- ORIGINAL ARTICLE -

Legal Information Retrieval System with Entity-Based Query Expansion: Case study in Traffic Accident Litigation

Sistema de Recuperación de Información Legal con Expansión de la Consulta Basada en Entidades: Caso de Estudio en Litigios por Accidentes de Tránsitos

Joel Catacora[1], Ana Casali[1,2], and Claudia Deco[1,3]

[1] *Facultad de Cs. Exactas, Ingeniería y Agrimensura,*
Universidad Nacional de Rosario (UNR) Argentina
joelcatacora@gmail.com, {acasali,deco}@fceia.unr.edu.ar
[2] *Centro Int. Franco Argentino de Cs. de la Información y de Sistemas, CIFASIS (CONICET-UNR)*
[3] *Facultad de Química e Ingeniería del Rosario, Universidad Católica Argentina (UCA)*
cdeco@uca.edu.ar

Abstract

This article describes an information retrieval system with entity query expansion by relevance feedback. The performance of the system is tested assuming its usage as a support tool for lawyers constructing a legal framework for a case. The objective is to improve the precision of results when searching for relevant jurisprudence. For this, the entities belonging to a knowledge base are used as a means to expand the query. The expansion can be done using either an automatic or an interactive mechanism. This second approach suggests to the user concepts related to the query, which might improve the search experience. An ontology and a knowledge base, called LegalOnto and LegalBase, respectively, were developed. The ontology includes concepts not addressed by existing legal ontologies, and the knowledge base integrates LegalOnto with the thesaurus of the Argentine System of Legal Information (Sistema Argentino de Información Jurídica: SAIJ), enriched in the subject of traffic accidents.

Quantitative experimentation is carried out upon a set of court documents that are used to populate the knowledge base. Preliminary results are encouraging.

Keywords: Information retrieval, relevance feedback, query expansion, legal knowledge base

Resumen

En este artículo se presenta un sistema de recuperación de información con expansión de la consulta basada en entidades mediante la retroalimentación por relevancia. Se propone una herramienta para los abogados que facilite la construcción del marco legal de un caso. El objetivo del sistema de búsqueda es mejorar la precisión de los resultados en la búsqueda de jurisprudencias relevantes. Para esto, se utilizan las entidades pertenecientes a una base de conocimiento como medio para reformular la consulta. La expansión puede realizarse mediante mecanismos automáticos o interactivos. Esta última opción puede sugerirle al usuario conceptos relacionados a su consulta, lo cual puede mejorar su experiencia de búsqueda. Para esta aplicación se construyeron una ontología y una base de conocimiento, llamadas LegalOnto y LegalBase respectivamente. La ontología incluye conceptos que no se encuentran en ontologías legales existentes y la base de conocimiento integra a LegalOnto junto con el tesauro del Sistema Argentino de Información Jurídica (SAIJ), enriquecido con conceptos pertenecientes al ámbito de los accidentes de tránsito. Se realizaron evaluaciones cuantitativas de los modelos de búsqueda propuestos sobre un conjunto de sumarios, los cuales también fueron utilizados para poblar la base de conocimiento. Los resultados preliminares obtenidos son alentadores.

Palabras claves: Recuperación de información, retroalimentación por relevancia, expansión de la consulta, base de conocimiento legal

1 Introduction

Information Retrieval aims at returning relevant documents to the user in response to a keyword query. Retrieval models provide a formal representation binding queries and documents. One approach is to use a knowledge base for query expansion. This mechanism adds other words to the original query that capture the user's intention or simply produce a new query that allows retrieving more relevant documents. The entities or concepts belonging to the knowledge base can be a means to expand a query.

Particularly, in the task that a lawyer performs when wanting to write, for example, a claim for a case, he/she needs to look for legal documents such as jurisprudence and doctrine that are relevant to the writing of his legal document. In our country, when a lawyer uses a system specialized in Argentine law, such as the SAIJ, this system does not always suggest the best legal concepts related to his/her search and if it does, the concepts can be presented in a hierarchical structure that is difficult to explore.

The SAIJ is a documentary database that contains legislation, jurisprudence and doctrine, both national and provincial. It offers searches by facets from the SAIJ Thesaurus of Argentine Law. The user can enter a Thesaurus topic as a query to retrieve documents that are classified with that topic. There are certain limitations in the search by facets and searches by keywords, for example, the user is in charge of finding the closest terms to his/her information need from a large faceted tree and can only express searches by keyword conjunctions.

For this reason, in this work a query expansion model was developed that suggests concepts related to the user's initial query, taking into account the legal scenario. In this expansion method, the suggested terms are semantically related to the query and allow the search to be specified, either to deepen it or to redirect it. All the proposed models and their extensions are unsupervised, so they do not require relevance judgments or query logs.

It is worth mentioning a number of related work in Argentina law systems. In [1] is proposed a legal recommendation system of legislation, useful for the semiautomatic development of a legal matrix, that can be seen as a set of laws of interest for an organization. This system uses the Support Vector Machine algorithm applied to national laws, labeled with the concepts of the SAIJ thesaurus. Another legal information retrieval model based on ontologies and semantic distances is proposed in [2] where legal and general vocabularies (ConceptNet, WordReference, Bank of Argentine Legal Vocabularies) are used to expand the query and rank the documents using similarities based on Normalized Google Distance. In [3] automatic language processing techniques are applied to a set of national laws, legal entities are identified, using the supervised Stanford NER algorithm and manual rules.

In this work we propose to expand the query using relevance feedback that exploits information extracted from a knowledge base. Instead of expanding the query with terms of the relevant documents, terms of the entities found in that database are used. We evaluate the search algorithms proposed in the legal field, to assist lawyers in their profession, for example, to develop a defense strategy. Two sources of information were used for this: a domain thesaurus and a collection of documents indexed thematically with the thesaurus. The sources of information with which the experimentation was carried out come from the SAIJ.

This paper is an extension of the work [4] where we proposed an information retrieval system and it was applied to the legal domain. In this system we use an iterative information retrieval model. We experimented with a document collection composed of SAIJ documents and a legal knowledge base based and the SAIJ thesaurus. Now, in this paper, the process of building a legal knowledge base and the ontology are detailed, as well as how groups of concepts in a thesaurus can be used for information retrieval. In addition, the ranking used to retrieve entities is extended, using the similarities of types and entities. On the one hand, the groups of concepts are expressed as ontological classes to be used in the type ranking method. While the entities named in the query, identified through entity linking, are incorporated for the computation of entity similarity.

The structure of this paper is as follows. In Section 2 the preliminary concepts are introduced. In Section 3 the architecture of the proposed search system is described, in Section 4 the construction process of the knowledge base used in the experimentation is detailed. Section 5 shows the proposed query expansion retrieval models and Section 6 presents the experimentation. Finally, Section 7 presents some conclusions.

2 Preliminary concepts

The uncertainty associated with the relevance of a document against a query has been probabilistically modeled in different ways, among them are the language models that show a probability distribution over text strings representing a given language. Among the retrieval models based on language models is the Query Likelihood Model (QL) [5]. The language models (LM) define a probabilistic distribution over the text, this distribution is called a language. The most simple language model is the unigram model. In this model where terms with conditional and positional independence are assumed, that is the model is a bag of words model. The Query Likelihood define an unigram language model θ_d for every document d in the collection. The score of d is the probability that $q = \langle w_1, \ldots, w_n \rangle$ is generated by the θ_d language model:

$$P(q|\theta_d) = \prod_{i=1}^{n} P(w_i|\theta_d). \tag{1}$$

The language model θ_d is estimated from the document d with the Maximum Likelihood Estimation (MLE). In this case is used the Jelinek-Mercer smoothing technique [6]:

$$P(t|\theta_d) = \lambda \frac{\mathrm{tf}_{t,d}}{|d|} + (1-\lambda)\frac{\mathrm{cf}_t}{|c|}, \tag{2}$$

where $\lambda \in [0,1]$ is the smoothing parameter, $\mathrm{tf}_{t,d}$ is the t term frequency in d document, cf_f is the t frequency in the collection and $|d|$ is the amount of terms in d.

Useful extensions of the QL can incorporate document fields, e. g. title, author. The Mixture Language Model (MLM) [7] incorporate weights to the fields as model parameters, these weights measure the field importance.

Let $\mathscr{F}$ be the set of fields or parts of a document, we annotate with f_d to the field $f \in \mathscr{F}$ of a document d. The MLM model combines the fields of a document according to their associated weights, that is:

$$P(t|\theta_d) = \sum_{f \in \mathscr{F}} \alpha_f P(t|\theta_{f_d}), \qquad (3)$$

where α_f is the weight or importance of the field f, such that $\sum_{f \in \mathscr{F}} \alpha_f = 1$, and θ_{f_d} is the language model of field f in document d. This value can reflect a priori knowledge of the domain, or it can be calculated from a training set.

The Probabilistic Retrieval Model for Semistructured Data (PRMS) [8] is based in the previous model and proposes an unsupervised estimations of the field weights.

The weight α_f in the Equation 3, is replaced with the probability of mapping the term t to the field f, $P(f|t)$, that is:

$$P(t|\theta_d) = \sum_{f \in \mathscr{F}} P(f|t) P(t|\theta_{f_d}), \qquad (4)$$

where $P(f|t)$ is defined by Bayes' rule and the Total Probability Theorem, as follows:

$$P(f|t) = \frac{P(t|f)P(f)}{P(t)} = \frac{P(t|f)P(f)}{\sum_{f' \in \mathscr{F}} P(t|f')P(f')}.$$

The probability $P(f)$ can incorporate prior knowledge or be left uniform. Whereas, the probability $P(t|f)$ is estimated using the language model of the entire collection in that field, i.e. $P(t|f) \cong P(t|\theta_{f_c})$. This model assumes a collection where fields can be characterized by distinctive term distributions and only works if that condition is met.

Feedback techniques have been investigated extensively based on LM, among which, the relevance model (RM1) [9] is a well-known example that empirically performs well.

The **RM1** model assumes that the terms in the query and in the relevant documents are independent samples and identically distributed from the relevance model:

$$P(w|R) \propto \sum_{\theta_d \in \Theta} P(\theta_d) P(w|\theta_d) \prod_{i=1}^{m} P(q_i|\theta_d), \qquad (5)$$

where Θ represents the set of smoothing document language models. In general, $P(\theta_d)$ is considered uniform and is ignored. Whereas, the expression $\prod_{i=1}^{m} P(q_i|\theta_d)$ is the QL model score (see Equation 1) for the document d.

Table 1: Entity description for "seat belt".

Field	Value
names	seat belt, belt.
related entities	traffic rules, motor traffic, motor vehicles, vehicles, transportation, ...
court documents texts	The omission of the belt... The objection will not prosper...
court documents titles	Damages, ... Appeal of unconstitutionality, ...
catch-all	seat belt, belt. traffic rules, motor traffic, motor vehicles, vehicles, transportation, ... The omission of the belt... The objection will not prosper... Damages, ... Appeal of unconstitutionality, ...

In general, when the Relevance Model is estimated, it is combined with the original language model of the query through an interpolation, this prevents the query from changing too much with respect to the original intention of the user, this is a problem known as *query drift*:

$$P(w|\hat{\theta}_q) = (1 - \lambda_q) P(w|\theta_q) + \lambda_q P(w|R), \qquad (6)$$

where λ_q is a parameter that controls the influence of the expanded model of the query, $P(w|R)$ is the estimation of the relevance model of Equation 5 and $P(w|\theta_q)$ the MLE estimate of the initial query language model, that is,

$$P(w|\theta_q) = \frac{c(w,q)}{l_q},$$

where $c(w,q)$ is the number of times w appears in q and l_q is the total number of terms in q.

The retrieval model that performs the RM1 expansion and uses the Kullback-Leibler divergence is denoted as **RM3** [10]:

$$score(d,q) = \sum_{w \in V} P(w|\hat{\theta}_q) \; log \; P(w|\theta_d),$$

where V is the vocabulary and $\hat{\theta}_q$ is the model of the expanded query (see Equation 6).

The third version of the relevance model (RM3) is widely regarded as a state-of-art model for pseudo relevance feedback.

The entities are particular units of recovery. The purpose of retrieving entities is to respond to queries through a ranked list of entities, for example "countries bordering Argentina". The entities are identifiable objects, with names, attributes and relations with other entities, e. g Argentina, Brazil, Peru are entities. They can be defined in a knowledge base modeled by ontologies. For entity search the entities can be ranked using traditional information retrieval models on entity representations as documents [11, p. 51]. For every entity in the knowledge base is created a document called entity description. This document contains all the information in the knowledge base about the entity that represent. The entity descriptions are constructed from the information of neighbour entities, this process is called *predicate folding* [12, p. 69]. Table 1 shows the description of an entity structured in 5 fields.

Document retrieval models that improve their effectiveness by incorporating knowledge bases can be classified into three groups [12, p. 293]: (1) *based on expansion*, entities are used as sources to expand the query with new terms, (2) *projection-based*, where the relevance between the query and documents is calculated by projecting them to an entity space and (3) *entity-based*, builds an entity space from the explicit semantic information that documents and queries have, for example documents annotated with named entities, to increase the number of terms.

In this work, the first of these approaches is used: models based on the expansion of the query through entities. Although these models achieve lower performance than the other approaches [11], there are unsupervised search algorithms that can be used with the type of data available. The expansion model proposed in this work is based on the Conceptual Language Model [13]. The expansion of the query is formulated as a double translation process, first the set of relevant entities is obtained and then, the vocabulary of terms associated with the entities of this set is used, as possible terms to estimate the expanded model of the query. It can be formalized as follows:

$$\begin{aligned} P(t|\hat{\theta}_q) &\propto \sum_{e \in \mathscr{E}} P(t|e,q)P(e|q) \\ &\approx \sum_{e \in \mathscr{E}} P(t|e)P(e|q), \end{aligned} \tag{7}$$

where $\mathscr{E}$ is the set of entities. This model is called the Conceptual Language Model.

The first probability $P(t|e)$, called **selection of terms**, indicates which are the most important terms of the entity e. The second probability $P(e|q)$, called **entity selection**, identifies the entities to be used in the query expansion, which are the relevant entities for q. In practice, only the k terms with the highest score of the Equation 7 are taken into account to form the model of the expanded query $\hat{\theta}_q$, with the probabilities renormalized such that $\sum_t P(t|\hat{\theta}_q) = 1$.

In this model, $P(t|e,q) \approxeq P(t|e)$ is considered, that is, conditional independence between the term t and the query q is assumed. The reason $P(t|e)$ is preferred is because it can be estimated in an unsupervised way. A simple way to compute $P(t|e)$ is through the language models of entities, i.e. $P(t|e) = P(t|\theta_e)$. While $P(e|q)$ can be estimated in various ways [12, p. 271].

On the other hand, ontologies have been used in the legal field for a wide variety of applications, we are interested in highlighting those that have been developed for standardization of the content of normative texts and their metadata. The adoption of web standards such as XML, RDF, SPARQL and the publication of information as linked data gave way to the design of vocabularies and ontologies of legal documents. Among the ontologies promoted by an international organization, *European Legislation Identifier* (ELI)[1] stands out, designed with the aim of unifying the national laws of the European Union. In addition, we find initiatives that extend and reuse the ELI ontology [14, 15, 16].

In particular, thesauri can be modeled using ontologies. A thesaurus is used to aid in information retrieval by guiding the choice of terms for indexing and searching [17]. One of its main applications is subject indexing, where the indexer assigns a set of descriptors to the documents and the user can use them to formulate a query. In the thesaurus standard ISO 25964 [18] the thesaurus is defined as "a controlled and structured vocabulary, where concepts are represented by terms, organized in such a way that the relationships between the concepts are explicit, and whose preferred terms are accompanied by synonyms". The SKOS (*Simple Knowledge Organization System*) [19] is a W3C recommendation designed for the representation of knowledge organization systems, such as thesauri, in the Semantic Web.

3 Proposed information retrieval system

We propose an architecture of a semantically enriched search system and two unsupervised search models that use the information contained in a domain knowledge base to expand the query. The expansion is done through relevance feedback based on entities [13, 12].

On the one hand, a Relevance Model with Entities (RE) is developed that automatically generates entities that are expected to be relevant to the query, through documents that describe the entities. Then, the query is expanded with the most important terms of the generated entities. The other method, the Iterative Model of Relevance with Entities (IRE) requires the assistance of the user so that he/she selects among the entities suggested in one or more iterations, those that he/she considers relevant.

The proposed architecture is shown in Figure 1, where the user enters a keyword query, transforms it into index terms and from these terms the search algorithm is applied. Then, the system returns a ranked list of documents in response to the query (*Module: Interaction with the user*). To expand the query it is necessary to reformulate it by the expansion model, for later execute the search algorithm on the reformulated query.

The core of the search engine, the implementation of the retrieval model and query expansion are carried out in the modules: *Query expansion based on entities* and *Document ranking*. The expansion can be done automatically or iteratively. If the IRE Model

[1] `https://op.europa.eu/en/web/eu-vocabularies/eli`

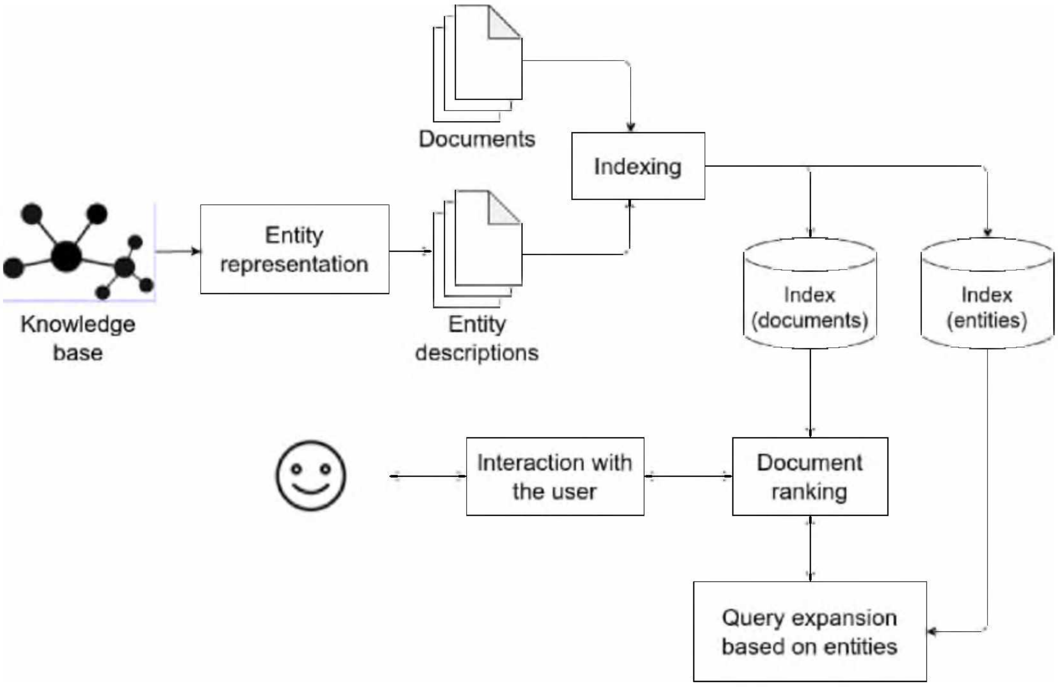


Figure 1: Proposed Architecture.

is used, a ranked list of entities is generated in each iteration of the model and the user must judge if they are relevant to his initial query and then proceed to the reformulation of the query. This system has two sources of information: the *Collection of documents* and the *Knowledge base*. The expansion models do not directly query the knowledge base but perform their computations on the descriptions of the entities. These descriptions are created from the knowledge base using the predicate folding procedure. This phase correspond to the *Entity representation* module. We need two indices, once for the *Documents collection* and other for the *Entity descriptions*. The resulting vocabulary of search system terms is the intersection of the vocabularies of each index. The expansion model is in charge of querying the entity index and the document retrieval model searches in the document index.

In this work we experiment with a legal knowledge base designed for this work and a collection of court documents collected from the SAIJ are used as a source of information.

4 A legal knowledge base

For the evaluation of the proposal, a legal knowledge base (LegalBase) and a collection of texts (Sumarios20) were developed, made up of 45556 SAIJ documents belonging to Argentina Civil Law of national scope and the province of Santa Fe, along with 10 information needs.

The legal knowledge base is based in the SAIJ thesaurus. We decided to enrich this thesaurus in a particular subject, the car accidents. For this purpose we create a car accident thesaurus. First, we identify keywords that a lawyer uses to represent different information needs from car accident cases. It is defined in SKOS format and it is composed with 91 concepts and 18 groups of concepts. Between them, there are 11 concepts that do not belong to the SAIJ thesaurus, which does not provide enough information for the entity descriptions construction. The entity descriptions are constructed from the information of neighbour entities and these are described by few words (prefered name, synonyms and related terms). So, we decided to use labels of documents collections with SAIJ thesaurus concepts. We develop a small ontology on legal documents, LegalOnto, which includes legislation, jurisprudence and doctrine, since pre-existing legal ontologies such as ELI, do not include all these types of documents, especially those that correspond to jurisprudence.

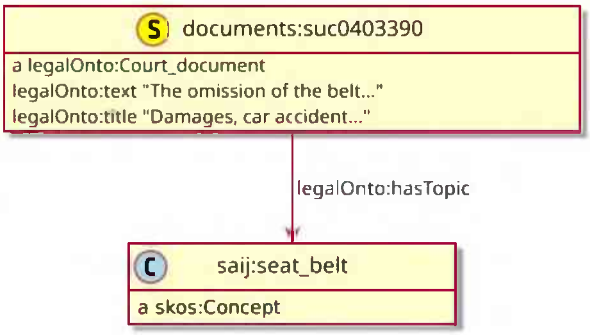
The car accident thesaurus, SAIJ thesaurus and LegalOnto are integrated in one single knowledge base: LegalBase. The integration was done using `owl:import` directive in a way to avoid inconsistencies. The LegalBase is populated with the collection of cataloged court documents. In this way we increase the information about the legal entities and consequently, the performance of the proposed search system. The concepts are ontology instances and are related to documents by the property `LegalOnto:hasTopics`, see Figure 3.

To include the court documents in the LegalOnto ontology we needed to translate them into RDF format and this could be done easily, since the semi-structure provided by the document fields allows each of these fields to be seen as a triplet, it was only necessary

Field	Value
@ids	http://www.semanticweb.org/ documents#suc0403390
@type	http://www.semanticweb.org/ documents#Court_document
text	The omission of the belt...
title	Damages, car accident,...
subject	http://vocabularios.saij.gob.ar/ saij/xml.php?skosTema=5259...

Field	Mapping
text	http://www.semanticweb.org/ legalOnto#text
title	http://www.semanticweb.org/ legalOnto#title
subject	http://www.semanticweb.org/ legalOnto#hasTopic

(a) A document and its semantic mapping.


S documents:suc0403390
a legalOnto:Court_document
legalOnto:text "The omission of the belt..."
legalOnto:title "Damages, car accident..."
legalOnto:hasTopic
C saij:seat_belt
a skos:Concept

(b) RDF graph of the document.

Figure 2: Example of a document converted to RDF format.

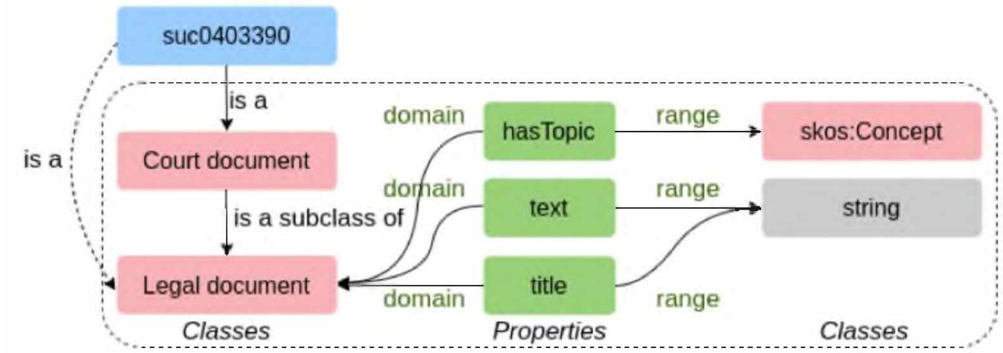


Figure 3: Example of LegalOnto schema.

to define a mapping that associates each field of the document with the desired property. Figure 2a shows a SAIJ document, where the field *topics* was obtained by applying entity linking to the topics expressed in natural language with which it was catalogued. Its transformation to RDF format is shown in Figure 2b.

4.1 Representing concept groups as classes

In the thesaurus of traffic accidents, groups of concepts were used. The grouping of concepts by topic is usually done if the thesaurus has a wide range of domains. In the ISO 25964 thesaurus standard, groups of concepts can have their own hierarchical relationship, through the properties *hasSubGroup*/*hasSuperGroup*. In SKOS there are constructors similar to the concept groups, these are: `skos:Collection` and `skos:ConceptSchema`, which are defined as sets of concepts. In this paper we use `skos:Collection`, since SKOS collections can be nested using the `skos:member` property similarly to *hasSuperGroup*. Instead, `skos:ConceptScheme` objects cannot be related. The way to add concepts to collections is done using the `skos:member` property. From here on, we will treat collections and groups of concepts as synonyms.

Figure 4 shows an extract from the RDF graph of the traffic accident thesaurus. You can see that we have the concept *regla de tránsito* and the collection *regla de tránsito*. It is necessary to maintain this distinction, because by SKOS integrity rule S37, the classes `skos:Concept` and `skos:Collection` are disjoint.

The `skos:member` property is not transitive in the SKOS standard. Transitivity would allow us to say that concepts are not only "similar" to concepts declared in their own group but also to concepts declared in ancestor groups. We decided to express transitivity in an implicit way, using a design pattern, we refer to the Class-Individual Mirror proposed in [20, p. 345]. This pattern allows to represent an individual as a class. The main reason for using this pattern is the possibility of including classes, which can be used by entity search algorithms that operate not only on the individuals of the ontology but also on their types (classes).

In Figure 5, the inferences that can be made about the "traffic light" concept, after having applied the pattern, are shown with dotted lines. The *Class-Individual*

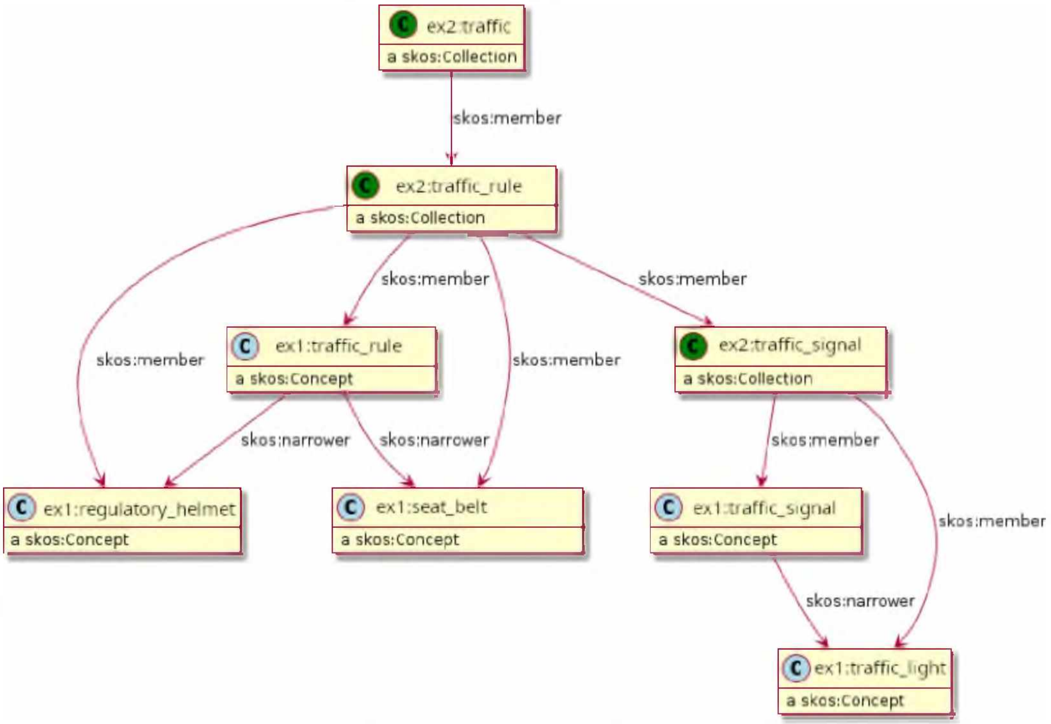


Figure 4: Use of concept groups.

Mirror pattern models situations where it is unclear whether something should be modeled as a class or an individual [20, p. 365]. This pattern models both situations and keeps them in sync. Additional representations on the individual side are reflected on the class side and vice versa. This is used to express the transitivity of `skos:member` without explicitly declaring it to be transitive, and also allows groups of concepts to be viewed as classes.

5 Query expansions

In this paper we propose two models of query expansion, one that performs automatically and the other, where the user interacts with the system in an iterative way.

5.1 Automatic query expansion

In the query expansion model "Conceptual Language Model" two components must be estimated: the selection of terms $P(t|e)$ and the selection of entities $P(e|q)$. We propose estimating $P(t|e)$ with the entity language model, that is $P(t|e) = P(t|\theta_e)$, and $P(e|q)$ with the QL model applied to entity retrieval:

$$\begin{aligned} P(t|\hat{\theta}_q) &\approx \sum_{e\in\mathscr{E}} P(t|e)P(e|q) \\ &\propto \sum_{e\in\mathscr{E}} P(t|e)P(q|e)P(e) \\ &\approx \sum_{\theta_e\in\Theta} P(\theta_e)P(t|\theta_e)\prod_{i=1}^{n} P(q_i|\theta_e). \end{aligned}$$

It can be seen that the proposed expansion model is equivalent to RM1, using entities instead of documents. For this reason, we call it the Relevance Model with Entities (RE).

We assume a uniform distribution of entities ($P(e)$ is ignored), then:

$$P(t|\hat{\theta}_q) \propto \sum_{\theta_e\in\Theta} P(t|\theta_e)\prod_{i=1}^{n} P(q_i|\theta_e), \qquad (8)$$

Then, the documents are ranked with the Kullback–Leibler divergence:

$$score(d,q) = \sum_{w\in V} P(w|\hat{\theta}_q)\ log\ P(w|\theta_d). \qquad (9)$$

5.2 Iterative query expansion

If in the Conceptual Language Model, we consider interactive and uniform entity selection, then the query expansion would be:

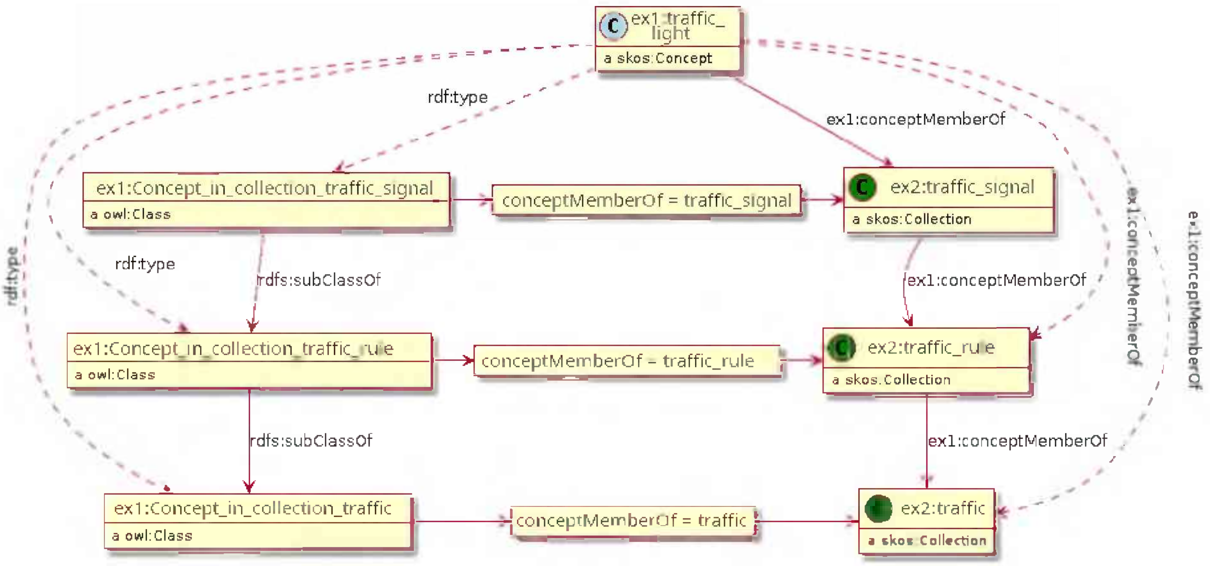


Figure 5: Application of *Class-Individual Mirror* pattern.

$$P(t|\hat{\theta}_q) \approx \sum_{e\in\mathscr{E}} P(t|e)P(e|q) \approx \frac{1}{|\mathscr{E}_q(k)|}\sum_{e\in\mathscr{E}_q(k)} P(t|\theta_e), \qquad (10)$$

where $\mathscr{E}_q(k)$ is the set of the first k entities of QL model that have been selected by the user. Also, $P(e|q)$ is a uniform probability in that set.

We propose an iterative query expansion model based in the iterative relevance model. In each iteration is shown to the user entities that can be selected. From these selected entities are extracted terms to reformulate the user query. Next, we explain the candidates entities generation.

Let $e \in \mathscr{K}$, an entity from the knowledge base and q the initial user query, the candidates entities showed to the user in the i iteration are ranked according to the Kullback-Leibler divergence:

$$score(e,q^{(i)}) = \sum_{t\in V} P(t|\theta_q^{(i-1)})\ log\ P(t|\theta_e), \qquad (11)$$

where $q^{(i)} = (q, \theta_q^{(i-1)})$ is composed by the initial query and the language model of the expanded query in the previous iteration $\theta_q^{(i-1)}$, defined according to the Equation 12. The entities showed in previous iterations are removed from the results. The first k entities of the score are showed to the user in the i iteration, with $1 \leq i \leq n$, n is the total of iterations considered in the query expansion. The candidates entities are added to a relevant entities collection $\mathscr{E}_q^{(i)}(k)$:

$$E_q^{(i)} = E_q^{(i-1)} \cup \mathscr{E}_q^{(i)}(k).$$

This collection maintains the relevant entities known until the i iteration. The collection can be initialized by *entity linking* in the query: $E^{(0)} = entityLinking(q)$.

The query expansion in the i iteration is defined by the following equation:

$$P^{(i)}(t|\hat{\theta}_q) = \frac{1}{|E_q^{(i)}|}\sum_{e\in E_q^{(i)}} P(t|\theta_e). \qquad (12)$$

The terms are extracted from the expanded query language model $\hat{\theta}_q$, these terms are extracted from the entity language model from the relevant entities collection. This equation is obtained from the Conceptual Language Model (see Equation 10) assuming uniform entity selection. The query reformulation in the i iteration is:

$$P^{(i)}(t|\theta_q) = \begin{cases} (1-\lambda_q)P^{(0)}(t|\theta_q) + \lambda_q P^{(i)}(t|\hat{\theta}_q) & i \geq 1 \\ \frac{\mathrm{tf}_{t,q}}{|q|} & i = 0 \end{cases}, \qquad (13)$$

where $\lambda_q \in [0,1]$ is the interpolation parameter that control the query expansion influence. In the next iteration $i+1$, the composed query has the form:

$$q^{(i+1)} = (q, \theta_q^{(i)}).$$

In the last iteration $i = n$, is obtained the query reformulation $\theta_q^{(n)}$. This one can be used in the document ranking score:

$$score(d,q) = \sum_{t\in V} P(t|\theta_q^{(n)})\ log\ P(t|\theta_d). \qquad (14)$$

In the Figure 6 we show the timeline of an example of the proposed iterative query expansion, for the query “traffic accident, right of way” where in two iterations, 5 entities are suggested per iteration.

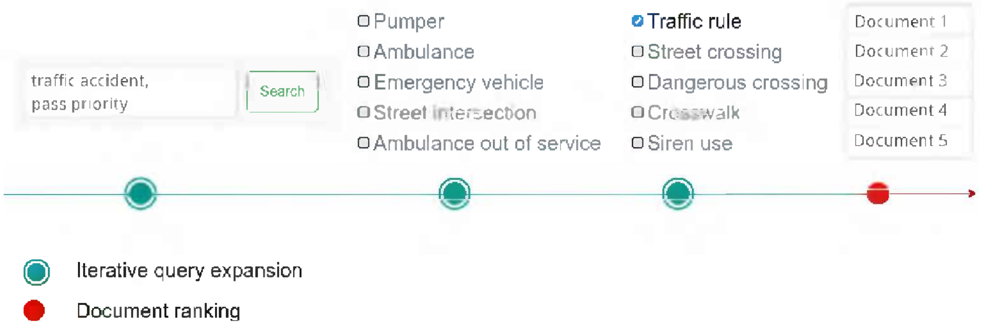


Figure 6: Timeline for the query "traffic accident, right of way".

Using the IRE model, the query expansion is made up of the terms of the original query and the most important terms of the entities considered relevant to the user. Following the example above, since the user only selected the entity "traffic rule" in the second and last iteration of the algorithm, then the final query expansion consists of the terms: "traffic accident", "pass priority" and the terms associated with the entity "traffic rule".

5.3 Extensions to the model

The semi-structure of the entity descriptions can be incorporated into the query expansion process. This structure is taken into account for the estimation of the language model of the entity $P(t|\theta_e)$. This is achieved using the following models: Mixture Language Model (MLM) and Probabilistic Retrieval Model for Semistructured Data (PRMS).

If MLM is used, then both the term selection $P(t|e)$ and the entity selection $P(e|q)$ from automatic query expansion ($P(t|\theta_e)$ and $P(q_i|\theta_e)$ in Equation 8) are estimated based on the Equation 3, we will note this model as RE+MLM. Whereas if PRMS is used, the above components of RE depend on the Equation 4, and we note it as RE+PRMS. In experimentation, these models are estimated with the Jelinek-Mercer interpolation.

The Equations 11 and 12 would be affected by the new estimation of $P(t|\theta_e)$ with the MLM or PRMS models. From here on, these models are noted as follows: IRE+PRMS, IRE+MLM.

Furthermore, in [12] it is suggested to enrich the query so that it does not depend only on the terms that describe it. If you have a rich query $\tilde{q} = (q, E_q, T_q)$, where E_q and T_q are the entities and types relevant to the query q, then $score(e,\tilde{q})$ can be the result of a linear combination of different search algorithms:

$$\begin{aligned} score(e,\tilde{q}) = {} & \lambda_{term}\, score_{term}(e,q) \\ & + \lambda_{entity}\, score_{entity}(e,E_q) \\ & + \lambda_{type}\, score_{type}(e,T_q) \end{aligned} \quad (15)$$

Now we describe each of the different types of ranking. We consider $score_{term}$ as the ranking based on terms (e.g. Equation 11, Equation 9), $score_{entity}$ is the ranking based on the similarity between entities (Equation 16) and $score_{type}$ is the similarity ranking between types (Equation 17).

Let e be an entity and $\mathscr{E}_q$ be the set of entities relevant to the query, similarity is defined as follows:

$$score_{entity}(e,\mathscr{E}_q) = \frac{1}{|\mathscr{E}_q|} \sum_{e' \in \mathscr{E}_q} \mathrm{sim}(e,e'). \quad (16)$$

The similarity between pairs of entities can be computed in several ways, in the experimentation the Vector Space Model similarity is used.

Let $\theta_{\mathscr{T}_q}$ and $\theta_{\mathscr{T}_e}$ be the type language models of the query and an entity, respectively. Type-based similarity can be calculated using the KL divergence [21]:

$$\begin{aligned} score_{type}(e,\mathscr{T}_q) = \max_{e' \in \mathscr{E}} (& KL(\theta_{\mathscr{T}_q}||\theta_{\mathscr{T}_{e'}}) \\ & - KL(\theta_{\mathscr{T}_q}||\theta_{\mathscr{T}_e})), \end{aligned} \quad (17)$$

where the maximum is used to convert the KL divergence to a similarity. Type language models are represented as a multinomial distribution.

Then, the following enriched query is proposed:

$$\tilde{q} = (q, entities(q), types(q)),$$

where $entities(q)$ consists of applying *entity linking* to the query, while $types(q)$ are the types related to the query:

$$types(q) = \{\mathscr{T}_e \mid e \in entities(q)\}.$$

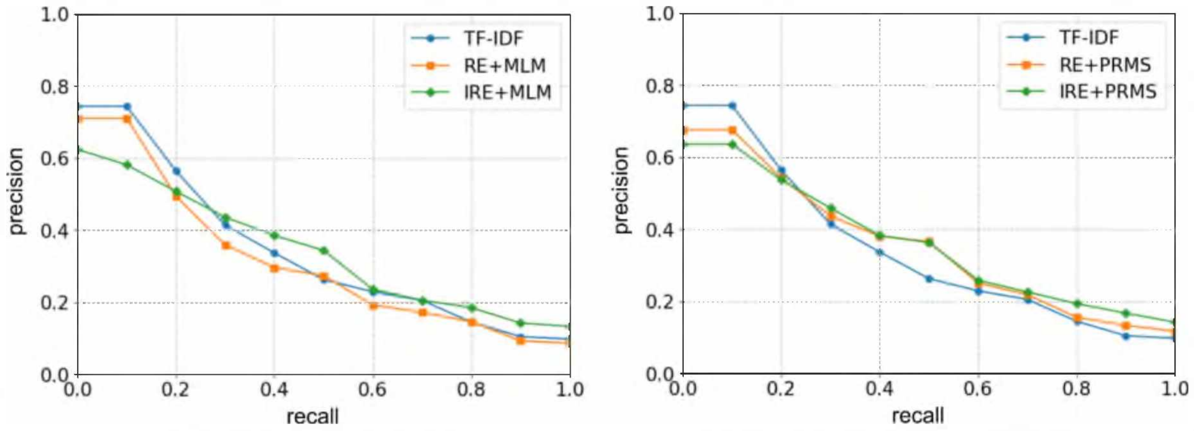
In experimentation, exact match is used as the *entity linking* method and no disambiguation is applied. We will denote with RE* and respectively with IRE* when the RE and IRE models use an enriched entity ranking. Note that when $\lambda_{term} = 1$ and $\lambda_{entity} = \lambda_{type} = 0$, we have the original model.

6 Experimentation

We perform the evaluation of the search algorithms on the Sumarios20 collection. The collection is made up

Table 2: Evaluation of retrieval systems with query expansion. The robustness index is calculated over TF-IDF.

Collection	Metric	RM3	RE	RE+MLM	RE+PRMS	IRE	IRE+MLM	IRE+PRMS
Sumarios20	MAP	0.313	0.310	0.299	0.335	0.318	0.312	**0.336**
	P@10	**0.280**	0.270	0.250	0.250	0.250	0.260	**0.280**
	P@20	0.200	0.195	0.180	0.210	**0.215**	0.205	0.209
	J@20	0.994	0.845	0.875	0.880	0.980	0.985	0.990
	RI	**0.10**	-0.10	-0.20	**0.10**	0	0	**0.10**



(a) TF-IDF, RE+MLM, IRE+MLM.

(b) TF-IDF, RE+PRMS, IRE+PRMS.

Figure 7: Precision-recall of the different models (MLM on the left and PRMS on the right).

of 45.556 court documents of the SAIJ belonging to the national scope and the province of Santa Fe, along with 10 information needs.

In this first evaluation, these retrieval models focused on entities are considered: RE, RE+MLM, RE+PRMS, IRE, IRE+MLM, IRE+PRMS, together with the RM3 algorithm (with JM interpolation). The search algorithms are evaluated using *4-fold cross-validation*, the parameters are adjusted with *grid search* and *freezing ranking* is used to evaluate the iterative model following what is proposed in [22].

We were able to optimize a given set of variables due to the limited computing power of the available hardware. The optimized parameters are: the JM smoothing interpolation, when estimating $P(t|\theta_e)$ in the proposed models and $P(t|\theta_d)$ in RM3, and the weights of the description fields for the PRMS and MLM models, with values in the range $[0,1]$ and steps of 0.25. The interpolation of the initial query with the expanded query model is $\lambda_q = 0.25$ in IRE (Equation 13) and RM3, and $\lambda_q = 0.5$ in RE. The first 15 terms were considered, the first 10 entities (RE), 10 documents (RM3) and 2 iterations of 10 entities, 2×10 (IRE).

Table 2 compares the different retrieval systems implementing different query expansion models presented in this paper. The comparison is carried out using the following metrics: Mean Average Precision (MAP), Precision at 10 (P@10), Precision at 20 (P@20), Judge at 20 (J@20 [23, p. 29]), Robustness Index (RI). The algorithms that use the PRMS model achieve the highest yields in terms of MAP, superior to the MLM models. According to the robustness index, MLM-based models have a higher query bias than those using PRMS. That is, the query expansion terms generated by the RE+MLM and IRE+MLM models would have a lower relation to the initial query compared to the terms produced by the RE+PRMS and IRE+PRMS models. Precision and recall curve are shown in Figure 7. We use Apache Lucene's TF-IDF model implementation.

In the evaluation, no significant differences were found between the RE model and the IRE model in the collection Sumarios20. Finding which entities are relevant to a query, as the IRE model does, does not present a difference in this experimentation compared to the automatic query expansion of the RE model. Both outperform the RM3 algorithm, which is a promising result. We understand that the interactive selection of entities did not provide great benefits in this experimentation compared to automatic expansion, since the entities selected by the user did not provide better terms to expand the query. According to the process of creating the entities description, the terms that expand the query are obtained from documents that deal with the topics that interest the user, but they are not necessarily the best terms to expand the query.

Also, extensions to these models were tested: MLM and PRMS. The use of the PRMS model, both in the RE and IRE algorithms, achieves higher performance than MLM. The dynamic field weight (PRMS) is

Table 3: Comparison of query expansion systems with enriched entity ranking.

Collection	Metric	RE	RE*	IRE	IRE*	IRE$_e^*$
Sumarios20	MAP	0.331	0.372	0.317	0.317	0.313
	P@10	0.250	0.310	0.290	0.290	0.270
	P@20	0.200	0.205	0.225	0.225	0.210
	J@20	0.820	0.890	0.975	0.934	0.985

better than the static field weight (MLM). Therefore, in this experimentation, the search algorithms with entity-centered query expansion that achieve the highest performance are those that use the PRMS model.

Another experiment was conducted. It was tested to include the types of the entities in the ranking as proposed in [24, 12]. In this way, the classes derived from the groups of concepts generated for the traffic accident thesaurus were used. That is, if you have an enriched query $\tilde{q} = (q, E_q, T_q)$, where E_q and T_q are the relevant entities and types of the q query, then $score(e, \tilde{q})$ (Equation 15) can be the result of a linear combination of different search algorithms. The evaluation was not performed by dividing the data into training set and test set as k-fold cross validation does, therefore the results obtained cannot be generalized.

Despite this, it was found for the Sumarios20 collection that the best performance of the query iterative expansion algorithm is not reached when the largest number of relevant entities is suggested to the user. Perhaps this is due to characteristics of the legal knowledge base. Higher quality entity descriptions might provide better terms to expand the query.

During this experimentation only a few parameters were optimized: the Jelinek-Mercer smoothing variable (λ_e) in the entity language models, with values in the interval $[0,1]$ and steps of 0.25, the enriched entity ranking variables (λ_{term}, λ_{entity}, λ_{type}), with values in the interval $[0,1]$ and steps of 0.05. The rest of the parameters were assigned the following values: $\lambda_d = 0.75$, $\lambda_q = 0.25$ (IRE), $\lambda_q = 0.5$ (RE). The first 15 terms, the first 10 entities (RE), 10 documents (RM3) and 2 iterations of 10 entities (IRE) were considered.

In Table 3 it can be seen that the use of enriched entity ranking improves documents retrieval, but only when using automatic query expansion (RE*) and not when performing iterative expansion (IRE*). So, in the collection *Sumarios20*, the expansions using the model IRE*, which are obtained from using the enriched entity ranking, do not give better results than the expansions generated by the IRE model, which uses a ranking of entities based on terms.

In general, the expansion of the query from the pseudo-relevance feedback, that is, adding terms to the query from documents that are obtained automatically and that are assumed to be relevant to that query, would have a lower performance than the relevance feedback that expands the query with terms belonging to the documents selected by the user. However, when these feedback mechanisms are applied to entity descriptions, the above advantage is no longer true for the *Sumarios20* collection. We believe that terms associated with a relevant entity may not be as good as the terms of a relevant document. The entities that would improve the query in the IRE model would not necessarily be the entities that the user understands as relevant, there would be a certain difference between the entities that the user considers as relevant and those entities that would be useful to expand the query. But, perhaps this is due to characteristics of the legal knowledge base, having higher quality entity descriptions could provide better terms to expand the query.

In addition, in Table 3 we can see the model IRE$_e^*$, which has the parameter configuration that shows the most relevant entities for the user. As can be seen, this does not imply a better reformulation of the query in the Sumarios20 collection.

7 Conclusions

In this work, we propose the architecture of a search system with query expansion to improve document retrieval. In addition, a traffic accident thesaurus was developed, reusing the SAIJ thesaurus, together with the LegalOnto ontology, which models the semantic indexing of legal documents. These sources of information and the SAIJ thesaurus were integrated to create the knowledge base LegalBase. We populated this knowledge base with documents of the collection Sumarios20. Therefore, LegalBase contains documents of the SAIJ, its semantic indexes and all the information from the two legal thesauri, their instances, properties and classes. It should be noted that all these semantic resources: the thesaurus, the legal ontology and the knowledge base, developed in this project are available for use in other applications.

This search system incorporates the knowledge base as a source of information to expand the query from terms associated with the entities relevant to the query. Two unsupervised search algorithms based on the Conceptual Language Model were implemented in this architecture. One is the Relevance Model with Entities (RE), which performs the expansion automatically. The other, is the Iterative Relevance Model with Entities (IRE), which requires user assistance. These models were evaluated in the legal domain using the legal knowledge base (LegalBase) and a test collection (Sumarios20) developed especially for this work. In the first evaluation no significant differences were found between these models. Furthermore, MLM and PRMS extensions were tested, obtaining better results with PRMS.

In the second evaluation, it was obtained that only the automatic expansion model improves its performance when a rich selection of entities is used. As the entities considered relevant by the user in this experimentation were not the best terms to expand the query, the interactive expansion model did not provide, in this case, better performance than the automatic expansion. These results may be due to the limitations of the knowledge base. For better performance it is desirable to have entities with better descriptive terms.

It should be taken into account that the results shown are subject to parameters that were not completely optimized since we were restricted by hardware limitations and another issue, is that the test collection Sumarios20 is too small to be representative. In any case, the results using these semantic expansion models are encouraging. With this semantic search system, it is expected to help lawyers in retrieving documents that are useful for drafting a legal claim.

As future work, it is necessary to experiment with larger collections and extend search support to other branches of law. It is proposed to include in the expansion model a selection of terms that depends on the query and to incorporate unsupervised language models such as word embeddings.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

AC and CD conceived the problem idea and led the project. JC conceived the different approaches for the solution, wrote the program and conducted the experiments. AC, CD and JC analyzed the results and wrote and revised the manuscript. All authors read and approved the final manuscript.

References

[1] L. Perezzini, A. Casali, and C. Deco, "Sistema de soporte para la recuperación de normativas en la ingeniería legal," in *SID 2020 - 49 JAIIO*, 2020.

[2] G. A. Dehner, K. B. Eckert, J. M. Lezcano, and H. J. Ruidías, "Modelo de recuperación de información jurídica basado en ontologías y distancias semánticas," in *XIX Simposio Argentino de Informática y Derecho (SID 2019)-JAIIO 48 (Salta)*, 2019.

[3] F. Cardellino, C. Cardellino, K. Haag, A. Soto, M. Teruel, L. Alonso i Alemany, and S. Villata, "Mejora del acceso a infoleg mediante técnicas de procesamiento automático del lenguaje," in *XVIII SID-47 JAIIO*, 2018.

[4] J. Catacora, A. Casali, and C. Deco, "Sistema de recuperación de información con expansión de la consulta basada en entidades," in *XXVII Congreso Argentino de Ciencias de la Computación (CACIC)(Modalidad virtual, 4 al 8 de octubre de 2021)*, 2021.

[5] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, (New York, NY, USA), p. 275–281, Association for Computing Machinery, 1998.

[6] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, (New York, NY, USA), p. 334–342, Association for Computing Machinery, 2001.

[7] P. Ogilvie and J. Callan, "Experiments using the lemur toolkit," in *TREC*, vol. 1, pp. 103–108, 2001.

[8] J. Kim, X. Xue, and W. B. Croft, "A probabilistic retrieval model for semistructured data," in *Advances in Information Retrieval* (M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, eds.), pp. 228–239, Springer Berlin Heidelberg, 2009.

[9] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (New York, NY, USA), p. 120–127, 2001.

[10] Y. Lv and C. Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1895–1898, 2009.

[11] R. Reinanda, E. Meij, M. de Rijke, *et al.*, "Knowledge graphs: An information retrieval perspective," *Foundations and Trends in Information Retrieval*, vol. 14, no. 4, 2020.

[12] K. Balog, *Entity-Oriented Search*, vol. 39 of *The Information Retrieval Series*. Springer, 2018.

[13] E. Meij, D. Trieschnigg, M. de Rijke, and W. Kraaij, "Conceptual language models for domain-specific retrieval," *Inf. Processing & Management*, vol. 46, no. 4, pp. 448–469, 2010.

[14] V. R. Doncel and E. M. Ponsoda, "Lynx: Towards a legal knowledge graph for multilingual europe," *Law in Context. A Socio-legal Journal*, vol. 37, no. 1, pp. 1–4, 2020.

[15] A. Oksanen, J. Tuominen, E. Mäkelä, M. Tamper, A. Hietanen, and E. Hyvönen, "Semantic finlex: Transforming, publishing, and using finnish legislation and case law as linked open data on the web," *Knowledge of the Law in the Big Data Age*, vol. 317, pp. 212–228, 2019.

[16] I. Chalkidis, C. Nikolaou, P. Soursos, and M. Koubarakis, "Modeling and querying greek legislation using semantic web technologies," in *European Semantic Web Conference*, pp. 591–606, Springer, 2017.

[17] S. G. D. Clarke, "The information retrieval thesaurus," *KO KNOWLEDGE ORGANIZATION*, vol. 46, no. 6, pp. 439–459, 2019.

[18] "ISO 25964-1:2011: information and documentation - Thesauri and interoperability with other vocabularies. Part 1, Thesauri for information retrieval," standard, International Organization for Standardization, 2011.

[19] A. Miles and S. Bechhofer, "SKOS simple knowledge organization system reference," W3C recommendation, W3C, Aug. 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/.

[20] D. Allemang, J. Hendler, and F. Gandon, *Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL*. New York, NY, USA: Association for Computing Machinery, 3 ed., 2020.

[21] K. Balog, M. Bron, and M. De Rijke, "Query modeling for entity search based on terms, categories, and examples," *ACM Trans. Inf. Syst.*, vol. 29, pp. 22:1–22:31, Dec. 2011.

[22] K. Bi, Q. Ai, and W. B. Croft, "Revisiting iterative relevance feedback for document and passage retrieval," *arXiv preprint arXiv:1812.05731*, 2018.

[23] J. Lin, R. Nogueira, and A. Yates, "Pretrained transformers for text ranking: Bert and beyond," *arXiv preprint arXiv:2010.06467*, 2020.

[24] K. Balog, M. Bron, and M. De Rijke, "Query modeling for entity search based on terms, categories, and examples," *ACM Transactions on Information Systems (TOIS)*, vol. 29, no. 4, pp. 1–31, 2011.

Citation: J. Catacora, A. Casali and C. Deco. *Legal Information Retrieval System with Entity-Based Query Expansion: Case study in Traffic Accident Litigation.* Journal of Computer Science & Technology, vol. 22, no. 2, pp. 151–163, 2022.
DOI: 10.24215/16666038.22.e12
Received: March 31, 2022 **Accepted:** June 13, 2022.