Modeling Legal Documents as Typed Linked Data for Relational Querying

Nada Mimouni

LIPN, Paris 13 University – Sorbonne Paris Cité & CNRS (UMR 7030)
F-93430, Villetaneuse, France
nada.mimouni@lipn.univ-paris13.fr

Abstract. Access to legal knowledge is particularly challenging to information retrieval systems. Not only is legal knowledge usually expressed in linguistically complex forms, but it is also structurally sophisticated (e.g pieces of legislation applicable to a case, version in force of a legal document, other related sources). Modeling the collection of documents in such complex domains requires taking into account the semantic content of the documents as well as their relational structure since documents are usually related to each other by various types of links. In this paper we describe two approaches for modeling and querying a collection of interlinked legal documents. The first approach is based on Formal Concept Analysis and Relational Concept Analysis to model and query the collection of documents. The second approach uses semantic web techniques (RDF, OWL and SPARQL). Different types of relational queries are discussed.

Keywords: Information Retrieval, Linked documents, Relational queries, Formal Concept Analysis, Relational Concept Analysis, Ontology, Semantic web.

1 Introduction

A collection of documents is usually represented as a set of documents. This is a very simplified view since in reality documents are get in a set of intertextual relations that condition their interpretation : a document should not be interpreted solely but with reference to the texts it cites, from which it derives or which derive from it. In the legal domain, documents are linked to each other with amendment, transposition, complementation, jurisprudence relations, etc. These links are not only made for documentary purposes. They also determine the legal validity of documents. For example, in French law, codification is the strong process of structuration of information and the links between documents must be explicitly expressed [1]. Consolidation requires that a legal act makes explicit reference to its successive amendements. Legal information access tools should take into account this complexity of legal material.

XML based standards have been defined to normalize the structure of legal texts, in order to facilitate the access and management of these data. The trend

is to use those standards in the process of legal drafting so as to solve the interoperability issue, which usefulness is obvious. In parallel, open government data initiatives are increasing (e.g. UK Government Linked Data) and many legal information access portals offer querying and search features on this data. However, the data made available are often underused.

Accessing such complex data, characterized by the extra abundance of cross references between legal texts (regulations, laws), requires a querying model integrating both semantic features and intertextual links. Our requirement analysis showed that the need for relational querying is critical from a legal point of view ("find by which texts a given order have been applied?", "what are the local texts that talk about noise that are valid at a given date?", "what are the texts that modify a given text?").

In this work, we propose two approaches which allow representing and querying in a unified manner the semantic content of documents, their structure and their intertextual relations. The proposed approaches are based on Formal and Relational Concept Analysis (FCA, RCA), and on semantic web techniques applied to documentary objects.

The paper first reviews the existing solutions (Sec. 2) and explains the requirements for relational querying (Sec. 3). Sections 4.1 and 4.2 describe the proposed approaches and show how the collections and queries are modeled. Those approaches are finally discussed in Section 5.

2 Related Work

In most specialized domains, documents, such as regulations or laws in the legal domain, must not be interpreted in isolation but in relation with other documents, with which they form "a collection of documents". Legal documents are linked to each other through various types of relations (*e.g.* amendment, transposition, implementation, etc.) and these links often determine their legal validity. We define a collection as a set of documents with semantic descriptors, associated metadata and various types of semantic links between them. Law corresponds thus to a large and highly interconnected network of documents. IR systems should make full use of the afforded richness when processing such complex data, thus exploiting the links, the documents structure as well as their semantic content.

Many efforts have been made to take intertextual links into account in an IR process. Semantic and relational search is handled by both general search engines and specialized legal IR systems in different ways: classic IR on semantic content then navigation with hyperlinks, boolean IR on semantic content then filtering according to links or semantic and intertextual queries.

2.1 Intertextuality in Existing IR Systems

Suppose we have a relational query of the form "what are documents (d') having a given type of link (l) with a document (d) talking about a given subject (s)?". Let's consider how the above types of systems deal with such a query.

- Generalist IR systems such as Google use the most trivial way to deal with intertextuality. The query is treated into two steps: a simple query on the semantic content (s) returns the document (d) and the user can then navigating the hyperlinks according to the type of link (l) to find the set of answers (d'). This category of systems do not allow for relational queries.
- In the second category we classify all systems that allow relational queries via attributes in the query such as XML native databases (queried with XPath, XQuery) and RDF data (queried with SPARQL). The query is treated in a first step as a boolean query on the semantic content (s) to find the set of d, then a filtering step is performed according to the XML elements specified in the query (for XML native repositories) or the set of constraints (in the case of SPARQL queries).
- The third category of systems consists of relational systems such as relational databases and relational concept analysis. Both types of systems allow encoding the references between documents in the model level and also formulating relational queries. The originality of this approach is that the documents collection is structured prior to being queried. For instance, in the case of RCA, a set of conceptual structures (called a relational lattice family) is build upon the semantic content of the documents and the links they hold between them. Then the query is executed against these relational structures to find relevant answers. The advantage of this approach is to allow for navigating within the created lattices to specialize or generalize the query if no exact answer is found.

2.2 Legal IR Systems

Legislative portals or legal access systems (*e.g.* Legifrance[1]) exist in most countries to enable a large and public access to the law. Based on XML standards, they offer rich functionalities such as hyperlink navigation, point-in-time access to historical and repealed documents versions[2], interactive generation of user-oriented up-to-date information[3].

However, so far, legal links between documents have been exploited in a limited way by IR systems. For instance, in Legifrance, explicit links are mainly dealt with manually. Some of them are included in the content of the data base (hyperlinks) and others are implemented as document attributes when the data base content is managed[4]. The UK Legislation site allows to search for changes made in the legislation since 2002[5]. The user can query the database either by specifying the modified legislation or the legal source that introduces the change. Whereas the system treats the general link "modifies/modified-by" as a relation

[1] www.legifrance.fr
[2] *e.g.* UK legislation (http://www.legislation.gov.uk/search/point-in-time)
[3] *e.g.* New South Wales legislation website (http://www.legislation.nsw.gov.au).
[4] Force (V), With force term (VT), Delayed effect (VD), Repealed (Ab), Canceled (A), Disjoint (D), Modified (M), Implied repeal (P), Substituted (S), Transferred (T).
[5] http://www.legislation.gov.uk/changes

between documents, more specific types of modifications are represented as document attributes. The Italian website Normattiva enables point-in-time access to legislation too[6], allowing thus to retrieve versions of a document in force at different dates.

The analyzed systems do not exploit explicit links between legal documents to their full potential. In order to illustrate this point we can think of a *continuum* from less to more operational representations of links in legal IR systems:

- Links are represented as strings in the text of the document: usually they appear in the final part of the document and are added manually (by an editorial team).
- Hyperlinks between documents: links are references that point to objects in the collection (other legal documents or fragments of those documents).
- Links are queryable as attributes: legal relations between documents are represented as attributes of the linked documents.
- Relational query: links are modeled as relations between documents in the collection. This allow for relational querying.

If we compare this continuum to the categories presented in the section 2.1, we notice that systems of the first and the second items belong to the first category, the third item systems belong to the second category and the last one corresponds to the third category. Our goal is to exploit the further end of this continuum, namely, the representation of *various types* of legal links as relations between documents in the collection. It is our assumption that such representation mirrors more precisely the way legal professionals conceive the network of legal provisions and will thus enable a more natural interaction between the user and the system.

3 Requirements for Relational Querying

Legal expert common queries show that it is important to distinguish and exploit different types of inter-document links. The query may deal with the case of application of a law text (for example: "find all application cases of a given order"), a validity date (for example : "which local texts deal with noise and are valid in a given date?") or modification links (for example: "which are the texts that modify another text?"). Table 1 give more examples of relational queries. To overcome such limitations, legal IR system should deal with the rich typology of relations linking the documents of a legal collection in order to enable relational querying.

4 Proposed Approaches

To meet these requirements, we propose two different approches for relational modeling and querying. The proposed approaches allow answering simple and

[6] http://www.normattiva.it/ricerca/avanzata/vigente

What conventions implement the recommendations that talk about termination ?
Which recommandation about benzene are implemented by conventions on occupational Cancer ?
Does a law text has been applied? and in which cases (give examples of case law) ?
What recommandations are implemented by conventions on air pollution ?
Given an order, what are the legal texts that it develops ?

Table 1. A sample of relational queries expressed by legal experts

relational queries on a collection of linked legal documents. This work is part of the LEGILOCAL project [7]. The collections of legal documents we are dealing with are characterized by:

- Different types of documents (laws, codes, editorial documents, etc.).
- A specific internal structure for each document type (sections, paragraphs, etc.).
- Various types of links between the different types of documents.
- Semantic descriptors annotating the documents w.r.t a semantic resource.

The first approach [2], based on FCA and RCA, creates classes of documents using their semantic contents and the links between them. Despite its consistency from a formal point of view, a major limit of this solution is the size of the created conceptual structures when applied to a big collection of documents. To tackle this problem, we proposed a second solution [3, 4], based on semantic web techniques (RDF, OWL, SPARQL), which is scalable and nevertheless adresses the problem of relational querying.

4.1 Conceptual Classification based on FCA and RCA

Figure 1 gives an overview of our approach, composed of four main steps:

Semantic content modeling: the semantic content of the documents is annotated and binary contexts are extracted based on those annotations allowing formal concept lattices to be build.

Intertextual structure modeling: the links between documents are identified and relational contexts are extracted based on those links allowing enriched relational lattices to be build.

Relational querying: the user creates a query, possibly as a combination of semantic descriptors and cross-references constraints.

Search and results: the search algorithm analyses the query and looks for relevant answers on the lattices. The user can get traditional list or graphs of result documents. Alternatively, he can directly visualize results in the lattice structure which can be further explored to get approximated results.

[7] LEGILOCAL is an FUI project 2010-13. See http://www.mondeca.com/fr/R-D/Projets/LegiLocal-Projet-FUI-9-Cap-digital-2010-2013.

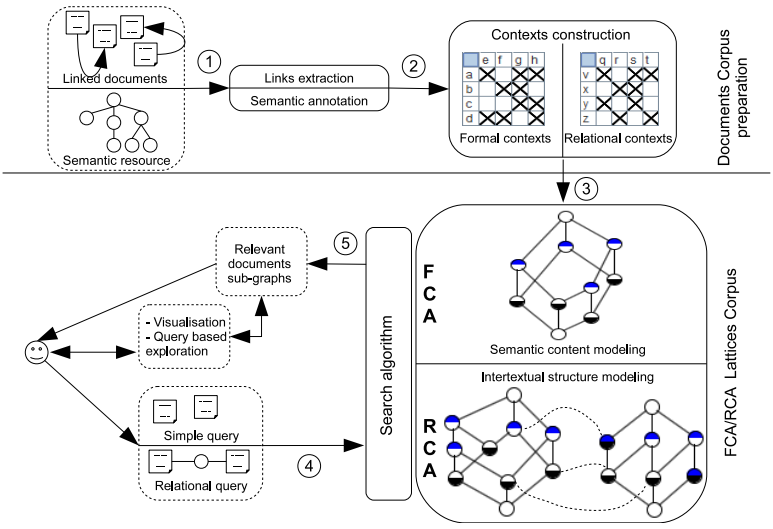

Fig. 1. Overview of the relational IR approach

The semantic content of documents is first modeled as a formal context which describes a binary relation between a set of objects and a set of attributes *(object x attributes)*. The objects correspond to documents. The attributes correspond to the semantic descriptors characterizing the content of these documents. In an information retrieval (IR) perspective, the lattice built by the FCA on binary contexts gathers all possible combinations of documents attributes. These combinations are represented by the intentions of concepts having as extensions all the documents sharing these properties. To answer a query, the search algorithm identifies the class of documents sharing the maximum number of attributes with the query.

We use RCA, the relational extension of FCA, to take into account the cross-references dimension in the modeling of the collection. The approach builds a family of relational contexts, from binary contexts *(documents* $\times$ *semantic descriptors)* and a relation represented separately in a new context defining a type of relation between documents *(documents* $\times$ *documents)*. This family of contexts forms the starting point for the creation of corresponding conceptual structures called Relational Lattice Family [5]. RCA is able to take into consideration different types of links, which are represented by different relational contexts.

Simple Queries We call "simple queries" the queries that are expressed as a set of semantic descriptors. For example *"Which orders talk about abnormally annoying noise (bag) and sound disturbance (ns)?"*. The key words "abnormally annoying noise" and "sound disturbance" are considered as semantic descriptors

annotating the documents which type is "orders". The initial lattice built with FCA represents the set of all the simple queries based on semantic descriptors combination which are satisfiable, i.e. return orders (all descriptors combinations associated to a non null extension). If the query corresponds to the intension of a concept having an extension, documents of this extension are returned as an answer to the query. If the query corresponds to an intension without a proper extension, we can propose a specialization or generalization of the query: this is the advantage of the proposed approach of relational information retrieval.

Relational Queries Our model allows also to answer relational queries. Relational queries contain not only a set of semantic descriptors but also relational indicators between documents. The relational indicators express one or different types of cross references between one or more types of documents. For example *"Which orders talk about abnormally annoying noise (bag) and make reference to decrees talking about soundproofing (ip)?"*. The key words "noise" and "acoustic pollution" are considered as the semantic descriptors annotating respectively the documents which are of type "orders" and "decrees". Different types of relational queries can be handled: legal text to legal text relational query, legal text to semantic category relational query and semantic category to semantic category relational query. Answers to these types of queries are graphs of linked documents.

4.2 An Ontology of Legal Documents Collection

In the second approach we propose an ontology based document model to support the sharing of documents of French local administrations. This ontology has been designed on the basis of Legilocal requirements analysis. It allows to represent all information on legal documents: 1) the structure of a document (sections, paragraphs, etc.), 2) the time frame in which it is registered, 3) the semantic description of its content using concepts or entities in the considered domain, 4) its type (law, decree, etc.) and 5) its relationships with other documents (modification, repeal, transposition, etc.). Our document ontology is structured into three main modules which allow to model the above properties : document module (properties 1 and 2), the semantic module (property 3) and the collection module (properties 4 and 5). Details of each module are given in the following.

Types and Structures of Documents In the Metalex ontology, resources are typed according to the FRBR convention as work, expression, manifestation and item. In our model, we focus on the two upper levels, namely work and expression, in order to represent the different versions of articles and documents. Moreover, those documents have different types (French legislation, court decisions, local acts as well as editorial documents). These various documents have different structures and are characterized by different metadata. Indeed, to prepare a municipal act on a particular subject, local administrators have to investigate national legislation and case law on the same subject. In order to

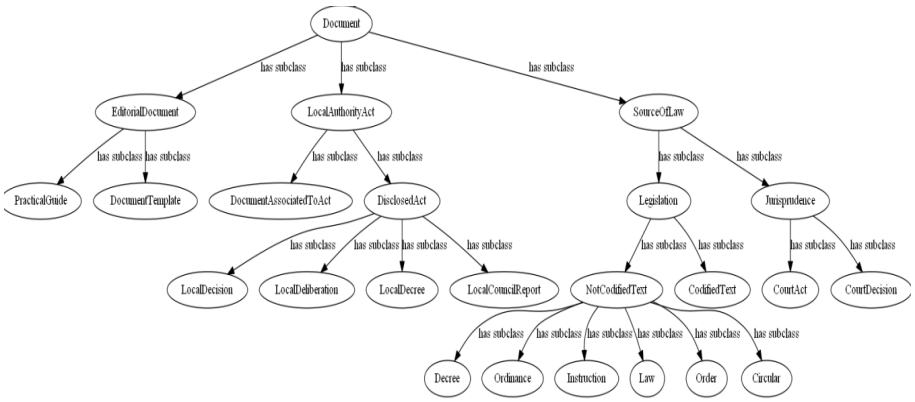

Fig. 2. Hierarchy of documents types

help them, our approach aims to provide semantic search in national legislation and case law, as well as in local acts of other municipalities on the same subject and even some editorial documents. These semantic search facilities require that the documents be annotated with both topics and interdependencies.

As Metalex ontology was firstly designed to model legislation, we extend it with a document typology (see figure 2) that enables us to describe specific properties for each type of documents. For example, we want to be able to specify the structures for certain local acts in order to check their conformance, and some related properties such as the local organization and the person in charge of the document which are specific for each local act. We propose a fine-grained description for legislation text in which the basic unit is the article (which has an independent life-cycle, and could be cited and returned as answer to a user query). On the contrary, for local acts, we do not go through fine-grained description and keep coarse decomposition.

Documents Relationships We want to answer queries such as: "*What are the judgments that implement articles 4 and 5 of the law on minor work?*" or "*Which amendments are made to the article7 of law1955?*"). To reach this goal we propose to model a collection of documents as a semantic network based on a fine description of the types of citations. Our reference model differs from the Metalex one in two respects. On the one hand, we refine the generic reference notion. A broad distinction opposes the citations that refer to a textual object and the semantic annotations that refer to non-textual objects, but we also introduce various semantic citation subtypes (see Fig. 3). On the other hand, we simplify the event-based model of Metalex by encoding references as direct relations, that are directly exploitable for search and visualization purposes.

In our model, each type of reference property is associated to specific domain and range, which allows to specify not only to which types and parts of texts (for

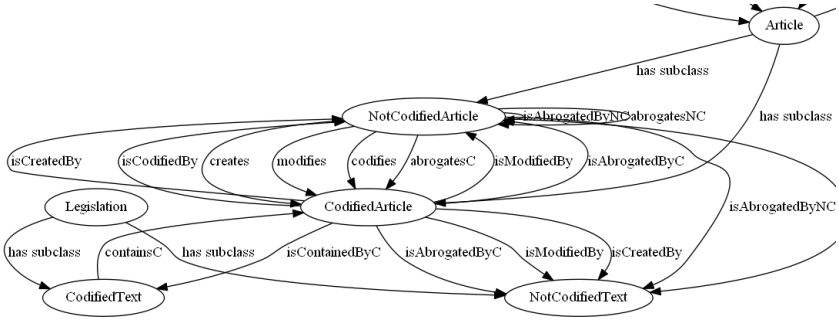

Fig. 3. Various types of links between different types of documents

citations) or semantic classes (for annotations) it refers but also in which types of texts and parts of texts it may appear. Actually, we introduce an opposition between document fragments and units to distinguish the document parts that are citable (units or `CitableBibliographicObject` in Metalex ontology) from those that are not (mere fragments). For instance, we consider whole documents and articles as units but not the preamble of a law. The same opposition holds for the search results: only graphs of document units can be returned to answer a relational query. On the contrary, semantic annotations can be attached to any fragment of text.

Semantic Annotations In this work, the term "semantic annotation" denotes the references that are not citations. We define semantic annotations as references referring to ontological entities that do not represent documents or parts of documents.

5 Discussion

In this work we have introduced a solution to the problem of the complexity of legal sources. Using semantic content descriptors, documents typology and cross references between documents, we have introduced two approaches to model and search within a collection of interlinked legal documents. This allows to answer relational queries and return graphs of linked documents. The first approach is based on FCA and RCA to model the collection as conceptual structures. We have experimented relational queries to explore and query this relational model and return relevant documents or graphs of documents. A more operational solution based on semantic technologies (RDF modeling and SPARQL querying) is introduced as a second approach. We propose an ontology-based model to tackle the complexity of legal sources and to model collections of interlinked legal documents. Beyond traditional legal search, those models already support fine-grained semantic and relational IR functionalities.

Adopting an integrated document model to encode the structure of the documents, their semantic annotations and the semantic structure of the collection enables to process complex queries combining structural, intertextual and content search criteria. For instance, if a local administrator wants to find examples of local acts dealing with "rural roads" and based on a particular decree d, he can express a query combining constraints on semantic annotation (`refers to` the class `chemin rural`) and document references (`cites` the decree d). Our future research will include conceiving user friendly interfaces, allowing to easily create a relational query based on the collection characteristics (semantic descriptors, documents types, references), and also to display results returned as graphs of documents.

References

1. Bourcier, D., Mazzega, P.: Codification, law article and graphs. In: Lodder, A., (eds.), L.M. (eds.) Legal Knowledge and Information Systems, JURIX. pp. 29–38. IOS Press (2007)
2. Mimouni, N., Fernàndez, M., Nazarenko, A., Bourcier, D., Salotti, S.: A relational approach for information retrieval on xml legal sources. In: Francesconi, E., Verheij, B. (eds.) ICAIL. pp. 212–216. ACM (2013)
3. Mimouni, N., Nazarenko, A., Salotti, S.: Une ontologie documentaire pour la recherche d'information relationnelle. In: Actes du 5ème Atelier Recherche d'Information SEmantique (IC-RISE 2013). Lille, France (Juillet 2013)
4. Mimouni, N., Salotti, S., Paul, E.: Modeling collections of french local administration documents. In: Proceedings of Jurix 2013 (accepted). Bologna, Italie (December 2013)
5. Rouane, M.H., Huchard, M., Napoli, A., Valtchev, P.: A proposal for combining formal concept analysis and description logics for mining relational data. In: ICFCA, pp. 51–65. Springer (2007)