ISBN 978 -987-633-574-4

Sistema de Recuperación de Información con Expansión de la Consulta Basada en Entidades

Joel Catacora[1], Ana Casali[12] y Claudia Deco[13]

[1] Facultad de Cs. Exactas, Ingeniería y Agrimensura,
Universidad Nacional de Rosario, Rosario, Argentina
[2] Centro Int. Franco Argentino de Cs. de la Información y de Sistemas
(CIFASIS: CONICET-UNR)
[3] Universidad Católica Argentina, Rosario, Argentina
{acasali,deco}@fceia.unr.edu.ar

Resumen Se propone un sistema de búsqueda semántico que expande la consulta del usuario mediante la retroalimentación por relevancia. Este sistema es aplicado al dominio legal, en particular al derecho civil, para mejorar la precisión de los resultados que puede encontrar un abogado en su búsqueda de jurisprudencia relevante para la construcción del marco legal de un caso. Para esto se utilizan las entidades pertenecientes a una base de conocimiento como medio para reformular la consulta. Se presentan dos modelos de expansión: uno automático y otro interactivo que le sugiere al usuario conceptos relacionados a su búsqueda inicial. El sistema de búsqueda propuesto se prueba a partir de un conjunto de sumarios del Sistema Argentino de Información Jurídica (SAIJ), utilizando una base de conocimiento que incluye una ontología legal desarrollada para este trabajo y el Tesauro SAIJ de Derecho Argentino. Con este sistema se pretende mejorar la experiencia de búsqueda del usuario y la precisión de los resultados.

Palabras claves: Recuperación de Información; Expansión de la Consulta; Retroalimentación por Relevancia; Base de Conocimiento; Tesauro

1. Introducción

La Recuperación de Información consiste en mostrarle al usuario documentos relevantes ante una consulta de palabras claves. Los modelos de recuperación representan formalmente el proceso de correspondencia entre la consulta y el documento. Uno de los enfoques utilizados consiste en incorporar en los modelos de búsqueda la información de una base de conocimiento mediante la expansión de la consulta. Este mecanismo añade a la consulta original otras palabras que capturan la intención del usuario o simplemente producen una consulta que permite recuperar documentos más relevantes. Las entidades o conceptos pertenecientes a la base de conocimiento pueden ser un medio para expandir una consulta.

En este trabajo se propone expandir la consulta utilizando la retroalimentación por relevancia, incorporando información semántica disponible en una base de conocimiento. En lugar de expandir la consulta con los términos de los documentos relevantes, se utilizan los términos de las entidades que se encuentran en dicha base. Se evalúan los algoritmos de búsquedas propuestos en el ámbito legal, donde

ISBN 978 -987-633-574-4

se puede asistir a los abogados en su profesión, por ejemplo, para elaborar una estrategia de defensa. Para esto, se tienen dos fuentes de información disponibles: un tesauro del dominio y una colección de documentos indexados temáticamente con el tesauro. Las fuentes de información con los que se realizó la experimentación provienen del Sistema Argentino de Información Jurídica[4] (SAIJ). El SAIJ es una base de datos documental que contiene legislación, jurisprudencia y doctrina, tanto nacional como provincial. Ofrece búsquedas por facetas a partir del Tesauro SAIJ de Derecho Argentino[5]. El usuario puede ingresar como consulta un tema del Tesauro para recuperar los documentos que se encuentran clasificados con dicho tema. Hay ciertas limitaciones en la búsqueda por facetas y búsquedas por palabras claves, por ejemplo, el usuario es el encargado de hallar los términos más cercanos a su necesidad de información y sólo puede expresar búsquedas por conjunciones de palabras claves.

Con respecto a trabajos relacionados, en Argentina hay investigaciones recientes sobre la búsqueda o recomendación de información legal y el reconocimiento de entidades nombradas en documentos legales. En [9] se propone un sistema de recomendación de normativas para construir de manera semi-automática la matriz legal de una empresa, se utiliza el algoritmo Support Vector Machine sobre leyes nacionales catalogadas con conceptos del Tesauro del SAIJ. Otro modelo de recuperación de información jurídica basado en ontologías y distancias semánticas se propone en [4] donde se utilizan vocabularios legales y generales (ConceptNet, WordReference, Banco de Vocabularios Jurídicos Argentinos) para expandir la consulta y ranquear los documentos a partir de similitudes basadas en Normalized Google Distance. En [3] se aplican técnicas de procesamiento automático del lenguaje a un conjunto de leyes de nacionales, se identifican entidades legales, utilizando el algoritmo supervisado Stanford NER y reglas manuales.

En este trabajo, se plantea un sistema de expansión semántica de búsqueda a partir de una base de conocimiento y documentos catalogados temáticamente, se lo aplica a la recuperación de información jurídica argentina. Se presentan dos modelos de expansión: uno automático y otro interactivo que le sugiere al usuario conceptos relacionados a su búsqueda inicial. El sistema propuesto se prueba sobre un conjunto de sumarios del SAIJ y una base de conocimiento integrada por una ontología legal desarrollada para este trabajo y el Tesauro SAIJ de Derecho Argentino.

La estructura de este artículo es la siguiente. En la Sección 2 se presentan conceptos preliminares. En la Sección 3 se detallan la arquitectura del sistema de búsqueda y los modelos de expansión de la consulta propuestos y en la Sección 4 se expone la experimentación realizada. Finalmente, en la Sección 5 se tienen las conclusiones.

2. Conceptos preliminares

La incertidumbre asociada a la relevancia de un documento frente a una consulta ha sido modelada probabilísticamente de diferentes maneras, entre ellas se destacan los modelos del lenguaje, estos definen una distribución de probabilidades sobre cadenas de texto representando un determinado lenguaje. En la Recuperación de

[4] http://www.saij.gob.ar

[5] http://datos.jus.gob.ar/dataset/tesauro-saij-de-derecho-argentino

ISBN 978 -987-633-574-4

Información suelen utilizarse los modelos del lenguaje más simples, los unigramas o modelos del lenguaje con distribución multinomial, donde se asume términos con independencia condicional y posicional, es decir, un modelo *bag of words*. Entre los modelos de recuperación basados en modelos se lenguaje se tiene el Query Likelihood Model (QL). Para cada documento d de la colección se define un modelo de lenguaje θ_d, el cual describe el tema del documento o las palabras claves que el usuario ingresaría si quisiera recuperar dicho documento. El puntaje del documento d es probabilidad de que la consulta $q = \langle w_1, \dots w_n \rangle$ sea una muestra o se genere de acuerdo a cada uno de estos modelos del lenguaje $P(q|\theta_d)$:

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

El modelo del lenguaje θ_d se estima a partir del documento d, mediante Maximum Likelihood Estimation y técnicas de suavizado. Entre los métodos de suavizado se tiene la interpolación de Jelinek-Mercer (JM):

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

donde $\lambda \in [0, 1]$ es un parámetro del suavizado, $\mathrm{tf}_{t,d}$ es la frecuencia del término t en el documento d, cf_t es la frecuencia del término t en toda la colección de documentos y $|d|$ la longitud del documento.

Además, se han desarrollado adaptaciones que permiten incluir la estructura de los documentos (sus campos), p. ej., título, autor, en los modelos de recuperación, esto potencia el rendimiento de las búsquedas. Para el modelo QL, se tienen dos extensiones: el Mixture Language Model (MLM) [8], donde se añaden pesos como parámetros del modelo que miden la importancia de cada campo del documento, el Probabilistic Retrieval Model for Semistructured Data (PRMS) [5], donde a partir del modelo anterior se propone una forma no supervisada de estimar los pesos de los campos.

Un tipo particular de objeto que puede recuperarse es una entidad. La recuperación de entidades tiene como objetivo responder a las consultas mediante una lista ranqueada de entidades, por ejemplo "países limítrofes de Argentina". Las *entidades* o conceptos son objetos unívocamente identificables, con nombre, atributos y relaciones con otras entidades, por ejemplo, personas, localizaciones y organizaciones. Estas pertenecen a un *catálogo de entidades*, un diccionario con los nombres de las entidades junto a sus identificadores. Este catálogo puede ser una base de conocimiento modelada por una ontología. Un enfoque para ranquear a las entidades consiste en utilizar los algoritmos de recuperación tradicionales sobre representaciones documentales de las entidades, aplicando sin modificaciones los modelos diseñados para ranquear documentos. Para esto, es necesario, crear un documento, llamado *descripción de la entidad*, para cada entidad del catálogo, el cual mantiene toda la información de la entidad en la base de conocimiento. La técnica que se utiliza para la construcción de las descripciones de las entidades se denomina *predicate folding* [1, p. 69].

ISBN 978 -987-633-574-4

3. Modelo de Expansión de Consulta y Sistema Propuesto

Se propone una arquitectura de un sistema de búsqueda enriquecido semánticamente y dos modelos de búsquedas no supervisados que utilizan la información contenida en una base de conocimiento del dominio para expandir la consulta. La expansión se realiza mediante la retroalimentación por (pseudo) relevancia basada en entidades [7,1]. Por un lado se desarrolla un Modelo de Relevancia con Entidades (RE) que genera de forma automática entidades que se esperan sean relevantes a la consulta, mediante documentos que describen a las entidades. Luego, se expande la consulta con los términos más importantes de las entidades generadas. El otro método, el Modelo Iterativo de Relevancia con Entidades (IRE) requiere la asistencia del usuario para que seleccione entre las entidades sugeridas en una o más iteraciones, aquellas que considere relevantes. La arquitectura propuesta se muestra en la Figura 1, donde el usuario ingresa una búsqueda de texto libre, se la transforma en términos índice y a partir de estos términos se aplica el algoritmo de búsqueda. Luego, el sistema retorna una lista ranqueada de documentos como respuesta a su consulta (módulo *Interacción con el usuario*). Para expandir la consulta es necesaria su reformulación por parte del modelo de expansión, para luego ejecutar el algoritmo de búsqueda sobre la consulta reformulada.

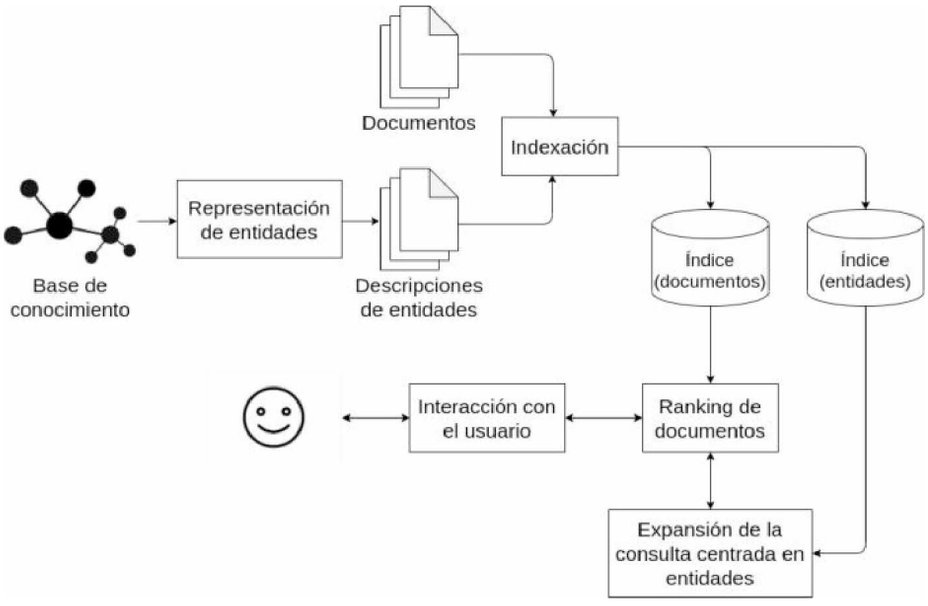


Figura 1: Arquitectura propuesta.

En este caso, el algoritmo de expansión utiliza las entidades pertenecientes a una base de conocimiento, en lugar de los documentos de la colección. El núcleo del motor de búsqueda, la implementación del modelo de recuperación y expansión de la consulta se encuentran en los módulos: *Ranking de documentos* y *Expansión de la consulta centrada en entidades*. La expansión puede realizarse de manera automáti-

ISBN 978 -987-633-574-4

ca o iterativa. Si se utiliza el Modelo IRE se genera una lista ranqueada de entidades en cada iteración del modelo y el usuario debe juzgar si son relevantes a su consulta inicial y luego se procede a la reformulación de la consulta. Este sistema tiene dos fuentes de información: la colección de documentos y la base de conocimiento. Los modelos de expansión no consultan directamente a la base de conocimiento sino que realizan sus cálculos sobre las descripciones de las entidades. Estas descripciones se crean a partir de la base de conocimiento mediante el procedimiento *predicate folding*. El módulo *Representación de entidades* implementa dicha transformación. Para realizar los cálculos del modelo de recuperación de manera eficiente se requiere precomputar ciertos valores estadísticos de la colección de documentos, los cuales se mantienen en estructuras de datos llamadas índices (módulo *Indexación*). En este caso, es necesaria la construcción de dos índices, uno por cada colección de documentos. El vocabulario de términos del sistema de búsqueda es la intersección de los vocabularios de cada índice. El modelo de expansión se encarga de consultar el índice de entidades y el modelo de recuperación de documentos consulta al índice de los documentos. Se utiliza como fuente de información una base de conocimiento legal diseñada para este trabajo y una colección de sumarios recolectados del SAIJ.

En la Tabla 1 se puede ver la descripción de la entidad "cinturón de seguridad" construida a partir de la base de conocimiento LegalBase, donde los campos *sumarios* y *sumarios-títulos* contienen todos los documentos catalogados con el concepto "cinturón de seguridad".

Campos	**Valor**
nombres	cinturón de seguridad, cinturón.
entidades-relacionadas	reglas de tránsito, tránsito automotor, automotores, vehículos, Transporte, ...
sumarios	La omisión de empleo del cinturón... La impugnación no ha de prosperar...
sumarios-títulos	Daños y perjuicios, ... Recurso de inconstitucionalidad, ...
catch-all	cinturón de seguridad, cinturón. reglas de tránsito, tránsito automotor, automotores,vehículos, Transporte, ... La omisión de empleo del cinturón... La impugnación no ha de prosperar... Daños y perjuicios, ... Recurso de inconstitucionalidad, ...

Tabla 1: Descripción de la entidad "cinturón de seguridad".

Desarrollo de una base de conocimiento legal

Para la evaluación de la propuesta se desarrollaron una base de conocimiento legal (LegalBase) y una colección de test, Sumarios20 (compuesta por 45.556 sumarios del SAIJ pertenecientes al Derecho Civil, del ámbito nacional y de la provincia de Santa Fe, junto con 10 necesidades de información). Para la creación de LegalBase se definió un tesauro de Accidentes de Tránsito reutilizando el Tesauro SAIJ y la ontología LegalOnto, la cual expresa la indexación semántica de documentos legales. Se inició la creación del tesauro identificando las palabras claves que utilizaría un abogado para describir su necesidad de información si afronta un caso

ISBN 978 -987-633-574-4

sobre accidente de tránsito. Las palabras claves se obtienen de casos ficticios y del Tesauro SAIJ. Esto se trabajó con la asistencia de un abogado. Al igual que el Tesauro SAIJ, el tesauro Accidentes de Tránsito sigue el estándar SKOS[6], define un total de 91 conceptos y 18 grupos de conceptos, de los cuales 11 son conceptos nuevos que no pertenecen al tesauro SAIJ. Por ejemplo, ante un caso de accidente de tránsito, el abogado de la víctima debería probar la existencia de los siguientes tópicos: Daño; Antijuridicidad; Relación de causalidad; Reproche o factor de atribución sobre el demandado. A partir de este ejemplo, notamos que algunos conceptos del tesauro podrían agruparse en los temas anteriores, por ej., el concepto "lucro cesante" podría formar parte del grupo "daño". Es decir, ciertos temas podrían tratarse como conjuntos de conceptos. Además, se decidió desarrollar una ontología (LegalOnto) que modele la indexación temática de documentos legales (legislación, jurisprudencia y doctrina) con conceptos SKOS. A partir de estas fuentes: el tesauro de Accidentes de Tránsito, la ontología LegalOnto y el tesauro SAIJ, se crea la base de conocimiento LegalBase para reunir el conocimiento disponible sobre las entidades legales. La integración se implementó con la directiva `owl:import` y se realizó de forma de no producir inconsistencias. La ontología LegalBase fue poblada con los sumarios de la colección Sumarios20.

Búsqueda de documentos

Se proponen dos modelos de expansión: uno que realiza una expansión automática, sin intervención del usuario, y el otro donde el usuario interviene en forma iterativa.

Expansión automática: El modelo de expansión de consultas Conceptual Language Model [7] se define de la siguiente manera para una consulta q:

$$P(t|q) \approx \sum_{e \in \mathcal{E}} P(t|e)P(e|q), \tag{2}$$

donde $\mathcal{E}$ es el catálogo de entidades. Este modelo asume que la probabilidad de seleccionar un término sólo depende del concepto una vez que se ha seleccionado ese concepto para la consulta. Deben estimarse dos componentes: la selección de términos $P(t|e)$ y la selección de entidades $P(e|q)$. Proponemos estimar: $P(t|e)$ con el modelo del lenguaje de la entidad, es decir $P(t|e) = P(t|\theta_e)$, y $P(e|q)$ con el modelo QL aplicado a la recuperación de entidades. Luego, como en [7] los documentos se ranquean con la divergencia de Kullback–Leibler (KL).

Se puede ver que el modelo de expansión propuesto es equivalente a Relevance Model [6], o también llamado RM1, con entidades en lugar de documentos, considerando todas las entidades igualmente probables ($P(e)$ se ignora). Por esto, lo llamamos Modelo de Relevancia con Entidades (RE).

Expansión interactiva: Se propone el Modelo Iterativo de Relevancia con Entidades (IRE), basado en el Iterative Relevance Model [2]. Las entidades candidatas

[6] `https://www.w3.org/2009/08/skos-reference/skos.html`

ISBN 978 -987-633-574-4

que se muestran al usuario en la i-ésima iteración se obtienen de aplicar KL:

$$score(e, q^{(i)}) = \sum_{t \in V} P(t|\theta_q^{(i-1)}) \; log \; P(t|\theta_e), \tag{3}$$

donde V es el vocabulario, $q^{(i)} = (q, \theta_q^{(i-1)})$, con $\theta_q^{(i-1)}$ la consulta expandida y reformulada de la iteración anterior, definida por la Ecuación 5, siendo $1 \leq i \leq n$. Las entidades que ya han sido mostradas en iteraciones anteriores se remueven del ranking. Las primeras k entidades de la Ecuación 3 que son juzgadas como relevantes, en la i-ésima iteración, conforman el conjunto $\mathcal{E}_q^{(i)}(k)$. Estas entidades son añadidas al conjunto de todas las revisiones hechas por el usuario hasta la i-ésima iteración $E_q^{(i)}$, o sea:

$$E_q^{(i)} = E_q^{(i-1)} \cup \; \mathcal{E}_q^{(i)}(k),$$

donde $E^{(0)} = entities(q)$ (*entity linking* sobre la consulta).

Luego, la expansión de la consulta de la i-ésima iteración es la siguiente:

$$P^{(i)}(t|\hat{\theta}_q) = \frac{1}{|E_q^{(i)}|} \sum_{e \in E_q^{(i)}} P(t|\theta_e). \tag{4}$$

Esta fórmula se obtiene de la Ecuación 2, considerando una selección de entidades por parte del usuario con distribución uniforme. En la práctica, solo se tienen en cuenta a los términos con los puntajes más altos para formar el modelo de la consulta expandida, con las probabilidades renormalizadas de modo tal que $\sum_t P^{(i)}(t|\hat{\theta}_q) = 1$. La reformulación de la consulta en la i-ésima iteración es:

$$P^{(i)}(t|\theta_q) = \begin{cases} (1-\lambda_q)P^{(0)}(t|\theta_q) + \lambda_q P^{(i)}(t|\hat{\theta}_q) & si \;\; i \geq 1 \\ \frac{\mathrm{tf}_{t,q}}{|q|} & si \; i = 0 \end{cases}, \tag{5}$$

donde $\lambda_q \in [0, 1]$ controla la influencia del modelo de expansión. En la próxima iteración, se ranquean las nuevas entidades candidatas (Ecuación 3) con $\theta_q^{(i)}$. Si $i < n$, entonces:

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

En la última iteración, $i = n$, se obtiene la reformulación de la consulta final, $\theta_q^{(n)}$. Luego, esta consulta expandida puede ser utilizada para la recuperación de documentos mediante el modelo KL. Se muestra en la Figura 2 un ejemplo de búsqueda iterativa, se sugieren 5 entidades por iteración, 5×2 (Entidades-Iteración).

En resumen, se mostraron dos modelos semánticos de búsqueda: el Modelo RE y el Modelo IRE. Ambos, utilizan una base de conocimiento para expandir la consulta a partir de los términos asociados a las entidades relacionadas a la consulta. El modelo IRE sugiere entidades ante una consulta en lugar de documentos, las cuales pueden revisarse por el usuario más rápidamente es decir, las sugerencias se relacionan semánticamente con la consulta y permiten precisar la dirección de la búsqueda ya sea para profundizarla o para moverse hacia otros aspectos de la búsqueda. Estos modelos pueden incorporan la semi-estructura de las descripciones de las entidades, mediante MLM y PRMS. Todos los modelos propuestos junto a sus extensiones son no supervisados, no requieren juicios de relevancia o *query logs*.

ISBN 978 -987-633-574-4

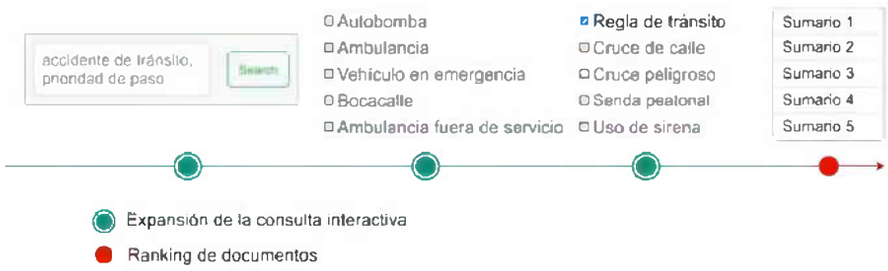


Figura 2: Línea de tiempo sobre la búsqueda "accidente de tránsito, prioridad de paso", con expansión de consulta iterativa.

4. Experimentación

Se realiza sobre la colección Sumario20 con los siguientes modelos de recuperación centrados en entidades: RE, RE+MLM, RE+PRMS, IRE, IRE+MLM, IRE+PRMS, junto con el algoritmo RM3 (con interpolación JM).

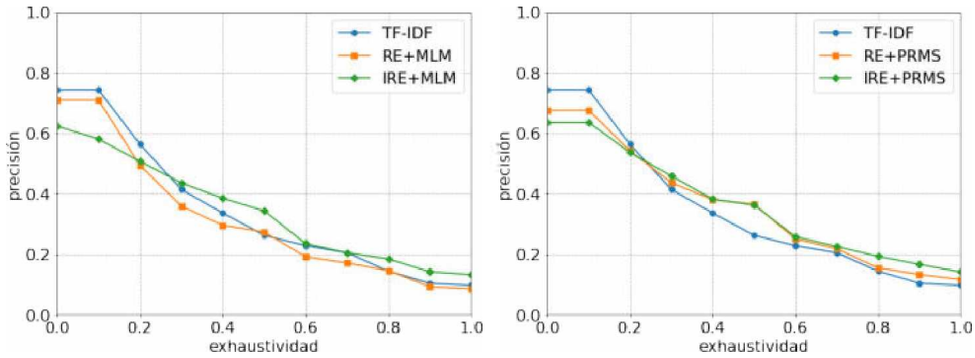
La evaluación de los algoritmos de búsqueda se realiza mediante *4-fold cross-validation*, los parámetros se ajustan con *grid search* y se utiliza *freezing ranking* para evaluar el modelo iterativo siguiendo lo propuesto en [2]. Se optimizó un conjunto determinado de variables debido al limitado poder de computo del hardware disponible. Los parámetros optimizados son: la interpolación del suavizado JM, al estimarse $P(t|\theta_e)$ en los modelos propuestos y $P(t|\theta_d)$ en RM3, y los pesos de los campos de las descripciones para los modelos PRMS y MLM, con valores en el intervalo $[0, 1]$ y saltos de 0,25. La interpolación de la consulta inicial con el modelo de la consulta expandida es $\lambda_q = 0{,}25$ en IRE (Ecuación 5) y RM3, y $\lambda_q = 0{,}5$ en RE. Se consideraron los primeros 15 términos, las primeras 10 entidades (RE), 10 documentos (RM3) y 2 iteraciones de 10 entidades, 2×10 (IRE).

En la Tabla 2 se comparan los modelos con expansión de la consulta. Los algoritmos que utilizan al modelo PRMS, alcanzan los mayores rendimientos en términos de MAP, superiores a los modelos MLM. De acuerdo a los índices de robustez, los modelos basados en MLM tienen un mayor desvío de la consulta que aquellos que usan PRMS. Es decir, los términos de expansión de la consulta generados por los modelos RE+MLM e IRE+MLM tendrían una menor relación a la consulta inicial en comparación a los términos producidos por los modelos RE+PRMS e IRE+PRMS. En la Figura 3 se muestran dos curvas de precisión y exhaustividad, donde se los compara con la implementación del modelo TF-IDF de Apache Lucene[7].

En la evaluación no se encontró diferencias significativas entre el modelo RE y modelo IRE en la colección Sumarios20. Saber cuáles son las entidades relevantes a una consulta, como lo hace el modelo IRE, no presenta en esta experimentación una diferencia frente a la expansión automática de la consulta del modelo RE. Ambos superan el rendimiento del algoritmo RM3. Entendemos que la selección interactiva de entidades no ofrece grandes beneficios ante la expansión automática, ya que las entidades seleccionadas por el usuario no aportan mejores términos para expandir la consulta. De acuerdo al proceso de creación de las descripción de las entidades, los

[7] `https://lucene.apache.org/`

ISBN 978 -987-633-574-4



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

Figura 3: Curvas de precisión y exhaustividad.

Colección	Métrica	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**

Tabla 2: Evaluación de sistemas de recuperación con expansión de la consulta. El índice de robustez se calcula sobre TF-IDF.

términos que expanden la consulta se obtienen de documentos que tratan sobre los temas que le interesan al usuario, pero no necesariamente son los mejores términos para expandir la consulta. Además, se probaron extensiones a estos modelos: los modelos MLM y PRMS. El uso del modelo PRMS, tanto en los algoritmos RE como IRE, alcanza un mayor rendimiento que MLM. Luego, en esta experimentación los algoritmos de búsqueda con expansión de la consulta centrada en entidades que alcanzan el mayor rendimiento son los que utilizan al modelo PRMS.

5. Conclusiones

En este trabajo se propone la arquitectura de un sistema de búsqueda con expansión de la consulta para mejorar la recuperación de documentos. Este sistema permite incorporar como fuente de información a una base de conocimiento, de modo tal de expandir la consulta a partir de los términos asociados a las entidades relevantes a la consulta. Se implementaron en dicha arquitectura dos algoritmos de búsqueda no supervisados basados en el Conceptual Language Model. Uno es el Modelo de Relevancia con Entidades, el cual realiza la expansión de manera automática. El otro es el Modelo Iterativo de Relevancia con Entidades, que requiere la asistencia del usuario. Estos modelos se evaluaron en el dominio legal, utilizando una base de conocimiento legal (LegalBase) y una colección de test (Sumarios20) desarrolladas especialmente para este trabajo. Esta base de conocimiento puede

ISBN 978 -987-633-574-4

utilizarse para otras aplicaciones futuras. En la evaluación no se encontraron diferencias significativas entre estos modelos para la colección Sumarios20. Además, se probaron las extensiones MLM y PRMS sobre estos modelos, siendo PRMS el de mayor rendimiento. Se debe tener en cuenta que los resultados mostrados están sujetos a parámetros que no fueron totalmente optimizados ya que estuvo restringido por limitaciones en el hardware y que la colección de test Sumario20 es muy pequeña para ser representativa. De todas maneras los resultados mediante estos modelos de expansión semántica resultan alentadores.

A través de este sistema de búsqueda semántico se espera ayudar a los profesionales del derecho en la recuperación de documentos que les sean útiles para la redacción de una demanda. Como trabajo futuro es necesario trabajar sobre colecciones más grandes y extender el soporte a la búsqueda para otras ramas del derecho. Se propone incluir en el modelo de expansión una selección de términos que dependa de la consulta e incorporar modelos de lenguaje no supervisados como los *word embeddings*.

Referencias

1. Balog, K.: Entity-Oriented Search, The Information Retrieval Series, vol. 39. Springer (2018), `https://eos-book.org`
2. Bi, K., Ai, Q., Croft, W.B.: Revisiting iterative relevance feedback for document and passage retrieval. arXiv preprint arXiv:1812.05731 (2018)
3. Cardellino, F., Cardellino, C., Haag, K., Soto, A., Teruel, M., Alonso i Alemany, L., Villata, S.: Mejora del acceso a infoleg mediante técnicas de procesamiento automático del lenguaje. In: XVIII SID- 47 JAIIO (2018)
4. Dehner, G.A., Eckert, K.B., Lezcano, J.M., Ruidías, H.J.: 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)
5. Kim, J., Xue, X., Croft, W.B.: A probabilistic retrieval model for semistructured data. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) Advances in Information Retrieval. pp. 228–239. Springer Berlin Heidelberg (2009)
6. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 120–127. New York, NY, USA (2001)
7. Meij, E., Trieschnigg, D., de Rijke, M., Kraaij, W.: Conceptual language models for domain-specific retrieval. Inf. Processing & Management 46(4), 448–469 (2010)
8. Ogilvie, P., Callan, J.: Experiments using the lemur toolkit. In: TREC. vol. 1, pp. 103–108 (2001)
9. Perezzini, L., Casali, A., Deco, C.: Sistema de soporte para la recuperación de normativas en la ingeniería legal. In: SID 2020 - 49 JAIIO (2020)