Legal Knowledge and Information Systems
J. Savelka et al. (Eds.)

doi:10.3233/FAIA241256

Combining Network and Text to Provide Legal *Pincites*

Nicolas GARNEAU [a], Henrik PALMER OLSEN [b,1], Antoine CORDUANT [c] and Fabien TARISSAN [c]

[a] *Computer Science Department, University of Copenhagen*
[b] *University of Copenhagen and University of Bergen*
[c] *Université Paris-Saclay, CNRS, ENS Paris-Saclay, ISP, France*

Abstract. The task of legal precedent retrieval is essential yet challenging for legal professionals, as it involves identifying relevant past cases that can inform current legal decisions. Building on previous work that integrates citation networks and text similarity analysis, we apply these techniques to a dataset comprising paragraphs from cases decided by the Court of Justice of the European Union (CJEU). While paragraph citation retrieval is way more challenging than case citation retrieval, we show that a careful combination of network and text signals improves computational efficiency without sacrificing performances. More precisely, our experiments first reveal the limitations of network analysis at the paragraph level due to the sparse connectivity of the data. We then explore a novel approach to this task by combining network analysis at the case level and natural language processing at the paragraph level, which we refer to as "pincites'.

Keywords. case law, network analysis, semantic analysis, link prediction, pincites

1. Introduction

One major task of legal professionals who has to prepare formal decisions in law is to know what precedents are relevant to the new case at hand. However finding relevant precedent may be time-consuming and therefore expensive, which has led to efforts in using various computational approaches to retrieve relevant precedent. In this paper, we set out to test a recently proposed approach to precedent retrieval (Bhattacharya et al., 2022), which combines network analysis and natural language processing i.e. semantic similarity. Much prior citation network analysis considers citations only to prior cases (also called precedents). See for example Kumar et al. (2011), Minocha et al. (2015) and Mones et al. (2021). These approaches miss important signals inherent in Statutes (written laws of a jurisdiction) and Bhattacharya et al. (2022) attempt to remedy this by augmenting a Precedent network with a heterogeneous network of Statutes, creating a bipartite network. They also introduce the use of text-based similarity information from a Doc2Vec model (Le and Mikolov, 2014) trained over legal case documents which they combine with the bi-partite network mentioned above. In this paper, we similarly com-

[1] Corresponding Author: Henrik Palmer Olsen, hpo@jur.ku.dk.

bine text and network information, but contrary to Bhattacharya et al. (2022) however, we focus on paragraph level retrieval, which, as explained in Palmer Olsen et al. (2023), is more closely tailored to real-world retrieval needs (both the court itself and litigating lawyers cite specific paragraphs in cases - not whole cases), but also more difficult. Moreover, for the present experiment, we do not include information from statutory law to enhance the citation network analysis. We aim to isolate the impact of combining network and text signals on precedent information alone. We have two reasons for this. First of all, the dataset we consider is the one introduced by Palmer Olsen et al. (2023). It is well known that much of CJEUs case law relates to general principles of law that are not specific to a single identifiable statutory provision but have a much broader scope. Secondly, Palmer Olsen et al. (2023) has shown that the keywords that are listed as metadata for each CJEU judgment do not align well with the positive results from their base model. Keywords are supposed to represent the main legal content of the cases in which a judgment is rendered and are added as information to the judgment in the EU legal information system EUR-Lex. They found that judgment paragraphs that were similar in regards to their legal content, could belong to cases that varied quite significantly in regards to the keywords that were attached to the different cases from which the paragraphs originated. This indicates that even if keywords are correctly assigned to cases, there can still be a discrepancy between paragraph similarity and case similarity and that case-level information about citations to statutory provisions may create more noise than signal when the task is paragraph retrieval.

2. Applying the Mones et al. (2021) Method at the Paragraph Level

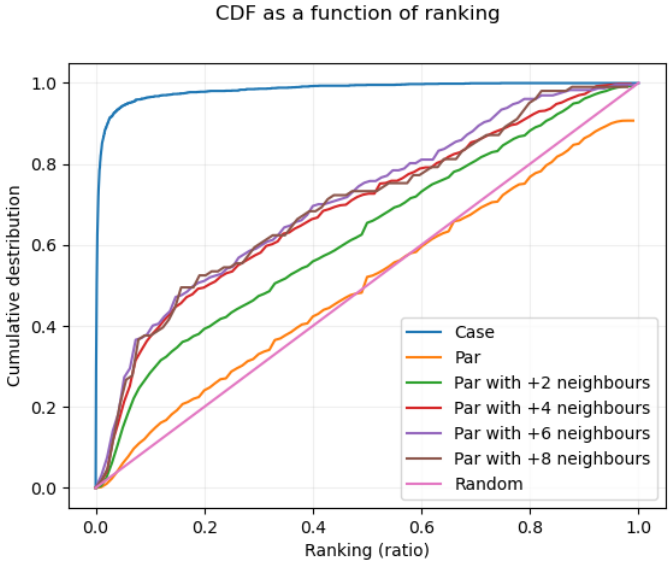
As presented in the introduction, the approach adopted for studying citation networks is based on their representation as graphs. In this model, paragraphs are nodes of the graph, while the edges indicate citation relationships from one to another. To predict citations between paragraphs, the chosen method relies on the characteristics of these links (e.g. who cites whom? Is the cited paragraph important in the studied jurisprudence? etc.). Similarly, one can examine the characteristics between two paragraphs that are not connected. By learning the differences in the characteristics of these two cases (connected and unconnected), a model can determine whether a citation link exists or not. The six features used in our model to predict if a link exists between paragraphs A and B are the same as used in Mones et al. (2021): Time difference; TF-IDF; Preferential attachment; Adamic-Adar; common neighbors and Common referrers. Again, similar to Mones et al. (2021), we used a Random Forest as the classification model and use the dataset introduced by Palmer Olsen et al. (2023) to apply it at the paragraph level. This dataset presents an interesting challenge where paragraphs are weakly connected, compared to cases. For the training dataset, we used each possible pair of nodes (i.e., $\frac{n\times(n-1)}{2}$ combinations, with n being the number of nodes). We used paragraphs before 2018 as training, and 2018 and onwards as testing. It is important to note that the features are calculated without considering the possible link between the two nodes, as doing so would introduce bias into certain characteristics. This dataset is extremely large, containing billions of combinations for the paragraph graph. A remark can be raised regarding the fact that the vast majority of these combinations describe non-existent links, as only 1 in 20,000 links are actual citations, compared to 1 in 1,000 for the cases. We follow the same evaluation

procedure as Mones et al. (2021) to evaluate the model's performance in the task of predicting citation links, we rely on the model's raw output score to perform a ranking evaluating the performance of predicting an existing link. This method allows the model to be evaluated not only on its global performance beyond the top ranking but also provides the opportunity to implement a sort of recommendation algorithm for paragraphs. To visualize these results, we use the cumulative distribution function (CDF) of the scores for all the nodes in the graph, allowing for a quick identification of the proportion of nodes with a score below a given value. The rankings are, in these visualizations, expressed as a ratio to allow comparison between rankings with a different number of competitors. Results are presented in Figure 1. The cumulative score distribution curve clearly shows the difference, perceptible through the ROC curve, between the model trained on the cases (Mones et al., 2021) and this new model trained on the paragraphs. Despite an encouraging initial peak, the latter seems incapable of outperforming randomness. A plausible explanation for this result can be deduced from the analysis of the datasets used, highlighting a significant distinction between the two graphs, primarily in their density. The median in-degree is also very enlightening. While this value is 8 for in the cases' graph, it drops to only 2 at the paragraph level. Similarly, the top 5% cited cases have 31 or more citations while the top 5% cited paragraphs have only 7 or more references. Since the features used to predict the links are mainly computed based on the network structure, few citations make it quite challenging to obtain an effective classifier. This is even strengthened by the evaluation method that remove links for testing predictions, leading 37% nodes to have 0 citations. To clarify these observations, Figure 1 presents the baseline results on the cases as a whole, accompanied by curves representing the cumulative distribution of scores for nodes with more than 2, 3, 5, or 8 neighbors. This approach aims to limit the classifier to cases where it is more likely to have significant features. It is observed that for these higher-degree nodes, the model demonstrates an improvement over the general case. However, despite this improvement, its performance remains significantly lower than that of the model classifying links between cases. This leads us to conclude that network analysis at the paragraph level remains significantly lower than at the case level.

3. Modifying the Bhattacharya et al. (2022) method for the Paragraph Level

Since we could not obtain satisfying results using network analysis at the paragraph level, we hereby combine network analysis at the case level (Mones et al., 2021) and semantics at the paragraph level (Westermann et al., 2021; Palmer Olsen et al., 2023), using SimCSE (Gao et al., 2021) and LexLM (Chalkidis et al., 2023), thus modifying the Bhattacharya et al. (2022) method. We introduce two ways of doing it. The first one uses the case-to-case link prediction model introduced by Mones et al. (2021) to *filter* the top-k cases that could potentially be cited by case A. We selected three different values of k being either 100, 500, and 1000. This is motivated by their finding that with just as few as 100 cases, they obtained a cumulative distribution function of around 95%. Then, as a second step, we apply the paragraph-to-paragraph retrieval method of Palmer Olsen et al. (2023) to select the top n paragraph amongst the selected cases. The second method *weights* the similarity of a particular paragraph by multiplying the probability of a given case to be cited provided by the case-to-case model to perform a re-ranking over all the

Figure 1. Cumulative distribution funtion (CDF) of the scores on the cases, the paragraphs, and the paragraphs with a higher degree (2, 4, 6, 8).



possible paragraphs in the test set. It is important to note that there is a lot of computation involved, either at the case or the paragraph level. The paragraph-level computation is mitigated in the first method by using the top-k cases predicted by the case-to-case model. However, the model proposed by Mones et al. (2021) uses a huge dataset which is heavily imbalanced (1 for 1000). Hence, we introduce a variant of this model where we sample their training set to obtain a class imbalance of only 1 for 100 reducing considerably the training time for this step. We follow the evaluation protocol of Palmer Olsen et al. (2023)[2] to evaluate the different scenario, which uses the Mean Average Precision (MAP) as the target metric. Results are displayed in Table 1. We can see that using the model trained on the full dataset to weigh the similarity of the paragraphs does not perform well, especially compared to the sampled version. The method that filters before computing paragraph similarity offers interesting performance compared to the original model proposed by Palmer Olsen et al. (2023). With only 100 selected cases, we achieve 0.294 MAP, while using up to 1,000 selected cases we achieve a very close performance of 0.314. On average, cases in this dataset have around 10 relevant paragraphs. Concretely, this means that for each of the selected k, we considered on average 1,000, 5,000, and 10,000 paragraphs respectively. The original version of Palmer Olsen et al. (2023) uses the 10,000 cases, which compares 100,000 paragraphs every time.

[2]The original dataset in Palmer Olsen et al. (2023) contains paragraphs with their associated number. We found out it leaked important information for the link prediction evaluation since the citing paragraph is often referring to the cited one by its number. After removing this, we obtained lower performance (0.325 instead of 0.489).

Table 1. Mean Average Precision on the test set of the dataset introduced by Palmer Olsen et al. (2023). We compare the incorporation of the case-to-case Random Forest classifier using either the full, imbalanced dataset, or a sub-sampled one. We also analyze how weighing or filtering the cases impacts the Mean Average Precision overall. We reproduced the original results in the last column.

	Weighted	**Filtered**			
		100	**500**	**1K**	Palmer Olsen et al. (2023)
Random Forest – Full	0.144	0.294	0.311	0.313	0.325
Random Forest – Sampled	0.276	0.294	0.311	0.314	

4. Conclusion and Future Research

Building a paragraph level recommender for case law retrieval remains an important and difficult task. Paragraph level networks are much more fragmented than case level networks, leading to significantly lower prediction precision for paragraph level prediction than for case level prediction. We make the following two observations in relation to paragraph retrieval:

1. Incorporating network information does not necessarily improve paragraph retrieval for our particular dataset compared to Palmer Olsen et al. (2023).
2. However, using only a 10th of the cases does not necessarily degrade the results. There are serious computational gains to be made there.

Compared to previous research it is noteworthy that the increase in performance gained by combining network and language in the context of case level retrieval (Bhattacharya et al., 2022) is not achieved in paragraph level retrieval (on our dataset). We also note that it is computationally expensive to calculate predictions based on network analysis compared to training a recommender with SimCSE. Still we believe there may be other ways to achieve better results. We suggest that future research should replace the TF/IDF feature used in the random forest classification model with a trained classifier. We hypothesize that this would provide a stronger language signal in the classifier and thereby help overcome the highly fragmented profile of the network. We also suggest that a re-ranking where more recent paragraphs and paragraphs with a higher in-degree is given more weight could be a way to improve predictive accuracy. Finally it is worth noting that case law retrieval is not only about being able to correctly predict paragraph-to-paragraph citations in a dataset of legal judgments. Being able to correctly identify a link between two historical paragraphs is ultimately an instrument towards being able to identify what paragraphs may be relevant to cite in a new case that has not yet been decided. The aim, in other words, is to be able to identify relevant sources of law (precedent) given a new legal problem. It will be most relevant therefore to not only focus on link finding optimisation, but also to focus on link finding relevance (validity and suitability); for example, lawyers will also be interested in knowing whether a recommended paragraph is still valid law or whether it has been overturned. Studying the legal content of recommended paragraphs, will therefore, from a legal perspective, be more relevant in assessing the performance of a paragraph recommender system, than the Mean Average Precision score in isolation.

References

Bhattacharya P, Ghosh K, Pal A, Ghosh S. Legal case document similarity: You need both network and text. Information Processing & Management 2022;59(6):103069. https://www.sciencedirect.com/science/article/pii/S0306457322001716.

Chalkidis I, Garneau N, Goanta C, Katz D, Søgaard A. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) Toronto, Canada: Association for Computational Linguistics; 2023. p. 15513–15535. https://aclanthology.org/2023.acl-long.865.

Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 6894–6910. https://aclanthology.org/2021.emnlp-main.552.

Kumar S, Reddy PK, Reddy VB, Singh A. Similarity analysis of legal judgments. In: Proceedings of the fourth annual ACM Bangalore conference; 2011. p. 1–4.

Le Q, Mikolov T. Distributed representations of sentences and documents. In: International conference on machine learning PMLR; 2014. p. 1188–1196.

Minocha A, Singh N, Srivastava A. Finding relevant indian judgments using dispersion of citation network. In: Proceedings of the 24th international conference on World Wide Web; 2015. p. 1085–1088.

Mones E, Sapieżyński P, Thordal S, Olsen HP, Lehmann S. Emergence of network effects and predictability in the judicial system. Scientific reports 2021;11(1):2740.

Palmer Olsen H, Garneau N, Panagis Y, Lindholm J, Søgaard A. Re-Framing Case Law Citation Prediction from a Paragraph Perspective. In: Legal Knowledge and Information Systems IOS Press; 2023.p. 323–328.

Westermann H, Savelka J, Benyekhlef K. Paragraph similarity scoring and fine-tuned BERT for legal information retrieval and entailment. In: New Frontiers in Artificial Intelligence: JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers 12 Springer; 2021. p. 269–285.