

SEBI Knowledge Graph

Deepti Saravanan

Abstract—Securities and Exchange Board of India (SEBI) is the regulatory body for securities and commodity market in India under the jurisdiction of Ministry of Finance, Government of India. SEBI has framed a set of regulations that companies are expected to abide by. In addition, it also regularly publishes additional documents that clarifies company's queries, amplifies the regulations with the purpose and exceptions and also discusses the financial market activities. The large volume of data that are highly interconnected to each other on the basis of diverse sets of relationships calls for a higher level structure that could store the information along with the relationships between them. Advances in NLP, coupled with the vast availability of digital text data have yielded impressive results in the legal domain. Domain specific knowledge graph is an active area of research where it is critical to collect all the true facts about the domain and study the relationships between them. In this paper, we construct a knowledge-aware documents representation with dynamic granularity of information for SEBI documents that could be utilised for various higher-level graph-based applications such as graph embedding. The nodes represent the documents and sub-regulations while the edges connecting them provides with the context/relationship between them.

I. INTRODUCTION

Knowledge graphs play an important role in many applications, such as data integration, natural language understanding and semantic search. Recently, there has been some work on constructing legal knowledge graphs from legal judgments as discussed in detail in the literature review section II. However, they suffer from some problems. First, existing work intends to build legal knowledge graph only at the level of legal terms. Second, existing graphs are static with a fixed granularity of information provided. To solve these problems, in this paper, we propose a framework for constructing a knowledge graph for SEBI from regulatory documents and legal judgments. We build a three-layered knowledge graph in the legal domain that provides increasing granularity of information with layers. In addition, it is also interactive for user navigability and activity around the document of interest, where the users can click on the nodes and edges for more information about the corresponding documents. This functions as a single hub for all SEBI-related information.

The novel properties of the multi-layered SEBI knowledge graph and our contributions can be summarized as follows:

- **Dynamic granularity of information with layers**

The three-layered knowledge graph proposed provides varying depths of information at different layers. This enables the user to view the corresponding suitable layer depending on their requirement by navigating around the graph.

- **Interactivity and Navigability**

The proposed knowledge graph enables users to move across layers, as explained in section IV-A. The nodes and edges when clicked redirects the user to the corresponding document's page for more information. The user can also navigate around nodes to understand various relationships and connectivities.

Roadmap. The paper is divided into the following section - A literature survey discussing a few existing solutions, Semantics extraction tasks from documents, our proposed multi-layered interactive SEBI knowledge graph, SEBI Amendment Documents reconstruction and finally the conclusion derived from this exploration along with the future scope.

II. LITERATURE REVIEW

Owing to the overwhelming amount of data available on the internet, many methods have been devised to automate various mundane works for auto-compliance check. Most of the legal related data are in the text format and hence is an active region of research in the field of Natural Language Processing. There have been various attempts of developing AI systems that could automate tasks in ensuring compliance of companies with regulatory bodies like SEBI and SEC, which otherwise would take a lot of legal hours by lawyers and other legal experts.

One unique aspect of legal domain is that it has different kinds of documents that are highly interconnected to each other. This provides a great potential to store these relations in the form of Dense Knowledge Graphs for processing by machines and humans. [1] proposes an approach to represent the legal data (legal norms and court decisions) of Austria and shows how this data can be used to build a legal knowledge graph that is usable in various applications for lawyers, attorneys, citizens or journalists.

[2] built a RDF Knowledge Graph for industrial use case using a semantic model for representing legal documents together. They have also compared and discussed the use and performance comparison of four different state-of-the-art mapping engines for the given sample of required use cases in the legal domain.

Existing literature explores various use cases where AI systems could automate the task of ensuring compliance. In this paper, we introduce a layered Semantic Knowledge Graph with dynamic depth of information with layers as opposed to the existing single layer of constant depth of information that we built for the Securities and Exchange Board of India (SEBI) documents.

III. SEMANTICS EXTRACTION FROM DOCUMENTS

The documents involved in the knowledge graph and a short description about them are given by table II. These documents are of unstructured text format.

The SEBI regulatory documents consists of both regulatory and non-regulatory sentences (for instance, additional explanations, footnotes and schedule clauses). Hence, it is important to identify regulatory sentences and extract them. Analysing the content across documents, it was observed that the words 'shall' and 'may', tagged as rule-relevant words, could be found majorly in sentences that are regulations. On this basis, the regulations were extracted with an accuracy of 85 percent.

Regulations are composed of two or more sub-regulations. Since the other SEBI documents refer to these regulatory documents in the sub-regulation level, the nodes in the knowledge graph represent the sub-regulations and not the regulations on the whole. In addition, the cross-references found in the SEBI regulatory documents occur between two sub-regulations only. This wouldn't be captured by nodes representing the regulations on the whole, hence such nodes are not included in the knowledge graph.

The semantic relationship between these sub-regulations are extracted via LDA topic modeling. The topics extracted from the sub-regulations are compared with another. Higher overlap indicates that the sub-regulations are semantically very similar.

Paragraph-level semantics were also extracted to define the context of the mention of sub-regulations in the additional SEBI documents (Concept Papers, Informal Guidance, Legal Cases). The context of the concept papers is defined by the topics discussed in them. These topics could be extracted from the titles of these documents. Each concept paper covers individual regulatory documents on the whole which are also extracted for mapping the corresponding documents with them.

Informal guidance includes the company queries, the facts based on the sub-regulations that are relevant to the query and an expanded answer for the queries by SEBI. The context of the sub-regulations mentions could be extracted from the keywords present in these facts. This is because the keywords generally describe the entities involved and the actions/situations, which function as the crux of the discussion in the document considered.

Allegations made, facts, company background, decision and penalty are a few subjects covered predominantly in legal case file documents. The paragraphs describing the allegations made against the actor(s) represent the context of sub-regulations mentioned in the case documents, as they mention the sub-regulations and the actions that violated them.

These extracted semantics are used to define the relationship between the document and the corresponding sub-regulations in the knowledge graph.

IV. SEBI KNOWLEDGE GRAPH

The semantic relationships between the actors/entities involved define the concept of regulations. The knowledge is stored in the form of directed graphs with nodes representing

the entities of the regulations and the edges defining the transaction between these entities as relationships between these nodes. This provides a simple approach to investigate the problem space with greater expressiveness and higher cognitive adequacy, since the relationships between the entities involved could be comprehended better by analysing the connectivity of the knowledge graph.

Figure 1 illustrates the three-layered knowledge representation and what each layer involves.

A. Knowledge Graph Construction

A three-layered knowledge graph was constructed such that the granularity of information increases with the depth of the level.

The level one consists of list of regulatory documents with two possible branches to layer two - sub-regulations and amendment timeline respectively.

Upon clicking the amendments option, the corresponding amendment timeline of the regulatory document is displayed. This timeline is a linearly constructed graph where the nodes represented the individual reconstructed regulatory documents that were constructed based on the proposed amendments by SEBI (refer to section V for more details). The edges between two nodes representing the two versions of the regulatory documents provide a detailed information about the amendments made. In addition, nodes representing concept papers are connecting to the corresponding version of regulatory documents they refer to. The edges between the concept papers and the regulatory documents represent the semantics extracted from the concept papers, as discussed in section III.

On the other hand, upon clicking the sub-regulations option, the other branch of layer two is displayed where there are three kinds of nodes - sub-regulations, informal guidelines and legal cases. The edges between the informal guidance and sub-regulation nodes provide the key words that define the context of the sub-regulation in the former while the edges between legal cases and sub-regulations represent the allegations made on the basis of the sub-regulation. The edges between two sub-regulation nodes are divided into three kinds:

Cross-reference: When one sub-regulation is directly referred to within another.

Similar Chapter: When the two sub-regulations occur within the same chapter of the regulatory document.

Similar topics: When the topics extracted from two sub-regulations using LDA are highly similar.

Further clicking on the sub-regulation node redirects the user to layer three which provides a detailed representation of it. Layer three consists of directed graphs with nodes representing the subjects and objects of the regulations and the edges defining the relationships between these nodes. The sub-regulations can be classified into two classes, based on the structural information. Class 1 regulations are simple sentences while class 2 regulations are composed of multiple sub-points of short phrases or sentences. The triplets extraction from Class 1 is straight-forward, as opposed to class 2 where the sub-points provide additional relationships with respect to the

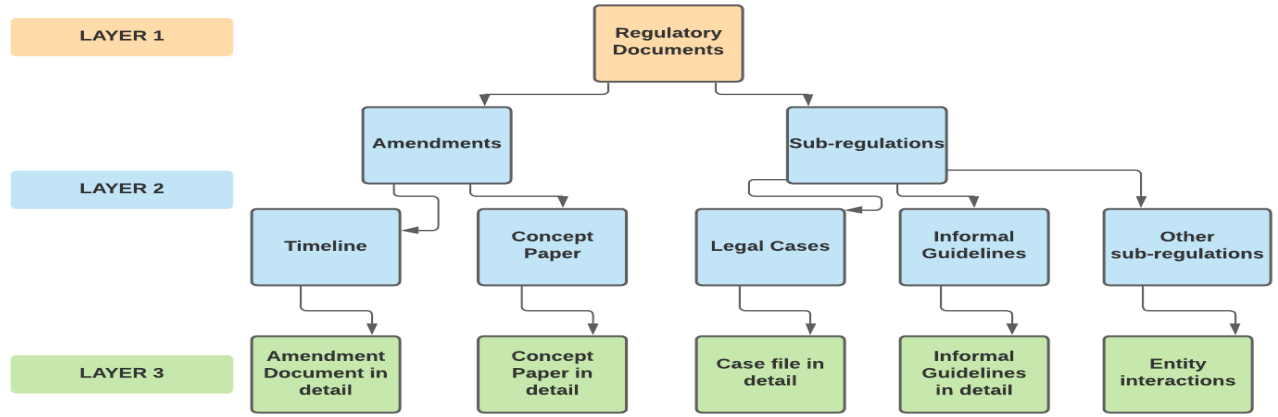


Fig. 1: Knowledge Graph Representation

TABLE I: SEBI Amended Regulations Reconstruction

Proposed Amendment	Before Amendment	After Amendment
after the words and symbol “key service providers,” insert “terms of reference of the committee constituted for approving the decisions of the Alternative Investment Fund,”	... tenure of the Alternative Investment Fund or scheme, conditions or limits on redemption, investment strategy, risk management tools and parameters employed, key service providers, conflict of interest tenure of the Alternative Investment Fund or scheme, conditions or limits on redemption, investment strategy, risk management tools and parameters employed, key service providers, terms of reference of the committee, conflict of interest ...
before the words “the network”, the words “ensure that” shall be inserted	... “sound track record” shall mean the sponsor should - the network is positive...	... “sound track record” shall mean the sponsor should - ensure that the network strategy, risk management tools and parameters is positive...

TABLE II: SEBI document types

Document	Description
Regulatory Documents	SEBI Regulations
Reconstructed Regulatory Documents	Intermediate Regulatory Documents constructed from Amendment Documents.
Concept Papers	Discussion on a topic wrt regulations
Informal Guidance	SEBI clarifications for Company queries
Legal Cases	Orders discussing allegations, violations and penalty

TABLE III: Rationale for special cases in Semantic Extraction

Sentence	Rationale
An insider shall be entitled to formulate a trading plan and present it to the compliance officer for approval.	When the model reaches ‘and’, it traverses back to ‘insider’ where forking takes place. Hence, ‘insider’ will be connected to ‘formulate’ and ‘present’.
For purposes of sub-regulation (3), the board of directors shall require the parties to execute agreements to contract confidentiality and non-disclosure obligations on the part of such parties	The phrase ‘such parties’ refers to the ‘parties’ that occurs prior to this reference. Hence, the edge should direct back to the original node from ‘obligations’.

entities to which they refer to. Hence, the sub-regulations are categorised into these two classes and parsed accordingly. Around 60 percent of these regulations fall under class 2 while the remaining 40 percent of these regulations fall under class 1.

A few notable cases where the identification of source and target nodes might get tricky are presence of conjunctions and words defining semantic references like ‘such’. In the former case, the model has to traverse back to the node where forking is supposed to happen. In the latter case, the model should be able to replace the semantic references with the referred subject/object phrase. Table III provides rationale for the above cases in detail.

A video demo of the navigation through the various layers of the knowledge graph could be found here: Knowledge Graph Demonstration Video.

B. Inferences

The following could be inferred from the layer three based on the semantic relationships:

- **Degree of the node** : Denotes the number of distinct contexts involving the object where each connection denotes an action in which the entity represented by the node is involved in. [3]
- **Keywords** : Based on Centrality Theory, higher importance scores are associated with nodes with higher degrees as they are connected to a lot other phrases or tokens within the regulations. And these are the keywords that represent the corresponding regulations. This is useful for keywords extraction. [4]
- **Clustering** : Regulations could be clustered based on similar sub-graph structures and analyzed by profiling the clusters formed. For instance, node2vec algorithm could be employed that converts the graphs to vector form. These vectors, each representing a sub-regulation, could then be clustered and analyzed.

- **Graph edit distance** : The degree of difference between the old and updated versions of amended regulations, or two closely related regulations could be measured by counting the number of updates (additions, deletions and modifications) to be performed to convert the first graph to the second.

V. AMENDED REGULATORY DOCUMENTS RECONSTRUCTION

The documents related to SEBI regulations are of two forms - the base regulatory document and amendment document. The amendment document has all the information about the changes made to the former both at word-level and phrase-level. The base regulatory document is reconstructed after a couple of years by incorporating all the amendments proposed till then. Hence, the new base document could have the changes incorporated with respect to one or more amendment documents released in between, leading to lack of uniformity in the number of amendment documents referred for reconstruction. This calls for uniform Reconstruction of intermediate regulatory documents after the release of every amendment document.

The document reconstruction process involves a two-stepped approach:

A. Extraction of Amendment information

The amendment document has text data in an unstructured format. The changes could be of three categories - insertion, replacement and modification. By observing the unique pattern and keywords used for each category, this unstructured data is converted into a structured format. The final data contains the following attributes - Regulatory Document name, Year of Amendment, Regulation and sub-regulation id, Action category and content to be updated (where and what).

B. Mapping and updating

This is the reverse process - structured data is converted back to unstructured format. The regulation and sub-regulation IDs are mapped into the immediate previous version of regulatory document. The change locations are figured and the content is updated accordingly. This final text document after incorporating all the changes is the reconstructed regulatory document for the year extracted in the previous step.

A few examples of reconstructed sub-regulations are illustrated in table I.

VI. CONCLUSION

In this paper, we propose a knowledge graph of various SEBI-related resources in a layered approach where the connections define different kinds and levels of relationships extracted. Layer one displays the regulatory documents and their amendment options. Layer two has two branches. The first branch displays the amendment timeline of the corresponding regulatory document. The concept papers, if exist, are connected to the corresponding versions of the reconstructed regulatory documents. The other branch of Layer two displays

the semantic relationships between the sub-regulations and the corresponding additional documents. The sub-regulations are also connected amongst themselves in three ways - cross-references, common chapter and common topics, where the topics for each sub-regulation is extracted using LDA topic modeling. Level three is displayed upon clicking a sub-regulation where the interplay between the entities involved in the sub-regulation is illustrated. For future work, we will cluster the various relationships in the knowledge graph in every level which could be helpful for similarity analysis and higher level understanding. This graphical information could also be used to study regulations proximity across countries via link prediction and graphical comparison.

REFERENCES

- [1] E. Filtz, "Building and processing a knowledge-graph for legal data," 05 2017.
- [2] A. C. Junior, F. Orlandi, D. O'Sullivan, C. Dirschl, and Q. Reul, "Using mapping languages for building legal knowledge graphs from xml files," 2019.
- [3] A. Veremyev, A. Semenov, E. Pasiliao, and V. Boginski, "Graph-based exploration and clustering analysis of semantic spaces," *Applied Network Science*, vol. 4, pp. 1–26, 2019.
- [4] S. Anjali, N. M. Meera, and M. Thushara, "A graph based approach for keyword extraction from documents," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, 2019, pp. 1–4.