Artificial Intelligence and Law
https://doi.org/10.1007/s10506-025-09482-6

ORIGINAL RESEARCH

LegisSearch: navigating legislation with graphs and large language models

Andrea Colombo[1] · **Anna Bernasconi**[1] · **Luigi Bellomarini**[2] · **Luigi Guiso**[3] · **Claudio Michelacci**[3] · **Stefano Ceri**[1]

Received: 26 April 2025 / Accepted: 2 September 2025

Abstract

Navigating and retrieving relevant excerpts of legislation is challenging, requiring time and effort, especially to fine-tune appropriate input search queries. Furthermore, the continuously growing, heterogeneous body of laws, combined with a deep interconnection among normative acts, adds a layer of complexity: some potentially relevant rules may be hidden in articles that, through multiple citations and references, might be relevant for the input query. Traditional search systems, based on keywords or more sophisticated approaches as BM25 or TF-IDF, do not support such flexible exploration, being ineffective at handling contextual information. To address these challenges, recent research proposed using graph data models for legislative knowledge management, introducing a straightforward approach to handling network complexity. They adopted the Property Graph data structure, demonstrating how it provides semantics and navigation power, supporting advanced querying tools for legislative acts, and implemented it on the Italian legislation. In this paper, we build on recent results on legislative knowledge management with graphs by proposing LegisSearch, an effective navigation system that, combining the graph data model with pre-trained Large Language Models and universal text embeddings, allows users to conduct powerful searches within a legislative system. We implement LegisSearch within the Italian graph of national laws, and we test its performance across multiple domains by comparing its search results with those provided in specific thematic areas by Italian ministries on their official websites, demonstrating its superior retrieval performance over traditional search systems and testing the contribution of each component.

Keywords Law · Search systems · Knowledge graph · Large language models

Extended author information available on the last page of the article

Published online: 06 October 2025

1 Introduction

A country's legislation comprises large amounts of complex documents, i.e., laws composed of various articles connected through references, a widely used approach to recall, in various forms, previous relevant legislation. Retrieving such large, unstructured collections of documents is a long-standing challenge for information retrieval or recommendation and search systems, whose task is to suggest relevant items to users based on an input query (Van Meteren et al. 2000; Thorat et al. 2015).

Traditional applications of information retrieval systems have been developed for datasets of news (Raza and Ding 2022), websites, or books (Mathew et al. 2016), with relatively fewer applications in the legislative area (Bellandi et al. 2022; Wehnert et al. 2024), especially when considering more modern retrieval systems, based on large text embeddings from large language models (LLMs) (Deng 2022; Kanwal et al. 2021) and/or graph technologies, which demonstrated their power in enhancing retrieval results (Zhang et al. 2018). Text embeddings are used to calculate distances between the initial input query and each point in the embedding space – representing textual documents – offering the closest point as a recommendation to the user. Knowledge graphs provide additional context in the retrieval task, allowing the system to have more context in detecting whether a document is relevant to the input query.

Recently, a comprehensive and high-quality knowledge graph of the Italian legislation has been presented in Colombo et al. (2025) and shared in a Zenodo repository. The graph is built on top of an internationally adopted standard for representing legal documents, i.e., the Akoma Ntoso XML standard (Barabucci et al. 2009), which was adopted by the Italian parliament (Palmirani 2021). It adopts the Property Graph data model in the legislative domain, allowing an efficient navigation approach through queries adopting the recently standardized Graph Query Language (International Organization for Standardization 2024). Using this resource as a foundational use case, we propose LegisSearch, a powerful search system that leverages the graph's semantics and structure to enhance information retrieval for laws. In this system, we combine state-of-the-art universal text embeddings with graphs and Large Language Models, which can play a critical role in expanding the user query (Wang et al. 2023b), especially considering in a highly specialized domain as the legal one, whose documents can be significantly more complex than news, articles, or books (Matsyupa et al. 2022).

To empower LegisSearch, we first enrich the graph with node embeddings that capture neighborhood information, leveraging the graph semantics to inject such knowledge within the vector representation and using more modern embedding models, built upon LLMs architecture as the *multilingual E5 model* (Wang et al. 2022), and a natural language template strategy (Liu et al. 2024). Then, given an input textual query, such as a short abstract, a title, or a set of keywords, we use Large Language Models to characterize the query and expand the list of topics that may be relevant for answering the query. We also compute context-aware vector representation for the prompted text and we derive recommendations – relevant pieces of legislation – by adopting cosine similarity search over the graph nodes. For this purpose, we leverage

the graph to create a customized score combining information from laws and their article nodes, accounting for the domain features.

We evaluate the performance of LegisSearch on various thematic areas with datasets collected from the official website of Italian ministries, representing the most relevant laws in diverse subjects. These areas are highly heterogeneous, including pensions, chemical substances, fuels, nuclear energy, ozone substances, plant protection regulations, and golden powers. We demonstrate that our system significantly outperforms traditional retrieval approaches based on BM25 (Robertson and Zaragoza 2009) and TF-IDF (Fautsch and Savoy 2010), also showing how each component provides a notable contribution across standard retrieval quality metrics (i.e., average precision, recall, and the discounted cumulative gain).

To demonstrate the advantage of using LegisSearch, we consider a practical problem: understanding and monitoring the implications of Italy's "golden power" regulations — these are rules governing state intervention in corporate transactions critical to national interests. These regulations, often dispersed across various legislative texts and evolving through amendments, can be challenging to locate and interpret. A traditional keyword-based search might fail to capture the relationships between legal documents. With LegisSearch, the analyst can explore an interconnected graph of legislative documents to uncover relevant provisions, amendments, and cross-referenced laws, thereby improving the search process and the user's productivity.

Our contributions can be summarized as follows.

- We present LegisSearch, an intelligent search system that is based on recent advancements in graph data models and large language models for enhancing legislative information search
- We implement the system for the Italian legislation, whose graph modelling is available on Zenodo
- We experiment with our system on real-world datasets, demonstrating the higher retrieval performances compared to traditional methods
- We discuss LegisSearch in a practical scenario, demonstrating its additional value in providing more useful insights within legislation

The rest of the paper is structured as follows: Section 2 reviews related work; Section 3 recalls the feature of the KG of the Italian legislation and discusses how we enrich it with context-embeddings; Section 4 presents the Graph-based Legislative Retrieval System, whose performances are tested in Section 5; finally, Section 6 concludes the paper.

2 Related work

Many recent efforts have focused on modeling legal knowledge concepts into machine-readable ontologies. These efforts enable automatic reasoning and artificial intelligence use in legislative applications (e.g., lawmaking).

The spread of internationally adopted standards for representing textual legal documents, such as laws or bills, has encouraged the possibility of having a uni-

fied representation of legal documents on an international level, offering specific semi-structured data models to represent texts and their structure (Lupo et al. 2007). Examples include the Legal Knowledge Interchange Format (LKIF) (Hoekstra et al. 2007), LegalRuleML (Athan et al. 2013), and Akoma Ntoso (AKN) (Barabucci et al. 2009; OASIS 2018). The latter is an XML-based flexible standard, recently officially adopted by many international and national bodies (UN System Chief Executives Board for Coordination 2017; European Union Publications Office 2023; Palmirani 2019). Such advancements have facilitated the possibility of representing factual legal knowledge in a more structured data model, allowing easy access to documents and their interconnections, e.g., using RDF-based Knowledge Graphs (Anelli et al. 2022; Angelidis et al. 2018; Rodríguez-Doncel et al. 2018).

Legal knowledge extraction strongly relates to ontologies and legal concepts. In Ren et al. (2022), authors develop an ontology for extracting legal facts from Chinese legal texts, then using a deep neural network based on LSTM to extract facts. Another example is the Automated System for Knowledge Extraction (ASKE) (Castano et al. 2024), which employs a combination of embedding models and zero-shot learning techniques to discover concepts and classify legal texts, intending to derive appropriate concepts for documents. It first constructs a graph of concepts, which is then used to classify documents at a paragraph-level granularity. It outperforms other topic modeling approaches, such as BERTopic and Zero-Shot TM (ZSTM) (Bianchi et al. 2021), and does not require a predefined number of topics in input. Although the primary goal of ASKE was topic modeling, it has also been exploited to develop a knowledge-based approach for legal document retrieval from a repository of Italian court decisions (Bellandi et al. 2022). Here, the main goal is to retrieve past court decisions; the input is sentences, definitions, and excerpts of articles. A pertinence score is provided as a result and represents the percentage value of similarity between the input query and the document. ASKE was used to derive multiple labels for the query through a query expansion mechanism and by computing the embedding distance between the cluster of query terms and the document chunks. Its performance was validated by a team of law experts who judged the relevance/pertinence of the retrieved result. While some application cases were quite successful, others were less satisfactory since the ASKE cycle, as highlighted by the authors, might fail to fully capture the contextual meaning of portions of texts.

Another work on legal knowledge extraction that focuses on case-law decisions is the CRIKE (CRIme Knowledge Extraction) framework, which leverages an ontology-based approach to support the extraction of legal knowledge from collections of legal documents. It is built upon a reference legal ontology called LATO (Legal Abstract Term Ontology), which formalizes legal abstract terms as concepts and defines relations among them. CRIKE aims to detect concrete applications of these abstract terms in case-law decisions, thereby explaining how judges apply legal concepts in their reasoning (Castano et al. 2019a, b, 2022). A similar effort has been LawV (Griffo et al. 2020), a visual symbolic representation for legal statements to facilitate an intuitive and more accessible understanding of legal content.

While legislative and legal documents are drafted according to precise rules, most countries either do not publish such data in a native machine-readable format or have

only recently started to do so (Colombo and Cambria 2025). This led to more direct application of deep learning approaches, without relying on ontologies or intermediate steps that facilitate the information extraction task.

Works in the literature here focused on different aspects, such as text classification, information extraction, and information retrieval (Winkels et al. 2014; Sansone and Sperlí 2022; Chalkidis and Kampas 2019; Huang et al. 2021; Yelmen et al. 2023; Wehnert et al. 2024). All of them provide benefits in supporting the activities of legal professionals (Zhong et al. 2020). For instance, information retrieval focuses on extracting valuable information from texts that can be used for tasks such as Legal Judgment Prediction (LJP), where the goal is to predict the judgment results from fact description and statutory articles (Zhong et al. 2018).

LexDrafter is a recently introduced framework that provides insights about existing definitions, helps define new terms based on a document's context, and aims to support a harmonized legal definition across different regulations to avoid ambiguities. LexDrafter assists in drafting Definitions articles for legislative documents using retrieval-augmented generation (RAG) over existing term definitions present in different legislative documents (Chouhan and Gertz 2024).

3 The graph of the Italian legislation

In this section, before presenting the architecture LegisSearch, we first provide an overview of the Property Graph of the Italian legislation that we will use as our main data source, as presented in Colombo et al. (2025), in which the acts of national legislation retrieved from the official Italian website, i.e. Normattiva (Istituto Poligrafico e Zecca dello Stato 2024), were transformed into a graph structure. Then, we discuss text embeddings and our technique to enrich the graph and its nodes with vector representations that we will use to implement LegisSearch. Finally, we provide an overview of how the temporal dimension of laws is managed through the use of graphs.

3.1 Graph schema of the legislation

Laws, as document objects, are modelled as nodes in the graph, with their articles and attachments being distinct nodes connected through a parthood relationship to the law node. Law-relevant metadata are assigned as properties directly to the law node, including title, publication date, and law domain (i.e., a property denoting the ministries responsible for the law's content).

Articles, instead, are nodes whose properties describe the portion of the law they refer to, including the text of the article, the article-specific heading or title, and a list of policy-relevant topics regulated by the specific article. The same applies to attachments, whose role in a law document differs from that of the articles. Nodes are interlinked via a set of directed edges that represent the relationship (and the type of relationship) connecting a pair of nodes, namely an *is legal basis of* citation, an *amends*, an *abrogates*, or generic *cites* edges. Figure 1 depicts the graph schema.

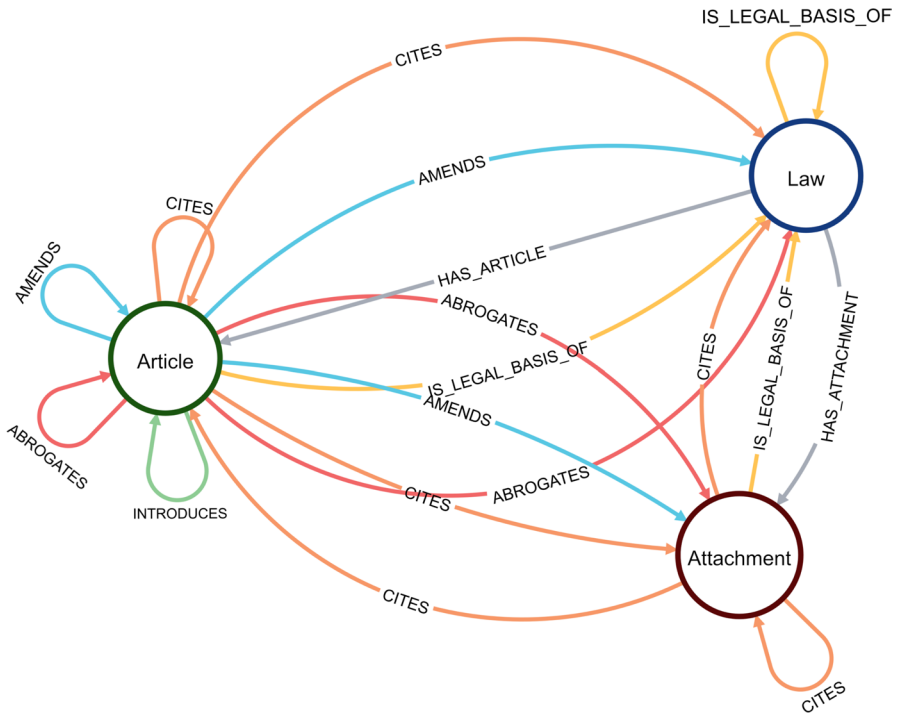

Fig. 1 Graph Schema of the Knowledge Graph of the Italian Legislation

3.2 Graph dimensions

The graph of the Italian legislation consists of law nodes representing the national Italian legislation since 1948 – the entry-into-force date of the Italian Constitution. The graph is stored in a Neo4j database; at the time of writing, it comprises 74k law nodes, 318k article nodes, 127k attachment nodes, 107k preamble citations, 64k abrogations, 80k amendments, and 228k generic citations. In the rest of the paper and for this work, we will assimilate attachment nodes to article nodes and refer only to "articles" for both. Preamble citations are laws or articles that, in the Italian legislation, are cited before the actual text of the law and represent legal dependencies of a law; thus, they are represented in the graph as *is legal basis of* edges, highlighting an explicit dependency to another law or article.

Topic of laws and articles In the used database, graph nodes (laws, articles, or attachments) have metadata as their properties and are enriched with content information, including the actual text as a property and the subjects regulated within the article. The construction of the LLM-assisted pipeline to generate the graph has been presented in Colombo et al. (2025). To derive topics and assign them as node properties, we adopt an LLM-guided strategy: a lightweight LLM (Mistral-7B (Jiang et al. 2023)) was fine-tuned on a dataset of high-quality text–topic pairs and applied to the Italian legislation corpus. Then, lemmatization and stemming were performed to

extract root words, enabling the generation of topics that balance novelty with sufficient generality to support exact string matching. This two-step approach accounts for synonyms and variations in topic naming, and it allows us to derive related topics, which were not explicitly mentioned, enabling an enhaned topic-related user queries.

A *topic* of a graph node corresponds to a list of keywords that capture the content of a text and possibly generalize it to one or more words that have broader semantics and/or are more frequently used. The graph can be conveniently filtered by leveraging topics (to focus on specific portions) or citations (to detect the neighborhood of the portions). Note that *topics* assigned to nodes are not ranked; hence, it is not possible to associate each law with its main topic. In some legislation, law topics are provided as metadata, like in the case of the US bills provided by the Library of Congress (Library of Congress 2025); however, a ranking is not provided.

In the Italian graph, we counted over two million topics, as each node is assigned to about five topics on average. We extracted 72k distinct topics. Their distribution is skewed since its median is 2 and its maximum frequency is 24k (*education*). This is a natural behavior of topics, which include both very generic topics, such as *education* or *budget*, and specific ones, such as *debt collection* or *agricultural consortium*.

3.3 Text embeddings for document retrieval

Text embeddings are a powerful technique in natural language processing that converts textual data into dense, fixed-length vectors, enabling the representation of semantic meaning in a continuous, high-dimensional space. By capturing the main relationships between words and phrases, textual embeddings are the driver of modern document retrieval systems, with first applications also in the legal domain (Chalkidis and Kampas 2019).

Recently, with the rise of LLMs, more modern embedding models have been developed. One of these approaches is the E5 model (Wang et al. 2022) and its multilingual - Mistral-7B extensions (Wang et al. 2023a, 2024), which are part of one of the first families of embedding models that go beyond the English language and are explicitly trained for information retrieval tasks.

A challenge when encoding long text as laws is the token limit of the models, which often does not allow us to capture the content of a law. Indeed, the average number of words of the full text – considering the published laws after 1992[1] – is 9k words, whereas state-of-the-art proprietary text embedding models are typically limited to 8,191 tokens[2], which approximately corresponds to 6k words. To tackle this, we can leverage topics and metadata available in the graph instead of the full text to capture the main content within the token limit. In addition, we can use the title of the law or the article as an essential summary of the documents.

This paper does not address the specialization of an LLM-based embedding model to the legislative domain, as we do not have a training dataset. Instead, we

[1] In 1992 a major change in Italian politics resulted in a fundamentally different legislative process, with a lower law production

[2] text-embedding-3-large

make use of recently introduced text embedding models, such as the *multilingual-E5-large* (ME5), to create the vector numerical representations (Wang et al. 2022); for instance, the ME5 model features 24 layers, an embedding size of 1024 (>768, i.e., the BERT-based embedding dimension), and is trained on a multilingual corpus. These embeddings are trained to be "universal" text embeddings, supporting multiple downstream tasks and languages (Li et al. 2022); they are highly specialized on tasks such as semantic search and information retrieval.

3.4 Graph-enhanced embeddings

The synergy of text embeddings with an underlying Knowledge Graph is a powerful approach that combines structured knowledge with semantic flexibility (Wang et al. 2018; Syed et al. 2022). In our context, the graph can be leveraged to examine the relationships connecting an article or a law, thus complementing the often vague or implicit meanings embedded in raw text that may cite other laws (resp. articles), hence potentially becoming relevant for a target subject.

In this work, we propose to create graph node embeddings by leveraging graph query tools to create context-aware representations of articles and laws. Since we use textual embeddings, we adopt a natural language template technique, as introduced in Liu et al. (2024), which utilizes natural language labels to segment distinct fields before creating the embeddings.

In particular, we create embeddings for law nodes by querying the Property Graph to derive granular topics for each article that composes the law and by navigating its legal foundation laws and articles to extract additional context. This query can be easily encoded in Cypher (Francis et al. 2018), the query language used to access the Property Graph, as:

```
MATCH (l:Law)-[:HAS_ARTICLE]->(a:Article|Attachment)
OPTIONAL MATCH (b)-[:IS_LEGAL_BASIS_OF]->(l)
RETURN l.title as TitleLaw, l.topic AS LawTopics,
       COLLECT(a.topic) as ArticleTopics, COLLECT(b.topic) AS Context
Topics
```

We then adopt the following natural language template, parsing the result of the query in the corresponding template positions:

```
<Law Title:> ...
<Law Topics:> ...
<Article Topics:> ...
<Context Topics:> ...
```

The combination of both law and article topics is essential for better encoding law nodes. Indeed, law titles and topics might be too generic to represent the actual full content of the law. For instance, a 'reorganization law' of a ministry is a bill that modifies many aspects of the structure and the duties of a ministry, and a typical generic title such as "Reorganization of the Defense Ministry" would not be enough

to understand the specific areas of intervention of its articles. By querying the graph, we can retrieve knowledge referring to each article, specific topics, and titles, such that we can compose them and embed them within the token limit.

A more advanced strategy is used for article nodes. Here, we query the neighborhood of an article to derive its context-aware embedding representation. In Cypher, the query can be encoded as:

```
MATCH (a:Article|Attachment)<-[:HAS_ARTICLE]-(l:Law)
OPTIONAL MATCH (a)-[:ABROGATES|AMENDS|INTRODUCES|CITES]->(a2:Article
|Law)
RETURN a.title AS Title, l.topic AS LawTopics,
       a.topic AS ArticleTopic, a2.topic AS ContextTopics
```

Similarly, for article (and attachment) nodes, we adopt the following template:

```
<Article Title:> ...
<Article Topics:> ...
<Context Topics:> ...
<Law Topics:> ...
```

Many variations of such queries to create the templates are indeed possible and facilitated by the graph data model. If needed, we could leverage recursive patterns to detect larger neighborhoods. However, an in-depth analysis determining which combinatory approach would be the most beneficial one is out of the scope of this work; here, instead, we aim to demonstrate the utility of a graph on a more general level.

3.5 Temporal aspects of the graph

Legislative systems continuously evolve, producing new laws that amend or repeal existing ones. Each modification gives rise to a new version of the article or the law, reflecting the updated text at the time of change (Colombo et al. 2025). The used graph-based model, together with its schema, provides a powerful way to fully capture the temporal evolution of legislation. Specifically, it stores the original textual version of each article of law as a node property and records the amended versions introduced by subsequent laws as edge properties. These queries exploit publication dates encoded in law nodes, ensuring that the appropriate version of the law can always be inferred.

The graph-based approach ensures efficient data management while providing a structured and temporally-aware framework for querying legislative data. In particular, abrogated articles or repealed laws can be straightforwardly excluded from query results. For example, when navigating the system, one can simply filter out articles by exploiting the *abrogates* edge. Conversely, there are scenarios in which it is essential to retrieve the exact set of articles that were in force at a specific timestamp, e.g., during a legal controversy. In these cases, the temporal information encoded in the graph enables the reconstruction of the applicable legal framework at any given moment.

4 LegisSearch architecture

Figure 2 provides a visual representation of the architecture of LegisSearch, our graph-based legislative retrieval system.

LLM-based user input expansion (1) Our system inputs a textual query from a user looking for relevant legislation in a certain thematic area. One of the most promising applications of Large Language Models in information retrieval systems is data augmentation (typically of users' inputs), with additional, strongly related knowledge, which is useful for searching over the vector space (Liu et al. 2024). This is especially beneficial in specific domains employing specialized words and terms highly related to the input but rare/uncommon for the general users.

To this aim, we adopt an LLM-based *understanding and expansion* approach, inspired by the recent work from Wang et al. (2023b), that aims to enrich the textual user input with additional/derived content helpful to a more effective search. We designed a two-step LLM intervention: first, we derive the main topics from the text input by the user. Then, a second LLM expands the list of topics by adding highly related topics. For both cases, we rely on a pre-trained state-of-the-art Large Language Model, LLama-3-70B-Instruct (Dubey et al. 2024), combined with few-shot learning for task specialization. Few-shot learning is a powerful and effective approach where an LLM receives directly as input a small amount of training data – pairs of task-specific question-answers – that contribute to *instruct* the model to perform the task (Wang et al. 2020; Parnami and Lee 2022). For the topic extraction task, we provide – as the system prompt for the LLM – the following task description: *You are an assistant that derives topics from a text. Topics must have a few words. Reply in Italian.*

Table 1 lists manually crafted examples to instruct the LLM in the topic extraction task.

Then, the pre-trained Llama-3-70B model is parameterized to perform the topic expansion task. In other words, the generative model is now asked to expand the list of topics extracted in the previous step by deriving similar and correlated topics. Similarly to the topic extraction task, we adopt the following system prompt: *You are an assistant that expands a list of topics with some similar and related ones. Reply in Italian.*

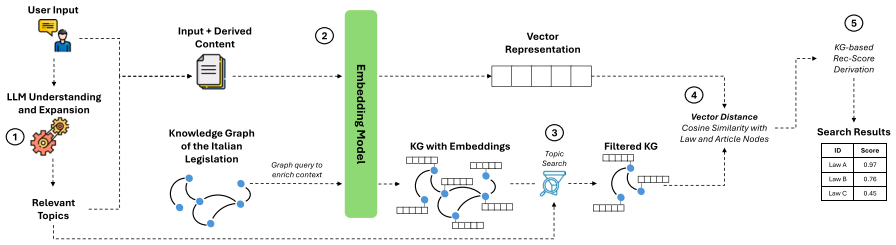

Fig. 2 Architecture of LegisSearch. First, based on a user query, we use an LLM as a query expander to derive additional content and related information (1) (i.e., topics). Then, a universal embedding model computes a vector representation of the enriched input (2), which is also used to compute graph node embeddings. Topics are also used to query only relevant nodes (3). Search results are derived by adopting the cosine similarity (4) and combining law and article distance values to compute a recommendation score (5)

Table 1 Few-shot learning user-assistant pairs provided to the Llama-3-70B model as examples of the topic *extraction* task

	User Prompt	Assistant Reply
1	Extract topics from this title: Refund of the general tax on exported products to the United States.	exports, taxation, USA
2	Extract topics from this title: Provisions for implementing the EU legislation concerning the Common Market Organisation (CMO) of wine.	EU legislation, CMO, wine
3	Extract topics from this title: Procedures for appointing teachers to tenure.	hiring, teacher, school

We perform few-shot learning, as shown in Table 2. By splitting it into two tasks, we let the model focus on more targeted tasks that help it achieve better final results. While LLMs have no theoretical guarantee of performing a given task, we observed that, for such simple tasks, we could rely on them to extract topics, as also demonstrated in previous research (Maragheh et al. 2023).

Embedding the User Input (2) The same embedding model used to create the graph embeddings (the ME5) is used to derive a vector representation of the user input. In this case, we adopt the following natural language template (with explicit reference to the list of expanded topics derived using the LLM):

```
<User Input:> ...
<Topics:> ...
```

Graph filtering (3) Before performing a vector similarity distance search between the newly computed embeddings and the graph nodes, we adopt a filtering technique that simplifies the search by focusing on specific portions of the graph. In particular, we use the same list of topics to query the nodes – articles and laws – that are potential candidates as relevant search results. A *potential candidate* is a graph node that shares at least one topic with the topics extracted from the user input. In the filtering, we also consider the node's neighborhood, as we did with the graph embedding computation in Section 3.4, thus checking whether a topic is in a node within a hop of a node. As suggested in Section 3.5, here laws or articles nodes can be filtered out based on their validity on the desired timestamp, which can be retrieved with a graph query.

Table 2 Few-shot learning user-assistant pairs provided to the Llama-3-70B model as examples of the topic *expansion* task

	User Prompt	Assistant Reply
1	Expand the following list of topics: exports, taxation, USA	exports, taxation, USA, united states, imports, trade, international trade
2	Expand the following list of topics: EU legislation, CMO, wine	EU legislation, CMO, wine, agriculture, european union, common market
3	Expand the following list of topics: hiring, teacher, school	hiring, teacher, school, high school, selection procedures, school staff

Embedding vector distance (4) Relevant laws are derived by computing the distance between embedding vectors. We calculate the distance by adopting the *cosine similarity*, a widely used method in information retrieval to compute the distance between the embedding vectors (Melville and Sindhwani 2010), which measures the angular distance between vectors Salton (1989). A higher cosine similarity indicates greater alignment, suggesting that the items are more relevant to the user's query.

Computing the recommendation score (5) As described in Section 3, the graph of the Italian legislation contains both laws and articles/attachments as nodes, reaching a deep level of detail in the graph-based representation of legislative systems. We can benefit from this fine granularity to provide more precise recommendations. While embedding distance can be computed for both law and article nodes, we consider these values jointly to enhance the precision of the recommender system; indeed, a high similarity with a law node embedding captures theme-specific laws, i.e., regulations that are widely dedicated to a specific area. Conversely, a high similarity with the article node embedding can help find laws that might be more generic, such as budget regulations. Still, it may have one or more portions of text highly relevant to a user. By balancing the two values, we can provide context-aware recommendations that can both 1) account for the big picture, i.e., *the general law theme*, and 2) by looking at article nodes, detect more "surprising" items, i.e., laws whose general theme is not strictly in line with the user input but relevant to the thematic area of interest.

To capture this feature, we build the final recommendation score for a law *l* as:

$$RecScore_l = \mathrm{CS}(\bar{e}_i, \bar{e}_l) + \max_a \mathrm{CS}(\bar{e}_i, \bar{e}_a) \tag{1}$$

where CS is the cosine similarity function, $\bar{e}_i$ is the vector of the embeddings of the textual (expanded) input *i*, $\bar{e}_l$ is the vector of the embeddings of the law node *l*, and *a* is the set of articles of law *l*, queried from the Knowledge Graph. In other words, we weigh a law by considering its similarity and the similarity of its articles (the maximum possible one) w.r.t. the user input. Intuitively, we favor laws that contain highly similar specific articles. By sorting laws based on their recommendation score, we derive a ranking of items (laws) suggested to the user.

5 Experiments and results

LegisSearch aims to indicate relevant legislation based on custom textual input from a user, be it a series of keywords or a short text. The system's output is a set of ranked results whose first entries are the most relevant laws related to the input text. We aim to demonstrate that:

A. Our search system, based on graphs and LLMs, performs significantly better than traditional benchmark approaches.

B. The choice of a state-of-the-art universal embedding model based on LLMs significantly contributes to increasing the quality of the search results.
C. As a query expander, the LLM step allows the system to achieve a higher degree of recall since it can expand the search space.

To evaluate (A), we compare our approach with traditional information retrieval approaches, namely a BM25 (Robertson and Zaragoza 2009) algorithm and a TF-IDF approach for document retrieval (Fautsch and Savoy 2010) (see Section 5.3). Then, we conduct an ablation study to answer (B) and (C) (see Section 5.4). In particular, we consider two embedding models, GloVe and a pre-trained Transformer Language Model, specifically designed for Italian Law, the Italian Legal-BERT (Licari and Comandè 2022, 2024)[3], and we test our search system's results, excluding the LLM steps.

Finally, we showcase LegisSearch in practice by demonstrating its additional value in the concrete use case of a financial analyst aiming to discover and analyse relevant legislation related to Golden Power. To this aim, we discuss how the graph visualization also contributes to highlighting additional patterns that allow users of LegisSearch to have more insights from its search.

Implementation We implemented LegisSearch as a Python module and we used a Neo4j instance to store the graph of the Italian legislation. We used an external provider for the LLM calls, which has to be specified within the module functions. We run LegisSearch on an Ubuntu 22.04 LTS server equipped with an Intel Xeon 5118 CPU @ 2.30GHz processor and 376GB of RAM,

5.1 Datasets and search queries

We collected a set of ground truth lists of laws describing relevant legislation about a specific theme. Such datasets were manually retrieved by navigating the websites of the Italian ministries. We considered documents from the ministry's webpage as our reference because they represent the most authoritative and reliable collections of laws, being curated and maintained by government officials. While other legal documents or articles could potentially be relevant to the thematic areas of our study, they may vary in quality, completeness, or alignment with the official legislative framework. By relying exclusively on the ministry's publications, we ensured that our evaluation was based on the highest-quality, government-validated references, minimizing the risk of including outdated, unofficial, or misclassified documents.

We selected the following thematic areas: pensions, chemical substances, jobs and occupations, fuel usage, nuclear energy, ozone substances, plant protection products regulations, and golden power. Such categories were chosen according to the availability of annotated datasets on the Italian ministries' websites. In addition, we selected categories whose lists included at least five relevant (national) laws.

We observed that such datasets only consider recent legislation, thus not reporting old– but still in force– laws that might be relevant for these areas.

[3] In detail, we employ its main version, available at https://huggingface.co/dlicari/Italian-Legal-BERT

To account for this, we limited the search system to run on legislation produced after 1994, a significant shift in Italian politics, which started the so-called Second Republic. An exception to this rule was made for thematic areas whose relevant laws had an older law, namely 1962 for *Nuclear Energy* and 1993 for *Ozone Substances*.

In Table 3, we illustrate the details of the dataset and show examples of potential user textual inputs related to the thematic areas under analysis. Such texts have been derived and adapted directly from our data sources (i.e., the ministry webpage) titles/subtitles where we collected the ground truth lists They represent ideal user input when trying to navigate legislation about certain thematic areas.

5.2 Evaluation metrics

We adopt widely used strategies in the literature of recommender systems and information retrieval (Shani and Gunawardana 2011; Avazpour et al. 2014; Ong et al. 2020). In the use application scenarios at hand, we do not deem a ranking useful, as typically, all the reported results will need to be inspected by the users. As a first metric, we consider the *Average Precision at k* (*AP@k*) to evaluate the **precision performance** of the system, i.e., if and how the system is capable of suggesting relevant items in the top positions. This is derived from the "precision at *k*" (Cormack and Lynam 2006), defined as the fraction of relevant items in the top-k recommendations. Formally, it is computed as:

Table 3 Thematic areas, corresponding search queries, and relevant true regulations. Appendix A, indicates the data sources from which we retrieved the evaluation datasets

Thematic Area	Search Query	Regulations
Pensions Regulation	Legislation related to pensions and retirement benefits.	42/2006, 78/2010, 216/2011, 98/2011, 95/2012, 102/2013, 201/2011, 208/2015, 4/2019, 147/2013, 78/2009, 138/2011, 243/2004, 183/2011, 228/2012, 147/2014, 247/2007
Chemical Substances	Legislation related to the regulation and management of chemical substances.	133/2009, 27/2014, 281/1997, 200/2011, 124/2016, 145/2008
Fuels	Legislation related to fuel production and distribution for civil, industrial, and maritime uses.	205/2007, 128/2005, 66/2005, 152/2006, 51/2017, 112/2014
Nuclear Energy	Legislation related to nuclear energy and radioactive waste management.	1860/1962, 101/2020, 23/2009, 45/2014, 314/2003, 282/2005, 1450/1970, 239/2004, 99/2009, 1/2012, 100/2011, 31/2010
Ozone Substances	Regulations, information, and obligations for those who produce, use, and possess ozone-depleting substances.	35/2001, 56/1996, 549/1993, 147/2006, 179/1997, 409/2000, 108/2013, 91/2014, 179/2002
Plant Protection Products	Regulations on plant protection products for the control of any organism harmful to cultivated plants.	194/1995, 55/2012, 150/2012, 290/2001, 69/2016
Golden Power	Legislation on Golden Power.	179/2020, 148/2017, 187/2022, 21/2012, 85/2014, 180/2020, 23/2020, 108/2014, 22/2019, 86/2014, 35/2014, 133/2022

$$P@k = \sum_{i=1}^{k} \frac{rel(i)}{k} \tag{2}$$

where k is the number of ascending ranked documents to be considered and

$$rel(i) = \begin{cases} 1 & \text{if the document i is relevant} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The Average Precision at k is computed as the average of precision values at all the relevant positions within k:

$$AP@k = \frac{1}{\min(m,k)} \sum_{i=1}^{k} P@i * rel(i) \tag{4}$$

where m is the total number of relevant items, i.e., laws of the thematic area of interest. The AP value has been widely used for evaluating retrieval systems (Buckley and Voorhees 2017; Aslam et al. 2005), with a higher value corresponding to a higher probability of seeing relevant documents in the top ranks.

We evaluate the **overall ranking quality** by adopting the *Discounted Cumulative Gain (DCG)* (Järvelin and Kekäläinen 2002) at k, where each position in the ranking is assigned a score that is penalized according to its position. Here, we adapt the DCG to the binary scenario (i.e., no relevance importance is available), which still provides an intuitive measure showing the gain of different systems (Kekäläinen 2005). The derived score rewards systems that place relevant items higher in the ranking, with less focus on precision. It is computed as:

$$DCG@k = \sum_{i=i}^{k} \frac{rel(i)}{\log_2(i+1)} \tag{5}$$

and it measures the usefulness, or gain, of an item based on its position in the result list, with the premise that relevant results appearing earlier in the list are more valuable.

To complement both metrics and to provide a more interpretable performance measure of our recommendation system, we also consider the *Recall at k* (*R@k*), which assesses the fraction of actually relevant laws (according to the ground truth selection) detected by our system. It is computed as:

$$R@k = \sum_{i=1}^{k} \frac{rel(i)}{m} \tag{6}$$

where m is the total number of relevant items, which, in our case, are the relevant laws for a specific thematic area.

Choice of *k* values We evaluate average precision and recall at three thresholds: $k = 5$, 20, and 50. Values such as 5 and 20 are commonly reported in the literature and reflect realistic scenarios where human users review only the top-ranked documents. The inclusion of a higher value, $k = 50$, may exceed what a human would typically examine; however, it is justified in the context of AI-assisted workflows. In such scenarios, an AI agent can efficiently process larger sets of retrieved documents, selecting relevant items to address specific queries without requiring manual review of each result. Therefore, evaluating performance at $k = 50$ provides insight into how the system behaves in automated or semi-automated retrieval settings, which are increasingly relevant in practical applications (Zhang et al. 2024).

5.3 Retrieval performances

To evaluate the performance of LegisSearch, we compare it against two widely used baseline models in information retrieval: BM25 and TF-IDF.

While we know other domain-specific document retrieval approaches, we think they are hardly adaptable to the specific domain. For instance, ASKE (Bellandi et al. 2022) has been adapted to information retrieval in Italian legal court decisions, but (i) its primary aim is multi-label classification, and (ii) it relies on a first step of Elasticsearch (Elasticsearch 2025) repository queries, which should be manually customized according to the thematic area of interest of the user.

BM25 The Best Matching 25 framework (BM25) (Robertson and Zaragoza 2009) is a probabilistic retrieval model that ranks documents based on the query terms appearing in each document, regardless of their proximity. It extends the probabilistic relevance framework and incorporates term frequency, document length normalization, and inverse document frequency. BM25 effectively handles long documents and has become a de facto standard in modern search systems.

TF-IDF Term Frequency-Inverse Document Frequency (TF-IDF) (Fautsch and Savoy 2010) is a statistical measure used to evaluate the importance of a term in a document relative to a collection of documents (the corpus). The TF-IDF score increases proportionally with the number of times a term appears in a document but is offset by the frequency of the term in the corpus. TF-IDF is simple, interpretable, and computationally efficient, making it a common choice for baseline comparisons in text retrieval and classification tasks.

Overall search performances For our benchmarks, both BM25 and TF-IDF, we consider laws as our ‘documents’, thus (i) without an underlying graph supporting the retrieval and (ii) directly on the input query from the used. Table 4 illustrates the evaluation metrics we considered, aggregating across all our datasets.

The results demonstrate that LegisSearch outperforms BM25 and TF-IDF across most evaluation metrics, making it the most effective method for searching legislative acts. Our system achieves the highest recall (R@5, R@20, and R@50) and discounted cumulative gain (DCG@5, DCG@20, and DCG@50), indicating its

Table 4 Average Precision, Recall and Discounted Cumulative Gain obtained at different thresholds k by the baseline models and our LegisSearch system. Our search system significantly outperforms both benchmarks in most of the considered metrics

	BM25	TF-IDF	LegisSearch
AP@5	0.77	**0.81**	0.72
AP@20	0.45	0.56	**0.62**
AP@50	0.33	0.48	**0.56**
R@5	0.14	0.20	**0.35**
R@20	0.33	0.39	**0.59**
R@50	0.48	0.43	**0.71**
DCG@5	0.93	1.17	**1.63**
DCG@20	1.28	1.66	**2.40**
DCG@50	1.55	1.78	**2.63**

superior ability to retrieve relevant items and rank them effectively. While TF-IDF performs slightly better in the average precision at shorter thresholds, as reflected by its highest AP@5, LegisSearch is more consistent at higher thresholds. BM25 lags behind both methods, with lower precision, recall, and ranking quality scores. Thus, LegisSearch is optimal for scenarios requiring high recall and ranking effectiveness, as it was the goal of our designed architecture.

5.4 Testing the role of LLMs and embeddings in LegisSearch

We conducted a second set of experiments to test each component's contribution to our search system's results. In detail, we aim to quantify the contribution of employing universal pre-trained embeddings instead of domain-specific ones based on simpler models and of LLMs as a query expander. First, to test the contribution of the embedding model, we run our search system by considering a baseline generic model, GloVe, and a BERT model designed explicitly for Italian legal knowledge.

GloVe is a widely used technique for obtaining document or sentence embeddings, which involves computing the average word embeddings in the sentence (Pennington et al. 2014). GloVe provides pre-trained word vectors that capture semantic relationships between words based on co-occurrence statistics. The individual word vectors for each word in the sentence are retrieved from the GloVe model to generate a sentence-level embedding. These vectors are then averaged element-wise to produce a single, fixed-size vector representing the entire sentence. This method assumes that the semantic content of a sentence can be effectively captured by averaging the embeddings of its constituent words, which is computationally efficient but may oversimplify the sentence's structure and ignore word order and syntax.

BERT is a transformer-based model that is widely used to process natural language. One of its key applications is text embedding, which generates fixed-size, dense vector representations of documents, allowing an accurate semantic similarity search (Reimers and Gurevych 2019). Its token limit is 512, which means approximately 400 words. Legal-BERT models, deriving from the BERT architecture, are a family of models designed for legal applications and have achieved state-of-the-art performance across legal tasks, including document synthesis, contract analysis, argument extraction, and legal prediction (Chalkidis et al. 2019, 2020, 2021). Recently, its Italian version was introduced (Licari and Comandè 2022), thus allowing us to benchmark against a domain and language-specific model appropriately.

The all-MiniLM-L6-v2 (Wang et al. 2020) is a Sentence-Transformers (SBERT) (Reimers and Gurevych 2019) model and a variant of the Sentence-BERT architecture with 6 transformer layers and a 384-dimensional hidden state. It creates dense vector representations that preserve semantic similarity and are well-suited to be used in applications such as clustering, semantic search, and information retrieval (Reimers and Gurevych 2019). It differs from other models in that it is fine-tuned with contrastive learning, which increases its efficiency and accuracy. The all-MiniLM-L6-v2 variation supports multiple languages, allowing it to generalize beyond English and making it applicable in our scenario of Italian legislative data. Nevertheless, this model is not specifically trained for legal text.

We compute GloVe, Italian-BERT, and all-MiniLM-L6-v2 embeddings, as described in Section 3.3, and assign them to the corresponding graph nodes. For GloVe, we use the averaging method of word embeddings to derive a vector representation for law nodes. For all the models, embeddings are computed on the same input data, i.e., the expanded natural language query.

Contribution of the embedding model Table 5 presents a comparison of different embedding models adopted within our graph, thus allowing us to test the contribution of adopting a more advanced yet not domain-specific model, as the multilingual E5 model that we implemented in our approach.

The results highlight a superior performance of the M5E model across all evaluated metrics compared to the baseline models, GloVe and the two BERT models, especially in the *AP*@50, *R*@50, and *DCG*@50 values, demonstrating its ability to retrieve and rank relevant items, with a comparable result only in the *AP*@5, indicating a good quality in the top results. With respect to the SBERT-based model, the M5E achieves a comparable average precision at 50, although its higher recall makes the system more performant in retrieving additional relevant results.

Contribution of the LLM as query expander We tested the contribution of the LLM in our system by comparing the Average Precision achieved when we exclude from our architecture the LLM steps (i.e., no user input query expansion in step 1 and potential candidates detection in step 3 of Fig. 2). Consequently, we can measure the impact of expanding the input query to retrieve more relevant laws that are relevant to the input query, thus accounting for non-trivial concepts that only an LLM could infer. In Table 6, we illustrate the recall values comparing results with/without the integration of the LLM-related steps across various thematic areas at recall thresholds $R@5$, $R@20$, and $R@50$.

Table 5 Comparison with other embedding models, when replaced within our retrieval system

Model	Metric					
	AP@5	*AP*@50	*R*@5	*R*@50	*DCG*@5	*DCG*@50
GloVe	0.11	0.07	0.02	0.09	0.05	0.25
LegalBERT-IT	0.68	0.38	0.18	0.37	0.81	1.26
all-MiniLM-L6-v2	0.68	0.56	0.18	0.25	0.78	1.02
M5E	**0.72**	**0.56**	**0.35**	**0.71**	**1.63**	**2.63**

Table 6 Results of our experiment to test the contribution of the LLM steps to the Recall at the selected k thresholds and for each thematic area

Thematic Area	w/o LLM Steps			LegisSearch		
	R@5	R@20	R@50	R@5	R@20	R@50
Pensions	0.0	0.06	0.12	0.05	0.12	0.18
Chemical Substances	0.33	0.33	0.33	0.33	0.50	0.50
Fuels	0.16	0.67	0.83	0.50	0.67	0.83
Nuclear Energy	0.41	0.58	0.67	0.41	0.58	0.75
Ozone Substances	0.33	0.55	0.67	0.44	0.67	0.89
Plant Products	0.60	0.80	0.80	0.20	0.80	0.80
Golden Power	0.50	0.67	0.75	0.50	0.91	1.00
Average	**0.25**	**0.53**	**0.59**	**0.35**	**0.60**	**0.71**

From an analysis of results, we observe that the impact of LLM-based query expansion varies across thematic areas. Substantial gains are achieved in domains such as *Fuels*, *Ozone Substances*, and *Golden Power*, where recall improvements at higher thresholds indicate that semantic enrichment helps overcome terminological gaps and domain-specific jargon. By contrast, in areas like *Plant Products* and *Nuclear Energy*, improvements are limited, suggesting that standardized terminology reduces the need for expansion. *Chemical Substances* shows moderate benefits at higher recall levels, indicating that the LLM primarily broadens the candidate pool rather than improving early precision. Overall, these heterogeneous effects suggest that the contribution of LLMs is domain-dependent, with the largest gains in areas characterized by complex or specialized vocabularies. In the case of pensions, the high number of legislation produced – annually – limits the power of search, which we think should be conducted by adding a filter on the specific timespan of interest. We observe that including LLM steps overall improves recall across thematic areas (with only a few exceptions), thereby highlighting the importance of expanding the query input to increase the search space. Note that the higher recall is also a requirement for performing re-ranking approaches: by having a broader set of results during the initial retrieval phase, the system provides a larger candidate pool for re-ranking algorithms to operate effectively. Additionally, the improvement in recall underscores the capability of LLM-augmented processes to handle semantic issues, potentially addressing challenges in queries involving ambiguous or complex thematic contexts.

5.5 LegisSearch in action: retrieval of golden power laws

Using graphs to enhance search and retrieval systems can offer more insights and add value to the search results, since such output can be plotted directly within the graph to understand the relationships among the search output. As a practical example, we recall the use case we introduced in Section 1 related to the retrieval and monitoring of legislation related to Golden Power. We asked an expert legal professional in the Golden Power topic to validate the results we got with our system. In detail, we asked the expert to provide a list of relevant laws and split them into groups based on the laws' specific use of Golden Powers. We then compared the list with the results obtained from the graph, confirming that (i) we retrieved mostly all relevant laws and (ii) our graph clusters resembled the ones of the groups.

In Fig. 3, we plot the retrieved laws for the topic, providing a comprehensive overview of the legislation about Golden Power. The graph structure highlights key nodes representing the most central laws—those with the highest degree of connectivity. Furthermore, the clusters of nodes reveal thematic groupings, such as those associated with emergency responses during the COVID-19 pandemic or policies linked to the National Recovery and Resilience Plan (PNRR). The clusters can allow policymakers and analysts to identify emergent patterns or gaps in regulation. For instance, the connectivity within the COVID-19 cluster underscores the legislative focus on safeguarding strategic assets during times of crisis. Similarly, the PNRR-related cluster informs on new regulations to foster economic recovery and ensure compliance with EU directives.

Using graph-based insights exemplifies how LegisSearch facilitates both the retrieval of relevant laws -as demonstrated by our experiments- and their interpretation, providing straightforward visualizations of the results.

5.6 Discussion

The results achieved by our legislative retrieval system demonstrate the significant contribution of adopting a graph- and LLM-based approach toward legal document retrieval, which allows us to obtain more precise results with higher levels of recall. By adopting a graph, we outperform traditional retrieval methods. Our approach con-

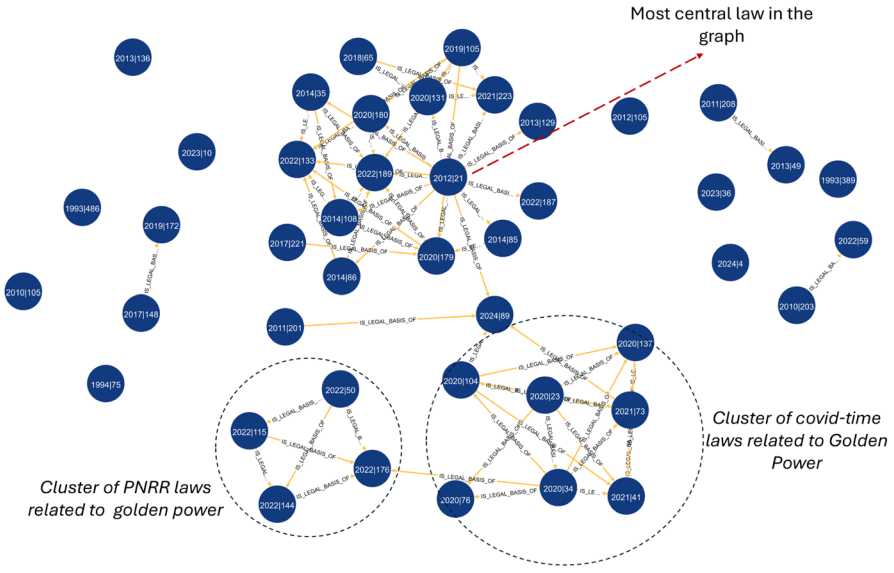

Fig. 3 Graph visualization of relevant laws suggested by LegisSearch, focusing on legal foundation connections. Nodes are annotated with the law identifier (i.e., year and law number, respectively). By plotting the search result, we can also derive more insights about laws related to Golden Power rules, such as identifying the most central law or clusters that characterize periods, such as Covid- or PNRR-related (i.e., the recovery Italian plan) Golden Power legislation

siders the problems of dealing with the interconnections among laws by leveraging the graph relationships to retrieve a context. We are not affected by the token limit of embedding models, which is often lower than the law length, since we exploit the semantics and the properties provided by the Property Graph structure instead of working on a document level.

While our evaluation datasets of true, relevant acts for the thematic areas are of high quality, as they are collected and produced by the Italian legislator, we do not have a ranking of such laws. In fact, the data source rarely indicates which law is the most important for each area and never ranks laws based on their relevance. This forced us to adapt the evaluation metric (i.e., the DCG) to the binary scenario at a cost in terms of the interpretation of the results.

This also applies to the textual input of the system (Table 3), which was collected directly from the source of the evaluation datasets (i.e., the title and subtitle of the website pages containing the list of relevant laws), while potential users might write input queries of lower quality.

6 Conclusions

LegisSearch is an advanced search system that combines Knowledge Graph technology with state-of-the-art embedding models and LLMs to navigate a complex legislative system, such as the Italian one. We demonstrated the overall excellent quality of our LegisSearch system, which is capable of suggesting relevant laws in multiple and diverse thematic areas.

Our system is based on graphs, allowing us to build context-aware retrieval systems, leveraging graph traversal semantics and features to compute more useful embeddings.

LLMs also play an important part, as they can act as query expanders, interpreting the often partial initial input the user provides. This step supports the embedding model in creating a better vector representation of the legislation. At the same time, the LLM output can be used as a filtering agent to identify potential laws that might be of interest, upon which the embedding model can perform the retrieval task.

In future work, we aim to further deepen the role of a graph data model in legislative search, for instance, by investigating the search strategies enabled by graph traversals and/or network theory; they could play a crucial role in ranking the search results, yielding to a search system that guarantees precision in addition to recall. In addition, we aim to apply the same system to other legislations or consider multiple levels of legislation at the same time, such as the European legislation (above the national one) or the regional/local legislation (below the national one).

Resources LegisSearch has been implemented as a Python module and can be found at https://github.com/andreac0/LegisSearch. The graph database of the Italian legislation is available at https://doi.org/10.5281/zenodo.13375510.

Appendix A Data Sources

Table 7 lists the official sources from which the datasets were retrieved.

Table 7 Official data sources for the evaluation datasets

Category	Source Link
Pensions Regulation	https://www.lavoro.gov.it/temi-e-priorita/previdenza/Pagine/Normativa
Chemical Substances	https://www.mase.gov.it/pagina/reach-normativa-nazionale
Jobs and Occupation	https://www.lavoro.gov.it/temi-e-priorita/occupazione/Pagine/Normativa
Fuels	https://www.mase.gov.it/pagina/combustibili-uso-trazione-normativa-nazionalehttps://www.mase.gov.it/pagina/combustibili-uso-civile-industriale-e-marittimo-normativa-nazionale
Nuclear Energy	https://www.mase.gov.it/pagina/normativa-di-riferimento-0
Ozone Substances	https://www.mase.gov.it/pagina/normativa
Plant Protection Products	https://www.mase.gov.it/pagina/normativa-prodotti-fitosanitari

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

References

Anelli VW, Brienza E, Recupero M, Greco F, De Maria A, Di Noia T, Di Sciascio E (2022) Navigating the Legal Landscape: Developing Italy's Official Legal Knowledge Graph for Enhanced Legislative and Public Services. In: Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, CEUR Workshop Proceedings

Angelidis I, Chalkidis I, Nikolaou C, Soursos P, Koubarakis M (2018) Nomothesia: A linked data platform for Greek legislation. MIREL 2018 Workshop on MIning and REasoning with Legal texts

Aslam JA, Yilmaz E, Pavlu V (2005) The maximum entropy method for analyzing retrieval measures. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA, SIGIR '05, pp 27–34, https://doi.org/10.1145/1076034.1076042

Athan T, Boley H, Governatori G, Palmirani M, Paschke A, Wyner A (2013) OASIS LegalRuleML. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law. Association for Computing Machinery, New York, NY, USA, ICAIL '13, pp3–2. https://doi.org/10.1145/2514601.2514603

Avazpour I, Pitakrat T, Grunske L et al (2014) Dimensions and Metrics for Evaluating Recommendation Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 245–273. https://doi.org/10.1007/978-3-642-45135-5_10

Barabucci G, Cervone L, Palmirani M, Peroni S, Vitali F (2009) Multi-layer markup and ontological structures in Akoma Ntoso. International Workshop on AI Approaches to the Complexity of Legal Systems. Springer. Springer, Beijing, China, pp 133–149

Bellandi V, Castano S, Ceravolo P, Damiani E, Ferrara A, Montanelli S, Picascia S, Polimeno A, Riva D (2022) Knowledge-Based Legal Document Retrieval: A Case Study on Italian Civil Court Decisions. In: CEUR Workshop Proceedings, CEUR-WS, pp 1–13

Bianchi F, Terragni S, Hovy D et al (2021) Cross-lingual contextualized topic models with zero-shot learning. In: Merlo P, Tiedemann J, Tsarfaty R (eds) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, pp 1676–1683, https://doi.org/10.18653/v1/2021.eacl-main.143, https://aclanthology.org/2021.eacl-main.143

Buckley C, Voorhees EM (2017) Evaluating evaluation measure stability. ACM SIGIR Forum. ACM New York, NY, USA, pp 235–242

Sansone C, Sperlí G (2022) Legal Information Retrieval systems: State-of-the-art and open issues. Inf Syst 106:101967. https://doi.org/10.1016/j.is.2021.101967, https://www.sciencedirect.com/science/article/pii/S0306437921001551

Castano S, Falduti M, Ferrara A, Montanelli S (2019a) Law data science and ethics: the CRIKE approach. In: CEUR WORKSHOP PROCEEDINGS, CEUR-WS, pp 1–9

Castano S, Falduti M, Ferrara A, Montanelli S (2019b) The CRIKE Data-Science Process for Legal Knowledge Extraction. In: CEUR WORKSHOP PROCEEDINGS, CEUR

Chalkidis I, Androutsopoulos I, Aletras N (2019) Neural legal judgment prediction in English. In: Korhonen A, Traum D, Màrquez L (eds) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp 4317–4323. https://doi.org/10.18653/v1/P19-1424, https://aclanthology.org/P19-1424/

Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020) LEGAL-BERT: The Muppets straight out of Law School. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Chalkidis I, Fergadiotis M, Tsarapatsanis D, Aletras N, Androutsopoulos I, Malakasiotis P (2021) Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.22

Chalkidis I, Kampas D (2019) Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif Intell Law 27(2):171–198. https://doi.org/10.1007/s10506-018-9238-9

Chouhan A, Gertz M (2024) LexDrafter: Terminology Drafting for Legislative Documents Using Retrieval Augmented Generation. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024), pp 10448–10458

Colombo A, Bernasconi A, Ceri S (2025) An LLM-assisted ETL pipeline to build a high-quality knowledge graph of the Italian legislation. Inf Process Manag 62(4):104082. https://doi.org/10.1016/j.ipm.2025.104082

Colombo A, Cambria F (2025) LLM-assisted Construction of the United States Legislative Graph. In: VLDB 2025 Workshop: LLM+Graph, https://www.vldb.org/2025/Workshops/VLDB-Workshops-2025/LLM+Graph/LLMGraph-2.pdf

Colombo A, Cambria F, Invernici F (2025) Legislative knowledge management with property graphs. In: Proceedings of the workshops of the EDBT/ICDT 2025 joint conferenceco-located with the EDBT/ICDT 2025 Joint Conference, Barcelona, Spain, March 25, 2025, CEUR-WS. org, pp 1–8

Cormack GV, Lynam TR (2006) Statistical precision of information retrieval evaluation. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA, SIGIR '06, pp 533–540. https://doi.org/10.1145/1148170.1148262

Deng Y (2022) Recommender Systems Based on Graph Embedding Techniques: A Review. IEEE Access 10:51587–51633. https://doi.org/10.1109/ACCESS.2022.3174197

Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, Goyal A (2024) The llama 3 herd of models. arXiv:2407.21783

Elasticsearch BV (2025) ElasticSearch: The Leading Distributed Search and Analytics Engine. https://www.elastic.co/elasticsearch, Accessed: 2025-04-13

European Union Publications Office (2023) A Common Structured Format for EU Legislative Documents. Available online at: https://op.europa.eu/it/web/eu-vocabularies/akn4eu, last accessed on 01.05.2024
Fautsch C, Savoy J (2010) Adapting the tf idf vector-space model to domain specific information retrieval. In: Proceedings of the 2010 ACM symposium on applied computing, pp 1708–1712
Francis N, Green A, Guagliardo P, Libkin L, Lindaaker T, Marsault V, Plantikow S, Rydberg M, Selmer P, Taylor A (2018) Cypher: An evolving query language for property graphs. In: Proceedings of the 2018 international conference on management of data, pp 1433–1445
Griffo C, Teixeira MD, Almeida JP, Gailly F, Guizzardi G (2020) LawV: Towards an ontology-based visual modeling language in the legal domain. In: Proceedings of the XIII Seminar on Ontology Research in Brazil and IV Doctoral and Masters Consortium on Ontologies (ONTOBRAS 2020), CEUR Workshop Proceedings, vol 2728. CEUR-WS, pp 14
Hoekstra R, Breuker J, Di Bello M, Boer A (2007) The LKIF Core Ontology of Basic Legal Concepts. LOAIT 321:43–63
Huang Z, Low C, Teng M, Zhang H, Ho DE, Krass MS, Grabmair M (2021) Context-aware legal citation recommendation using deep learning. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. ACM, ICAIL '21, pp 79–88, https://doi.org/10.1145/3462757.3466066
International Organization for Standardization (2024) Information technology – Database languages – GQL (Graph Query Language). available online: https://www.iso.org/standard/76120.html
Istituto Poligrafico e Zecca dello Stato (2024) Normattiva Website. https://www.normattiva.it/ricerca/avanzata (Accessed on 05/15/2024)
Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20(4):422–446. https://doi.org/10.1145/582415.582418
Jiang D, Liu Y, Liu S, Zhao JE, Zhang H, Gao Z, Zhang X, Li J, Xiong H (2023) Mistral 7B. https://doi.org/10.48550/ARXIV.2310.06825, arXiv:2310.06825
Kanwal S, Nawaz S, Malik MK, Nawaz Z (2021) A Review of Text-Based Recommendation Systems. IEEE Access 9:31638–31661. https://doi.org/10.1109/ACCESS.2021.3059312
Kekäläinen J (2005) Binary and graded relevance in IR evaluations: comparison of the effects on ranking of IR systems. Inf Process Manage 41(5):1019–1033
Li R, Zhao X, Moens MF (2022) A Brief Overview of Universal Sentence Representation Methods: A Linguistic View. ACM Comput Surv 55(3). https://doi.org/10.1145/3482853
Library of Congress (2025) Congress.gov API. https://api.congress.gov/, (Accessed on 09/18/2024)
Licari D, Comandè G (2022) Italian-legal-bert: A pre-trained transformer language model for italian law. In: Symeonidou, Danai and Yu, Ran and Ceolin, Davide and Poveda-Villalón, María and Audrito, Davide and Caro, Luigi Di and Grasso, Francesca and Nai, Roberto and Sulis, Emilio and Ekaputra, Fajar J. and Kutz, Oliver and Troquard, Nicolas (ed) Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, CEUR Workshop Proceedings, vol 3256. CEUR, Bozen-Bolzano, Italy, https://ceur-ws.org/Vol-3256/#km4law3, ISSN: 1613–0073
Licari D, Comandè G (2024) ITALIAN-LEGAL-BERT models for improving natural language processing tasks in the Italian legal domain. Comput Law Sec Rev 52:105908. https://doi.org/10.1016/j.clsr.2023.105908
Liu Q, Chen N, Sakai T, Wu XM (2024) ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, NY, USA, WSDM '24, pp 452–461. https://doi.org/10.1145/3616855.3635845
Lupo C, Vitali F, Francesconi E, Palmirani M, Winkels R, de Maat E, Boer A, Mascellani P (2007) General XML format (s) for legal Sources. Technical report, IST-2004-027655 ESTRELLA European project for Standardised Transparent Representations in order to Extend Legal Accessibility
Mathew P, Kuriakose B, Hegde V (2016) Book Recommendation System through content based and collaborative filtering method. In: 2016 International conference on data mining and advanced computing (SAPIENCE), pp 47–52, https://doi.org/10.1109/SAPIENCE.2016.7684166
Matsyupa KV, Maksimova SY, Ilyukhin NI (2022) Legalese or lawspeak – diversity within the unity. In: European proceedings of social and behavioural sciences. European Publisher
Melville P, Sindhwani V (2010) Recommender systems. Encyclopedia. Mach Learn 1:829–838
OASIS (2018) Akoma Ntoso Version 1.0 becomes an OASIS Standard. Available online at: https://www.oasis-open.org/news/announcements/akoma-ntoso-version-1-0-becomes-an-oasis-standard/, last accessed on 01.05.2024

Ong K, Haw SC, Ng KW (2020) Deep Learning Based-Recommendation System: An Overview on Models, Datasets, Evaluation Metrics, and Future Trends. In: Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems. Association for Computing Machinery, New York, NY, USA, CIIS '19, pp 6–11, https://doi.org/10.1145/3372422.3372444

Palmirani M (2021) Lexdatafication: Italian Legal Knowledge Modelling in Akoma Ntoso, Springer International Publishing, pp 31–47. https://doi.org/10.1007/978-3-030-89811-3_3

Palmirani M (2019) Akoma Ntoso for making FAO resolutions accessible. Knowledge of the Law in the Big Data Age Frontiers in Artificial Intelligence and Applications 317:159–169

Parnami A, Lee M (2022) Learning from few examples: A summary of approaches to few-shot learning. arXiv:2203.04291

Pennington J, Socher R, Manning C (2014) GloVe: Global vectors for word representation. In: Moschitti A, Pang B, Daelemans W (eds) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1532–1543. https://doi.org/10.3115/v1/D14-1162, https://aclanthology.org/D14-1162

Maragheh RY, Fang C, Irugu CC, Parikh P, Cho J, Xu J, Sukumar S, Patel M, Korpeoglu E, Kumar S, Achan K (2023) LLM-TAKE: Theme-Aware Keyword Extraction Using Large Language Models. In: 2023 IEEE International conference on big data (BigData). IEEE Computer Society, Los Alamitos, CA, USA, pp 4318–4324. https://doi.org/10.1109/BigData59044.2023.10386476

Raza S, Ding C (2022) News recommender system: a review of recent progress, challenges, and opportunities. Artif Intell Rev 55(1):749–800. https://doi.org/10.1007/s10462-021-10043-x

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui K, Jiang J, Ng V, et al (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 3982–3992. https://doi.org/10.18653/v1/D19-1410, https://aclanthology.org/D19-1410

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 3982–3992

Ren Y, Han J, Lin Y, Mei X, Zhang L (2022) An Ontology-Based and Deep Learning-Driven Method for Extracting Legal Facts from Chinese Legal Texts. Electronics 11(12):1821. https://doi.org/10.3390/electronics11121821

Robertson S, Zaragoza H (2009) The probabilistic relevance framework: BM25 and beyond. Found Trends® Inf Retrieval 3(4):333–389

Rodríguez-Doncel V, Navas-Loro M, Montiel-Ponsoda E, Casanovas P (2018) Spanish Legislation as Linked Data. In: Proceedings of the 2nd Workshop on Technologies for Regulatory Compliance co-located with the 31st International Conference on Legal Knowledge and Information Systems (JURIX 2018), Groningen, The Netherlands, December 12, 2018. CEUR-WS.org, Groningen, The Netherlands, CEUR Workshop Proceedings, pp 135–141

Salton G (1989) Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc, USA

Shani G, Gunawardana A (2011) Evaluating recommendation systems. Recommender systems handbook pp 257–297

Castano S, Ferrara A, Furiosi E, Montanelli S, Picascia S, Riva D, Stefanetti C (2024) Enforcing legal information extraction through context-aware techniques: The ASKE approach. Comput Law Sec Rev 52:105903. https://doi.org/10.1016/j.clsr.2023.105903, https://www.sciencedirect.com/science/article/pii/S0267364923001139

Castano S, Falduti M, Ferrara A, Montanelli S (2022) A knowledge-centered framework for exploration and retrieval of legal documents. Inf Syst 106:101842. https://doi.org/10.1016/j.is.2021.101842, https://www.sciencedirect.com/science/article/pii/S0306437921000788

Syed MH, Huy TQ, Chung ST (2022) Context-aware explainable recommendation based on domain knowledge graph. Big Data Cognitive Comput 6(1):11

Thorat PB, Goudar RM, Barve S (2015) Survey on collaborative filtering, content-based filtering and hybrid recommendation system. Int J Comput Appl 110(4):31–36

UN System Chief Executives Board for Coordination (2017) Akoma Ntoso for the United Nations. Available online at: https://unsceb.org/unsif-akn4un, last accessed on 01.05.2024

Van Meteren R, Van Someren M (2000) Using content-based filtering for recommendation. In: Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop, Barcelona, pp 47–56

Wang H, Zhang F, Wang J, Zhao M, Li W, Xie X, Guo M (2018) RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, New York, NY, USA, CIKM '18, pp 417–426. https://doi.org/10.1145/3269206.3271739

Wang L, Yang N, Huang X, Jiao B, Yang L, Jiang D, Majumder R, Wei F (2022) Text embeddings by weakly-supervised contrastive pre-training. arXiv:2212.03533

Wang L, Yang N, Huang X, Jiao B, Yang L, Jiang D, Majumder R, Wei F (2023a) Improving text embeddings with large language models. arXiv:2401.00368

Wang L, Yang N, Huang X, Jiao B, Yang L, Jiang D, Majumder R, Wei F (2024) Multilingual e5 text embeddings: A technical report. arXiv:2402.05672

Wang L, Yang N, Wei F (2023b) Query2doc: Query Expansion with Large Language Models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Assoc Computat Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.585

Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M (2020) MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Proceedings of the 34th international conference on neural information processing systems. Curran Associates Inc., Red Hook, NY, USA, NIPS '20

Wang Y, Yao Q, Kwok JT, Ni LM (2020) Generalizing from a Few Examples: A Survey on Few-shot Learning. ACM Comput Surv 53(3). https://doi.org/10.1145/3386252

Wehnert S, Padmanabhan V, De Luca EW (2024) Hybrid Legal Norm Retrieval: Leveraging Knowledge Graphs and Textual Representations. IOS Press. https://doi.org/10.3233/faia241245

Winkels R, Boer A, Vredebregt B, Van Someren A (2014) Towards a legal recommender system. In: Legal knowledge and information systems. IOS Press, p 169–178

Yelmen I, Gunes A, Zontul M (2023) Multi-Class Document Classification Using Lexical Ontology-Based Deep Learning. Appl Sci 13(10):6139. https://doi.org/10.3390/app13106139

Zhang Y, Liu Z, Wen Q, Pang L, Liu W, Yu PS (2024) AI Agent for Information Retrieval: Generating and Ranking. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, New York, NY, USA, CIKM '24, pp 5605–5607. https://doi.org/10.1145/3627673.3680120

Zhang Z, Wang L, Xie X, Pan H (2018) A Graph Based Document Retrieval Method. In: 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)), pp 426–432. https://doi.org/10.1109/CSCWD.2018.8465295

Zhong H, Guo Z, Tu C et al (2018) Legal judgment prediction via topological learning. In: Riloff E, Chiang D, Hockenmaier J, et al (eds) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 3540–3549. https://doi.org/10.18653/v1/D18-1390, https://aclanthology.org/D18-1390

Zhong H, Xiao C, Tu C et al (2020) "How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence". In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for computational linguistics, Online, pp 5218–5230, https://doi.org/10.18653/v1/2020.acl-main.466, https://aclanthology.org/2020.acl-main.466

Authors and Affiliations

Andrea Colombo[1] · Anna Bernasconi[1] · Luigi Bellomarini[2] · Luigi Guiso[3] · Claudio Michelacci[3] · Stefano Ceri[1]

✉ Andrea Colombo
andrea1.colombo@polimi.it

Anna Bernasconi
anna.bernasconi@polimi.it

Luigi Bellomarini
luigi.bellomarini@bancaditalia.it

Luigi Guiso
luigi.guiso55@gmail.com

Claudio Michelacci
c.michelacci1968@gmail.com

Stefano Ceri
stefano.ceri@polimi.it

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Giuseppe Ponzio, 34, Milan 20133, Italy

[2] Dipartimento di Informatica della Banca d'Italia Centro Donato Menichella, Largo G. Carlo 1, Frascati 00044, Italy

[3] EIEF — Einaudi Institute for Economics and Finance, via Sallustiana, 62, Rome 00187, Italy