available on the internet. We have to explore ways to increase the speed of extraction of entities and relations using machine learning techniques [13].

References

1. Yunrong Yang, Zhidong Cao , Pengfei Zhao, Dajun Daniel Zeng, Qingpeng Zhang, Yin Luo, “Constructing public health evidence knowledge graph for decision-making support from COVID-19 literature of modelling study,” Journal of Safety Science and Resilience, Elsevier, 2021.
2. Zhixue Jiang, Chengying Chi and Yunyun Zhan, “Research on Medical Question Answering system based on knowledge graph,” IEEE Access 2021.
3. Huaxuan Zhao, Yueling Pan, and Feng Yang, “Research on Information Extraction of technical documents and construction of Domain Knowledge Graph,” IEEE Access, 2020.
4. Rui Zhang, Dimitar Hristovski, Dalton Schutte, Andrej Kastrin, Marcelo Fiszman, Halil Kilicoglu, “Drug repurposing for COVID-19 via knowledge graph completion,” Journal of Biomedical Informatics, Elsevier,2021.
5. Min Zhang, Feng Li, Yang Wang, Zequn Zhang, Yanhai Zhou, and Xiaoyu Li, “Coarse and Fine Granularity Graph Reasoning for interpretable Multi-Hop Question Answering,” IEEE Access,2020.
6. Colby Wise, Vassilis N. Ioannidis, Miguel Romero Calvo, Xiang Song, George Price, Ninad Kulkarni, Ryan Brand, Parminder Bhatia, George Karypis, “COVID-19 Knowledge Graph: Accelerating Information Retrieval and Discovery for Scientific Literature,” arXiv preprint arXiv:2007.12731, 2020.
7. Fang Miao, Xueting Wang, Pu Zhang, Libiao Jin, “Question-Answering system based on the Knowledge Graph of Traditional Chinese Medicine,” 11th International Conference of Intelligent Human-Machine Systems and Cybernetics(IHMSC), IEEE,2019.

8. Jianpeng Zhang, Yulong Pei, George Fletcher, Mykola Pechenizkiy, “Evaluation of the sample Clustering Process on Graphs,” IEEE Transactions On Knowledge And Data Engineering,2018.
9. Daifeng Li, Andrew Madden, “Cascade Embedding Model for Knowledge Graph Inference and Retrieval,” Information Processing and Management, Elsevier, 2019.
10. Kanyadara Saakshara, Kandula Pranathi, R.M. Gomathi, A. Sivasangari, P. Ajitha, T. Anandhi, "Speaker Recognition System using Gaussian Mixture Model", 2020 International Conference on Communication and Signal Processing (ICCSP), pp.1041-1044, July 28 - 30, 2020
11. R. M. Gomathi, P. Ajitha, G. H. S. Krishna and I. H. Pranay, "Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862048.
12. Pasquale De Meo, Mark Levene, Fabrizio Messina, and Alessandro Provetti, “A General Centrality Framework based on Node Navigability,” IEEE Transactions on Knowledge and Data Engineering, 2019.
13. Yunyeong Yang and Sangwoo Kang, “Common Sense-Based Reasoning using External Knowledge for Question Answering,” IEEE Access,2020.
14. Subhi Issa , Onaopepo Adekunle, Fayçal Hamdi , Samira Si-Said Cherfi, Michel Dumontier, and Amrapali Zaveri, “Knowledge Graph Completeness: A Systematic Literature Review,” IEEE Access, 2021.
15. Mercy Dol, Angelina Geetha, “A Learning Transition from Machine Learning to Deep Learning: A Survey,” IEEE International Conference on Emerging Techniques in Computational Intelligence (ICETCI), DOI: 10.1109/ICETCI51973.2021.9574066,2021.
16. A Sivasangari, P Ajitha, RM Gomathi, "Light weight security scheme in wireless body area sensor network using logistic chaotic scheme", International Journal of Networking and Virtual Organisations, 22(4), PP.433-444, 2020
17. Sivasangari, A., Nivetha, S., Pavithra,, Ajitha, P., Gomathi, R.M. (2020),” Indian Traffic Sign Board Recognition and Driver Alert System Using CNN”, 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020, 2020, 9315260
18. Sindhu K, Subhashini R, Gowri S, Vimali JS, "A Women Safety Portable Hidden camera detector and jammer", Proceedings of the 3rd International Conference on Communication and Electronics Systems, ICCES 2018, 2018, pp. 1187–1189, 8724066
19. Ancy S, Kumar R, Ashokan R, Subhashini R ,"Prediction of onset of south west monsoon using multiple regression",Proceedings of ICCCS 2014 - IEEE International Conference on Computer Communication and Systems, 2014, pp. 170–175, 7068188
20. Akshaya, R., N. Niroshma Raj, and S. Gowri. "Smart Mirror-Digital Magazine for University Implemented Using Raspberry Pi." In 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), pp. 1-4. IEEE, 2018.
21. Gowri, S. and Divya, G., 2015, February. Automation of garden tools monitored using mobile application. In International Confernce on Innovation Information in Computing Technologies (pp. 1-6). IEEE.
22. Ajitha, P.Sivasangari, A.Gomathi, R.M.Indira, K."Prediction of customer plan using churn analysis for telecom industry",Recent Advances in Computer Science and Communications,Volume 13, Issue 5, 2020, Pages 926-929.