Modelo de Recuperación de Información Jurídica basado en ontologías y distancias semánticas

Gabriel A. Dehner[1a], Karina B. Eckert[1b], Juan M. Lezcano[2c], Héctor J. Ruidías[1d]

[1]Departamento de Ingeniería y Ciencias de la Producción
[2]Departamento de Ciencias Jurídicas y Sociales
Universidad Gastón Dachary, Posadas, Misiones, Argentina

[a]dehner.gabriel@ugd.edu.ar, {[b]karinaeck, [c]juanmanuellezcano, [d]chandra149}@gmail.com

Resumen. En el ámbito jurídico, el proceso de búsqueda y clasificación de documentos legales, incide en gran manera en el desempeño profesional de los especialistas en Derecho. En este trabajo se propone un modelo de recuperación y clasificación de documentos jurídicos, utilizando ontologías y distancias semánticas, a fin de mejorar la relevancia de los documentos obtenidos en este dominio. Con este objetivo, la validación se llevó a cabo en dos escenarios, y el modelo fue evaluado en base a las métricas de precisión, exhaustividad y F-Score. Los resultados examinados señalan que el modelo propuesto tuvo un mejor desempeño que la búsqueda de ocurrencias literales, obteniendo un valor de F-Score promedio de 95%.

Palabras Claves: Recuperación de Información Legal, Ontologías, Búsqueda Semántica, Distancia Semántica, Sentencias.

1 Introducción

En la actualidad el acceso a información confiable, de calidad y en el tiempo oportuno juega un rol crucial para el desarrollo de la sociedad moderna, principalmente atendiendo a un aspecto fundamental que es el de la toma de decisiones. Es por ello que la calidad de la información obtenida no debe estar reñida con la oportunidad, por lo que es fundamental contar con métodos, técnicas y herramientas de Recuperación de Información (RI, en inglés Information Retrieval) que se orientan a la obtención de información relevante para los fines propuestos, en alineación con las necesidades de información de los usuarios.

Además, el acceso a información relevante se ve limitado por el volumen de datos que constantemente se encuentra en aumento. Según informa la Internacional Data Corporation (IDC), los datos globales aumentarán de 33 ZB[1] en 2018 a 175 ZB para el 2025 [1]. Se deduce entonces que por más que un usuario disponga del acceso a

[1] Un zettabyte equivale a 10^{21} bytes

toda esta información (exhaustividad), la misma será inútil ya que dicho usuario deberá realizar el trabajo de búsqueda, lo cual resulta inviable, debido a un volumen tan elevado de información. Por este motivo, la precisión en los motores de búsqueda y modelos (técnicas, enfoques) de RI son vitales para la obtención de resultados que satisfagan las necesidades del usuario; e indefectiblemente se requiere de la utilización de nuevos modelos y estrategias capaces de abordar el crecimiento exponencial de dicha información.

En el dominio jurídico, la Recuperación de Información Legal (RIL, en inglés Legal Information Retrieval) incide en gran manera en el desempeño profesional de los especialistas en Derecho. La búsqueda y RI en este ámbito se diferencia de otros dado que se considera muy compleja y sus procesos dependen en gran medida de la interpretación del conocimiento por un experto humano [2]. Dicha complejidad está dada por la estructura sintáctica y la terminología dependiente del dominio [3].

En contraste a otras disciplinas las taxonomías rara vez son inherentes a la ley. Los vocabularios legales contienen términos abiertos, siendo intrínsecamente dinámicos y cuyas aplicaciones dependen de las interpretaciones de los especialistas, resultando en términos legales semánticamente ambiguos [4], [5].

Por lo tanto, las búsquedas booleanas basadas en la aparición y repetición de los términos de la consulta ingresada por el usuario, pueden resultar ineficientes dado que no consideran un análisis global en el cual se ponderen los aspectos semánticos de dicha consulta. Servicios como Westlaw, LexisNexis y Findlaw también implementan este tipo de búsquedas.

Este método para calificar la relevancia se ve limitado al hecho de que la consulta del usuario posea una ocurrencia exacta de los términos y a la repetición de los mismos en los documentos. De este modo, pueden existir palabras no pertenecientes a la consulta, como ser sinónimos, hiperónimos, hipónimos, jerarquías entre términos, ontologías y tesauros, las cuales a pesar de que se presenten en el documento y estén fuertemente relacionados con dicha consulta, no inciden en la relevancia del mismo ya que no se atienden las relaciones entre vocablos (ontologías) [6]; es decir, este método de relevancia no contempla la semántica entre el documento y los términos ingresados para la recuperación – y menos aún la distancia semántica entre los mismos –.

Por otro lado, si los términos de una consulta no pertenecen al dominio –como ser términos en relación con aspectos culturales, folksonomías, medios de comunicación, redes sociales, etc.–, existen escasas probabilidades de recuperar documentos pertinentes. Esto también se ve reflejado con la aparición de nuevas figuras legales (homicidio agravado, femicidio, entre otros) que no tienen lugar en los documentos anteriores a las mismas de modo que, al no contemplar la semántica del texto, la búsqueda no recupera –o lo hace en forma ineficaz– documentos relevantes.

No considerar los factores semánticos y los aspectos propios del dominio legal pueden conducir a búsquedas cuyos resultados presenten documentos no relevantes, clasificados como relevantes o documentos relevantes clasificados como irrelevantes, no respondiendo así a las necesidades de los usuarios.

En este trabajo se diseña e implementa un modelo basado en ontologías –tanto legales como genéricas y de propósito general – y distancias semánticas para mejorar la

consulta del usuario y obtener sentencias relevantes. Esto se fundamenta en el hecho de que las ontologías aseguran una recuperación eficiente al permitir inferencias basadas en el conocimiento del dominio [3].

Se utilizan diversas ontologías con dos objetivos principales. El primer objetivo consiste en establecer una relación entre una consulta ingresada por el usuario – que puede no estar expresada con vocablos relativos al ámbito legal – y términos legales. Como segundo objetivo se procura expandir la consulta del usuario y, mediante la utilización de distancias semánticas generar un ranking – dado que la exhaustividad resulta inviable – para la presentación de los resultados al usuario.

Este modelo procura apreciar no sólo la ocurrencia/aparición y repetición de términos en los documentos, sino también la semántica entre la consulta ingresada por el usuario y los documentos sobre los cuales se realiza la búsqueda.

El resto de este artículo se organiza como se describe a continuación. La sección 2 presenta un breve marco conceptual referido a las ontologías, relaciones semánticas, expansión de consultas y antecedentes. Luego, en la sección 3, se describe el modelo propuesto. Seguidamente, en la sección 4, se exponen las pruebas y experimentos realizados. Finalmente, en la sección 5 y 6, se detallan las conclusiones y los lineamientos futuros, respectivamente.

2 Ontologías y Relaciones Semánticas

2.1 Expansión de Consultas

La utilidad de la tarea de expansión de consultas se ve manifiesta en que las palabras claves de una consulta pueden no encontrarse textualmente en los documentos, pero sí sus propiedades y/o relaciones definidas en las ontologías. Por lo tanto, se procura expandir una consulta de búsqueda a partir de las palabras claves, mediante la obtención de sinónimos y términos relacionados (hiperónimos, hipónimos, merónimos) en las ontologías. En el ámbito de la RI, las ontologías permiten expandir las consultas [7]. Análogamente sucede con la RIL y las ontologías legales [8].

Como ejemplo de esta utilidad, Saravanan y otros [8] utilizaron una ontología legal referida al control de alquileres, impuestos y otros. Al recibir una consulta de usuario que contiene como palabras claves “rental arreas” (atrasos de alquiler), el Sistema de Recuperación de Información Legal (SRIL) recopilaría términos relacionados en la ontología, incluyendo “rent in arrears” (renta atrasada), “default of payment of rent” (incumplimiento del pago de la renta), y otros términos y frases relacionadas. De esta manera no sólo se busca el patrón exacto sino también otras formas derivadas de palabras o frases.

En este trabajo se utilizan distintas ontologías y redes semánticas; algunas de éstas con relaciones propias del ámbito legal y otras de propósito general dado que los usuarios pueden expresar sus consultas con términos que no estén referidos específicamente al ámbito jurídico. Por lo tanto, se busca establecer relaciones semánticas entre los términos, que permitan corresponder la consulta del usuario con los vocablos propios de los documentos legales.

2.2 Medidas Semánticas

En diversos ámbitos de aplicación se utilizan cálculos de medidas semánticas para poder determinar cómo y en cuánto se relacionan dos o más palabras [9], [10].

Para este trabajo, se emplea la Normalizad Google Distance (NGD) para los cálculos de distancias y relaciones semánticas.

La métrica NGD fue propuesta por Cilibrasi y Vitanyi [11] y se basa en los conteos de páginas de Google (total de documentos con un término) para determinar la distancia semántica entre palabras.

Ahora bien, debido a que NGD [12] no tiene en cuenta el contexto en el que coexisten las palabras, tiene inconvenientes:

- El análisis de conteo de páginas ignora la posición de una página web. Por lo tanto, aunque aparezcan dos palabras en una página, es posible que no estén realmente relacionadas.
- El recuento de páginas de una palabra polisémica (una palabra con varios sentidos) puede contener una combinación de todos sus sentidos. Por ejemplo, los conteos de páginas para apple contienen recuentos de páginas para apple como fruta y apple como compañía.
- A su vez, dada la escala y el ruido en la Web, las palabras pueden coexistir en las páginas sin estar realmente relacionadas.

2.3 Antecedentes

Improving legal information retrieval using an ontological framework [8]

En este artículo se propuso un marco ontológico para mejorar la consulta del usuario para la recuperación de juicios legales relevantes. Las ontologías aseguraron una recuperación eficiente al permitir inferencias basadas en el conocimiento del dominio.

Los resultados empíricos dados en esta publicación demuestran que las búsquedas basadas en ontologías generan resultados significativamente mejores que los métodos de búsqueda tradicionales. Cabe destacar que no pudieron hacerse inferencias significativas con términos ajenos al dominio legal.

Research on information retrieval model based on ontology [13]

En esta publicación se presenta un modelo basado en ontologías de un dominio. Dicho modelo incluye el procesamiento y recuperación de los documentos. Además, se utilizó un algoritmo genético como parte del modelo de recuperación.

Para la creación de ontologías y las pruebas, se utilizó un corpus de 1000 artículos científicos y documentos de la IEEE (Institute of Electrical and Electronics Engineers). Éstos se dividieron en 10 grupos, los cuales tenían 100 artículos relacionados a un tema o consulta. El criterio para evaluar la RI considera la similitud de cada artículo con cada palabra de la consulta.

El modelo de recuperación basado en ontologías muestra una mejor precisión y tasa de recuperación, comprendiendo en mayor medida los requisitos de los usuarios.

3 Modelo Propuesto

El modelo propuesto para la recuperación y clasificación de documentos legales se encuentra representado en la Fig. 1. Mediante este modelo se busca establecer la relevancia de los documentos, otorgando como resultado final un ranking a partir del cual el usuario puede vislumbrar un orden basado en este enfoque de búsqueda.

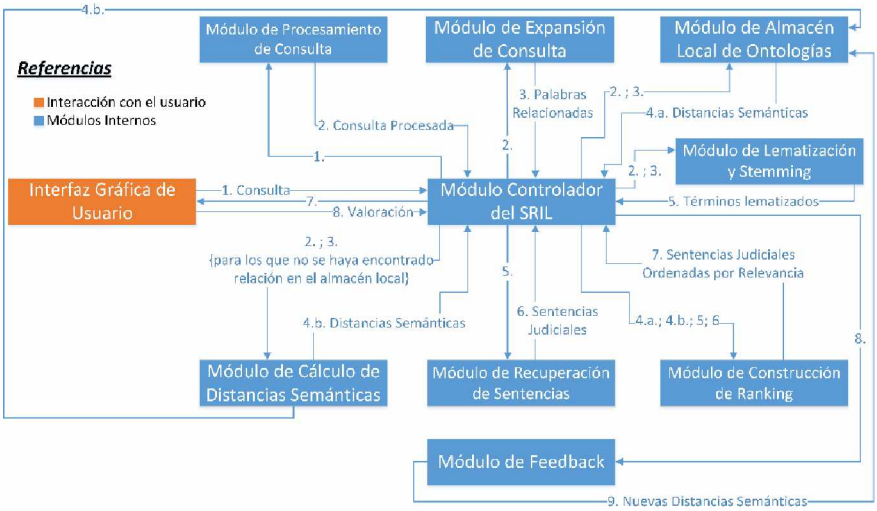

Fig. 1. Módulos e interacciones del SRIL

El funcionamiento global del modelo comienza con la *Consulta* que ingresa al *Módulo Controlador del SRIL* y es enviado al *Módulo de Procesamiento de Consulta* para tratar los términos obteniendo la *Consulta Procesada*. Luego, mediante el *Módulo de Expansión de Consulta* se recuperan las *Palabras Relacionadas*, a partir de las cuales se calculan las *Distancias/Relaciones Semánticas* con el *Módulo de Cálculo de Distancias Semánticas* y el *Módulo de Almacén Local de Ontologías*.

Seguidamente, el *Módulo de Lematización y Stemming* retorna los *Términos Lematizados* que son utilizados por el *Módulo de Recuperación de Sentencias* para obtener las *Sentencias Judiciales*.

A continuación, el *Módulo de Construcción de Ranking* es utilizado a fin de recuperar las *Sentencias Judiciales Ordenadas por Relevancia*. Éstas últimas son presentadas al usuario –mediante la *Interfaz Gráfica de Usuario*–, quien además puede emitir una *Valoración* de los resultados, la cual es recibida por el *Módulo de Feedback* para calcular las *Nuevas Distancias/Relaciones Semánticas* que serán actualizadas en el almacén local.

Para llevar a cabo la implementación del modelo se utilizó un entorno de ejecución para JavaScript conocido como Node [14], debido a la gran compatibilidad con diferentes tecnologías y APIs utilizadas en este trabajo.

Más específicamente para la interacción con el usuario (frontend) se utilizó el framework conocido como Angular [15] y para el desarrollo interno del backend, Express [16].

Las funciones de los módulos de la Fig. 1 son descritas a continuación.

3.1 Módulo de Procesamiento de Consulta

Este módulo se encarga de adecuar los términos de la consulta ingresada por el usuario.

Al recibir la consulta, se identifican los términos y/o frases que forman parte de la misma, proceso similar a la tokenización. Por cada aparición del carácter especial "+" los vocablos (y frases) son separados para evaluarse individualmente. Sucesivamente, para cada uno de éstos, se realiza la corrección ortográfica (en español) – proceso conocido como Spelling – de los tokens involucrados en la consulta. Además, se eliminan signos y caracteres tales como "!", "@", ";", "#", entre otros.

Este procesamiento incide en otros módulos dado que la modificación de los vocablos repercutirá en las búsquedas de relaciones con otros términos y cálculos de distancias semánticas.

3.2 Módulo de Expansión de Consulta

La tarea de este módulo consiste en recuperar las relaciones existentes con los términos de la consulta. Para ello se utilizan las siguientes ontologías y redes semánticas:

- ConceptNet [17]: donde se obtienen relaciones como "Synonym" (sinónimos), "RelatedTo" (relacionado con), "IsA" (es un), "DerivedFrom" (derivado de), "EtymologicallyRelatedTo" (etimológicamente relacionado con), entre otras.
- WordReference [18]: donde se obtienen sinónimos.
- Banco de Vocabularios Jurídicos Argentinos (BVJA) [19]: donde se obtienen relaciones como "usado por" (UP), "término general" (TG), "término específico" (TE), "término relacionado" (TR), entre otros.

A continuación se reflejan gráficamente, en la Fig. 2, las interacciones entre las APIs descritas anteriormente.

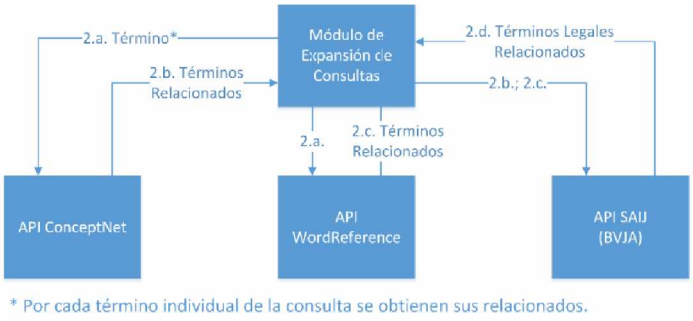

Fig. 2. Interacción del Módulo de Expansión de Consultas con APIs externas.

Por cada *Término* –o frase individual– se buscan sus relacionados tanto en la *API ConceptNet* como en la *API WordReference*. Luego, el *Módulo de Expansión de Consultas* recibe los *Términos Relacionados*. A partir de estos últimos, se recupera de la

API del Sistema Argentino de Información Jurídica (BVJA) los *Términos Legales Relacionados* a los anteriores.

Esta tarea aporta flexibilidad dado que las redes semánticas van cambiando conforme avanza el tiempo, por lo cual se obtiene una mayor versatilidad al no depender de términos locales estáticos en el tiempo.

3.3 Módulo de Almacén Local de Ontologías

Además, se gestiona una base de datos con lenguaje de consultas SPARQL [20] – a través de la *Plataforma de Gestión Stardog* [21] (Fig. 3) – que incluye el almacén de los términos anteriormente recuperados (sección 3.2) y a su vez la distancia/relación semántica existente entre la consulta de búsqueda y los términos recuperados (sección 3.4).

Por cada token o frase individual de la consulta se recurre a esta base de datos (*BD SPARQL*) para la obtención de la *Distancia Semántica* entre vocablos, la cual será utilizada posteriormente para la construcción del ranking y orden de relevancia (sección 3.7).

Por otro lado, este módulo también se comunica con el feedback (sección 3.8) para poder ir modificando/ajustando –por medio de *Query (SPARQL)*– las distancias entre términos en función de las valoraciones y calificaciones otorgadas por el usuario respecto de la relevancia de los documentos. De esta manera se proporciona una gran flexibilidad dado que las distancias entre términos pueden ir variando conforme avanza el tiempo.

Fig. 3. Interacción de Módulo de Almacén de Ontologías con Stardog.

3.4 Módulo de Cálculo de Distancias Semánticas

En caso de que no se posea almacenada localmente la distancia semántica entre dos términos, se utiliza la NGD como valor estimado de la relación. A continuación se observa en la Fig. 4 la interacción del módulo con el buscador de Google.

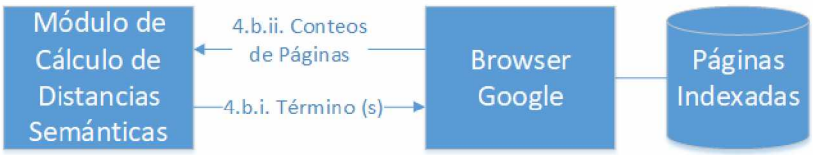

Fig. 4. Interacción del Sistema con el Buscador de Google.

El *Módulo de Cálculo de Distancias Semánticas* peticiona al *Browser Google* –y éste busca en su base de *Páginas Indexadas*– los *Conteos de Páginas* de cada *Término/s.* A partir de los *Conteos de Páginas* se computa el valor de NGD, y se calcula una estimación inicial de la relación semántica entre dos términos.

Esta aproximación inicial se va adaptando mediante el módulo de feedback definido en la sección 3.8.

3.5 Módulo de Lematización y Stemming

Este módulo se ocupa de la tarea relacionada a la conversión de los vocablos a su lema (raíz) eliminando los plurales, géneros y formas verbales. De esta manera se logra una normalización lingüística reduciendo el término a una única forma común a la cual se la conoce como stem o lema. El objetivo de este proceso refiere a poder identificar – mediante el módulo de la sección 3.6 – coincidencias con palabras que posean igual raíz, sin limitar a la aparición exacta de las mismas en los documentos de búsqueda [4].

3.6 Módulo de Recuperación de Sentencias

Este módulo gestiona la interacción con la base de datos que posee los documentos jurídicos. Específicamente, dichos documentos corresponden con Sentencias Judiciales de la Tercera Circunscripción Judicial extraídas del vigente Sistema de Publicaciones del Portal Oficial, provenientes del Superior Tribunal de Justicia de la Provincia de Misiones, con carácter confidencial.

Una vez obtenidos los lemas de las distintas palabras obtenidas (módulo de la sección 3.5), se busca cada sentencia que posea al menos una ocurrencia de dichos lemas/raíces.

Todas las sentencias recuperadas son comunicadas al módulo de la sección 3.7 para la elaboración del ranking de sentencias basado en la distancia/relación semántica entre términos.

3.7 Módulo de Construcción de Ranking

Este módulo se ocupa de determinar el orden de las sentencias, estableciendo a su vez, la relevancia cualitativa de las mismas (Muy Relevante, Relevante, Suficientemente Relevante, Poco Relevante, Irrelevante). Para ello, a cada sentencia se le asigna un puntaje de acuerdo a la cantidad de ocurrencias de las palabras y las relaciones/ponderaciones semánticas.

A su vez, se deduce que cuanto más cercano se encuentren, o mayor similitud posean los vocablos de la consulta y sus relacionados, mayor valor en el ranking tendrá un documento – sin olvidar la ocurrencia de los términos –.

Por lo tanto, un término puede poseer grandes cantidades de ocurrencias y mayor frecuencia que otros vocablos, pero si no posee estrecha relación con la consulta, no aportará significativamente al valor (puntaje) del ranking. Análogamente, a pesar de que un término pueda tener poca frecuencia en un documento, al poseer un valor ele-

vado de relación semántica puede aportar en gran manera para el valor del ranking, dado que dicho coeficiente está condicionado por la relación semántica entre los términos.

Las sentencias se organizan en orden decreciente, es decir, aquellas sentencias con mayores puntuaciones tendrán valores *Muy Relevante* y las siguientes irán descendiendo hasta *Irrelevante*.

3.8 Módulo de Feedback

Este módulo se utiliza con el objetivo de retroalimentar, mediante las valoraciones/calificaciones de los usuarios, el SRIL para modificar y ajustar las distancias entre los términos, fortaleciéndolas o debilitándolas – por medio de un coeficiente y de acuerdo al porcentaje de participación en el valor de relevancia –, en función de dichas valoraciones.

Si el SRIL determina el valor de *Relevante* para un documento *j*, y el usuario lo califica como *Poco Relevante*, se deduce que la relación semántica entre los términos ocurridos en el documento debería ser menor, dado que la calificación para dicho documento difiere de la establecida por el sistema.

Para ajustar las distancias/relaciones entre los términos se contempla el aporte que cada término otorga para el valor en el ranking, de manera que aquellos que mayor importancia tienen, sufrirán un mayor ajuste. Así, se obtendrán nuevas distancias entre los términos, las cuales permitirán corregir las anteriores, a fin de obtener un valor en ranking que tienda ser más acorde a la valoración del usuario.

3.9 Módulo Controlador del SRIL

Éste se ocupa de administrar interfaces para la interacción entre los distintos módulos permitiendo así realizar las funcionalidades necesarias para SRIL.

4 Pruebas y Resultados

A fin de comparar el modelo de búsqueda planteado (sección 3) – y variantes del mismo – con búsquedas basadas únicamente en la ocurrencia literal de las palabras, se determinan 2 escenarios, los cuales se encuentran definidos con temáticas referidas al ámbito jurídico.

En primer lugar, se definieron los distintos corpus de documentos, compuestos por conjuntos de sentencias judiciales.

Luego, se determinaron las claves (o consultas) de búsqueda a utilizar sobre cada corpus, para así obtener los resultados de cada uno de los diferentes modelos.

Todos los modelos fueron evaluados mediante métricas conocidas como precisión (ecuación 1), exhaustividad (ecuación 2) y F-score – como media armónica de las dos anteriores –.

$$Precisión = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ positivos} \quad (1)$$

$$Exhaustividad = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ negativos} \quad (2)$$

Se tomaron como relevantes los resultados clasificados con valores *Muy Relevante*, *Relevante*, *Suficientemente Relevante* y *Poco Relevante*, en caso contrario –*Nada relevante*– fue considerado como una sentencia irrelevante frente a la consulta de búsqueda utilizada en el conjunto de documentos.

Específicamente se utilizaron 3 modelos de recuperación/clasificación con 2 variantes cada uno. La primera de las variantes construye el ranking en intervalos de igual tamaño; en cambio, la segunda cuenta con la aplicación de la técnica de Clustering Jerárquico Aglomerativo para la conformación de dicho ranking.

Entre los modelos/estrategias a comparar se emplearon:

- La búsqueda de ocurrencia literal, es decir, la correspondencia exacta de los términos de la consulta.
- La RI basada en ontologías utilizando NGD para obtener la relación semántica.
- La RI basada en ontologías empleando una aproximación a NGD –considerando como fuente de conteos de búsqueda un conjunto de sentencias judiciales locales, en lugar de Google– a la cual se la denotará como Distancia Jurídica Normalizada (DJN). De esta manera se pretende solventar la inexactitud de NGD para palabras polisémicas dado que DJN considera únicamente los conteos de palabras en documentos jurídicos.

4.1 Prueba 1: "Accidentes de Trabajo"

En la primera prueba se determinó la temática de "*Accidentes de Trabajo*" para el corpus de documentos. Dicho corpus cuenta con 50 sentencias judiciales, las cuales se dividen en 35 relevantes y 15 irrelevantes para el tema en cuestión.

Se utilizaron simultáneamente 4 claves de búsqueda: *damnificación*, *obra*, *incapacidad*, *enfermedad*. Cada uno de estos términos refiere, al tema de las sentencias, pero no mencionan explícitamente *accidente de trabajo* o *accidente laboral*.

En la Tabla 1 se observan los resultados utilizando las variantes de modelos descritos anteriormente.

Tabla 1. Resultados de Prueba N° 1

Modelo/ Búsqueda	VCMR	CSRRC	TSRR	Precisión	Exhaustividad	F-Score
(1)	[9, 2]	8	9	0,8889	0,2286	0,3636
(2)		8	9	0,8889	0,2286	0,3636
(3)	[37,86-8,37]	6	6	1	0,1714	0,2927
(4)	[37,86-3,16]	30	31	0,9677	0,8571	0,9091
(5)	[59,43-12.90]	10	10	1	0,2857	0,4444
(6)	[59.43-2.88]	35	44	0,7954	1	0,8861

Referencias de Tabla 1 y Tabla 2.

VCMR: Valores cuantitativos máximo y mínimo de sentencias relevantes.

CSRRC: Cantidad de sentencias relevantes recuperadas correctamente.

TSRR: Total de Sentencias relevantes recuperadas.

(1) Búsqueda de Ocurrencia Literal; (2) Búsqueda de Ocurrencia Literal c/clustering; (3) Modelo con NGD; (4) Modelo con NGD c/clustering; (5) Modelo con DJN; (6) Modelo con DJN c/clustering.

Cabe destacar que los VCMR de la búsqueda de ocurrencia literal, son menores que las demás estrategias, dado que las ocurrencias de los términos son inferiores. Esto sucede porque dichas ocurrencias son únicamente con los vocablos de la consulta de búsqueda. Esta deficiencia es solventada, en estas pruebas, por los métodos semánticos propuestos.

Asimismo, el número de resultados correctos entre el número de todos los resultados devueltos (precisión) superan el 75%. En cambio, la exhaustividad difiere en gran manera por cada modelo.

Al estudiar los resultados obtenidos por los métodos semánticos (3) y (5), se detectaron que los primeros valores cuantitativos del ranking de las sentencias, eran demasiado altos (outliers) en relación a los demás valores. De esta manera, al dividir las escalas del ranking en igual tamaño, las sentencias de cada intervalo no eran representativas del mismo por causa de los valores altos –en relación a los demás– en el ranking. Por este motivo se utilizó la técnica de clustering para agrupar de una manera diferente las escalas de relevancia de los documentos. Los resultados mejoraron significativamente en más del doble de su F-Score. Además, se observa que la búsqueda de ocurrencia literal (1) no ofreció mejoría al utilizar clustering (2) para el ranking, dado que, mediante esta estrategia, todos los documentos clasificados como irrelevantes poseían valores prácticamente nulos. Esto es producto de que las ocurrencias fueron muy escasas pese a que la consulta de búsqueda se relacionó en gran manera con las sentencias. Es decir, el valor de recuperación tan bajo por parte de esta búsqueda, no tuvo relación con la construcción del ranking.

Por otro lado, al examinar el modelo (5) y (6) se percibió que la DJN no mejora la exactitud de la relación semántica entre términos dado que, el conjunto de sentencias que se poseían para realizar los conteos, fue reducido y, por ende, muchas de las co-ocurrencias entre términos fueron nulas.

Cabe mencionar que los modelos semánticos presentaron una desventaja, en cuanto al tiempo que demoraron los mismos en otorgar los resultados. Los modelos que más demoraron fueron el (3) y (4); en promedio, tres minutos.

Al resumir esta primera prueba, el modelo (4) obtuvo las mejores puntuaciones; cerca de 91% en F-Score.

4.2 Prueba 2: "Homicidio"

En esta prueba se definió, para el corpus de documentos, la temática de *"Homicidio"*. Dicho corpus cuenta con 50 sentencias judiciales, las cuales se dividen en 30 documentos relevantes y 20 irrelevantes para el tema establecido.

Al igual que en la prueba anterior, se utilizaron simultáneamente 4 claves de búsqueda: *muerte*, *asesinato*, *crimen*, *parricidio*. Cada término refiere, al tema de las sentencias, pero no mencionan exactamente el término *homicidio*.

En la Tabla 2 se observan los resultados utilizando las variantes de modelos descritos.

Tabla 2. Resultados de Prueba N° 2

Modelo	VCMR	CSRRC	TSRR	Precisión	Exhaustividad	F-Score
(1)	[10-3]	4	4	1	0,1333	0,2353
(2)		4	4	1	0,1333	0,2353
(3)	[74,21-15]	13	13	1	0,4333	0,6046
(4)	[74,21-6,50]	30	30	1	1	1
(5)	[85,44-17,62]	17	17	1	0,5667	0,7234
(6)	[85,44-8,30]	30	31	0,9677	1	0,9836

Se puede observar que los VCMR de la búsqueda de ocurrencia literal, son menores que las demás estrategias, por el mismo motivo descrito en la primera prueba.

Además, casi todas las estrategias –a excepción de (6)– poseen la más alta precisión (100%), en cambio, solo las búsquedas (4), (5) y (6) superan el 50% en la exhaustividad. En cuanto a los tiempos de demora en brindar resultados, los valores fueron similares a la primera prueba.

En resumen, nuevamente el modelo (4) obtuvo las mejores puntuaciones; el 100% en F-Score.

5 Conclusiones

En base a las pruebas realizadas, se puede afirmar que el modelo planteado es apto para la recuperación y clasificación de documentos jurídicos.

En términos generales corresponde mencionar que, en principio, los resultados del modelo semántico propuesto (y sus variantes) no aportaron mejoras significativas en la clasificación de las sentencias. Sus valores de F-Score no superaron el 50%, producto de que muchos de los documentos relevantes fueron tratados como irrelevantes – buena precisión, pero poca exhaustividad –.

Luego, a partir de la utilización del Clustering Jerárquico Aglomerativo para la construcción del ranking, se obtuvieron mejorías significativas de acuerdo a las métricas utilizadas para la validación. Cabe destacar que al utilizar clustering en las búsquedas de ocurrencia literal, no se percibieron cambios en los resultados.

Al examinar la primera prueba, se observa que DJN es una muy buena aproximación de distancia entre términos jurídicos dado que su variante correspondiente del modelo, obtuvo un 89% en F-Score. De esta manera, se obtuvo otra alternativa cuantificable de relación entre términos, sin la necesidad de realizar peticiones a Google para el cálculo de NGD. La ventaja radica en el menor tiempo que se requiere para el

cálculo local de DJN en contraste a NGD. Igualmente, cabe mencionar que el modelo utilizando NGD y clustering obtuvo el mayor puntaje; el 91% en F-Score.

En la segunda prueba, habiendo cambiado el corpus de sentencias, los resultados analizados evidencian similitudes con el escenario anterior dado que la mejor puntuación fue del modelo con NGD y clustering con un valor de 100% en F-Score.

Cabe mencionar que el modelo con DJN otorga valores menores de F-Score. Al examinar en detalle, se observó que, por falta de un mayor número de documentos, muchos de los pares de términos no poseían coocurrencias en las sentencias y, por ende, la relación entre éstos terminó siendo muy cercana a 0, no aportando a la recuperación de dichas sentencias.

Teniendo en cuenta la problemática planteada al inicio, referida a la necesidad de información relevante en el ámbito jurídico, se considera que este modelo basado en ontologías y distancias semánticas brinda mejoras notables para la búsqueda y recuperación de documentos de esta índole.

6 Lineamientos Futuros

En primer lugar, sería de gran valor aumentar la cantidad de sentencias utilizadas a fin de obtener una DJN más significativa entre vocablos y, por ende, mejores resultados.

Además, debido al tiempo requerido por el método de NGD para brindar resultados, sería necesario utilizar otra alternativa para el cálculo de la distancia entre vocablos. Para ello se podría optimizar la DJN, utilizada en las pruebas, ya que mejora el tiempo mencionado. Esto otorgaría una mayor eficiencia dado que se reducirían los tiempos de procesamiento, lo cual es fundamental considerando la usabilidad del sistema implementado.

Por otra parte, se podría realizar inferencias mediante las ontologías, con el objetivo de deducir relaciones significativas entre vocablos de búsqueda y no solamente una expansión de términos.

7 Referencias

1. Reinsel, D., Gantz, J., Rydning, J.: The Digitization of the World from Edge to Core. 28 (2018).
2. Peters, W., Sagri, M.-T., Tiscornia, D.: The structuring of legal knowledge in LOIS. Artificial Intelligence and Law. 15, 117–135 (2007). https://doi.org/10.1007/s10506-007-9034-4.
3. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, Cambridge (2011). https://doi.org/10.1017/CBO9781139058452.
4. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press, New York (2008).
5. Büttcher, S., Clarke, C.L.A., Cormack, G.V.: Information Retrieval: implementing and evaluating search engines. The MIT Press, Cambridge, Massachusetts London, England (2016).

6. Guarino, N., Oberle, D., Staab, S.: What Is an Ontology? In: Staab, S. and Studer, R. (eds.) Handbook on Ontologies. pp. 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_0.
7. Ashley, K.D.: Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge University Press, Cambridge (2017). https://doi.org/10.1017/9781316761380.
8. Saravanan, M., Ravindran, B., Raman, S.: Improving legal information retrieval using an ontological framework. Artificial Intelligence and Law. 17, 101–124 (2009). https://doi.org/10.1007/s10506-009-9075-y.
9. Gracia, J., Mena, E.: Web-Based Measure of Semantic Relatedness. In: Bailey, J., Maier, D., Schewe, K.-D., Thalheim, B., and Wang, X.S. (eds.) Web Information Systems Engineering - WISE 2008. pp. 136–150. Springer Berlin Heidelberg, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85481-4_12.
10. Lofi, C.: Measuring Semantic Similarity and Relatedness with Distributional and Knowledge-based Approaches. IMT. 10, 493–501 (2015). https://doi.org/10.11185/imt.10.493.
11. Cilibrasi, R.L., Vitanyi, P.M.B.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering. 19, 370–383 (2007). https://doi.org/10.1109/TKDE.2007.48.
12. Bollegala, D., Matsuo, Y., Ishizuka, M.: A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. IEEE Transactions on Knowledge and Data Engineering. 23, 977–990 (2011). https://doi.org/10.1109/TKDE.2010.172.
13. Yu, B.: Research on information retrieval model based on ontology. EURASIP Journal on Wireless Communications and Networking. 2019, 30 (2019). https://doi.org/10.1186/s13638-019-1354-z.
14. Node.js, F. de: Node.js, https://nodejs.org/es/.
15. Google: Angular, https://angular.io/.
16. StrongLoop, Inc. y otros colaboradores de expressjs.com: Express - Infraestructura de aplicaciones web Node.js, https://expressjs.com/es/.
17. Luminoso Technologies, Inc: ConceptNet, http://conceptnet.io/.
18. English to French, Italian, German & Spanish Dictionary - WordReference.com, http://www.wordreference.com/.
19. Sistema Argentino de Información Jurídica: Banco de Vocabularios Jurídicos de Argentina | SAIJ, http://vocabularios.saij.gob.ar/portalthes/acerca.php.
20. SPARQL Query Language for RDF, https://www.w3.org/TR/rdf-sparql-query/.
21. Union, S.: Stardog: The Enterprise Knowledge Graph Platform, https://www.stardog.com/.