Methods for Computing Legal Document Similarity: A Comparative Study

Paheli Bhattacharya [a] Kripabandhu Ghosh [b] Arindam Pal [c] and Saptarshi Ghosh [a]
[a] *Indian Institute of Technology Kharagpur, India*
[b] *Tata Research Development and Design Centre (TRDDC) Pune, India*
[c] *Data61, CSIRO, Australia*

Abstract. Computing similarity between two legal documents is an important and challenging task in the domain of Legal Information Retrieval. Finding similar legal documents has many applications in downstream tasks, including prior-case retrieval, recommendation of legal articles, and so on. Prior works have proposed two broad ways of measuring similarity between legal documents – analysing the precedent citation network, and measuring similarity based on textual content similarity measures. But there has not been a comprehensive comparison of these existing methods, on a common platform. In this paper, we perform the first systematic analysis of the existing methods. In addition, we explore two promising new similarity computation methods – one text-based and the other based on network embeddings – which have not been considered till now.

1. Introduction

In countries following the Common Law system, there are two primary sources of law – Statutes (established laws) and Precedents (prior cases). When a case comes to a legal expert, he has to go through a huge number of legal documents to understand and analyse which ones are relevant to the current case. He prepares his legal reasoning citing or referring these relevant or similar cases. But with the advancement of the Web and huge amount of legal content being made available everyday, it is now becoming intractable for legal practitioners to find these relevant legal documents. This calls for the need of automating the search of similar legal documents. Automatically retrieving similar documents will also help legal academicians who wish to know about an area of law.

In this paper, we focus on the task of computing similarity between two legal documents. The notion of similarity is domain-specific and not completely defined; ultimately, two legal case documents are considered similar if legal experts judge them to be similar. The challenge is to automate this similarity computation.

Existing automatic methodologies for finding similar legal documents can be broadly classified into two categories – (i) network-based methods (e.g., [1]), which rely on citations to prior case documents, and (ii) text-based methods (e.g., [2, 3], which use the content/textual information of the documents. We refer to these two types of similarity measures as *Precedent Citation Similarity* and *Textual Similarity* respectively. But instead of evaluating the methods on a common dataset of legal documents, different prior works have developed their own set of documents. This situation poses a difficulty

in understanding which method is more efficient in finding legal document similarity and why.

In this paper, we aim to bridge this gap. Specifically, we reproduce several existing Precedent Citation Similarity and Textual Similarity methods on a common dataset introduced in [4] (which was also used in our prior work [3]). The dataset contains 47 pairs of Indian Supreme Court case documents, where similarity between each pair of documents is annotated on a scale from $0-10$ by law experts. We compare all the existing methods of finding legal document similarity on this dataset to ensure a fair comparison.

In addition, we propose two new methodologies for deriving legal document similarity – (i) a Precedent Citation Similarity-based method using a recent graph embedding approach (Node2Vec [5]) on the citation network, and (ii) a Textual Similarity method which finds the textual similarity between the different thematic segments (facts, arguments, ratio, ruling etc.) of case documents. We also combine various Precedent Citation Similarities and Textual Similarity approaches and analyse their performance.

To the best of our knowledge, this is the first work that (i) attempts to perform a fair comparison of existing methods for computing legal document similarity, and (ii) introduces additional notions of similarity – one using a network/graph embedding based approach and another using similarities between the different thematic segments in a document.

The rest of the paper is organized as follows. In Section 2, we describe how we prepare the training and test datasets. In Section 3, we explain the existing algorithms for computing similarity between legal documents. In Section 4, we report and analyse the results obtained from our experiments. In Section 5, we summarize the lessons learned in the paper and give some future research directions.

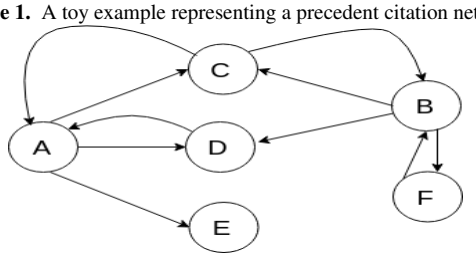
2. Dataset

In this paper, we consider legal judgments from the Supreme Court of India, crawled from the website of Thomson Reuters Westlaw India (`http://www.westlawindia.com`). We crawled $53,210$ case documents in total. Note that we use only the publicly available full text of the judgment. All other proprietary information had been removed before performing the experiments described in this paper.

Constructing a citation network: We construct the prior-case citation network of this document set to compute the Precedent Citation Similarity. The vertices of the network are the case documents. A directed edge exists between two vertices i and j if document i cites document j in its text. Consider an example graph shown in Figure 1. In our example, an edge exists from vertex A to E since A cites E. There were 53,210 nodes and 208,921 edges in total in this citation graph.

Data for Evaluation: For evaluation of various legal document similarity methods, we consider the dataset introduced in [4] (which was also used in our prior work [3]). It contains 47 pairs of Indian Supreme Court case documents. Each pair is assigned a score on a scale of $0-10$ where 0 represents that the document pair is not similar and 10 represents that the pair has maximum similarity. To ensure a fair comparison, we use the same dataset for comparing all the methods explored in this paper.

Figure 1. A toy example representing a precedent citation network.



3. Methods for Legal Document Similarity

In this section we describe the basic ideas behind the existing methodologies we reproduce in this work. As stated earlier, the existing methodologies can be broadly classified into two categories – (i) methods for precedent citation similarity, and (ii) methods for textual similarity.

3.1. Methods for Precedent Citation Similarity

We explain the Precedent Citation Similarity metrics using the running example in Figure 1, where we want to measure the similarity between vertices A and B. Apart from three existing network-based similarity measures, we also describe an additional Precedent Citation Similarity method which we newly explore in this work (Node2Vec).

- **Bibliographic Coupling [1]** : It is defined as the number of common precedent cases cited by a document pair, i.e, number of common out-citations. In the example graph, the set of out-citations of A is $A_{out} = \{C,D,E\}$ and the set of out-citations of B is $B_{out} = \{C,D,F\}$. The common out-citation is $A_{out} \cap B_{out} = \{C,D\}$. So, the bibliographic coupling based similarity $sim(A,B) = |A_{out} \cap B_{out}| = 2$. This value is normalized by the total number of distinct out-citations of A and B, which in this case is $|A_{out} \cup B_{out}| = 4$.
- **Co-citation [1]** : This metric is defined as the number of common incoming citations to each document of the document pair. In the example graph, the in-citations to A, come from the set $A_{in} = \{C,D\}$ and in-citations of B come from $B_{in} = \{C,F\}$. The common in-citation is $A_{in} \cap B_{in} = \{C\}$. So, the co-citation based similarity $sim(A,B) = |A_{in} \cap B_{in}| = 1$. This value is normalized by the total number of distinct in-citations to A and B, which in this case is $|A_{in} \cup B_{in}| = 3$.
- **Dispersion [6]** : Dispersion was originally used to identify romantic relationships in the Facebook network. In the context of legal document similarity, it aims to find to what extent the neighbours (out-citation documents) of two documents are themselves similar (occurs in the same community/cluster). We use the *NetworkX* implementation for this measure.[1]
- **Node2Vec**: We explore a novel approach based on graph embeddings on the same prior-case citation network on which the above metrics have been implemented. Node2Vec [5] is a state-of-the-art algorithm for learning embeddings of nodes in a homogeneous network (a network having same type of nodes, e.g. a network

[1] `https://networkx.github.io/documentation/networkx-1.9/reference/generated/networkx.algorithms.centrality.dispersion.html`

of documents citing each other, friendship network of people etc.). Given such a network, Node2Vec aims to map the vertices/nodes of the graph to a vector space such that nodes having similar neighbourhoods in the network have similar embeddings/representations. Given the vector representation of the vertices, we can compute cosine similarity between the vectors to understand the similarity between the two vertices. For our running example, if the vector of vertex A as given by Node2Vec is represented as $\vec{A}$ and vector of vertex B is $\vec{B}$, then $sim(A,B) = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\|\|\vec{B}\|}$.

3.2. Methods for Textual Similarity

Now we describe methods for computing textual similarity between legal documents. Along with reproducing methods from prior works, we also propose a new method for textual similarity, namely the similarity between the thematic segments of the case documents.

- **Paragraph Links [4]** : In this measure of similarity, a network is formed in which the nodes/vertices are the paragraphs of the documents. Links/edges are established between two vertices/paragraphs, if the similarity between the paragraphs (as measured by TF-IDF) is above a particular threshold. To measure the similarity of the two documents, bibliographic coupling on the above network is calculated.
- **FullText Similarity [3]**: Similar to Node2Vec, Doc2Vec [7] represents a whole document in a vector space. The vectors preserve the semantics of the document, such that semantically similar documents have similar vector representations. In our prior work [3] we explored a wide range of legal document representation methods (e.g., whole document, a summary of the document, reason for citation of prior-judgments) and also state-of-the-art techniques (e.g., TF-IDF, topic models, word embeddings and document embeddings) to calculate document similarity between various representations. We observed that document embedding using Doc2Vec on the whole document gave the best result in computing legal document similarity. In this paper, we utilise this technique developed in [3]. We train Doc2Vec on a large set of Indian Supreme Court case documents (the 53,201 documents stated in Section 2, excluding the documents in the dataset for evaluation). We then use the learned model to infer the vectors of the documents of the 47 pairs in the evaluation dataset. We then compute cosine similarity between the vectors of the documents to find the similarity.
- **Thematic Similarity**: It is known that a legal case document contains various themes/segments/functional parts like Facts, Arguments, Ratio of the decision, Final judgment and so on. While judgments of many countries reflect this segmented structure through section headings, Indian legal judgments are devoid of any such systematic structure. We recently proposed a Machine Learning-based method for thematic segmentation of Indian Supreme Court Case documents [8]. The method (based on deep neural networks) has been shown to work well for 7 rhetorical roles/themes (Facts, Arguments, Ratio of the decision, Statute, Precedent, Ruling by Lower Court, and Ruling by Present Court) over documents from 5 popular legal domains. The implementation of this segmentation method is publicly available at `https://github.com/Law-AI/semantic-segmentation`.

Table 1. A sample of the dataset showing the expert score and the similarity inferred by a subset of methods explored in this paper. The values nearest to the expert score are bold-faced.

Document Pair	Expert Score	Node2Vec	FullText Similarity [3]	Thematic Similarity (Avg)	Thematic Similarity (Max)
1992_47 & 1992_76	0	0.195	0.188	**0.154**	0.571
1979_110 & 1989_233	3	0.613	0.465	0.104	**0.415**
1953_24 & 1957_52	7	0.234	0.264	0.377	**0.757**
1983_37 & 1979_33	10	0.574	**0.711**	0.209	0.692

We use the same method to segment the two documents for which we want to calculate the similarity. After getting the segments (Fact, Argument, Ratio, etc.) from both the documents, segment-level similarities (e.g., Fact-Fact similarity, Argument-Argument similarity, Ratio-Ratio similarity, and so on) between the two documents are computed. Then an aggregated similarity is reported as the final similarity between the document pairs. For aggregation we try the following two ways – (i) *max*: the maximum similarity value between any segment of the document pairs is considered as the overall similarity, and (ii) *average*: the average similarity across the segments is considered as the overall similarity of the document pair.

In this section, we have described several Precedent Citation Similarity measures and several Textual Similarity measures for computing the similarity between two legal documents. We will compare these methodologies (and also their combinations) in the next section.

4. Results and Analysis

We now compare the performances of the various methods stated in the previous section, over the evaluation dataset of 47 document pairs (stated in Section 2).

4.1. Method for evaluation

Recall from Section 2 that the evaluation dataset contains 47 document pairs, where each pair has a similarity score in the range $[0, 10]$ that is assigned by legal experts. To judge the performance of a particular similarity method, we use the Pearson Correlation coefficient between the expert scores and the computationally obtained similarities from the said methodology.

Table 1 shows a sample of the evaluation framework. For each document pair, an expert similarity score in the range $0 - 10$ is available. Each of the methodologies Node2Vec, FullText Similarity, Thematic Similarity, etc assigns a similarity value to the same document pairs. We then compute an overall Pearson Correlation coefficient for each of the methodologies.

4.2. Analysis

Table 2 shows the results (Pearson Correlation) for each method. We find that among the Precedent Citation Similarity based methods, the graph embedding approach, Node2Vec, performs the best. Legal citation networks are very sparse. Only a few cases are cited for each document. Results suggest that simpler measures like co-citation and dispersion

Table 2. Pearson Correlation between the Expert Score and Similarity inferred by each method

Category	Method	Correlation
Prior-case citation network based measures	Bibliographic Coupling [1]	0.443
	Cocitation [1]	0.205
	Dispersion [6]	0.246
	Node2Vec	**0.487**
Text based measure	Paragraph Links [4]	0.33
	FullText Similarity [3]	**0.605**
	Thematic Similarity (avg)	0.523
	Thematic Similarity (max)	**0.599**

Table 3. Correlation Results after combining Precedent Citation Similarity methods and FullText Similarity

Combination	Methods Combined	Correlation
Max	Biblio + FullText Sim	**0.626**
	Cocitation + FullText Sim	0.582
	Dispersion + FullText Sim	0.600
	Node2Vec + FullText Sim	0.595
Average	Biblio + FullText Sim	0.575
	Cocitation + FullText Sim	0.432
	Dispersion + FullText Sim	0.507
	Node2Vec + FullText Sim	0.552

perform poorly in such scenarios. On the other hand, graph embedding based method could perform much better in comparison.

Among Textual Similarity measures, we find that document embedding on the full text performs the best.[2] Thematic similarity combined using the *max* shows very comparable performance. This observation suggests that even textually, there are certain themes in which two documents can be more similar than in others; for instance, two documents may be very similar in terms of their facts but different in their final judgment or arguments. It depends on the perspective through which a legal document being similar or dissimilar is judged. In our framework we find that the documents are most similar in Ratio (correlation 0.45), Precedent (0.42) and Fact (0.37). Detailed results are omitted due to space constraints.

Combining Textual Similarity and Precedent Citation Similarity: We also attempt to combine the Textual and Precedent Citation similarities. We combine FullText Sim (obtained using [3]) and Thematic Similarity – the two best-performing textual similarity measures – with the various Precedent Similarity measures. The results are shown in Table 3 (for combining FullText similarity) and Table 4 (for combining Thematic similarity) respectively. We try two aggregation functions namely *max* and *average*. Suppose for a document pair A and B, the network based similarity is s_1 and text based similarity is s_2. Then the aggregated similarity using *max* is $max(s_1, s_2)$. The aggregated similarity using *average* is $\frac{1}{2}(s_1 + s_2)$.

[2]Note that the prior work [3] reported a slightly different Pearson correlation value than what we are reporting here (for the same evaluation dataset), because the training data for the Doc2vec model is different in this work.

Table 4. Correlation Results after combining Precedent Citation Similarity methods with Thematic Similarity

Combination	Methods Combined	Correlation
Max	Biblio + Thematic Sim	0.604
	Cocitation + Thematic Sim	0.560
	Dispersion + Thematic Sim	0.586
	Node2Vec + Thematic Sim	0.530
Average	Biblio + Thematic Sim	0.565
	Cocitation + Thematic Sim	0.416
	Dispersion + ThematicSim	0.485
	Node2Vec + Thematic Sim	**0.615**

We find that the *max* performs consistently well in Table 3. The best performance is observed when we combine the bibliographic coupling based precedent citation similarity with the FullText similarity. On the other hand when we combine Thematic Similarity with the Precedent Citation Similarity measures, the average combination with the graph embedding based approach performs the best. In both cases, we get a higher correlation with expert scores by combining Textual Similarity and Precedent Citation Similarity, than what we obtained by using one type of methodology alone.

These results show that both textual similarity and precedent citation similarity are helpful in measuring the similarity between legal case documents.

5. Conclusion and Future Work

In this paper, we attempt to compare existing approaches for legal document similarity. We perform a systematic comparison of the methods, on a set of 47 document pairs. We also explore two new methods, one based on graph embeddings, and the other based on textual similarity between thematic segments.

We show that understanding legal document similarity is indeed a challenging task, and has various different facets, such as similarity in precedent citation, similarity in textual content, similarity of themes, and so on. Multiple facets of document properties contribute to the overall similarity of the documents. It may not be possible to capture all of them by a single methodology; rather, an aggregation of the perspectives is necessary.

There are several future directions of the work reported in this paper. One immediate future work is to develop better methods for judging similarity of legal case documents, which would agree more closely with the opinion of the legal experts. A detailed inspection of document-pairs that the experts judge to be highly similar may reveal other factors that need to be considered while judging the similarity of such documents. In this context, it is important to consider explainability of the similarity computation – ideally, experts can be asked to explain how they judge similarity between two legal documents, and automatic methods may be developed based on these explanations. We plan to explore these directions in future.

References

[1] S. Kumar, P. K. Reddy, V. B. Reddy, and A. Singh, "Similarity analysis of legal judgments," in *Proc. Annual ACM India COMPUTE Conference*, pp. 17:1–17:4, ACM, 2011.

[2] J. Landthaler, E. Scepankova, I. Glaser, H. Lecker, and F. Matthes, "Semantic Text Matching of Contract Clauses and Legal Comments in Tenancy Law," in *IRIS: Internationales Rechtsinformatik Symposium*, 2018.

[3] A. Mandal, R. Chaki, S. Saha, K. Ghosh, A. Pal, and S. Ghosh, "Measuring similarity among legal court case documents," in *Proc. Annual ACM India COMPUTE Conference*, pp. 1–9, 2017.

[4] S. Kumar, P. K. Reddy, V. B. Reddy, and M. Suri, "Similar legal judgements under common law system," in *International Workshop on Databases in Networked Information Systems*, 2013.

[5] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, 2016.

[6] A. Minocha, N. Singh, and A. Srivastava, "Finding relevant indian judgments using dispersion of citation network," in *Proc. International Conference on World Wide Web*, pp. 1085–1088, 2015.

[7] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proc. International Conference on Machine Learning*, 2014.

[8] P. Bhattacharya, S. Paul, K. Ghosh, S. Ghosh, and A. Wyner, "Identification of rhetorical roles of sentences in indian legal judgments," in *Proc. International Conference on Legal Knowledge and Information Systems (JURIX)*, 2019.