Legal Document Similarity using Triples Extracted from Unstructured Text

Akshay Minocha, Navjyoti Singh
IIIT-Hyderabad
Hyderabad, India
akshay.minocha@research.iiit.ac.in, navjyoti@iiit.ac.in

Abstract

This research is an attempt to address the complication of bridging the gap between the traditional systems and the future of legal systems. We discuss one of the facets of the process of legal understanding and decision-making in the court of law, as well as aim to increase general public comprehension on the topic of constitutional importance. We focus on a selected list of documents gathered through citation network analysis, and using the knowledge of the Sections in the Income Tax Act of India which govern them; after processing the proceedings identified, through the proposed technique. The resulting triples are used to evaluate the similarity of such legal documents.

Keywords: legal information retrieval, document similarity, ontology, knowledge graph, semantic triples

1. Introduction

With the development and countrywide acceptance of internet-centric applications which fall under the category of e-governance in India; the legal domain is one area of interest which deserves great mention. The availability of reasonable sources and legal data, help in the building of practical and assistive applications which would be of great use to the legal experts. To a domain practitioner, there are other detailed applications such as - document classification, legal knowledge discovery, legal information retrieval, predictive mechanisms, and so on.

Many efforts have been made in this field to facilitate faster and better legal help to legal practitioners, advocates, researchers and the non-domain people. Although the accessible legal resources in India have been recently made available and there is a lot of data which can be used to increase the efficiency of these services, yet this information by large remains unstructured. In our research, we aim to look at cases which belong to a specific category of cases adhering to finance and income tax. We have generated an ontology of the sub-domain of the legal area and try to align the cases which cite these Sections of the Act according to the triples made by the technique proposed in Section 3 . The similarity of these cases is evaluated based on a thematic scheme, and we then discuss the results in Section 5 where a complete state of the situation and further steps are mentioned.

2. Related Work

In (Kumar et al., 2011), (Kumar et al., 2013) the authors use statistical measures and connective properties in text to predict the similarity of legal judgments, and on the other front (Saravanan et al., 2009), (Saravanan et al., 2006), talk about a novel method of legal document summarizing and effective retrieval by suggesting that we approach the problem with an ontological perspective. For legal information retrieval, it is the objective to manufacture an intuitive data space to consequently outline content information to an adjustable ontology. Legal Ontological enquiry has been inspired by the work done by LKIF (Hoekstra et al., 2007), (Breuker et al., 2007) where they also come up with a legal information interchange format along with an ontology of basic legal concepts in Italian law. The work is really inspirational in terms of providing a movement towards a knowledge representation formalism in the legal domain.

A triple as the name suggests is a combination of three different sets of words, an atomic form of information which provides semantics to the situation or in our case the legal text in hand. Just put into words it is a subject-predicate-object expression. Just like we have specific grammar while writing computer programs we have to find out a way in which we can simplify phrases and sentences into a more machine-readable format. A sentence can be broken down into multiple triples according to its complexity. Triples are one of the many ways in which information from a judgment is presented in a less complicated manner with fewer relevant words.

Understanding the relational facts from understandable content has for quite some time been of enthusiasm for data extraction research. The critical issue is to adjust the exchange off between high precision, recall, and adaptability. With the rise of the Semantic Web and various ontologies, information combination has turned into an extra test. There has been a lot of research on semi-supervised strategies utilizing bootstrapping methods together with beginning seed relations to make extraction designs. Unsupervised methodologies have contributed to work in the legal domain by not requiring hand-tagged information. These methodologies have addressed efficiently versatility and accuracy factors when connected on web-scale corpora. A system like LODifier (Augenstein et al., 2012) is a cornerstone in the achievements towards triple extraction research. Our data is not as much tagged and linked to entities so that it can easily be mapped onto a very well developed knowledge base, as the (Exner and Nugues, 2012), we connect the extracted entities in an unsupervised way which in turn would bring form and structure to the legal domain knowledge base.

3. Methodology

3.1. Dataset

The Income Tax Act of India was authorized in the year 1961 and is the statute under which everything identified with tax collection is recorded. The Act incorporates levy, collection, organization, and recuperation of wage assessments. The act represents a constitutional reference for people seeking support from a consolidated set of rules identified with tax collection in the nation. Organized into over 23 chapters and many schedules the act covers a great deal of the laws which are to be followed by individuals, firms, partnership firms about their dealings and their functioning. Due to its large breadth, we wanted to cover a specific part of the act which would comprise of knowledge which is in some way self-sufficient. The method of identifying this group of Sections within the Act was to generate by scraping[1] all the cases that cite the individual parts of the act and then narrow down to the division which shows the highest coverage in terms of the ratio of cases which only deal with this part of the act. The citation connections between the legal documents were made in a similar way as implemented to find graph connective measures for various network properties in the legal domain (Minocha et al., 2015). This grouping of cases, would ensure an investigation on an independent group of the act where most of the cases can be categorized into and belong within the sub-domain of the legal knowledge. The reason for this was to come up with an ontology which is tending to complete on its own with little dependencies on other parts of the act. We chose the part of the Income Tax Act which deals with 'Changes in constitution, succession, and dissolution of firms and partnerships'. The Sections of the act which were of interest are Section 187, 188, 188A, 189, and 189A. The number of such cases until the day of investigation was close to 80, and this number also seemed perfect to experiment with the methodology discussed later on in the paper.

3.2. System Architecture

Figure 1, explains the different modules that are involved in the work-flow of the system to generate some tagging and triples for the documents so that they can be compared against each other for similarity.

- **Headnote extraction**: The legal proceedings and documents available usually have the facts and a summary related to the case in the initial part of the document. In most of the documents such is the case, as the court proceedings first comprise of the known and acknowledged facts that have been put forward as the basis of the case. Extracting this part is crucial for us since we do not want the remaining text which would include discussions related to different cited cases, references, and opinions that might not be facts yet. Such data in triples can be conflicting, and hence headnote is extracted heuristically from the proceedings for our research.

[1] https://indiankanoon.org/

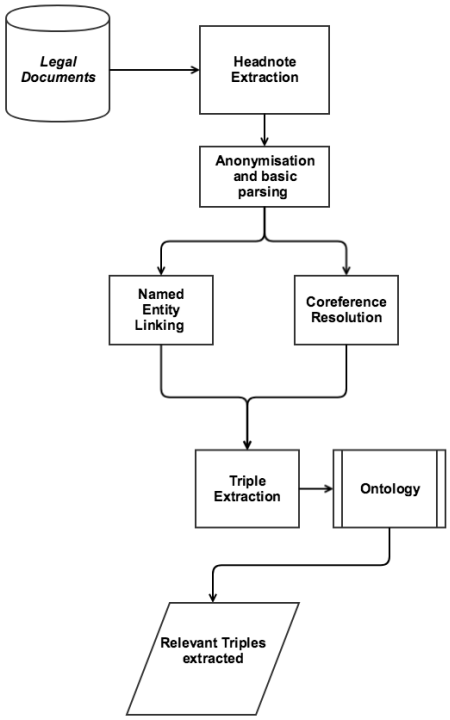

Figure 1: Triple Extraction from Legal Text

- **Anonymisation**: In this stage we try to anonymise the names of the partners and the firms so that more overlapping structures can be made while linking the triples.
- **Named Entity Linking and Co-reference Resolution**: Legal texts are lengthy, and constructing rules to extract triples becomes increasingly difficult, resulting in either very lengthy relations or issues in correct noun phrasal entity inclusion. To tackle this problem and to make the process efficient, we perform these text pre-processing tasks to help in obtaining relevant triples.
- **Triple Extraction**: For the triple extraction we use OpenIE (Fader et al., 2011),(Etzioni et al., 2008), a confidence score is obtained along with the extracted triple. We implement an instance of OpenIE in our work which in a single pass extracts a large set of relational tuples from the data. OpenIE does not require any human intervention in labelling or input
- **Ontology Mapping**: In this phase we map the triples to the common terms and actors which have been identified by describing the ontology of the legal domain in question, by doing this we create a set of triples which would have high overlap when the input legal cases are similar, due to their standardised nature.

4. Evaluation

To categorize or find similar legal proceedings which handle intricacies related to legal entities in a conceptual way.

We provided more than 80 random pairs of judgments from our dataset to three legal experts and asked them to rank these documents concerning similarity in the range of 1 to 10 (Raghav et al., 2015). Information about the dataset was given with the evaluation exercise. A similarity score of 10 would mean that the documents have a great deal in common and can be treated as a reliable reference by a legal practitioner while preparing arguments. The score does not represent a binary classification because of the nature of the extent of similarities which is planned to be used in case of future experiments. Although, we would be using this in a binary form for our analysis at hand, details of which follow.

Inspired by the work done for LODifier (Augenstein et al., 2012), our similarity measure is based on the distance and overlap between similar nodes. A short path indicates more relevant semantic information.

In Section 3.1. we described how that gold data which elucidates the similarity function, concerning scores are annotated for similarity by professionals and legal practitioners. However for comparison with other metrics and the extensions with the proposed changes we will use story link detection test, which when used initially, analyzed the information where two randomly selected stories to discuss the same news topic (Augenstein et al., 2012).

We have a score computed for each approach which is termed as sim, and like setting a base threshold, we have a similar limit here for which the following classification holds -

$$class(d_p, \theta) = \begin{cases} positive, & \text{if } sim(d_p) \geq \theta \\ negative, & \text{otherwise} \end{cases}$$

According to the above equation, a document is said to be similar if it has a similarity of θ or above, $theta$ being our threshold for the assertion. The central statement here lies in computing the θ parameter for each experiment. According to the investigation, we would use cross-validation to split our dataset into test and train classes. In this case θ would predict the training set in each iteration of the tuning (let's say $k = 100$), as well as possible. The distance of θ would be such that it would maximize the number of similar pairs.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left[\sum_{d_p \in pos\_train} min(0, sim(d_p) - \theta)^2 + \sum_{d_p \in neg\_train} min(0, \theta - sim(d_p))^2 \right]$$

We can then over more tuning iterations predict a better and more accurate value of θ.

$$prosim_{k,Rel,f}(G_1, G_2) = \frac{\sum\limits_{\substack{a,b \in Rel(G_1) \\ <a,b> \in C_k(G_1) \cap C_k(G_2)}} f(l(a,b))}{\sum\limits_{\substack{a,b \in Rel(G_1) \\ <a,b> \in C_k(G_1)}} f(l(a,b))}$$

We use a measure called $proSim$ which translates to path relevance overlap similarity (Augenstein et al., 2012). When, $f(l) = 1$ this counts the number of paths irrespective of the length. We do this since unlike the other tasks the graphs generated from the *headnote* are not massive and hence very long and complicated paths are not encountered. Accordingly, we also select the graph with more number of nodes since if the documents are somewhat similar G_1 will absorb the facts conveyed by G_1.

5. Results

We chose to see a more additional correlation, in light of the most limited ways between similar documents. This mirrors our instinct that a more informative structural source catering to two documents means a striking semantic connection and a similar theme between them as well.

The other methods against which the evaluation task has been held is a cosine similarity model - a standard evaluation metric in the domain of corpus-based document evaluation, used as a baseline in many situations. However, the disadvantage of this metric are that this is somewhat based on a bag-of-words (BoW) model. Therefore, it does not take into account the position of the word in the text, semantics, and co-occurrences. Nonetheless, the results of this metric are not very poor because the documents belong to one domain and have similar kind of terminologies mentioned in them; however much complex rules and semantic relations as discussed are not captured which makes this metric not credible concerning finding similarity for legal cases.

Technique	**Accuracy**	**F1-Score**
Our Method	73.17%	0.807
L_{mod}	59.7%	0.718
Cosine Similarity	54.87%	0.53

Table 1: Results from techniques mentioned in Section 5

The other metric is to compare with similar research which had taken place in the Indian legal context and is important research regarding Legal Ontology-based inquiry (Saravanan et al., 2009). The original extract of the legal ontology as mentioned is modified to deliver better results, L_{mod}. The initial extract was a workaround of all legal cases; we reduced the acts to be the Sections and also introduced primary events such as death, retirement, penalty, etc., so that the results are somewhat comparable. The results show that our design technique for the comparable metric is promising, but since the whole idea of designing a specific ontology is to get better results, a modification to a particular use case cannot do justice to the original purpose. We did not choose metrics related to co-citation networks since the dataset has been designed keeping in mind the same systems and hence the comparisons would not hold proper meaning.

6. Conclusion and Future Work

In this particular research, the work is related to validating the concept of the ontology based triples, and the same helping in the assessment of similarity of legal documents.

The number of proceedings positively affected the results concerning differences, and assessing more reports by including more divisions to the Income Tax Act would only heartily approve of the technique in future. Some things that we would want to point out are that the court upholds the law in the best possible way, by the rules defined in the Sections and the corresponding ontology.

On analyzing further, we saw that some inefficiencies in the results were due to some facts being furnished later on in the proceeding after discussions, or that there were some facts which at the time of being provided initially are wrong which is then rectified in the text. With more triples, we can even generate a triple-store for the cases and a query based information retrieval mechanism can help in finding out proper precedent for the situation at hand. Even though there were cases which were complicated, there were also cases which were similar and redundant; this just reflects the inefficiencies of the administration and the lack of information about the law amongst the public.

A natural clarification for the execution of our ontology-based framework is that it gives a knowledge base which has an enormous accumulation of terms and its connections and other related components which are utilized for better upgrades of query terms. Likewise, our basic structure can be extended with the expansion of conditions by including new archives, and from different subdomains in the future course of time.

We would like to conclude by saying, that the results are promising, and more efficient ontological rules across a broader spectrum of legal norms along with more efficient triple alignment techniques can help us further, with not only better document similarity metrics, but also in terms of a legal knowledge graph with untapped potential in terms of applications aiming to find precedents, similar judgments and understanding legal constraints along the way.

7. Bibliographical References

Augenstein, I., Padó, S., and Rudolph, S. (2012). Lodifier: Generating linked data from unstructured text. In *Extended Semantic Web Conference*, pages 210–224. Springer.

Breuker, J., Hoekstra, R., van den Berg, K., Rubino, R., Sartor, G., Palmirani, M., Wyner, A., Bench-Capon, T., et al. (2007). Owl ontology of basic legal concepts (lkif-core).

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Exner, P. and Nugues, P. (2012). Entity extraction: From unstructured text to dbpedia rdf triples. In *The Web of Linked Entities Workshop (WoLE 2012)*, pages 58–69. CEUR.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics.

Hoekstra, R., Breuker, J., Di Bello, M., Boer, A., et al. (2007). The lkif core ontology of basic legal concepts. *LOAIT*, 321:43–63.

Kumar, S., Reddy, P. K., Reddy, V. B., and Singh, A. (2011). Similarity analysis of legal judgments. In *Proceedings of the Fourth Annual ACM Bangalore Conference*, page 17. ACM.

Kumar, S., Reddy, P. K., Reddy, V. B., and Suri, M. (2013). Finding similar legal judgements under common law system. In *International Workshop on Databases in Networked Information Systems*, pages 103–116. Springer.

Minocha, A., Singh, N., and Srivastava, A. (2015). Finding relevant indian judgments using dispersion of citation network. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1085–1088. ACM.

Raghav, K., Reddy, P. B., Reddy, V. B., and Reddy, P. K. (2015). Text and citations based cluster analysis of legal judgments. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 449–459. Springer.

Saravanan, M., Ravindran, B., and Raman, S. (2006). Improving legal document summarization using graphical models. *Frontiers in Artificial Intelligence and Applications*, 152:51.

Saravanan, M., Ravindran, B., and Raman, S. (2009). Improving legal information retrieval using an ontological framework. *Artificial Intelligence and Law*, 17(2):101–124.