Vorname Nachname et al. (Hrsg.): 18. Fachtagung "Datenbanksysteme für Business, Technologie und Web" (BTW), 2019,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 1

Context Selection in a Heterogeneous Legal Ontology

Sabine Wehnert, Wolfram Fenske, Gunter Saake[1]

Abstract: Ontology building in the legal domain is subject to ongoing research. Taxonomic ontologies provide for instance concept hierarchies for term definitions, annotations, query expansion and support for inferences. However, the context-dependent application of statuatory legal texts is hard to model, often leading to a limited ontology scope and fixed terminology to avoid conflicts. In previous work, we presented a method to create a lightweight heterogeneous ontology from textbooks offering connections between laws, while avoiding an error-prone and costly ontology alignment step. In our ontology, laws are linked by common contexts. We propose a new data model, so that the context can be explored and selected by a user, which is necessary for many applications, such as recommender systems. To obtain the relevant user context, we added a mechanism to retrieve linked laws from our ontology, given a scope of user interest and context information for each law.

Keywords: Heterogeneous Ontologies, Legal Text Linking, Context Selection, Full-text Search.

1 Introduction

Nowadays, people are overwhelmed by the amount of legal regulations to consider. Especially for international companies, it is becoming increasingly difficult to ensure that decisions comply with all laws. Therefore, our greater research goal is to provide a decision support system which monitors legislation and informs companies about relevant regulatory changes so they can update their processes to ensure legal compliance[2]. It is not trivial to determine relevance though, since it depends on many factors (e.g., user context, conditioned law applications). There are two main approaches to incorporate domain knowledge into a system. First, expert systems define answers for manually pre-defined queries, which is very costly. Second, ontologies are an approach to ensure a common understanding of the concepts of a domain, such that a query can be answered by rule-based reasoning. Ontologies are built from terms of increasing abstraction level, forming a concept hierarchy. For the legal domain, they can fulfill several reasoning tasks, for example finding consequences of a prohibition or obligation, or determining analogies between legal cases [Na12]. There are many legal ontologies [Aj16; Bu16; Ho07; So07], but they are limited to a highly specific domain (e.g., national law) or too abstract to be used as a stand-alone knowledge representation. For laws only describing rules for abstract events, a subsumption to real-world situations is necessary to understand whether a law applies to a given scenario [Di07]. As a consequence, experts

[1] Otto von Guericke University Magdeburg, <firstname>.<lastname>@ovgu.de
[2] The work is supported by Legal Horizon AG, Grant No.:1704/00082

create new or extend existing ontologies to fulfill the requirements of the respective user. This is time-consuming and costly. We therefore propose a mechanism which we call *context selection*, that allows users identify the laws applicable to their business scenario. It is a challenging task to model this user context for all possible situations, so we seek to automate this step by using external sources. In previous work, we extracted concept hierarchies from legal textbooks which capture law application contexts [We18]. In particular, we annotated table of contents elements (*TOC*) and applied them as a hierarchy for any cited legal text within the respective book. Our process is depicted in Figure 1. From each sentence containing a reference (*REF*), we compute a so-called citation summary (*CS*), the reason for citing the legal text in the given section.

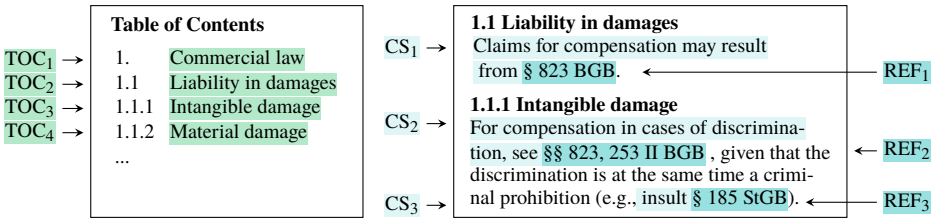

Fig. 1: Annotation process for references *(REF)*, citation summary *(CS)* and table of contents *(TOC)*.

For example, the citation summary for § 185 StBG covers the word *insult* as a reason for citing. Each book forms its own concept hierarchy from law citations within textbook sections. We follow the notion of a heterogeneous ontology by Visser and Cui [Vi98] who cluster concept hierarchies without any alignment. A cluster in our case can be a collection of similar books on one broad topic, such as banking or IT law. Clusters can be seen as knowledge modules which can be applied and queried separately. On one hand, books within each cluster may provide different perspectives on the same topic, and on the other hand they can enrich the knowledge base with their distinct content. Based on his or her interests, a user can select relevant concept hierarchies. In this work, we propose a mechanism for a user to navigate within the heterogeneous structure. Therefore, we develop a context selection method, based on two use cases:

For the first use case (a), the user searches for possible applications of a law. The user receives all occurrences of this law and context information from the concept hierarchy. Then, the user can select one context and determine the level of abstraction. After this context selection, all laws cited in the same context are retrieved. In the second use case (b), the user has a passive role and just receives an alert when a law from his or her context has changed. The context has to be selected beforehand, for example, by subscribing to one law and selecting one context description. By automatically expanding the subscription to laws referenced in the same context, users will also receive notifications about changes to relevant laws they were unaware of. To support these use cases, a suitable data model is needed. In this paper, we focus on the following aspects: First, we investigate an indexing method for law reference lookups. Second, we develop a data model for interactive graph traversal for knowledge extracted from legal textbooks with regard to the previously described use cases.

2 Context Selection

In this section, we describe the properties of our extracted data to choose appropriate search and data storage methods. Then, we explain our data model and methods to navigate within the data. Figure 2 illustrates the proposed workflow. We refer to the books which contain user-relevant contexts as the scope of interest S, which can be composed of several textbook clusters. The extracted concept hierarchies are stored in a graph database and replicated into a full-text search engine. Then, we consider one specific query for all occurrences of a reference to a law and its context. Given the query response, the user can select an appropriate context by choosing a cutoff point in the respective concept hierarchy. For instance, in Figure 1, the user can select the cutoff at section *1.1.1 Intangible damage*, so that any reference from *1.1.2 Material damage* will be excluded.

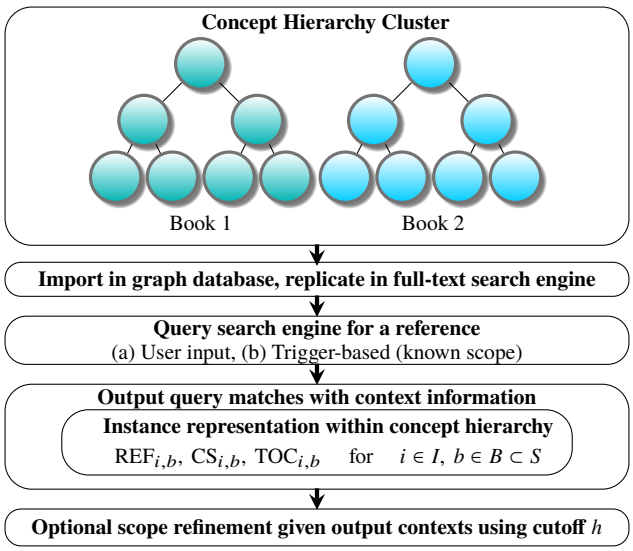

Fig. 2: Proposed workflow for context selection, given an output of matching instances i for a query. An instance consists of a reference *(REF)*, the citation summary *(CS)* and hierarchical context from the table of contents *(TOC)*. All instances I are drawn from textbooks B within the user-defined scope S. This scope is reduced by context selection: A user specifies a cutoff level h to prune higher and more abstract concept hierarchy levels and thereby removes connections to irrelevant laws.

2.1 Search Indexing

Using the context information, the user gets an overview of the concepts related to a legal text. We expect slight variations in law citations, such as roman or arabic numerals referring to a specific part of a law. Despite the need for approximate string matching, called fuzziness, the amount of variation needs to be controlled because the names of statute books can differ by a single letter, such as the German civil code *BGB* and the commercial code *HGB*.

Instead, we use exact matching for the query with references (while allowing for other present strings). In case no result is obtained (e.g., due to spelling mistakes), we employ fuzzy search. Despite the advantage of approximate string matching in full-text indexes, the data can also be stored as a graph. Graph data allow for connections regardless of hierarchy level and are optimized to process multiple outgoing relationships from one node. Later on, we plan to analyze the content of legal text documents, which can result in further links between laws, a so-called citation network. Graph data can be updated easily, however, the information from a book will not change, once it has been inserted into the ontology. Hierarchical storage of the data, for example in JSON format, is therefore also an option for search in hierarchical data, especially when approximate string matching is required. In our current prototype, we load all references to legal text as nodes into a graph database, create *PartOf* relationships with each corresponding table of contents element and replicate the data in hierarchical storage. Both systems have their own advantages - approximate string matching and graph traversal - and we can select for each query where to process it.

2.2 A Data Model for Graph Traversal over Linked Legal Texts

As a first step toward graph traversal, we transform the data which were previously extracted[3] into two separate `csv` files, one for the entities and one for their relationships. The data model is shown in Figure 3. We store all entities using the same `LABEL` *Node* in the graph and the search index. Furthermore, we define an additional field for each of them to preserve entity `TYPE` information (e.g., of type *Chapter*, *REF*). In the `FIELDSTRING`, we store the original text sequence from the book (see the highlighted sample text in Figure 1). There can be an additional `PROPERTY`, for example the statute book of each *REF* (e.g., BGB) or the noun groups within a *CS* which define the reason for citing (e.g., claims for compensation). A relationship connects two entities via an id pointer (`START_ID` for outgoing and `END_ID` for ingoing relationships). Relationships have a mandatory `TYPE` property. In our case, the relationship type is a *PartOf* relation indicating a bottom-up concept hierarchy, such that an entity of `TYPE` *Subsection* will be *PartOf* another entity of `TYPE` *Section*. References to legal text have only outgoing relationships, while the book instance at the top of the hierarchy just receives ingoing relationships. We directly access the `ID`s for scope definition and linked reference search. This data model supports our use cases as follows. Suppose a user is searching for possible applications of a law (a). The final output contains all references to that law, together with context information until the desired abstraction level. Likewise, a modified law may impact another law within the specified user context (b). An example for the latter case are changes in company size threshold values for German dismissal protection regulations (§23 Abs. 1 KSchG), which may be unknown to the user. By using graph traversal for our two given use cases, legal texts are retrieved which share the same concept hierarchy node with respect to a start reference. Nodes are traversed up to the user-chosen cutoff h by accessing the relationships to find the path to the next entity of type *REF*.

[3] An implementation of the first use case and previous work can be found at `https://github.com/anybass/HONto`

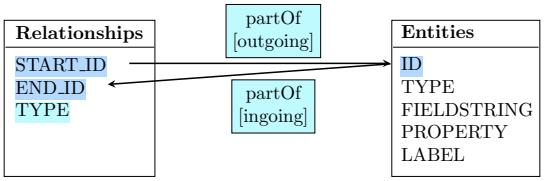

Fig. 3: Data model indicating the mapping between entity and relationship IDs in the csv files. The entity and relationship specifications contain their own TYPE information.

3 Related Work

First, we regard work in relation to legal ontology learning and second, we examine approaches for ontology-based query expansion. Legal ontology learning approaches are an automated way of constructing a legal ontology. We observe that it is possible to follow a combined top-down and bottom-up approach of ontology learning [Ag09; Ca08; El17; Fr10; Ho07; Pe07]. Unfortunately, these approarches either require expert input or use statistical language modeling methods which carry an inherent randomness and suffer from instability. Semi-automated approaches can assist in ontology population, but still face the challenge of transferring book knowledge - as it is - into an existing ontology without limiting the knowledge scope. While we identify context-dependent links between legal texts, a suitable domain ontology can be invoked for reasoning on the document level. Query expansion is a method to enrich user search terms with further related words - such as synonyms, hypernyms and homonyms - in order to retrieve more relevant items [Sc07]. Although this direction is promising, an understandable presentation of the higher amount of documents to the user is needed, while still accounting for a high recall. Using our context selection mechanism, a user can control and reduce the number of output documents by pruning irrelevant subtrees of a concept hierarchy.

4 Conclusion and Future Work

In this paper, we present a method to enable context selection in a heterogeneous lightweight ontology obtained from legal textbooks. Our ontology offers context-dependent relationships between legal texts. To this end, we propose a data model for storing the data in a graph database or in hierarchical format. We develop a context selection mechanism that helps a user navigate in our legal knowledge base and find different applications of a law, especially in two use cases: Ad-hoc search for a legal reference and a subscription service. For future work, we want to compare which storage option is better suited for other types of queries, such as topics. We will also examine how existing ontologies can be applied for document-level reasoning. Although preliminary results are promising, we will properly evaluate our approach in a user study with domain experts.

References

[Ag09] Agnoloni, T. et al.: A two-level knowledge approach to support multilingual legislative drafting. In: Proceedings of the 2009 conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood. 2009.

[Aj16] Ajani, G. et al.: The European Taxonomy Syllabus: A multi-lingual, multi-level ontology framework to untangle the web of European legal terminology. Applied Ontology 11/4, pp. 325–375, 2016.

[Bu16] Buey, M. G. et al.: The AIS Project: Boosting Information Extraction from Legal Documents by using Ontologies. In: Proceedings of the 8th International Conference on Agents and Artificial Intelligence. 2016.

[Ca08] Casellas Caralt, N.: Modelling Legal Knowledge Through Ontologies: OPJK: the Ontology of Professional Judicial Knowledge, PhD thesis, 2008.

[Di07] Dietrich, A. et al.: Agent Approach to Online Legal Trade. In (Krogstie, J.; Opdahl, A. L.; Brinkkemper, S., eds.): Conceptual Modelling in Information Systems Engineering. Springer Berlin Heidelberg, pp. 177–194, 2007.

[El17] El Ghosh, M. et al.: Ontology Learning Process as a Bottom-up Strategy for Building Domain-specific Ontology from Legal Texts. In: ICAART (2). 2017.

[Fr10] Francesconi, E. et al.: Integrating a bottom–up and top–down methodology for building semantic resources for the multilingual legal domain. In: Semantic Processing of Legal Texts. Springer, pp. 95–121, 2010.

[Ho07] Hoekstra, R. et al.: The LKIF Core Ontology of Basic Legal Concepts. In: Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT). 2007.

[Na12] Naik, V. M. et al.: Building a Legal Expert System for Legal Reasoning in Specific Domain-A Survey. International Journal of Computer Science & Information Technology 4/5, p. 175, 2012.

[Pe07] Peters, W. et al.: The structuring of legal knowledge in LOIS. Artificial Intelligence and Law 15/2, pp. 117–135, 2007.

[Sc07] Schweighofer, E. et al.: Legal Query Expansion using Ontologies and Relevance Feedback. In: Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT). 2007.

[So07] Soria, C. et al.: Automatic extraction of semantics in law documents. In: Proceedings of the V Legislative XML Workshop. 2007.

[Vi98] Visser, P. R. et al.: Heterogeneous Ontology Structures for Distributed Architectures. In: Workshop on Applications of Ontologies and Problem-solving Methods. 13th European Conference on Artificial Intelligence (ECAI-98), 1998.

[We18] Wehnert, S. et al.: Concept Hierarchy Extraction from Legal Literature. In: Proceedings of the ACM CIKM 2018 Workshops. CEUR Workshop Proceedings, 2018, URL: http://ceur-ws.org.