Information extraction for legal knowledge representation – a review of approaches and trends

Denis Andrei de Araujo[1]
Carolina Müller[2]
Rove Chisman[2]
Sandro José Rigo[1]

Resumo: Este trabalho apresenta uma introdução e um levantamento de sistema de Extração da Informação para a área jurídica. É analisado em especial as técnicas que tomam por base a representação do conhecimento jurídico como forma de alcançar um melhor desempenho, com ênfase nas técnicas que utilizam ontologias e linguística. São apresentadas algumas características das implementações dos sistemas, seguida por uma análise dos aspectos positivos e negativos de cada abordagem, visando levar ao leitor uma análise crítica das soluções estudadas.

Palavras-chave: Ontologias. Extração da informação. Recuperação da informação.

Abstract: This work presents an introduction to Information Extraction systems and a survey of the known approaches of Information Extraction in the legal area. This work analyzes with particular attention the techniques that rely on the representation of legal knowledge as a means to achieve better performance, with emphasis on those techniques including ontologies and linguistic support. Some details of the systems implementations are presented, followed by an analysis of the positive and negative points of each approach, aiming at bringing the reader a critical position regarding the solutions studied.

Keywords: Ontologies. Information extraction. Information recovery.

1 Introduction

In this work we provide an introduction to the Information Extraction (IE) systems and describe specific needs and possibilities for the legal area. In general, IE can be described as a process aimed at making the recognition and the extraction of certain types of information, as well as the recognition of relevant relations among them 16. One important aspect in IE systems is the kind of data manipulated, which remains in the natural language text group [2, 3]. As an example of an IE system devoted to the legal area, we can describe a system that processes a group of legal documents and extracts information regarding the relevant events, stated facts, individual references and date or time indicators. As a general approach, the IE systems are dependent on specific models that are determinant to guide its correct operation. Known models apply several approaches, such as extraction rules [4, 5, 6], linguistic analysis [7, 8], ontology or semantic information 16.

Some factors are important for the correct law enforcement, such as the knowledge of previous cases, relevant laws and correlated facts. The correct law application presents great difficulties when there is no adequate access to appropriate information and knowledge regarding the subject. Professionals in the legal area need to make exhaustive information search in order to take the correct decisions. That situation is related to the information overload. Various specific systems for processing legal information have been developed in order to fulfill these needs. Therefore, it can be observed the increasing interest in IE systems.

[1]Programa Interdisciplinar de Pós-Graduação em Computação Aplicada (PIPCA) – Unisinos – São Leopoldo- RS – Brasil.
{denis.andrei.araujo@gmail.com; rigo@unisinos.br}
[2]Programa de Pós-Graduação em Linguística Aplicada (PPGLA) – Unisinos– São Leopoldo-RS – Brasil
{muller.carolina@ymail.com; rove@unisinos.br}

http://dx.doi.org/10.5335/rbca.2014.3542

There is a close relation between IE systems and Information Retrieval (IR) Systems, in the sense that the results obtained from IE processes can also be incorporated in IR systems, in order to increase precision and expand recovery patterns. This relation can be also observed in different kinds of systems, such as text understanding systems or case based reasoning systems.

Systems that deal with legal knowledge are designed to accomplish specific necessities that can be very diverse. However, the IE results can be the most valuable to several cases. An initial example is the information retrieval, from which the user can find laws and case studies relevant to a trial, related to facts that describe a legal situation. Another example is legal reasoning and argumentation, where software systems can help to develop, enhance or identify a useful legal reasoning based on previous cases. Finally, we can mention the edition of legal texts, in which software systems provide support in the production of legal texts. Regardless of the specific area of action and objective, the success of legal information processing systems depends crucially on its legal knowledge representation [3, 10]. For instance, in IE systems, the aspects of knowledge representation available are determinant to the precision and recall of the results.

Currently, there are several initiatives underway to achieve standards to the representation of legal documents, therefore facilitating their automatic processing. In Brazil, the LEXML[3] project works towards this goal, as well as other projects in Europe, or in other contexts [11, 12]. In general, standards and schemes are used and implemented in document sets in flexible formats, such as XML, fostering the generation of various patterns for document annotation and the access initiatives for automatic processing. This trend in providing affordable computational formats is an important fact showing the correct positioning of the efforts to create automatic tools for the treatment of legal documents.

This work presents an overview of Information Extraction approaches and an analysis regarding its main characteristics and limitations. Due to the close relation between IE systems and IR systems, the work also presents a general overview of the IR process and some specific aspects of existing IR systems. Techniques that rely on the representation of legal knowledge are discussed in particular attention in order to achieve better results, with emphasis on techniques that explore ontologies aspects and linguistic support. Some details of the implementations are also presented, followed by an analysis of the positive and negative aspects of each approach, aiming at bringing to the reader a critical position regarding the techniques and the solutions available.

The text is organized as follows. In section 2, proposals of Information Retrieval Systems are reviewed, giving particular attention to those regarding the legal field. In section 3, the systems that use computational representation of legal knowledge are described, focusing on those that use ontologies. From this analysis, some issues and limitations inherent to the conceptual model of knowledge representation used in each examined solution are highlighted. In section 4, the details of the information extraction process in the analyzed systems are presented, clustering them by the applied technique. Section 5 is dedicated to the review and discussion regarding analyzed systems features. Finally, in section 6, present the conclusions of this work are presented.

2 Information retrieval in legal domain

Information Retrieval (IR) is a sub-area of Computer Science that studies the automatic document storage and retrieval [13, 14]. Their objectives are related to structure, analysis, organization, storage, search and information dissemination. An IR system is designed to provide a set of information to a specific population 16.

In the legal area, the search for information in databases, which are generally very large, is an essential task for the involved professionals. These professionals commonly deal with complex situations which must be examined in order to determine the rights and responsibilities of the involved ones. Therefore, they need to find laws and judicial decisions that have been applied previously in similar cases. The construction of IR systems for legal databases is extremely necessary, due to the importance of the availability of correct and relevant information for the legal activity. Besides, a great amount of legal information is currently stored in digital media. This situation can foster a broader dissemination of legal information.

In order to introduce the field of interest of this work, we present concepts regarding IR approaches and some legal domain specific necessities to these systems.

[3] http://projeto.lexml.gov.br

2.1 IR Systems based on keyword indexing

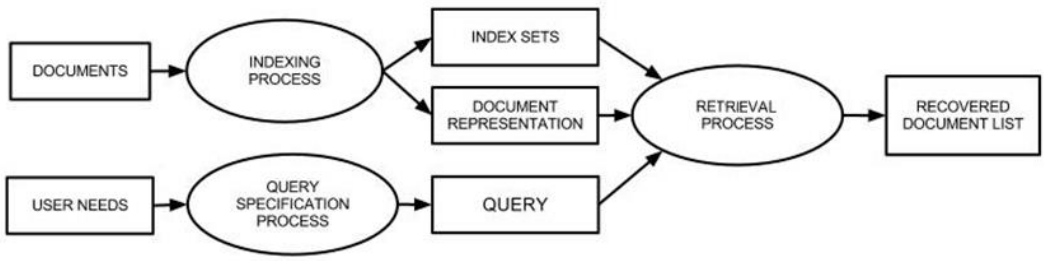
With the objective of reviewing concepts in this field, the classical IR approach is briefly described. This approach can be found in the first IR systems developed and its analysis provides important indications of its limitations.

The term "Information Retrieval" usually indicates the manipulation and organization of textual data, although, in a broader sense, it can also specifies the retrieval of records composed by different kinds of data. This is particularly true in Internet context, since the current technological infrastructure provides the access to a great amount of different kinds of data, such as images, audio, video, sensor data or structured data 16 Particularly for the legal domain, it is important the treatment of textual and structured data records. The first one is available in repository of documents and is composed mainly by free-form natural language text, given some formalism and textual structure observed in legal area. The second one is available in recent but increasingly available digital repositories that are mainly related to e-government initiatives [11, 12, 17].

Considering some collection of documents related to a specific knowledge field, the typical IR system should provide a mechanism that responds to information necessities formulated by a user or by a system, in an efficient and accurate way. The information needs are expressed in queries that can be more general or more specific. It is also possible that the queries can be accomplished by a great number of documents in the collection, which should be treated by the IR system in a process called ranking.

A classic example of architecture for IR systems based on word indexing can be seen in Figure 1. Initially, the system components deal with documents and user needs. The documents are used in an indexing process that generates index sets and documents representations, describing document relevant features for the next processing step, such as features to ranking generation. The index sets and the document representation are both used along with the query specification in the recovery process. At the end of this process, the system retrieves a list of documents considered to be relevant [14, 18].

Figure 1: General architecture of an information retrieval system



The models typically observed in classical IR process are the Boolean, Vector and Probabilistic 17. In general, document search strategies are relevant to a given kind query and necessity. These models assume that each document is described by a set of keywords, called indexing terms. Each indexing term in a document has a weight that quantifies the correlation between the term and the document. This correlation is used to calculate the relevance of the document to the query. The keyword indexing approach is commonly used for the implementation of search systems, and in legal domain some IR systems are implemented based on this approach. Some of those are commented in the next section.

2.2 Keyword indexing IR systems examples in legal domain

There are several examples of both public and private IR systems that present solutions based on words indexing. We do not intend to present a complete list in this paper but suggest related work on the subject [14, 16, 19]. Nevertheless, some of these IR systems aimed at the legal documents manipulation are briefly commented below. This list is not complete, but it presents some highly accessed web portals in the legal field, in order to provide some evidences of the IR systems widely adopted in this area, as well as an initial step to analyze some of its weaknesses.

The first portal[4] to be analyzed is LexML Web, which was officially released in June, 2009. It can be considered the main outcome of the project Brazil LexML-BR. This portal is specialized in legal and legislative Brazilian information. It aims at unifying, organizing and facilitating access to descriptive information, case law, legislation and doctrine of legislative propositions in public administration. The LexML is one of the key pieces of the federal Brazilian Electronic Government program[5]. Another example, the Web portal JusBrasil[6] is the most accessed Brazilian legal site, as stated in a research made by the Website Alexa[7] audience meter. This is a private web portal for hosting news, legislation and case law. Finally, the web portal FindLaw[8] is one of the largest legal information portals in the world. Founded in 1996, it provides access to information on legal cases and statutes, among other services. The portal is still today one of the most popular free legal information websites on the Internet, with more than four million hits per month.

Users of traditional IR systems, as the ones just cited above, have in common the fact that they must always think of your searches in terms of key words. This is not the most suitable way for information search and is considered as a negative aspect of index-based IR model 17. In addition, questions concerning the ambiguity and semantics of words are not properly treated in these systems, implying less accurate results [21, 22, 23].

It can be noted that in legal domain, the known problems of IR systems based on word indexing appear in the same manner than in another field. For instance, the problems in ranking a great number of documents obtained from a query or the word and term ambiguity have both the same origin. One important difference that can be noted, however, is related to the possible consequences of these problems. When a word indexing based IR system excludes from the results some relevant documents to a given query, the possible consequence is the poor law enforcement or a lower quality appeal, due to the lack of relevant jurisprudence. Besides that, the structured and concept-related nature of the legal documents represents an important aspect that carries the possibility of better indexing techniques application. Smarter recovery systems should make use of appropriate language for knowledge representation in order to obtain better and more accurate results. Some of these limitations identified in word indexing IR systems are overcome with intelligent systems for localization and recovery of legal information, which will be presented in the subsequent section.

3 IR systems based on legal information representation

In this section we analyze some solutions proposed and implemented for legal information retrieval based on more advanced models that make use of knowledge representation resources. The Knowledge Representation (KR) is a research area that aims at representing symbolically the knowledge in order to promote the computational and automatic processing from its elements, also allowing the generation of new knowledge. Research on knowledge representation involves the analysis of how to implement computational reasoning models accurately and effectively, as well as how to represent a set of specific domain facts 17.

The following sections present an analysis of some systems that use knowledge representation for the legal information processing. The sections are organized to indicate the main technical approach applied. Therefore, the first section described some Case Based Reasoning systems; the second section analyzes the Semantic Network approach, and the last section describes Ontology based systems.

The systems mentioned are developed under the assumption that knowledge representation can be an important element to achieve some improvements in IR. Although some results can show such evidences, the knowledge representation alone does not necessarily lead to a better retrieval performance and should be analyzed as an additional component to the other steps involved in the IR process.

3.1 Case based reasoning

The model of Case Based Reasoning (CBR) considers the solution of new problems from previous similar cases 17. Fundamentally, in the context of the CBR, a case is the formalization of a knowledge obtained

[4] http://www.lexml.gov.br
[5] https://repositorio.ufba.br/ri/bitstream/ufba/473/3/implantacao_repositorio_web.pdf#page=251
[6] http://www.jusbrasil.com.br
[7] http://www.alexa.com/topsites/countries/BR
[8] http://www.findlaw.com

from practical experience in a given context. CBR systems rely on recovery algorithms able to detect similar cases to each new situation presented. In this analysis, a previous case is considered useful when it is similar to the new situation in enough dimensions and aspects. Therefore, it can help achieving desired goals 17. The definition of the similarities between the cases is obtained by the application of specific knowledge by an expert. In most CBR systems the similarity among the cases can be established either by syntactical similarity or semantical attributes 17. Computational systems that make use of the CBR model must usually contain a database of cases and algorithms providing the search, comparison and classification of cases 17.

A lawyer, who is defending a given question, bases its reasoning on previous cases; or a judge, who decides on the basis of previous case law are both making use of the CBR model. This situation was important to inspire the CBR adoption in several initiatives aimed at modeling smart systems, in order to help the searching for legal documents. The detailed approach and objectives can be very different among the known examples. Some were designed to help students train pleading 17, some focused in legal argument identification 16 and others aim at searching for similar jurisprudence 17. These examples are analyzed below, in order to identify some important features that distinguish it from the traditional IR systems. This selection of examples is not intended to be complete.

A case-based system model is proposed in 17 to the smart recovery of jurisprudence of the 1st Region Federal Court (TRF1, Portuguese acronym for "Tribunal Regional Federal da 1a. Região"), which uses a new structure of representation, storage and retrieval of judged cases, combined with a CBR engine that aggregates the existing empirical knowledge. The proposal uses a system of syntactic similarity measure for the location of previous cases from the search terms informed by the user in the search system. The proposed approach uses a legal thesaurus to expand or restrict the search terms in the query, which are then classified in relation to the categories of knowledge expressed in the thesaurus. Each category has an associated weight, used for calculating similarity with the cases.

With respect to knowledge representation, the author describes that the knowledge categories are created from the text by legal experts. The weight assigned to the indexing categories are thus distributed: 0.35 for the fact, 0.35 to matter, 0.20 to the understanding and 0.10 for the argument. It was found that the query requests are concentrated mostly on facts and on legal matters and less on the understanding and arguments made in trials. Analyzing the author's choice by syntactic similarity over semantics, we realized that the approach proposed by 17 is very close to a ranking method of retrieved documents.

The CATO 17 system is an example of learning intelligent environment based on cases, designed to help students training legal reasoning and argumentation. The system generates the pros and cons arguments to the requests of the parts involved in some demand, therefore helping in the analysis of precedent cases. This analysis is performed in the system by the processing of data structures called FACTORs. The FACTORs are binaries information (either presented or not in the case) and can promote the complaint or the defendant of the case.

The SMILE project 16 aims at the automatic location of argumentation elements for the CATO system. Initially the SMILE system used machine learning techniques to fill the FACTORs data structures associated with the case being analyzed, making use of an approach based on symbolic processing of documents through the ID3 algorithm 17, combined with a legal thesaurus. After the initial experiments, it was found necessary a meaningful text representation. This was implemented with the application of a linguistic analysis of text sentences, combined with a better knowledge representation of the language used in the legal cases texts.

The authors indicates that the abstraction of persons and individual events, combined with the capture of actions expressed in more than one term and the recognition of negation, altogether can lead to a better representation of legal cases. These results are used to the automatic generation of the FACTORs on the basis of the analysis of texts that describe the legal cases. These improvements in information extraction are associated with greater representativeness of legal knowledge in the SMILE system. Representativeness would be increased by the use of natural language processing techniques, which could improve the capacity for analysis and extraction of SMILE. The system apply regular expressions and software tools such as Sundance and AutoSlog 17 for the analysis of texts.

The automatic extraction of information for the correct filling of FACTORs demands a greater understanding of the semantic meaning of the texts. This greater understanding can be provided by linguistic analysis. This would make it possible for the system SMILE to find the information in the extracted texts and use them for the filling of the FACTORs. The FACTORs, in turn, store knowledge extracted from texts by SMILE.

Therefore, this model establishes a relationship between implicit semantic, linguistic analysis, information extraction and knowledge representation.

3.2 Semantic network

A semantic network is a graphical notation consisting of interconnected nodes. Semantic networks can be either used for knowledge representation or as a tool to support automated inference systems 17. In a similar way, as demonstrated in the CBR systems, the Semantic Network approach can be considered suited to represent relevant aspects in legal documents. One of these can be the connection between concepts, or the causal aspect between facts and concepts. Both aforementioned situations can be modeled in Semantic Network systems. Some aspects are analyzed in examples of works with this approach. The SALOMON system 17 aims at developing legal criminal cases summaries automatically, extracting excerpts from texts to compose a summary of the case. These summaries help in the quick identification and location of relevant cases. Techniques based on discourse analysis are applied to extracting text snippets through the use of grammars. These excerpts, along with some attributes also identified by the grammar, are stored in data structures called frames. The frames are organized in semantic networks, related to each other hierarchically, sequentially or conditionally.

The authors indicate that texts produced in specific areas follow a speech pattern that, once identified, can be used for locating specific information. As well as experienced readers of a certain type of text, the extractors of information systems could find specific information on predetermined text snippets. Discourse analysis should be applied for identification of textual substructures and then the automated systems could find and extract the desired information. According to the experiments carried out by 17, experienced readers expect that the text has an overall structure or follow a writing template which allow them to predict, with a high degree of accuracy, where certain information can be found. In other words, writers and readers agree implicitly about the fact that the text has a general outline to be followed. The analysis of the texts would enable the identification of the signs at the beginning and end of these structures.

The identification and formalization of these layouts and signs would lead to valuable knowledge to the automatic location information in texts by computational systems through the development of specialized textual grammars.

According to the authors, the texts produced over the legal processes are typical examples of essays that follow an agreed schema. The SALOMON system uses textual grammars to categorize and structure the text, based on pre-established indicators, signals and superstructure. The signs may be sequences of letters or sets of words that signals the beginning and end of specific structures.

In the study conducted in SALOMON project, criminal cases were classified into seven main categories. Nine components or segments were identified as core components for documents: header, the accused and victim's identification, alleged offenses, stage of the proceedings, court opinion, legal reasoning, verdict and conclusion.

The textual grammar used in SALOMON describes the elements that make up the text and their relations. It can also specify the communication objectives of the case and its constituents, as well as define or categorize superficial linguistic indicators. The grammar is represented by a semantic network of frames.

The research conducted in the SALOMON project stated that the knowledge of the speech patterns is very useful for the extraction process. The approach is quite dependent on regional differences in the structures of discourse of criminal cases, indicating that it is necessary to refine the superstructure of legal documents, as well as linguistic indicators to represent the possible variations of the written text. It was also pointed out that the depth of the studies of legal discourse analysis is essential for the proposed model to have success in locating and extracting information in legal texts.

Once again it can be noted the approach of knowledge representation associated with natural language processing. The explicit relationship that the authors make between discourse analysis and location of information clearly suggests that the SALOMON system approach is fundamentally linguistic, implemented on the system by using grammars and textual patterns for the location of the information.

The use of semantic networks to formalize the relationships between text segments (frames) is also another point to be noted in SALOMON, because this choice also has the aim at increasing the capacity of knowledge representation in the legal system. It appears that the authors have set their minds on the idea of using

ontologies for representation of relationships, even though it is placed in doubt the expressivity capability of languages for ontologies formalization.

There is an explicit dependency of SALOMON system to the spelling system used in legal documents to be summarized. The use of the comparison of patterns of characters (words, word set or sentences) tends to speed obsolescence because the written language is dynamic and, even if it is suggested patterns for the drafting of legal documents, the inclusion of new terms in such indicators can promote errors in the process to find information.

3.3 Ontology

The term ontology has been borrowed from philosophy, which defines it as being a description of the nature of being. The meaning for Artificial Intelligence is a little different. The definition most commonly used is that ontology is "a specification of a conceptualization" 17. Generally speaking, ontologies are composed of concepts and their relationships, structuring an overview of entities 17, being increasingly important their aspects of reuse and sharing.

Since 90´s it can be observed an increasing amount of researches and legal ontologies implementations 17. These legal ontologies can be grouped by their use or role in the legal area: organization and structuring of information; reasoning and problem-solving; semantic indexing and search; semantic integration; domain understanding. Some interesting works related to legal ontology are described and analyzed bellow.

The work done in 16 covers the process of creation and use of knowledge in the legal domain for the areas of legislation, case law and sentences. It also suggests a model for organizing and structuring of information, indicating the possible applications of legal information retrieval model. The authors define that the legal ontology objective is to indicate how and what information will be used in the application. Information Entities (IEs) represents the basic units of information, while Case Scripts determines how a specific type of legal case will be composed by IEs.

The IEs are connected to each other by edges that establish similarity measures. A legal case is composed of a node that connects to a set of IEs by edges of relevance. The IEs are stored in an XML file 17, along with the Case Scripts. The Case Scripts specifies which IEs are relevant for specific types of legal cases (such as divorce or murder), and can also designate a clause of relevance to some IEs. Relevance clauses determine which IEs may be included to a case at any given time. It can be seen as a guiding mechanism defining what next steps can or should constitute a legal case. A third component of the system is the foundation of legal cases, which are texts, stored in relational databases. The repository of the IEs and case scripts is an XML file, which are defined by the authors as the ontology of the system.

With the information structured and stored in this way, it would be possible to implement search engines for locating cases. It is important to note that the main sources for the definition of IEs and case scripts are the legislation and experience of legal agents in examining previous cases. Search operations regarding case law, which are based on the location of similar cases, would also be possible by the combination of the proposed model with a reasoning system based on a comparison of the current facts and similar ones found in the database of cases.

It becomes apparent in this work that the ontology definition varies according to the vision that the authors have about its conceptualization. The author uses some of the theoretical definitions of ontology, but its implementation is physically done on pure XML files. It does not identify a formalization of concepts on information stored in the XML file, but a repository of information, composed by data, relationships and process flows. Another important issue to be highlighted concerns the reuse, one of the fundamental concepts when defining ontologies. There is a strong functional dependency between the ontology, as referred by the authors, and the application; such a relationship negatively impacts on their re-use.

However, it is undeniable that the information stored represents legal knowledge, serving primarily to the goal of organizing and structuring the information. The concept of ontology is independent of the model used to represent it. File patterns as OWL 18, DAML-OIL 18, RDF 18, XML, or even plain text, all can be used to store the ontology information.

In the work proposed by 16 it was chosen the language DAML-OIL for the representation of the ontology. This work proposes a semi-automatic methodology to transform a traditional IR system in a semantic search

system. It describes the application of this methodology in the web system of information retrieval for the Portugal´s Attorney General's Office (PGR). This methodology suggests a sequence of three steps for semantic search system accreditation:

- development of an appropriate semantic ontology;
- enrichment of text with semantic information;
- construction of an inference machine capable of rationalizing the semantic information.

In the authors' view, the construction of the ontology initially demands knowledge about the domain and also the indication of what semantic language is used to represent knowledge. They also suggest that, in the ontology creation step, it must be defined their structural and semantic objects. In the experiment conducted by the authors on PGR, the ontology's structural objects were identified by analysis of legal documents from the database. The definition of semantic objects of the domain involved linguistic analysis of documents, via the use of the parser developed in the VISL project by 18.

The semantic enrichment of the text phase contains two objectives: automatic insertion of semantic annotations 18 directly in the body of the document, relating the structural concepts from ontology to specific snippets of text; automatic population of ontology, from the actions identified in the document and semantically annotated.

The semantic inference machine aims to answer questions like:

- which are the documents where the property P is V
- which documents address the concept C
- which are the documents where the action A was performed

It is important to stress that the inference machine must be able to handle relations represented in the ontology. For example, in the question "which documents address the concept C" should mean "which documents address the concept C or any of its more specific sub concepts ". The process of inference should consider the knowledge represented in the ontology. The authors used the Prolog logic programming language 18 as inference machine on this project. The ontology and the legal documents, semantically annotated, are converted into Prolog rules to allow the inference.

It can be observed in 16 work that the ontology now is formalized in a particular language for this purpose, the DAML-OIL. Among the benefits of adopting an internationally standardized language for the formalization of the ontology it can be found the reuse of this ontology in other projects, as well as the ability to accomplish the goals of integration and semantic inter operation with other systems.

Again we see a relationship between the extraction of semantic information and Linguistics, made explicit by the use of the VISL project parser for the location of semantic objects in legal documents.

Another important point to note is the use of inference. Although the authors cite the inference only in respect to the expansion of the consultations to the concepts and relationships stored in ontology, its use brings many other benefits and application options. Currently languages for formalization of ontology support natively the use of inference engines and specific languages for ontology queries 18.

Finally, it is essential to highlight the use of the technique of inserting semantic annotations in text. This technique allows the integration of the ontology to documents with an objective of semantically improve the text. The semantic annotation of documents is one of the core aspects of the future semantic web advocated by 18.

Semantic annotations on documents and population of ontologies from linguistic based extractions is exactly the central theme of the work done in 16, where it is presented an architecture for the semi-automatic population of ontologies from documents. In the proposed architecture the resources are described in RDF format and OWL ontology is adopted for model description and the implementation of the knowledge base is done with Topic Maps specifications 18.

The knowledge base contains instances of concepts, properties, and relationships described in domain ontology. All knowledge pertinent to the field contained in documents will be captured for instantiation in the knowledge base and semantic annotations for inserting semantic annotations in documents. These notes can be shared, published, consulted or used for other purposes. The solution proposed in the work of 16 applies two software tools: the platform for knowledge management called “Intelligent Topic Manager” (ITM) and the linguistic analyzer “Insight Discoverer Extractor” (IDE). For the integration of these two systems, the authors suggest a process based on the following steps: 1) analysis of the conceptual tree resulting from linguistic

analysis; 2) acquisition rules definition of linguistic tags for ontology concepts; 3) the automated application of these rules in documents.

The authors point out that, in the project submitted, it appears that the linguistic tools and domain ontology can be modeled in a completely independent way and, to illustrate this claim, present the results of the implementation solution in the legal field. The representation modeling is of outmost importance for the knowledge capture and semantic annotation of documents. In this work it is emphasized the importance of a new component for IR system: the knowledge base. Mainly due to the performance issues, but also for reasons related to reuse, inter operation and integration, the semantic knowledge base is used for storing instances of the ontology. It is an essential resource for the implementation of IR semantic systems.

The author indicates that the ontology and the information extraction method are independent from each other. This reinforces the idea that the ontologies aim to be used in knowledge representation of a domain. Ontologies that really meet this goal are of course reusable. On the other hand, ontologies containing application data will have less reuse possibilities, creating a functional dependency between ontology and application. A good ontology structure is related to the success of semantic IR systems. More details about the elements involved in creating a legal ontology are presented below.

Among the various reasons for the creation of legal ontologies, such as democratization of knowledge acquisition, knowledge representation and reuse, perhaps the most important one is to foster the integration between Artificial Intelligence and legal theory.

Some of the pioneering research in computational legal area [48, 49], have generated the first legal ontologies: the Frame Based Ontology/FBO 18 and Functional Ontology/FOLaw 18, both formalized in ONTOLINGUA description language 18.

The FBO was designed to be a general and reusable legal ontology, with three primary classes: standards, actions and concepts; the FOLaw aims to help the organization and interconnection of legal knowledge, concerning particularly to conceptual information, showing recovery of six basic categories for knowledge representation: normative, meta-legal, real-world knowledge, responsibility, reactions and finally creative knowledge. The FOLaw was used in projects such as the "ONLINE", an architecture for case resolution 18, and CLIME/MILE, applied to the classification of ships and of maritime law 18.

The knowledge generated in the project FOLaw launched the theoretical bases for the E-Court project 18 and subsequently influenced the development of ontology LRI-Core 18, which was used experimentally on projects like E-Power 18 and DIRECT 18.

Another project to be quoted is the Lexical Ontologies for legal Information Sharing/LOIS 18, whose objective is to provide multilingual access facilitated to European legal databases for both legal professionals and laymen, which can make queries in the legal IR system in their own language and retrieve documents in other languages. The LOIS project relates to a series of other sub projects such as EuroWordNet 18, Jur-WordNet 18, DOLCE 18, among others.

The results obtained with the LOIS project are considered very important for the development of an advanced legal ontology 19. However, the initial perspectives with the LOIS project have not been achieved, mainly due to the time consumed by the project and difficulties encountered in the development of ontology. The amount of 5,000 descriptors obtained is considered not sufficient to a legal application. Despite these problems, the resulting thesaurus from the LOIS project is cited as an excellent starting point for the future Comprehensive Legal Ontology/CLO 19.

The structure of the CLO is based on conceptual models of FBO, FOLaw ontologies and LRI-Core, simplifying and modifying these templates when appropriate. The CLO is an ontology consisting of three representations: real world, legal and conceptual. The real world is described on the basis of three types of frames: people (agents), objects and processes. Agents can be either natural persons and legal entities or autonomous systems (agents or robots). Objects refer to existing things. Processes refer to the changes that occur in the real world.

The CLO authors propose that the legal thesaurus (as produced on LOIS project) can be extended to form a lexical ontology, composing the enhanced of legal concept ontology. The main component of CLO ontology consists of frames of concepts (ontology to conceptual entity). These structures contain information such as name, settings, connections with real-world knowledge and legal, as well as with the legal structural knowledge.

As defined by the authors, the CLO is hybrid knowledge ontology, i.e. its structure will also be composed by conventional legal information, based on documents digitally stored. There are several systems that perform semi-automatic linguistic analysis for location and extraction of semantic knowledge from texts, so they may effectively be integrated to the ontology under construction.

It can be observed the complexity of the process of building a legal ontology. The components and processes involved require a deep knowledge of the legal area for proper legal knowledge representation in ontology. It can also be noted the knowledge gained from the construction of legal ontologies already existent in previous projects, leveraging these experiences in creating new ontologies.

It can be indicated from the various works analyzed so far that they share one common aspect: the need to locate and extract information from documents containing text in natural language. Whether the objective is to compose the knowledge base, whether it is the construction of ontologies, all analyzed jobs present different techniques to achieve the goal of extracting semantic information from natural language documents. In the next section we will look at these works from the perspective of information extraction.

4 Information extraction from legal texts

The main objective of Information Extraction (IE) techniques is to identify types of predefined information in natural language texts. The interest in this research field is increasing, as a consequence of the observed growing amount of available textual data. In order to provide intelligent access to documents content, diverse IE techniques can be found to implement automatic or semi-automatic identification of information relevant to determined tasks 16. In general, the automatic recovery of relevant information in natural language style texts is obtained with the identification of a particular class of facts, events or their relation 16. This task is accomplished with the use of specific extraction rules.

In general, the IE systems consider the following steps of operation 16: a) textual analysis; b) extraction rules selection; c) extraction rules application and results identification. The first step can be done with techniques based on simple sentences segmentation, or with some specific Natural Language Processing technique. The second step is related to the identification of specific factors in the text that can indicate conditions to be verified as the confirmation of the localization of relevant information. The last step is dedicated to the results indication and annotation for future use. There are diverse approaches to the extraction rules description and one of the main difficulties in the area is the construction of rules that are able to represent the diversity observed in natural language texts 16.

These aspects observed in IE systems are related as the main motivation to the development of solutions that are domain specific and can increase the results and performance by observing some particular domain characteristics. This can be verified by remarking that in some areas the commonly applied vocabulary is in general more restrict and related to related concepts and typical processes. Also, there can be verified textual constructions that are frequent and associated with specific sense. Those aspects lead to the consideration of resources such as Natural Language Processing techniques, along with ontologies. Natural language understanding is crucial for most information extraction processes because more precise information can be identified when the conceptual roles and facts described in natural language texts are recognized 16. The relation between ontologies and IE relies in the fact that the information described in domain ontologies can be applied in order to help concepts and relations identification, in a more flexible and effective way 16.

Therefore, it is important to describe the specific characteristics observed in legal texts that can be exploited by IE techniques in order to achieve better results. Also it is important the description of the possible improvements obtained with ontology based approaches. These subjects are addressed in the following subsections.

4.1 Legal texts aspects relevant to IE approaches

The legal area is characterized by intensive use of the storage of information contained on texts expressed in natural language. The legal documents are in fact semi-structured information repositories, since they are composed of previously defined sections, each one observing information requirements. The automatic retrieval of information on these documents depends on the ability of the computational system to model and process the

specific information expressed in the documents. Considering the domain context properly modeled, then the IR systems should be able to find precise information and domain-specific relationships in the text.

Some aspects can be highlighted, regarding documents and texts in legal area. We assume these are key components to more precise IE systems. The most important ones are the legal institutional structure, then the related juridical processes, followed by the legal documents internal structure and, finally, the contextual use of concepts and sentence constructions.

Those general aspects are defined by specific laws and associated with each country´s constitution, besides the proper language dependence observed. Nevertheless, those elements define a framework of related concepts and associated processes that can be exploited in both representation with ontologies and extraction rules description. Those aspects are exemplified bellow.

The legal ontologies can already be found as a widely adopted resource to describe concepts and relations about actors and events [36, 37, 63]. One important aspect of its use is related to the open textured nature of the legal concepts. That highlights the importance contextual factors that influence the exact meaning of the concepts. The ontology approach can be flexible and complete enough to relate all the important contextual information regarding the correct interpretation of the legal concepts.

The extraction rules can benefit from the phrasal constructs typical to the legal area, because in general there is the necessity, in legal texts, to represent with precision and without ambiguity all the relevant components involved in the sentences. This can be seen in the example of typical phases, such as the ones describing that the Judge stated the condemnation of some defendant, associated with some specific evidence and therefore related to some specific law.

Some of the legal area characteristics contribute also to the combination of the aforementioned resources. This can be observed when the documents structure and the juridical processes are analyzed in the same context. Each juridical process is clearly defined and for each of its steps there are a set of documents designed to be applied. This information can be most valuable to IE systems, when integrated with some knowledge representation resource, such as a legal ontology. Along with the information of process context, each legal document is composed by specific section, each one containing some relevant and necessary information. This information in general can be identified in regularly used phrasal constructions, fact that is of great importance to the extraction rules description and application.

After the analysis of specific characteristics related to legal area documents, the next sections describe examples of techniques application in IE systems.

4.2 Regular expressions

The use of regular expression in IE for legal documents is related to the first approaches in the IE area, regarding texts in general. The main argument for this use is that the regular expressions can describe all the important structures and terms applied in the identification of relevant information. The results are very precise for the known extraction rules, but these systems fails when there are irregular textual expressions or differences in writing style. Some initiatives are directed to specific objectives 16 and therefore have better possibilities of integrate and describe the expression used in texts. Other are more general approaches 16, relying in concepts theory or corpus analysis to the description of the extraction rules. Some details of these systems are provided bellow.

In the project SMILE 16, the authors use as an example the analysis of disputes in the area of commercial law, more specifically in disputes involving copyright and industrial espionage. The legal disputes in this area are invariably composed by the following elements: a complainant, a defendant and an object of dispute, generally called product. The locations of the parties (claimant and defendant) is accomplished by applying pattern matching (regular expressions), while the products, due to the complexity involved, would be located by extraction patterns tool called AutoSlog. The AutoSlog is based on Sundance tool 17 for superficial linguistic analysis of text that is sufficient to achieve the goal of the system, that is to find information about the product.

The analysis that Sundance performs takes place as follows: given a sentence in natural language, the Sundance tool applies heuristics to break the sentence down into clauses and find its components: the subject, the verb, the object and prepositional phrases, providing a partial analysis of the sentence. The output of the system is then submitted to AutoSlog tool, which find occurrences of a word in text language contexts. The basic idea of

AutoSlog is that, given a set of characteristics of a target word, it is possible to locate other terms that refer or relate to it. It is exactly this property of AutoSlog to locate associated terms that makes AutoSlog a very practical tool for the localization of a product in the legal texts. A breakdown of the process described above is presented in 16, where it is demonstrated the application of techniques to a practical example.

In Addressing 16, the information extracted from texts of legal cases is stored in structures called Information Entity (IE). For the location of the information, the authors use the superficial analysis of sentences, because they argue that the PLN tools available at the time would not have the necessary efficiency to meet the needs of the project.

The first stage of the process of conversion of the sentence into IEs is based on a set of rules defined by node concepts theory of 19. A node concept becomes active when a particular word is found. Once activated the node concept, a set of conditions is applied to test whether the sentence structure is the required one. This analysis is based on search for sequences of characters. When a node concept is found, the relevant information in the sentence is extracted and originates one or more IEs. The Case Scripts are used to help in the conversion process, defining what IEs need to be found in the texts under consideration. In a second stage, a specialist examines the outcome of the case, fixing any problems.

4.3 Linguistic analysis

The Linguistic Analysis approach is related to the evaluation of regular expression approaches systems. One of the problems known in these systems based in regular expression is the open and diverse nature of textual styles. By applying a deeper linguistic analysis in some examples of the target texts, it is possible to identify more flexible textual structures description, in general related to the Part-of-Speech classification of the text components. Therefore, some initiatives rely on tools like text analyzers 16, from which the textual structures can be obtained and then used in diverse approaches. Other approaches 16 apply additional structures to map domain concepts with textual elements identified by IE procedures. Some details of these systems are provided bellow.

Along the process of creating the proposed legal ontology 16 it can be observed the use of information extraction techniques for localization of semantic objects, through linguistic analysis of documents from the Attorney General of the Republic of Portugal. To assist in linguistic analysis it is used in the project the text analyzer developed in project VISL 18. There is also a reference to a failed attempt to relate new concepts using the results of the Portuguese version of the Wordnet project, mainly due to its low number of registered terms. Although we have a superficial description of the process, it can be again noted the relationship between identification of the semantics of the text and the use of Linguistics, at this time applied to the human driven review process.

The work of 16 apply linguistic analysis to extract semantic information from documents that describe legal decisions issued by the Supreme French court in response to appeals on divorce and labor processes. The goal is to instantiate the concepts and relations of ontology, feeding it into the knowledge base. The process starts by submitting the document to the parser called “Insight Discoverer Extractor” (IDE), which generates a conceptual tree of each linguistic analysis performed on legal decisions reports. Each node of this tree is semantically marked and the text itself is included in brackets. This tree is then parsed with the goal of mapping the extracted information to a concept in domain ontology. In the mapping process the extracted information is either related to a topic, an attribute, an association or a role in the knowledge base. The analysis of the tree follows certain principles: a tree has a root, representing the document or your main subject; the tree is examined by the prefixed top-down methodology.

Two analysis are performed: the first one has the objective to capture the entities and their attributes; the second one has the objective to capture the associations among the entities roles. The knowledge acquisition rules were implemented in the Xpath[9] language. These rules automatically create instances of ontology concepts for each matching node in the conceptual tree. If a node can instantiate more than one concept, it will be checked for context and the parent, child and brothers nodes will be used to help in resolving the ambiguity.

The conceptual tree resulting from linguistic analysis is evaluated by all rules of acquisition and for each relevant node found it is created its instantiation in the knowledge base. To avoid creating multiple records in the

[9] http://www.w3.org/TR/xpath

same instance, the existence of the instance in the knowledge base is first verified. After finished the processing of the tree, the user can then visualize the results obtained through a validation interface, modifying or removing the instances set up by the system.

The corpora used in the experiment was composed of 36 reports of the Supreme Court of appeal. Only 4 documents were used for the definition of 72 acquisition rules related to 7 entity classes, 17 types of attributes, 1 association type and 2 role types, resulting in an average of 3 acquisition rules by type of concept. The 32 remaining documents were used to test the performance of the acquisition rules.

In the work developed by 16 it can be found the use of a technique for knowledge acquisition based on linguistic rules. This approach assumes that the knowledge, expressed in written form in documents, can be extracted based on the result of linguistic analysis performed automatically by a specialized program regarding syntactic and semantic classification of sentences. The location of several entities expressed in text, as well as events and entities relationships, can be achieved by the application of rules that refer to the concepts represented in the ontology. The treatment of situations of dubious interpretation can benefit from the application of the restrictions expressed in the ontology. Obviously, the use of rules does not solve all the problems inherent in information extraction process, but contributes significantly to the success in automation of semantic interpretation of written language.

5 General analysis

An IR system aims to assist in finding documents that contain information about a particular subject. Regardless the model adopted, IR systems involve a step in which the documents are processed for information extraction. In the case of IR systems based on indexing of words, the extraction phase of the information refers to the confection of reverse indexes that will associate the terms to documents in which they appear. This recovery model is efficient in what it proposes: quickly find the documents where certain terms appear.

However, the IR for indexing words presents problems of accuracy when applied to repositories with large amounts of documents. This problem is mainly associated with the fact that only the spelling of words it used be the model, in order to perform searches. In addition to spelling, words have a meaning related to the context in which they are inserted. The user, when doing a search, certainly has a specific meaning of the word in mind. To ignore this meaning implies in the recovery of documents that do not meet the information needs of the user. When users make a search in an IR system they are interested in a particular subject. In general the user is not interested in all the information related to the subject, but rather something that fulfill their information needs.

One of the solutions to the problem presented above involves the use of additional structures to represent, beyond words, their possible meanings within a given context, domain or area of expertise. These structures allow the representation of domain knowledge. The knowledge representation enables the computer system to better understand the meaning of words and thus respond more accurately to the information needs of the user.

In this work we analyze some approaches that make use of knowledge representation to obtain better answers specifically for the recovery of information in the legal area: Case-based reasoning, Semantic Network and Ontologies.

Case-based reasoning (CBR) is a technique used by 17 for optimizing the search system of the Federal Regional Court of first region, analyzes the legal texts and stores the data that characterize it. Another example of a case-based approach, the FACTORs used in the CATO system also has the function of store special features of legal texts. The CBR model has specific niches where can be applied, since their use depends mainly on the possibility of the knowledge representation in the form of cases. Although this is a feature that enables it to represent knowledge in the legal area, it can be observed some limited use of this approach, directed only to the jurisprudence. This restriction limits the use of the technique in the specific situations of recovery of legal cases, preventing the application of the model to more general information location within the legal domain.

A most comprehensive and generic approach can be found in the semantic network representation of legal knowledge. It is used in the approach chosen by 17 in the project in which we could see SALOMON, an example of the use of discourse analysis to the location of information and semantic network for the storage of knowledge extracted from texts, with the goal of automatic preparation of legal texts summaries. Although

SALOMON does not have the recovery of information as final goal, the authors point out that the resulting semantic information extraction process could be used in a legal semantic IR system.

The semantic network approach presents several similarities with the ontology approach. We can say that it has been built, already for a long time, a consensus on the use of ontologies as a widely accepted way for the representation of knowledge and to the development of applications based on these resources 19. In this work we presented some examples of research where the ontology is used specifically for the representation of legal knowledge, either to compose IR solutions 16, either for inserting semantic annotations in legal documents 16 or even to the conversion of traditional IR systems in semantic systems 16. The use of ontologies in the legal field has motivated several academic researches lately, primarily as a way to extend the capabilities of IR systems, in order to deal not only with the spelling of the words, but also with their meanings and related contexts.

The ontologies are a powerful and expressive choice to represent legal knowledge. It is related the complexity involved in creating a legal ontology in the work presented by 19, where the authors describe the possible processes for the construction of a Comprehensive Legal Ontology. In 16 are suggested the creation of legal ontologies, among other techniques, with the use of linguistic analysis to location of semantic objects. In these two works, linguistics complies with an important role in the acquisition of knowledge in the field of law.

When ontologies are used in IR systems, should be noted the importance of the phase of semantic information extraction from documents, in which the ontologies are used in the process of locating information in specific texts. The linguistic analysis is applied, this time by the use of systems Analyzers (parsers) combined with heuristic rules to assist in the correct identification of the elements by the systems of information extraction and their association to information expressed in legal ontology concepts.

The extracted information from texts can be stored directly in the ontology, as seen in 16 and 16 or in the form of semantic annotations inserted directly into documents 16 or be stored in the form of instances of components of ontology in specific repositories, commonly referred to as knowledge base 16.

With the use of legal ontologies, built with legal information semantically identified, the IR system will be able to carry out the search of information not only based on the highest level of spelling words, but related to the meanings and relationships of the words in legal domains, therefore filling some gaps in traditional search systems functional currently in use.

By the study of the cited examples and systems, it can be stated that the intensive use of storage and digitalization in the legal area allows the development of automatic information retrieval systems for these documents as well as information extraction systems. Both systems can benefit from some legal area characteristics, such as the well-structured and defined institutional organization, the well documented and defined juridical procedures, and finally, the document structure. In both approaches (IR or IE), the knowledge representation can be of importance to lead to better results, since it can overcome other approaches known approaches, such as the word indexing well-known problems.

5.1 Potential research questions

We state that it can be expected the legal area applications for IR and, in particular, for IE to grow in many different directions, allowing important research opportunities. The main reasons for that are the observed grown in the legal documents digitalization process, along with the relevance of these cited systems to the professional in that area. In this section we try to identify and briefly comment some major aspects in this context.

a) To improve the construction of legal ontologies: there are known approaches to build legal ontologies with expert's involvement and also with automatic or semi-automatic techniques. The improvement of the known techniques and the application of reuse and integration patterns can be of importance to the future systems.

b) To integrate IR and IE systems with legal domain document repositories: there is a grown in the constitution of large document repositories for the legal area documents. The integration of the documents in these repositories can lead to updated information.

c) To improve IE knowledge and linguistic based processes: there is a clear but recent trend in the use

6 Conclusions

This work presents the analysis of various proposals for Information Retrieval Systems, specifically for the legal area, describing briefly some techniques and approaches used in their implementation. Approaches to achieve better knowledge representation strategies can help in optimizing the performance of such systems. In particular the ontology-based approaches are described as an example of suitable options.

References

[1] RUSSEL, S., NORVIG, P. Artificial Intelligence, a modern approach. 2nd edition (Prentice-Hall, Englewood Cliffs). 2006.

[2] RILOFF, E. Information extraction as a stepping stone toward story understanding. In Ram, A. and Moorman, K., editors, Understanding language understanding: Computational models of reading. Cambridge: MIT Press, 1999. p. 435–460.

[3] WIMALASURIYA, D. C.; DOU, D., Ontology-based information extraction: an introduction and a survey of current approaches. Journal of Information Science, vol. 36, no. 3, p. 306–323, 2010.

[4] BRUNINGHAUS, S., & ASHLEY, K.. Improving the representation of legal case texts with information extraction methods. In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL ’01),. ACM Press, 2001. p. 42-51.

[5] COSTA, M.; NEVES, J. Practical knowledge management in the legal domain. In: Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. Proceedings. Fourth International Conference on. IEEE, 2000. p. 133-136.

[6] CHAKRABARTI, S. Mining the Web: Discovering knowledge from hypertext data. Morgan Kaufmann, 2003.

[7] SAIAS, J. & QUARESMA, P. Semantic enrichment of a web legal information retrieval system. In T. Bench-Capon (Ed.), JURIX'2002 - Fifteenth Annual International Conference on Legal Knowledge and Information Systems, London, UK, December 2002. IOS Press. 2002.

[8] AMARDEILH, F., LAUBLET, P., Minel, J-L. Document Annotation and Ontology Population from Linguistic Extractions. In Proceedings of Knowledge Capture (KCAP05), Banff, 2005.

[9] NÉDELLEC, C.; NAZARENKO, A.; BOSSY; R. Information Extraction. In: Handbook on Ontologies SE, Staab, S., Studer, R. (eds.). Second Edition. International Handbooks on Information System. Springer-Verlag, Berlin, Germany. 2009.

[10] CROSS G. R. Legal Knowledge. Proceedings of the IEEE. 101445-1450, 1986.

[11] BRIGHI, R.; PALMIRANI, M. Legal Text Analysis of the Modification Provisions: A Pattern Oriented Approach, Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL ’09, pages 238–239, ACM, New York, NY, USA, 2009.

[12] BIAGIOLI, C. E et al. Automatic Semantics Extraction in Law Documents, Proceedings of the 10th International Conference on Artificial Intelligence and Law, ICAIL ’05, pages 133–140, ACM, New York, NY, USA, 2005.

[13] FRAKES, W. B.; BAEZA-YATES, R. Information Retrieval Data Structures & Algorithms, Prentice Hall, 1992.

[14] BAEZA-YATES; R., RIBEIRO-NETO, B., Modern Information Retrieval -the concepts and technology behind search. ACM Academic Press, ISBN: 9780321416919, 2011.

[15] SALTON, G.; HARMAN, D. Information retrieval. In Anthony Ralston, Edwin D. Reilly and David Hemmendinger (Eds.), Encyclopedia of Computer Science 4th ed. (pp.858-863). Chichester: John Wiley and Sons Ltd, 2003. p. 858-863.

[16] GREENGRASS, E. Information Retrieval: A Survey, University of Maryland, Baltimore Cunty, 2000.

[17] LEXML, LEXML – Rede de Informação Legislativa e Jurídica. Relatório. 2008. Available at: <http://projeto.lexml.gov.br/documentacao/Apresentacao.pdf>. Accessed in: Dec. 2012.

[18] CARDOSO, O. N. P. Recuperação de Informação. Semana de Ciência da Computação da UFLA. Available at: <http://www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf>. Accessed: 6 Jun. 2012. 2000.

[19] MITRA, M., CHAUDHURI, B.B. Information Retrieval from Documents: A Survey. Springer Netherlands, v.2, n.2. 2000. p.141-163.

[20] KOWALSKI, G. Information Retrieval Architecture and Algorithms. New York: Springer. 2011.

[21] RESNIK, P., Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11 (1999) 95-130. 1999.

[22] VALLET-WEADON, D.; FERNÁNDEZ-SÁNCHEZ, M.; CASTELLS-AZPILICUETA, P. The quest for information retrieval on the semantic web. UPGRADE: the European Journal for the Informatics Professional. Monograph: The Semantic Web, v. 2005, n. 6, p. 19-23, 2005.

[23] SANCHEZ, D. et al. Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications: An International Journal. v. 39, Issue 9, July, 2012.

[24] SOWA, J. Semantic Networks. In Stuart C Shapiro. Encyclopedia of Artificial Intelligence. 1987. Available at: <http://www.jfsowa.com/pubs/semnet.htm>. Accessed in: June 6 2012.

[25] LENZ, M. et al. Case-based reasoning technology, from foundations to applications. Springer-Verlag, 1998.

[26] KOLODNER, J.; REASONING, Case-Based. Morgan Kaufmann. San Mateo, CA, p. 545-555, 1993.

[27] AAMODT, Agnar; PLAZA, Enric. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI communications, v. 7, n. 1, p. 39-59, 1994.

[28] DE MANTARAS, Ramon Lopez. Case-based reasoning. In: Machine Learning and Its Applications. Springer Berlin Heidelberg, 2001. p. 127-145.

[29] ALEVEN, V.; ASHLEY, K. D. Teaching case-based argumentation through a model and examples: Empirical evaluation of an intelligent learning environment. In: Artificial intelligence in education. IOS Press, 1997. p. 87-94.

[30] JÚNIOR, M. de S. B. Proposta de modelo RBC para a recuperação inteligente de jurisprudência na Justiça Federal. 2001. Tese de Doutorado. Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia de Produção.

[31] QUINLAN, J.. Ross. Induction of decision trees. Machine learning, v. 1, n. 1, p. 81-106, 1986.

[32] RILOFF, E.; PHILLIPS, W. An introduction to the Sundance and Autoslog systems. Technical Report UUCS-04-015, School of Computing, University of Utah, 2004.

[33] MOENS, M.; UYTTENDAELE, C.; DUMORTIER, J. Information extraction from legal texts: the potential of discourse analysis. International Journal of Human-Computer Studies, v. 51, n. 6, p. 1155-1171, 1999.

[34] DILLON, A. Readers' models of text structures: the case of academic articles. International Journal of Man-Machine Studies, v. 35, n. 6, p. 913-925, 1991.

[35] GRUBER, T. A translation approach to portable ontologies. Knowledge Acquisition, 5(2): p. 199-220. 1993.

[36] MOMMERS, L. A knowledge-based ontology of the legal domain. In: Second International Workshop on Legal Ontologies, JURIX. 2001.

[37] VALENTE, A. Types and roles of legal ontologies. In: Law and the semantic web. Springer Berlin Heidelberg, 2005. p. 65-76.

[38] BRAY, T. et al. Extensible markup language (XML). World Wide Web Journal, v. 2, n. 4, p. 27-66, 1997.

[39] MCGUINNESS, D. L. et al. OWL web ontology language overview. W3C recommendation, v. 10, n. 2004-03, p. 10, 2004.

[40] HORROCKS, Ian et al. DAML+OIL: a Description Logic for the Semantic Web. IEEE Data Eng. Bull., v. 25, n. 1, p. 4-9, 2002.

[41] LASSILA, Ora; SWICK, Ralph R. Resource description framework (RDF) model and syntax specification. 1999.

[42] BICK, E. The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus: Aarhus University Press, 2000.

[43] KIRYAKOV, Atanas et al. Semantic annotation, indexing, and retrieval. Web Semantics: Science, Services and Agents on the World Wide Web, v. 2, n. 1, p. 49-79, 2004.

[44] CLOCKSIN, William F.; MELLISH, Christopher S. Programming in PROLOG. 1984.

[45] WANG, X. H. et al. Ontology based context modeling and reasoning using OWL. In: Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on. IEEE, 2004. p. 18-22.

[46] BERNERS-LEE, Tim et al. The semantic web. Scientific American, v. 284, n. 5, p. 28-37, 2001.

[47] PARKER, Jack; HUNTING, Sam (Ed.). XML Topic Maps: creating and using topic maps for the Web. Addison-Wesley Professional, 2003.

[48] HAFNER, C. D. An information retrieval system based on a computer model of legal knowledge. 1978.

[49] CROSS, George R.; DEBESSONET, Cary G. Representation of legal knowledge for conceptual retrieval. Information Processing & Management, v. 21, n. 1, p. 35-44, 1985.

[50] VAN KRALINGEN, R. Frame-based Conceptual Models of Staute Law. Ph. D. diss, University of Leiden: The Hague, NL. 1995.

[51] VALENTE, A. Legal Knowledge Engineering. A Modeling Approach, IOS Press, Amsterdam, Dissertation, 1995.

[52] GRUBER, T. R. Ontolingua: A mechanism to support portable ontologies. Stanford University, Knowledge Systems Laboratory, 1992.

[53] VVINKELS, R.; BOER, A.; HOEKSTRA, R. CLIME: lessons learned in legal information serving. In: ECAI 2002: 15th European Conference on Artificial Intelligence, July 21-26, 2002, Lyon France: Including Prestigious Applications of Intelligent Systems (PAIS 2002): Proceedings. IOS Press, 2002. p. 230.

[54] BREUKER, J. et al. IT Support for the Judiciary: Use of Ontologies in the e-Court Project. In: Proceedings of the 10th International Conference On Conceptual Structures, Integration and Interfaces (ICCS 2002). 2002. p. 15-19.

[55] BREUKER, J. Constructing a Legal Ontology: LRI-Core. In: Proceedings of WONTO-2004, Workshop on Ontologies and Their Applications. Livro Rápido, Recife, Brazil. 2004.

[56] VAN ENGERS, T. M. et al. POWER: using UML/OCL for modeling legislation-an application report. In: Proceedings of the 8th international conference on Artificial intelligence and law. ACM, 2001. p. 157-167.

[57] BREUKER, J.; HOEKSTRA, R. DIRECT: Ontology-based Discovery of Responsibility and Causality in Legal Case Descriptions. Legal Knowledge and Information Systems. Jurix, p. 59-68, 2004.

[58] PETERS, W. et al. The LOIS project. In: Linguistic Resources Evaluation Conference. 2006.

[59] VOSSEN, P. EuroWordNet: building a multilingual database with Wordnets for European languages. The ELRA Newsletter, v. 3, n. 1, p. 7-12, 1998.

[60] SAGRI, M. T.; TISCORNIA, D.; BERTAGNA, F. Jur-WordNet. In: Proceedings of the 2nd International Global Wordnet Conference. 2004. p. 305-310.

[61] GANGEMI, A. et al. Sweetening ontologies with DOLCE. In: Knowledge engineering and knowledge management: Ontologies and the semantic Web. Springer Berlin Heidelberg, 2002. p. 166-181.

[62] SCHWEIGHOFER, E.; LIEBWALD, D. Advanced lexical ontologies and hybrid knowledge based systems: First steps to a dynamic legal electronic commentary. Artificial Intelligence and Law, v. 15, n. 2, p. 103-115, 2007.

[63] LAME, G. Using NLP techniques to identify legal ontology components: concepts and relations. In: Law and the Semantic Web. Springer Berlin Heidelberg, 2005. p. 169-184.

[64] BREITMAN, K. K. K.; CASANOVA, M. A.; TRUSZKOWSKI, W. Semantic web: concepts, technologies and applications. Springer, 2007.

[65] NARDI, D.; BRACHMAN, R. J. An introduction to Description Logic. In: BAADER, Franz; McGuinness, Deborah L.; NARDI, Danieli; PATEL-SCHNEIDER, Peter F. (eds.). The Description Logic Handbook: Theory, Implementation, and Applications. [S.l.]: Cambridge University Press, 2003. p. 5-44.

[66] LENZ, Mario; GLINTSCHERT, Alexander. On texts, cases, and concepts. In: XPS-99: Knowledge-Based Systems. Survey and Future Directions. Springer Berlin Heidelberg, 1999. p. 148-156.