Advancing Legal Information Extraction: A GNN-Based Retrieval Attractor Framework for Legal PDF Documents

Vitória S. Santos [1], Carina F. Dorneles[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88.040-900 – Florianópolis – SC – Brazil

vit2santoss@gmail.com, dorneles@gmail.com

Abstract. *Legal Information Extraction (IE) from PDF documents remains challenging due to the structural complexity of legal texts. Although Large Language Models (LLMs), Graph Neural Networks (GNNs), and Retrieval-Augmented Generation (RAG) have advanced legal document processing, retrieval approaches remain sensitive to retrieval noise and contextual misalignment. This doctoral research proposes a retrieval framework based on query-aware dynamical systems over graph representations of legal documents. The approach introduces retrieval attractors, defined as stable semantic regions emerging through iterative graph propagation, and investigates retrieval under Euclidean and Hyperbolic representations. Preliminary results suggest that graph-based retrieval attractors are a promising direction for legal information extraction.*

Resumo. *A Extração de Informação (EI) em documentos jurídicos em formato PDF permanece desafiadora devido à complexidade estrutural desses textos. Embora Grandes Modelos de Linguagem (LLMs), Redes Neurais em Grafos (GNNs) e Retrieval-Augmented Generation (RAG) tenham avançado o processamento de documentos jurídicos, abordagens de recuperação ainda apresentam sensibilidade a ruídos e desalinhamento contextual. Esta pesquisa de doutorado propõe um framework de recuperação baseado em sistemas dinâmicos guiados por consulta sobre representações em grafos de documentos jurídicos. A abordagem introduz atratores de recuperação, definidos como regiões semânticas estáveis emergentes de propagação iterativa em grafos, e investiga recuperação em representações Euclidianas e Hiperbólicas. Resultados preliminares sugerem que atratores baseados em grafos representam uma direção promissora para extração de informação jurídica.*

Informações Gerais. Nível: Doutorado. **Aluno:** Vitória Soares dos Santos. **Orientador:** Carina Friedrich Dorneles. **Programa:** Programa de Pós-Graduação em Ciências de Computação da Universidade Federal de Santa Catarina. **Admissão:** Agosto 2023. **Exame de qualificação:** Dezembro 2025. **Defesa prevista:** Fevereiro 2028. **Etapas concluídas:** Disciplinas obrigatórias, Seminário de Andamento do Doutorado e Exame de Qualificação do Doutorado.

1. Introduction

Information Extraction (IE) is a fundamental task in Natural Language Processing (NLP) that aims to automatically identify and structure relevant information from unstructured or semi-structured text [Singh 2018, Jurafsky and Martin 2023, Zhang et al. 2022, Manning et al. 2021]. In the legal domain, IE is particularly important due to the large volume of textual information produced daily, including judicial decisions, legal proceedings, contracts, and administrative documents. Efficient extraction mechanisms can support legal analytics, intelligent search, information integration, and decision support systems. From a Database perspective, IE also enables the transformation of unstructured legal documents into structured and queryable repositories, supporting downstream applications such as retrieval, knowledge discovery and Business Intelligence systems

The rapid progress of Natural Language Processing (NLP) has enabled the integration of deep learning models such as Large Language Models (LLMs) and Graph Neural Networks (GNNs), together with Retrieval-Augmented Generation (RAG) techniques, to enhance information extraction from PDF documents. Transformer-based architectures such as BERT and domain-specific adaptations like LEGAL-BERT have substantially advanced legal text processing tasks [Devlin et al. 2019, Chalkidis et al. 2020]. Despite these advances, legal PDF documents remain particularly challenging due to their structural and semantic complexity. Unlike standard textual formats, legal documents frequently contain multi-column layouts, footnotes, embedded tables, nested clauses, and extensive cross-references [Bommarito and Katz 2022]. Furthermore, legal reasoning often depends on long-range dependencies and hierarchical semantic structures that conventional retrieval strategies struggle to represent effectively.

Although RAG-based approaches improve contextual grounding through external retrieval, their effectiveness strongly depends on retrieval quality. Retrieval errors frequently propagate incomplete or irrelevant information into the generation stage, leading to hallucinations and unreliable outputs [Nguyen and Satoh 2024]. Similarly, Graph Neural Networks have shown promise in modeling structural dependencies among legal entities and documents, but challenges remain regarding graph construction complexity, computational cost, and scalability [Tang et al. 2024a, Tang et al. 2024b, Li et al. 2023].

Moreover, hyperbolic neural representations suggest that non-Euclidean spaces are better suited for modeling hierarchical and multi-relational structures than conventional Euclidean geometry [Tifrea et al. 2021, He et al. 2025]. Due to the exponential volume growth of negatively curved spaces, hyperbolic representations can preserve hierarchical relationships with lower distortion and have shown promising results in graph learning and retrieval tasks [Chami et al. 2019, Chen et al. 2021, Balažević et al. 2019]. Since legal documents naturally exhibit nested clauses and extensive cross-referenced structures, they present characteristics that may benefit from hierarchical geometric representations.

To address the limitations of current retrieval approaches while exploring these geometric opportunities, this research proposes a novel retrieval framework grounded in query-aware dynamical systems over graph representations of legal documents. The proposed approach introduces retrieval attractors, in which stable node representations emerge through iterative propagation within graph neural architectures. Retrieval is modeled as a dynamic process capable of progressively refining contextual information through graph interactions, aiming to improve retrieval robustness, contextual coherence, interpretability, and downstream information extraction quality. More specifically, this work investigates whether graph-based retrieval attractors can improve information extraction from legal PDF documents by modeling retrieval as a query-aware dynamical system over graph representations of document chunks. Additionally, the framework investigates retrieval behavior under both Euclidean and hyperbolic representations by transforming Euclidean semantic embeddings into the Poincaré space, enabling the analysis of how geometry influences semantic organization and retrieval effectiveness.

The main contributions of this doctoral research are summarized as follows:

- A novel graph-based retrieval attractor framework integrating query-aware propagation dynamics to improve retrieval robustness and contextual coherence.
- A theoretical formulation connecting graph retrieval mechanisms, dynamical systems, and semantic stability.
- An investigation of retrieval behavior under both Euclidean and hyperbolic similarity spaces for legal document processing.
- An empirical analysis of retrieval limitations in existing legal information extraction approaches.

2. Proposed Solution: GNN-Based Retrieval Attractor Framework

We introduce a retrieval framework inspired by the mathematical theory of *dynamical systems*, which describe how the state of a system evolves through repeated applications of an update rule [Strogatz 2015]. A central concept in this theory is the notion of an *attractor*, representing stable configurations toward which a system converges under iterative dynamics [Hale 1980]. Inspired by this concept, retrieval is modeled as a graph dynamical process in which document representations progressively evolve through semantic interactions.

Each legal document is divided into semantically meaningful chunks represented as nodes in a graph. Edges encode semantic similarity relationships among text segments, while the query is also represented as a node connected to all chunks. Through iterative propagation, node representations evolve according to both local graph structure and query influence, forming a dynamic system of semantic interactions.

Within this evolving structure, we define *retrieval attractors* as nodes whose representations stabilize after successive propagation steps using a Graph Neural Network (GNN). Unlike conventional retrieval approaches based solely on direct query similarity, attractors emerge from the interaction between semantically related nodes and query-aware reinforcement mechanisms. Consequently, retrieval prioritizes semantically coherent and contextually grounded regions of text. Figure 1 illustrates the proposed framework.

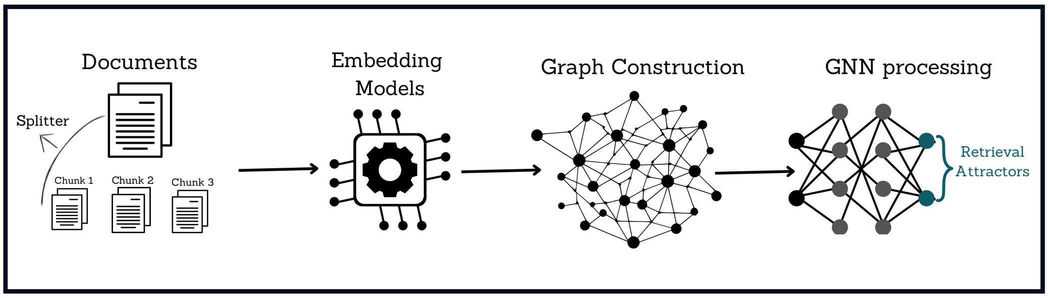

Figure 1. GNN-Based Retrieval Attractor Framework.

Formally, node representations are iteratively updated according to:

$$x_i^{(t+1)} = \sigma\left(\sum_{j\in N(i)} v_{ij}^{(q)} x_j^{(t)} + w_{iq} q\right),$$

where $x_i^{(t)}$ denotes the representation of node i at iteration t, and q is the embedding of the query. The coefficient $v_{ij}^{(q)}$ is a query-aware edge weight that modulates the contribution of neighbor j to node i, explicitly conditioned on the semantic relationship between each chunk and the query, and the term w_{iq} represents the direct influence of the query on node i. The map $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ is a non-linear activation function applied element-wise to ensure controlled propagation of information.

This recursive update defines a graph dynamical process in which node representations progressively evolve through semantic interactions. A node is considered a retrieval attractor when its representation stabilizes under iterative propagation, that is,

$$\|x_i^{(t+1)} - x_i^{(t)}\| < \varepsilon, \quad \forall t \geq t_0,$$

where ε denotes a convergence threshold and t_0 represents an iteration after which the representation remains stable. This criterion identifies nodes whose representations converge under graph propagation and encode contextually reinforced, query-relevant information.

To investigate the influence of geometry on retrieval behavior, the proposed framework operates under both Euclidean and hyperbolic representations. Hyperbolic retrieval is implemented through the Poincaré ball model, where Euclidean semantic embeddings are transformed into hyperbolic representations. This formulation enables the investigation of whether hierarchical structures commonly observed in legal documents can be represented more effectively through negatively curved spaces within the retrieval attractor framework.

3. Comparative Analysis of GNN-Based Methods

Following the formulation of the proposed retrieval attractor mechanism, we compare our framework with representative graph-based retrieval approaches from legal information extraction and retrieval literature. Since our proposal is grounded in query-aware propagation dynamics and semantic stability, the comparison focuses on mathematical formulations and key differences relative to our attractor-based update rule.

Table 1 summarizes representative approaches, contrasting their update equations and retrieval mechanisms with the proposed framework. As shown in the comparison, existing methods primarily rely on neighborhood aggregation, query modulation, or global equilibrium formulations, but do not jointly model retrieval through explicit query injection, iterative semantic convergence, and attractor-based stability.

4. Preliminary Evaluation

Preliminary experiments were conducted using a manually annotated collection of 100 Brazilian Civil Non-Prosecution Agreements (ANPCs) provided and annotated by specialists from the Public Prosecutor's Office of Santa Catarina (MPSC). The annotated

Table 1. Comparison with representative graph-based retrieval approaches

Paper	**Update Equation**	**Difference Relative to Our Framework**
CaseGNN: Legal Case Retrieval via Text-Attributed GNN [Tang et al. 2023]	$\mathbf{h}'_v = W_s\mathbf{h}_v + \frac{1}{K}\sum \alpha_k(W_n\mathbf{h}_u + W_e\mathbf{h}_{e_{uv}})$	Focuses on structural legal relations through neighborhood aggregation, without query-aware dynamical refinement.
CaseGNN++: Graph Contrastive Learning [Tang et al. 2024a]	$\mathcal{L} = \mathcal{L}_{InfoNCE} + \lambda\mathcal{L}_{graph}$	Improves graph robustness through contrastive objectives rather than retrieval dynamics and semantic stabilization.
Advancing Legal Case Retrieval with LLMs and GNNs [Tang et al. 2025]	$\mathbf{H} = GNN_\theta(\mathbf{X}, \mathbf{A})$	Uses document-level graph representations without chunk-level semantic selection.
GNN-RAG: Graph Neural Retrieval [Mavromatis and Karypis 2024]	$\mathbf{h}_v^{(l)} = \psi(\mathbf{h}_v^{(l-1)}, \sum \omega(q, r)m)$	Introduces query-conditioned propagation but lacks explicit query injection and stability-based retrieval.
GFM-RAG: Graph Foundation Model for RAG [Zhang et al. 2025]	$h_e^{(l+1)} = Lin(h_e + \sum MLP(h_e \odot q \odot h_u))$	Applies query modulation within graph retrieval but does not model semantic convergence.
Implicit Hypergraph Neural Networks [Li et al. 2025]	$Z = \phi(AZW + \tilde{X}_q)$	Solves globally for equilibrium. Our proposal differs by using the trajectory of convergence to filter irrelevant chunks.
Retrieval Attractors	$x_i^{(t+1)} = \sigma(\sum_{j\in N(i)} v_{ij}^{(q)} x_j^{(t)} + w_{iq}q)$ Stability: $\|x_i^{(t+1)} - x_i^{(t)}\| < \varepsilon$	Combines query injection, iterative graph dynamics, semantic convergence, and attractor-based retrieval.

dataset serves as a gold-standard reference for evaluating legal information extraction performance.

The experiments focused on extracting six legal attributes: *Inquiry Number* (*Número do Inquérito Civil*), *Reimbursement Value* (*Valor de Ressarcimento*), *Civil Fine* (*Multa Civil*), *Forfeiture Value* (*Valor de Perdimento*), *Movement Date* (*Data do Movimento*), and *District* (*Comarca*). Five retrieval and extraction configurations were compared: Direct LLM extraction, Standard RAG, Hyperbolic-RAG, Euclidean-GNN Retrieval Attractors, and Hyperbolic-GNN Retrieval Attractors. Experiments were performed using Gemma 3 (27B) and Llama 3.3 (70B) under standardized structured-output settings.

Table 2 summarizes the initial accuracy results. Retrieval-based approaches consistently improved over direct extraction, while graph-based retrieval attractors achieved competitive or superior performance across several attributes. In particular, the Hyperbolic-GNN configuration obtained the best results for *Inquiry Number*, *Reimbursement Value*, *Civil Fine* and *Forfeiture Value* suggesting that the combination of graph dynamics and hyperbolic representations may provide advantages for legal retrieval and

Table 2. Preliminary accuracy comparison across retrieval configurations

Attribute	LLM-only		Standard RAG		Hyperbolic-RAG		Euclidean-GNN		Hyperbolic-GNN	
	L	G	L	G	L	G	L	G	L	G
Inquiry Number	0.38	0.39	0.52	0.48	0.62	0.60	0.73	0.68	**0.80**	**0.73**
Reimbursement Value	0.61	0.62	0.71	0.70	0.86	0.84	0.81	0.78	**0.92**	**0.88**
Civil Fine	0.57	0.68	0.72	0.70	0.95	0.94	0.97	0.96	**0.98**	**0.97**
Forfeiture Value	0.93	0.95	0.86	0.93	0.93	0.96	0.95	**0.98**	**0.98**	**0.98**
Movement Date	**0.87**	**0.87**	0.43	0.42	0.65	0.62	0.80	0.77	0.80	0.78
District	**0.96**	**0.96**	0.92	0.86	**0.96**	0.95	**0.96**	0.93	**0.96**	**0.96**

L = Llama 3.3 (70B), G = Gemma 3 (27B)

information extraction.

5. Conclusions and Future Work

Preliminary findings indicate that combining graph-based retrieval mechanisms with semantic stability criteria represents a promising direction for legal information extraction. From a Database perspective, the proposed framework contributes to information extraction, semantic retrieval, and the transformation of unstructured legal PDF documents into structured and queryable repositories. Furthermore, the extracted information enabled the construction of a Business Intelligence (BI) dashboard for the Public Prosecutor's Office of Santa Catarina (MPSC), demonstrating the practical applicability of the proposed pipeline for supporting legal analytics and decision-making.

Future work will focus on formalizing the mathematical properties of the proposed dynamics, including existence, stability, and convergence conditions. Additional GNN variants will also be investigated, particularly dynamic attractor mechanisms with adaptive graph refinement and hyperbolic propagation. Finally, broader validation will include experiments on public datasets and benchmark collections to evaluate retrieval effectiveness and generalization across different domains.

References

Balažević, I., Allen, C., and Hospedales, T. (2019). Multi-relational Poincaré graph embeddings. Advances in Neural Information Processing Systems (NeurIPS).

Bommarito, M. and Katz, D. M. (2022). Machine Learning for Law: Text, Cases, and Analytics. Cambridge University Press.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school.

Chami, I., Ying, R., Ré, C., and Leskovec, J. (2019). Hyperbolic graph convolutional neural networks. Advances in Neural Information Processing Systems (NeurIPS).

Chen, W., Han, X., Lin, Y., Zhao, H., Liu, Z., Li, P., Sun, M., and Zhou, J. (2021). Fully hyperbolic neural networks.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186. Proceedings of NAACL-HLT.

Hale, J. K. (1980). Ordinary Differential Equations. John Wiley and Sons, New York, 2nd edition.

He, N., Madhu, H., Bui, N., Yang, M., and Ying, R. (2025). Hyperbolic deep learning for foundation models: A survey. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25).

Jurafsky, D. and Martin, J. H. (2023). Speech and Language Processing. Pearson, 3rd edition.

Li, H., Su, W., Wang, C., Wu, Y., Ai, Q., and Liu, Y. (2023). Thuir@coliee 2023: Incorporating structural knowledge into pre-trained language models for legal case retrieval.

Li, Y., Wang, G., and Yang, Y. (2025). Implicit hypergraph neural networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Manning, C. D., Raghavan, P., and Schütze, H. (2021). Introduction to Information Retrieval. Cambridge University Press, 2nd edition.

Mavromatis, C. and Karypis, G. (2024). Gnn-rag: Graph neural retrieval for large language model reasoning.

Nguyen, H.-T. and Satoh, K. (2024). ConsRAG: Minimize LLM Hallucinations in the Legal Domain.

Singh, S. (2018). Natural language processing for information extraction.

Strogatz, S. H. (2015). Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. Westview Press, 2nd edition.

Tang, Y., Qiu, R., and Huang, Z. (2025). Uqlegalai@coliee2025: Advancing legal case retrieval with large language models and graph neural networks.

Tang, Y., Qiu, R., Liu, Y., Li, X., and Huang, Z. (2023). Casegnn: Graph neural networks for legal case retrieval with text-attributed graphs.

Tang, Y., Qiu, R., Liu, Y., Li, X., and Huang, Z. (2024a). Casegnn++: Graph contrastive learning for legal case retrieval with graph augmentation.

Tang, Y., Qiu, R., Yin, H., Li, X., and Huang, Z. (2024b). Caselink: Inductive graph learning for legal case retrieval.

Tifrea, A., Becigneul, G., and Ganea, O. (2021). Hyperbolic deep neural networks: A survey. Journal of Machine Learning Research, 22(169):1–42.

Zhang, B., Liang, Y., and Wang, H. (2025). Gfm-rag: Graph foundation model for retrieval augmented generation.

Zhang, W., Liu, X., and Wang, Y. (2022). Advances in relation extraction for natural language processing. Computational Linguistics, 48(2):245–268.