La búsqueda de información jurídica: de los tesauros a la inteligencia artificial

Searching legal information: from thesauri to Artificial Intelligence

Ángel SANCHO FERRER, Carlos FERNÁNDEZ HERNÁNDEZ, Pierre BOULAT

Departamento de I+D de Wolters Kluwer España
{asancho|cafernandez|pboulat}@wke.es

Resumen

Se repasa la evolución de las tecnologías de búsqueda de información legal en soporte electrónico, a lo largo de sus 25 años de existencia. Se presta especial atención a las desarrolladas por el Departamento de I+D de Wolters Kluwer España que, basadas en la búsqueda en lenguaje natural, con expansión semántica y avanzados algoritmos de presentación de resultados por relevancia, se han situado a la cabeza de la tecnología mundial en este ámbito. Los últimos estudios sobre el comportamiento de los usuarios evidencian la necesidad de continuar avanzando en estos desarrollos para dar solución a los complejos problemas detectados, incorporando a los motores de búsqueda nuevas y mejores técnicas de inteligencia artificial.

Palabras clave: Recuperación de información. Información legal. Lenguaje natural. Expansión semántica. Inteligencia artificial.

Abstract

The evolution of search technologies in electronic legal information, along its 25 years of existence, is reviewed. Special attention is given to those developed by the R & D team of Wolters Kluwer Spain, which are based on natural language search with semantic expansion and advanced algorithms for the presentation of results by relevance, and have been at the forefront of the technology in this field worldwide. Recent studies on user behavior show the need for further progress in these developments to solve the complex problems detected, involving the improvement of the search engines with new and enhanced artificial intelligence techniques.

Keywords: Information retrieval. Legal information. Natural language. Semantic expansion. Artificial Intelligence.

> Search is a wicked problem of terrific consequences. As the choice of first resort for many users and tasks, search is a defining element of the user experience. […] Search is the source of endless frustration […] We find too many results or too few […] Search remains as noisy and irregular as language and communication […]
>
> (Morville & Callender, 2010)

1. Problemas que plantea la búsqueda de información legal en soporte electrónico

1.1. La dificultad de formular un problema jurídico

La mayoría de las bases de datos legales existentes en el mercado son de tipo documental, es decir, contienen un determinado número de documentos de origen o interés legal, que los profesionales del derecho necesitan consultar, típicamente para conocer la norma aplicable a una cuestión, la interpretación que de la misma han realizado los tribunales y el análisis que de una y otra han realizado los estudiosos.

Para localizar esa información se utiliza la búsqueda, la cual se dirige esencialmente a ideas y conceptos (Morville y Rosenfeld, 2006).

Un típico escenario de búsqueda se inicia con una necesidad de información que un usuario debe primero destilar en una consulta, generalmente breve, que se formula al motor de búsqueda. Posteriormente el usuario debe analizar los resultados obtenidos, identificar los documentos relevantes y seguidamente formular nuevas consultas basado en lo que ha aprendido con los documentos anteriores, repitiendo el proceso hasta confiar en que la búsqueda le permite identificar las decisiones a adoptar a continuación (Jackson y Al-Kofahi, 2010).

Toda búsqueda en una base de datos documental consiste en establecer una relación entre las palabras contenidas en una consulta y unos documentos que contienen esas mismas palabras. Y este problema presenta una serie de desafíos, el primero de los cuales es que frecuentemente la necesidad de información está insuficientemente descrita. Sea en el ámbito legal o no, los usuarios construyen sus búsquedas con conceptos ambiguos y desordenados. Y dado que esas búsquedas se dirigen a documentos llenos igualmente de conceptos, ambigüedad y desorden, la respuesta a la consulta,

si existe, es un objetivo igualmente ambiguo y en movimiento (Morville y Callender, 2006).

Utilizando un símil, podríamos identificar la búsqueda como el puente que une la necesidad de información del usuario con el documento o documentos que responden a esa necesidad. Y ese puente está formado por una combinación de palabras que, de no ser las adecuadas, no permitirán encontrar el documento necesario: "[the user] have to say the magic words in the right combination" (Jackson, 2010).

Como también han señalado Morville y Rosenfeld (2006, p. 149), la búsqueda es un proceso iterativo, no solo porque los resultados obtenidos puedan no ser satisfactorios en el primer intento, sino también porque frecuentemente dar con las palabras adecuadas para una consulta requiere varios intentos.

1.2. De los tesauros a la búsqueda libre

La búsqueda de información legal en soporte electrónico se planteó inicialmente intentando reproducir las técnicas de búsqueda del soporte en papel: los tesauros y los árboles de contenidos.

Como ya pusimos de manifiesto en Sancho Ferrer et al. (2008 y 2011), tales instrumentos no permiten obtener buenos resultados. Los tesauros sirven para clasificar información o para navegar por ella, pero no para buscarla en grandes repositorios. Manejar un tesauro en modo electrónico es mucho más complicado que en papel.

Para paliar esos problemas, se planteó la construcción de consultas mediante operadores booleanos (Y, O, NO,…), que indicasen la relación deseada entre las palabras utilizadas, pero este sistema resulta complejo y requiere unos conocimientos de sintaxis frecuentemente desconocidos por los usuarios.

Por eso el uso de operadores es muy residual en la práctica, como muestra el siguiente cuadro (Tabla I), construido a partir del análisis de 3.000 consultas, en el que claramente se aprecia que el operador más utilizado es "Y" (innecesario en la mayoría de ocasiones), seguido de la búsqueda literal en tan solo un 11% escaso de ocasiones.

Por ende, la aparición del buscador Google a finales de los años 90 pronto puso de manifiesto que, en el soporte Internet, era posible localizar resultados de muy buena calidad sin utilizar operadores booleanos, simplemente utilizando consultas compuestas con pocas palabras.

Este formato de búsqueda libre es el que ha acabado convirtiéndose en el standard de búsqueda de información legal.

Operadores	*% de utilización*
AND	32,67%
Literal	10,96%
NEAR	1,07%
OR	0,77%
NOT	0,1%
Near and before	0,03%

Tabla I. Porcentaje de uso de operadores booleanos en bases de datos legales

De ahí la radical importancia de la consulta como punto de partida para establecer esa relación.

2. La consulta como clave de la búsqueda

Como acabamos de señalar, un problema jurídico es difícil de plantear en términos de consulta a una base de datos.

La consulta está compuesta por palabras o números, o una combinación de ambos. Las palabras utilizadas pueden tener diferentes significados. Un conjunto de palabras puede formar una expresión (concepto) con un significado propio (que también, en ocasiones, puede ser polisémico). El significado de cada palabra o concepto está condicionado por el resto de palabras (y números) que constituyen la consulta, pues todas interactúan entre sí.

Ese conjunto de palabras constituyen la expresión con la que el usuario representa su necesidad de documentación. Es la forma en que el usuario intenta acercarse al lenguaje del documento que necesita, expresándose de un modo que considera comprensible por la máquina.

Una primera aproximación a las fases de la construcción de una consulta puede ser la siguiente:

1. El usuario se representa el problema jurídico para el que necesita información *(Thinking).*
2. El usuario construye mentalmente la formulación de la consulta *(Building).*
3. El usuario escribe la consulta *(Typing).*

En la práctica, esta secuencia típica puede ser objeto de una pausada reflexión en pos de la

solución óptima o, más habitualmente, producirse en unidad de acto, sin solución de continuidad entre las diferentes fases. Sin embargo cada una de ellas presenta una problemática y casuística propias.

Por eso el problema principal que se deriva de la consulta es la interpretación de la voluntad del usuario (el *User intent,* en terminología anglosajona), expresado por medio de las palabras introducidas en la caja de búsqueda.

Hoy en día, las tecnologías de búsqueda no son capaces de identificar la intención o necesidad del usuario, ni de descifrar el significado que el usuario ha atribuido al conjunto de palabras que constituyen la consulta. Solo pueden trabajar con las palabras utilizadas en la consulta y, en la medida de lo posible, facilitar al usuario la selección de la mejor combinación para expresar su necesidad, en la menor cantidad posible de intentos.

El esfuerzo actual de los investigadores en la materia consiste principalmente en facilitar la expresión o formalización de la necesidad de documentación, en la forma más útil para la máquina.

2.1. ¿Cómo buscan los usuarios?

A través del análisis de los regitros *(logs)* de búsqueda podemos conocer lo que los usuarios esperan de nuestras aplicaciones y cómo articulan o describen sus necesidades a través de la búsqueda (Morville y Rosenfeld, 2006). En Sancho Ferrer et al. (2008a; 2008b) ya presentamos unos primeros datos sobre el comportamiento de los usuarios al formular una búsqueda de contenido legal. Según nuestros estudios, esos datos no han experimentado variaciones sustanciales desde entonces.

Sintéticamente, podemos recordar ahora que los usuarios se inclinan por la opción más sencilla para formular sus consultas: la búsqueda libre sin operadores. Un 75% de las búsquedas se realizan a texto libre, con un promedio de tres términos por consulta, con muchos errores tipográficos, y con dos a tres intentos de reformulación por consulta, añadiendo, quitando o sustituyendo términos.

Y, como hemos recordado *supra,* del total de búsquedas, la mitad no utilizan ningún operador, cerca de un 40% utilizan el operador <Y> (lo que, en la práctica, significa no utilizar ninguno, pues <Y> es el operador utilizado por defecto entre todos los términos de una búsqueda por la mayoría de motores de búsqueda). Solo en un muy reducido porcentaje de casos se ha advertido la utilización de operadores booleanos avanzados (<O>, <NO>, paréntesis, truncamiento), pero, en la mayoría de las veces mal utilizados (frecuentemente los paréntesis no están correctamente colocados), por lo que las búsquedas no pueden proporcionar buenos resultados.

Un interesante dato final en este repaso: en la mayoría de los casos, los usuarios solo consultan la primera página de resultados, por lo que aquellos documentos que no se encuentren en las primeras posiciones, difícilmente serán localizados.

2.2. Tipología de las consultas

El trabajo con miles de registros de búsqueda de usuarios y con las tecnologías de búsqueda disponibles, evidencia que la variedad de consultas es muy grande (Figura 1).

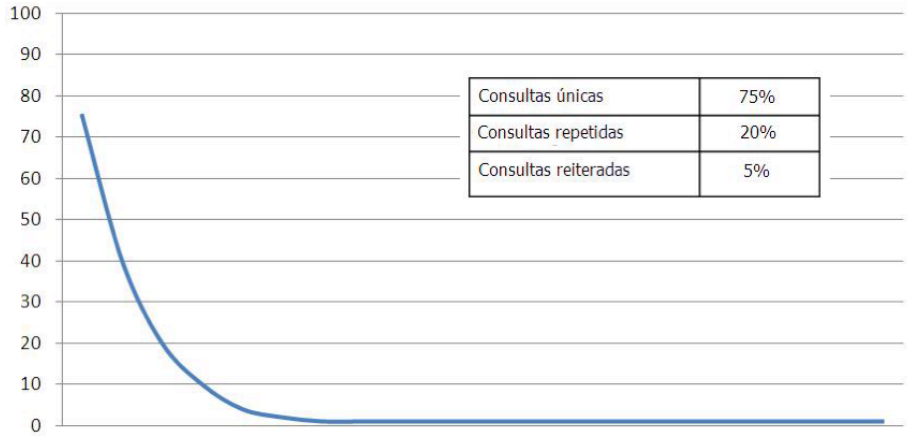

Fig.1. Porcentaje de consultas únicas a bases de datos legales

Son muy pocas las consultas repetidas y muchas las consultas únicas (siguiendo el clásico modelo de *long tail* definido por Anderson en 2004).

El análisis de las consultas que devuelven peores resultados nos ha permitido identificar los tipos de consultas a solventar más frecuentes:

1. *Consultas muy cortas:* De uno a dos palabras o conceptos. Por su propia naturaleza suelen ser muy genéricas y, por tanto, devuelven un muy alto número de resultados, lo que las hace poco útiles. Obligan a reformular varias veces la consulta, añadiendo nuevas palabras.
   - — Impugnación costas
   - — Embargo de bienes
   - — Compensación deudas
2. *Consultas muy largas:* Excesivamente minuciosas, plantean tantas restricciones al motor de búsqueda (por la necesidad de encontrar los términos buscados en una proximidad relevante) que no suelen devolver resultados. Obligan a reformular la consulta suprimiendo alguna o varias de las palabras utilizadas.
   - — Indemnización por incumplimiento de contrato en cuanto a los plazos de ejecución de determinadas obras de mejora de un inmovilizado
   - — Contrato de arrendamiento de local y fijación de la fecha de finalización cuando el cómputo de los años de duración establecido no es coincidente con la fecha de finalización fijada en el contrato
   - — Responsabilidad patrimonial de la administración negligencia médica por falta de atención con resultado de muerte
3. *Consultas destinadas a localizar un documento concreto:*
   - — Ley de aguas
   - — Ley del mercado de valores
   - — Ley de sociedades de capital
4. *Consultas engañosas:* Consultas que aparentemente representan de forma apropiada un problema jurídico concreto, pero que no permiten al buscador obtener respuestas de calidad (pese a que, frecuentemente, un humano sí podría construir dicha respuesta, trabajando en la materia, pues existen disposiciones legales que permiten ofrecer respuesta al problema planteado).
   - — Principio acusatorio hechos probados
   - — No contestación a la demanda cuando se reclama el pago de un pagaré
   - — Valor del silencio como manifestación contractual
5. *Consultas numéricas:* Compuestas exclusiva o mayoritariamente por números, representan una gran variedad de objetivos (normas, artículos, número de sentencia, número de recurso de sentencias, fechas, números oficiales de referencia…) (Tabla II)

Cada uno de los supuestos requiere un tratamiento tecnológico diferente.

Fecha	*Número de recurso*	*Número de norma o sentencia*	*Artículo de norma*
STS de 21 de septiembre de 2007	STS 12 de junio de 2003 sala 1ª rec 3172/1997	Directiva 2008/1/CE del parlamento europeo	Artículo 136 lrjap-pac
09/02/05 Audiencia Nacional, Sala de lo Contencioso	06/06/2007 nº 06/22650	Ley 2/1987 de 16 de marzo elecciones Extremadura	Disposición adicional 15 de la lopj 1/2009
08 de octubre de 2010 consentimiento familiar	0595-04 de 11 de marzo de 2004	Real Decreto 261/2002 de 8 de marzo	Artículo 1041 ley 7/1985

Tabla II. Ejemplos de búsquedas de documentos por datos numéricos

Sin embargo, a efectos prácticos, esta enumeración puede simplificarse en solo dos categorías:

1. *Buenas consultas:* Aquellas que proporcionan resultados con una alta o completa coincidencia con la literalidad del texto buscado en la base de datos y el significado previsiblemente atribuible al mismo por el usuario, satisfaciendo sus expectativas.
2. *Malas consultas:* Aquellas que proporcionan resultados con poca o ninguna coincidencia con la literalidad del texto buscado o el significado previsiblemente atribuible al mismo y, por tanto, no satisfaciendo las expectativas de los usuarios.

Un usuario experto (típicamente, un documentalista, que conoce el dominio jurídico, las colecciones documentales a consultar y la sintaxis de consulta), construirá su búsqueda con más precisión, utilizando las expresiones adecuadas a sus necesidades, junto con sus sinónimos, relacionándolas adecuadamente por medio de operadores booleanos (Sancho Ferrer et al., 2008; 2008b). Este comportamiento experto contiene elementos explícitos, mientras que otros elementos solo están tácitamente incluidos, lo que, nuevamente, conduce a la necesidad del análisis de

los comportamientos de búsqueda (Jackson y Al-Kofahi, 2010).

2.3. La búsqueda semántica

Como modelo para superar las dificultades derivadas de la búsqueda libre, ya en 2008 se presentó la metodología de búsqueda semántica aplicada a las bases de datos legales (Sancho Ferrer et al., 2008 a y b). Esta metodología y conjunto de funcionalidades constituyó en aquel momento una novedad mundial, entonces aplicada solo a un reducido número de productos españoles. Hoy día, esa misma metodología se aplica en productos de Francia y de Portugal, en todos los casos con excelentes resultados.

Sus principales elementos, en relación con la creación y ampliación del ámbito de la consulta son: (1) el reconocimiento de los conceptos utilizados en la búsqueda por el usuario; (2) la incorporación a la búsqueda, por medio de un diccionario creado al efecto, de los sinónimos, acrónimos e incluso erratas de tales expresiones; y (3) una funcionalidad de reconocimiento de erratas en el texto de las consultas, que permite la sugerencia de redacciones correctas en un muy alto porcentaje de casos. La figura 2 ilustra el modo en que estos sinónimos se incorporan a la búsqueda.

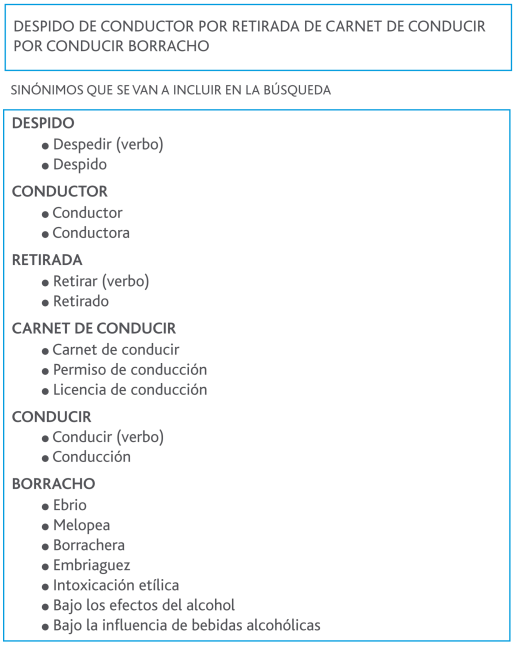

Fig.2. Búsqueda semántica en La Ley Digital

Estas nuevas tecnologías supusieron una importante mejora en la calidad de la búsqueda, en relación con los modelos existentes hasta entonces (Sancho Ferrer et al., 2008a).

2.4. El autocompletado de consultas (y encontrar antes de acabar de escribir)

Un nuevo paso en la ayuda a la formulación de buenas consultas lo constituye el autocompletado de consultas, una funcionalidad bien conocida. La construcción de este tipo de funcionalidad se basa en la combinación de dos elementos: (1) una fuente de sugerencias y (2) un algoritmo adecuado.

La fuente de sugerencias proviene, generalmente, de dos tipos de fuentes: (a) consultas formuladas por los usuarios y recogidas por el sistema y (b) un tesauro de materias. Las sugerencias basadas en tesauros ofrecen la ventaja de ser fáciles de construir, ser sistemáticas y resultar visualmente atractivas. Sin embargo, también presentan importantes limitaciones prácticas: su carácter es más limitado y su extensión más reducida, no ofrecen garantías de calidad, pueden producir cero resultados y requieren un mantenimiento manual, siempre complejo.

Por su parte, las búsquedas realizadas por otros usuarios presentan mucha mayor variedad y detalle que las procedentes de un tesauro, pero tampoco son *per se* suficientes para garantizar buenas sugerencias, pues un alto porcentaje de las mismas no permiten obtener buenos resultados, generalmente porque contienen cualquiera de los tipos de consultas difícilmente resolubles en la actualidad por un motor de búsqueda (consultas muy largas, o compuestas principalmente por números, o con erratas, o imprecisas).

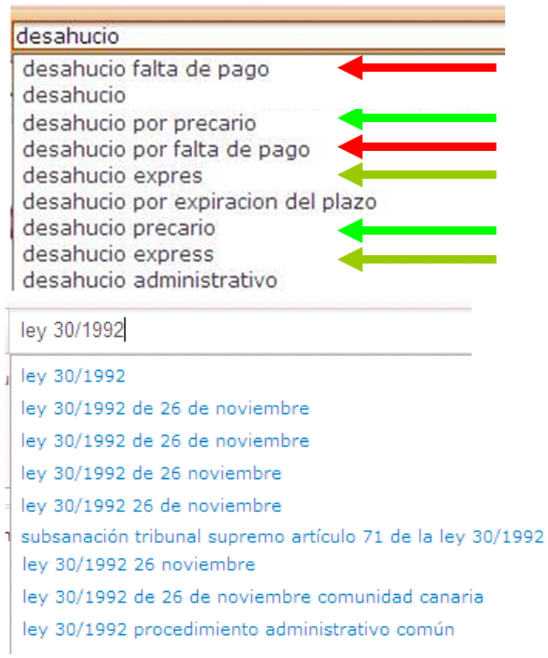

Fig. 3. Ejemplos de sugerencias de búsqueda ineficientes generadas a partir de búsquedas anteriores

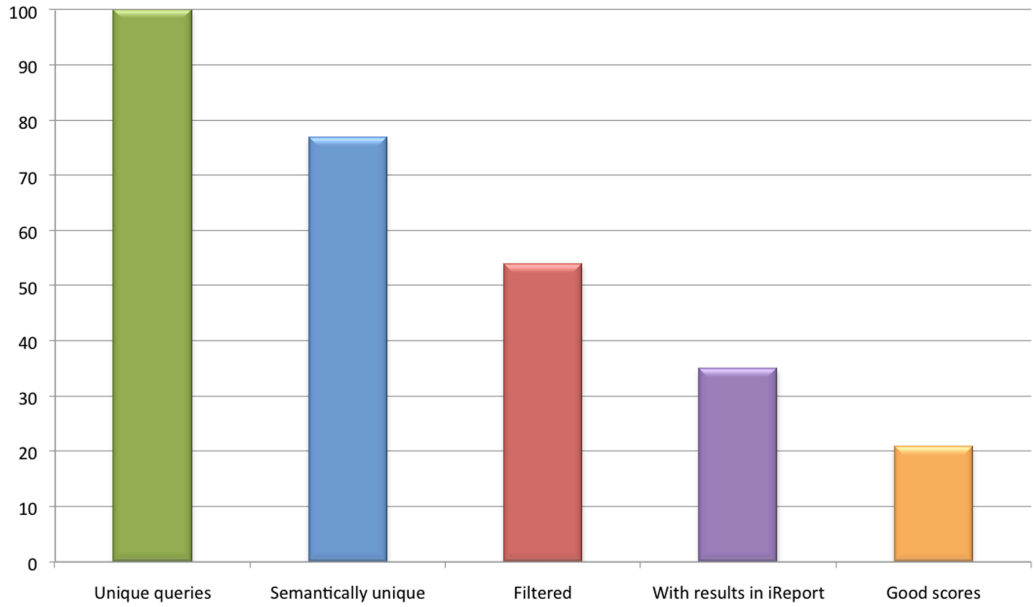

Fig. 4. Porcentaje de consultas restantes en el proceso de elaboración del índice de sugerencias

En consecuencia, para poder ofrecer una adecuada calidad en las sugerencias, es necesario someter al índice de las mismas a un elevado nivel de depuración, que elimine aquellas consultas que no pueden asegurar buenos resultados, así como las repetidas o similares (puede verse un ejemplo en la figura 3).

Para ese trabajo se requieren dos piezas esenciales: la posibilidad de expandir semánticamente el texto de las consultas formuladas (tanto para enriquecer el propio contenido del índice, valorar su calidad e identificar adecuadamente las consultas duplicadas), así como un algoritmo muy preciso, capaz de identificar, con el necesario nivel de fiabilidad y apoyándose en la expansión semántica, las consultas que sí dan buenos resultados.

Como ejemplo de la complejidad del proceso, podemos señalar que, en WKE se comenzó enriqueciendo un fichero inicial compuesto por casi 14 millones de consultas de usuarios con el diccionario de expansión semántica, pues este a priori asegura resultados de calidad (más de 100.000 entradas).

Posteriormente se procedió a eliminar todas aquellas consultas demasiado largas o compuestas solamente por números, pues también sabemos que son aquellas que peores resultados proporcionan.

Seguidamente se realizaron dos procesos sucesivos, uno de asignación genérica de interés a las consultas, contra el total de nuestra base de datos documental (obteniendo un score relativo pero fiable) y, otro, previo a la incorporación del índice de sugerencias a un producto, contra el contenido específico de este, a fin de dejar, en ambos casos solo aquellas consultas que, desde un punto de vista general y también particular de cada producto, aseguraban resultados de buena calidad.

El proceso permitió la sucesiva eliminación de "malas" consultas, en un proceso que gráficamente puede representarse como sigue (figura 4).

De este modo se obtuvo un índice de casi un millón de sugerencias (para un producto de tipo generalista) normalizadas, con alto nivel de calidad y granularidad, que evita en gran medida las sugerencias duplicadas así como las de escasa utilidad.

Paralelamente y sabiendo que un relativamente elevado número de consultas se orientaban a encontrar un documento preciso, y que el algoritmo de relevancia puede identificar la mejor respuesta para una búsqueda de ese tipo, se planteó la opción de sugerir directamente esos documentos inmediatamente debajo de la caja de búsqueda, a modo de una respuesta directa.

Las pruebas de calidad mostraron la eficiencia del sistema y su aplicabilidad a diversos tipos documentales (legislación y artículos doctrinales, principalmente, pero todavía no a la juris-

prudencia, por la dificultad de ofrecer solo una o dos sentencias que respondan a una consulta). Por lo que, dado que los tests realizados con usuarios confirmaron la buena acogida de esta nueva propuesta, en diciembre de 2011 pudo presentarse una nueva funcionalidad de sugerencias, tanto de consultas como de documentos, de alta calidad (figura 5).

Fig. 5. Sugerencia de consultas y documentos en La Ley Digital (2011)

3. La lista de resultados

3.1. Listas únicas y listas agrupadas

Dado que en las bases de datos legales, la modalidad de búsqueda predominante es la búsqueda universal sobre todos los contenidos del producto, a la hora de presentar los resultados pueden seguirse dos modelos diferentes: O bien una lista única con todos los documentos localizados, sin distinción de tipos, o bien una lista de resultados con los documentos agrupados según su tipo (Legislación, Jurisprudencia, Doctrinae, tc.).

La segunda solución presenta una importante ventaja: permite ordenar cada colección documental según su propia relevancia específica. Es decir, no puede compararse la importancia de una sentencia con la de una norma, sino la de una sentencia con otra sentencia (pues, en principio, una sentencia del TS tendrá más importancia que una de un tribunal inferior) y la de una norma con otra norma (pues, en principio, una Ley Orgánica tendrá más importancia que una Orden Ministerial).

3.2. La asignación de relevancia: cuando no basta el modelo TF*IDF

La búsqueda es un problema engañoso. Es fácil implementar un sistema de búsqueda en una base de datos que encuentre documentos, pero es difícil (incluso *muy* difícil) obtener un buen sistema de presentación de resultados según su importancia. Como muy gráficamente han señalado Morville y Callender (2010): una búsqueda es tan buena como lo son los primeros resultados presentados, porque solo los tres primeros reciben un 80% de la atención de los usuarios.

En efecto, como se ha señalado, generalmente los usuarios no pasan de la primera lista de resultados, lo que provoca que la calidad de los primeros documentos mostrados determina la calidad de la búsqueda para el usuario y que unos pocos primeros resultados pueden oscurecer la calidad del resto (Morville y Rosenfeld, 2006)

A ello se añade una cuestión en la que, llamativamente, coinciden los principales desarrolladores o estudiosos de la búsqueda: la dificultad de definir la relevancia de una lista de resultados en abstracto, algo que se refleja tanto en el seminal trabajo de Brin y Page de 1994, como en trabajos más recientes como los de Jackson y Al-Kofahi de 2010.

Paradójicamente, pese a esa dificultad, sí es fácil para los usuarios identificar un documento relevante en una lista de resultados.

El modelo tradicional para asignar importancia a un documento dentro de una colección, tras haber efectuado una búsqueda en virtud de unos términos, es el clásico TF*IDF *(term frequency–inverse document frequency),* que tiene en cuenta tanto el número de veces que el texto buscado aparece en el documento, como en el conjunto de documentos considerados.

Sin embargo, nuestra experiencia evidenció que, en el ámbito legal, es necesario tomar en cuenta otras consideraciones como la proximidad entre los términos, las zonas del documento en que aparecen las ocurrencias y, en especial, la importancia objetiva de cada documento en cada colección considerada (su “autoridad” desde un punto de vista legal).

Como hemos expuesto en Sancho Ferrer et al. (2008a; 2008b), para conseguir buenos resultados es necesario incorporar al algoritmo de búsqueda nuevos factores, tanto en tiempo de indexación (considerando metadatos relevantes y asignando un rango estático a cada documento), como en tiempo de búsqueda (considerando operadores lógicos entre diferentes zonas de los documentos, criterios de distancia entre términos…). La clave para ello está en identificar los mejores fragmentos *(Cluster)*, en las zonas más relevantes de los documentos *(FancyHits),* ponderando la importancia de cada documento

mediante un modelo de valores que representen su "autoridad" desde un punto de vista legal (el *RankModel* del que hablaremos a continuación).

Estas innovaciones en el algoritmo de relevancia pronto evidenciaron la mejora aportada a la calidad de la búsqueda.

3.3. El RankModel

Esta herramienta, desarrollada exclusivamente por WKE, ofrece significativas ventajas en la experiencia de búsqueda, por cuanto, al integrar el análisis de los documentos con los algoritmos de búsqueda, permite optimizar ésta teniendo en cuenta los metadatos asignados a cada documento y su importancia relativa en cada caso.

Simplificando mucho su definición, podemos decir que se trata de una adaptación del modelo de *PageRank* de Google al ámbito legal, aportando criterios objetivos y subjetivos de importancia desde el punto de vista jurídico a los documentos localizados.

El fundamento es que cada metadato asignado a un documento puede influir en la asignación de su relevancia en una búsqueda en un porcentaje y con un valor diferente.

Se trata de un modelo de datos dinámico, en el que los valores pueden modificarse siempre que sea necesario sin afectar a los procesos, y en el que cada producto o índice de contenidos puede tener su propio modelo de valores (figura 6).

```
<Meta name="RANGO" weight="30" default="-5" isMultivaluated="false">
  <Value value="LEY" weight="10" />
  <Value value="LEY ORGANICA" weight="10" />
  <Value value="ORDEN MINISTERIAL" weight="2" />
  <Value value="REAL DECRETO" weight="5" />
  <Value value="REAL DECRETO-LEY" weight="7" />
  <Value value="DECRETO LEGISLATIVO" weight="7" />
  <Value value="REAL DECRETO LEGISLATIVO" weight="7" />
  <Value value="DIRECTIVA CE" weight="9" />
  <Value value="CONVENIO INTERNACIONAL" weight="7" />
  <Value value="TRATADO INTERNACIONAL" weight="7" />
  <Value value="DECRETO" weight="1" />

  <Value value="CORRECCION DE ERRATAS" weight="-9" />
  <Value value="CORRECCION DE ERRORES" weight="-9" />
  <Value value="CORRECCION DE ERRORES Y DE ERRATAS" weight="-9" />
  <Value value="DECRETO FORAL" weight="-2" />
  <Value value="DECRETO FORAL LEGISLATIVO" weight="-2" />
  <Value value="DECRETO FORAL NORMATIVO" weight="-2" />
  <Value value="DECRETO FORAL-NORMA" weight="-6" />
  <Value value="DECRETO FORAL-NORMA DE URGENCIA FISCAL" weight="-6" />
  <Value value="DECRETO FORAL-RESOLUCION" weight="-6" />
  <Value value="LEY FORAL" weight="-1" />
  <Value value="NORMA FORAL" weight="-1" />
  <Value value="ORDEN FORAL" weight="-2" />
  <Value value="RECURSO DE INCONSTITUCIONALIDAD" weight="-8" />

  <Value value="RESOLUCION" weight="-1" />
</Meta>
```

Fig. 6. Rank model de legislación (vista parcial)

3.4. Resúmenes automáticos de los documentos

Presentar una lista de resultados muy precisa no es suficiente. Es necesario facilitar al usuario la rápida comprensión de la relevancia de los documentos ofrecidos, algo no siempre evidente. Tal como se detalló en Sancho Ferrer et al. (2008 a y b), ello es posible hacerlo de dos maneras: (1) mostrando el mejor fragmento de cada documento en la lista de resultados (el denomindo *Keyword in context*); y (2) incorporando sumarios dinámicos o resúmenes automáticos, construidos "al vuelo" sobre cada documento, en función de los términos de búsqueda utilizados.

"Construidos al vuelo" implica que la extracción de dichos fragmentos se realiza: (a) sobre el documento completo, (b) presentando frases completas y (c) mostrando solo aquellos fragmentos que contienen todos los términos de la consulta formulada.

Adicionalmente, estos fragmentos permiten saltar al punto concreto del documento del que provienen (figura 7).

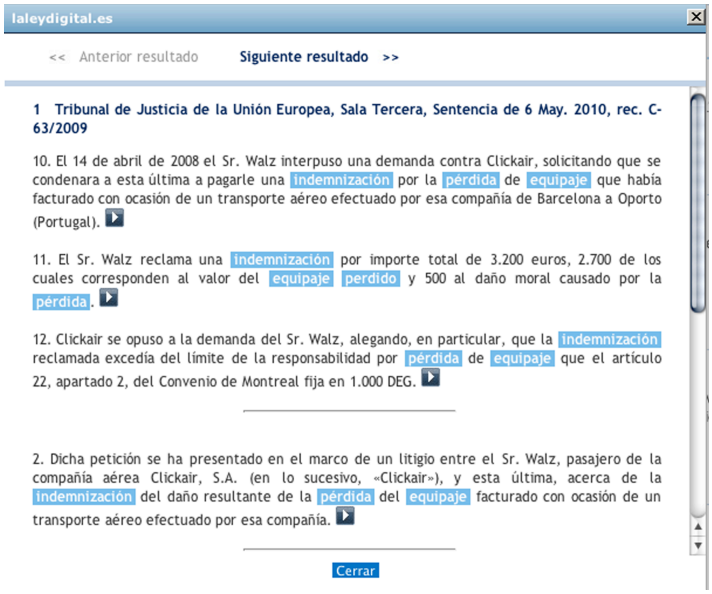

Fig.7. Sumario dinámico de un documento en La Ley Digital para la búsqueda "Indemnización por pérdida de equipaje"

3.5. El problema de las grandes listas de resultados

Una de las consecuencias paradójicas de la mejora de la calidad de búsqueda fue la obtención por los usuarios de grandes listas de resultados, frecuentemente de miles de documentos.

Pese a la calidad de los resultados que puedan contener esas listas y las facilidades para el filtrado de documentos y la reformulación de consultas que se ofrecen a los usuarios, las mismas "imponen" al usuario la carga de dedicar tiempo y trabajo a leer enormes volúmenes de información, con el temor, siempre presente, de haber olvidado algo importante por el camino.

3.6. Solo los mejores resultados

Para resolver esa necesidad se han propuesto dos alternativas: (1) mostrar solo los primeros documentos de la lista de resultados (modelo *Overview* de Westlaw); o, (2) confeccionar una nueva lista de resultados con los mejores documentos localizados, de forma que se presenten primero los mejores resultados, con independencia del orden fijo previsto en la lista de resultados.

Este segundo enfoque, propuesto por los productos de WKE, vuelve a requerir apoyarse en la expansión semántica y en algoritmos de determinación de la relevancia extremadamente precisos, en particular para la identificación y valoración de los fragmentos más significativos para la búsqueda realizada, en relación con el contenido del producto de que se trate.

A la hora de desarrollar esta funcionalidad se hizo evidente:

1. Que el orden de los tipos documentales a mostrar no debía ser estático, como en una lista de resultados clásica, sino que debía configurarse conforme los mejores documentos localizados en cada búsqueda concreta. De esta manera, estos documentos "subirían" posiciones para hacerse más visibles, incluso en aquellos casos en que no se hubiesen buscado expresamente.
2. Que, por ello, debía de poder excluirse de esa nueva lista aquellos tipos documentales que no tuviesen una relevancia suficiente. Incluso, podía llegar a presentarse un documento vacío o casi vacío, cuando solo un número muy reducido de documentos cumpliesen realmente el requisito de calidad fijado.
3. Que el número de documentos a incluir en cada tipo debía de ser necesariamente reducido, para ser rápida y fácilmente analizado por el usuario.
4. Que los fragmentos más destacados de cada documento debían ser resaltados casi a la manera en que lo haría un humano, con un lápiz en la mano, es decir, no simplemente "iluminando" las palabras buscadas.

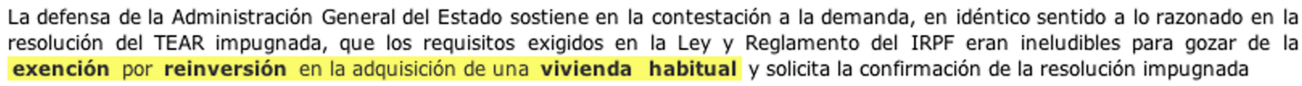

Fig. 8. Ejemplo de identificación de fragmento avanzado para la búsqueda "Exención por reinversión en vivienda habitual"

Fig. 9. Acceso a la funcionalidad "Informe de mejores resultados" en La Ley Digital

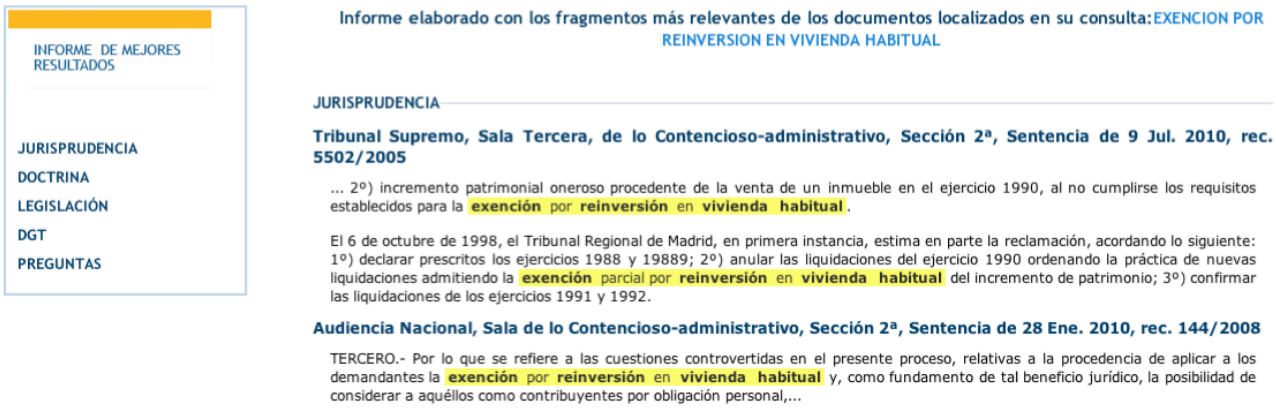

Fig. 10. Contenido de la funcionalidad "Informe de mejores resultados" para la búsqueda "Exención por reinversión en vivienda habitual"

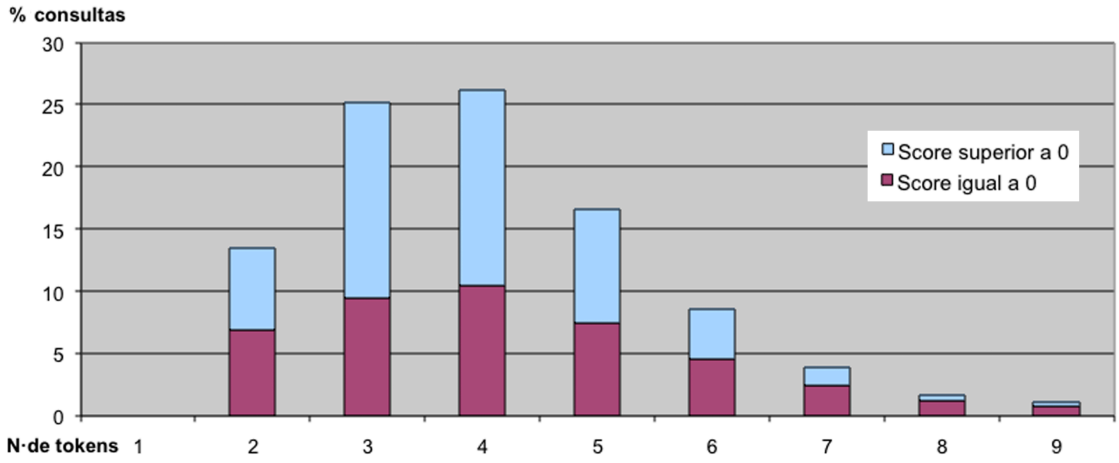

Fig. 11. Distribución de buenos resultados en relación al número de palabras utilizadas en la consulta.

5. Que esa nueva lista de resultados debía ser tan manejable como un documento elaborado por un experto, tener entidad propia y separada de la lista de resultados.

Para establecer esta relevancia, en efecto, no basta con la simple localización de las palabras buscadas en un documento, sino que también es necesario valorar la relación de proximidad existente entre ellas. Es necesario identificar los fragmentos de otra forma (figura 8, en la página anterior).

La consecuencia de este desarrollo ha sido la funcionalidad denominada “Informe de mejores resultados” (originalmente *iReport*) que, accesible junto al comienzo de la lista de resultados presentada tras una búsqueda, ofrece una re-elaboración de dicha lista en el orden y con los documentos estrictamente más pertinentes (figuras 9 y 10, en la página anterior).

4. Nuevos desafíos

El continuo estudio de las búsquedas de los usuarios ha puesto de manifiesto dos datos significativos: (1) Un porcentaje muy significativo de consultas (como mínimo un 50%, según nuestros estudios) no ofrece realmente buenos resultados. (2) La dificultad de obtener buenos resultados aumenta exponencialmente con cada nueva palabra añadida a la consulta (figura 11).

¿Qué consecuencias podemos obtener de estos datos? Sin duda que la búsqueda en bases de datos legales tiene aun un importante camino por delante, lleno de desafíos.

Dado que ningún algoritmo puede resolver una consulta mal formulada, estos desafíos probablemente conducirán a nuevas maneras de orientar la búsqueda de los usuarios, por medio de (a) diálogos, propuestos a través de nuevas técnicas de inteligencia artificial antes y después de su búsqueda, que le permitan precisar con el mayor nivel de detalle su interés concreto, ayudando al motor de búsqueda a realizar su trabajo; y (b) agentes de búsqueda, capaces de construir buenas consultas a partir de una mala.

5. Conclusiones

Los avances obtenidos en tecnologías de búsqueda legal, mediante la aplicación de técnicas de inteligencia artificial, han permitido a los usuarios realizar búsquedas en lenguaje natural, expandibles semánticamente a través de un diccionario de sinónimos y obtener resultados muy precisos gracias a algoritmos de relevancia optimizados para el ámbito legal.

La utilización de metadatos incorporados a los documentos ha permitido mejoras objetivas de calidad para cada colección documental, pues cada tipo documental debe ser valorado conforme a criterios específicos.

Nuevas herramientas desarrolladas en los últimos años han permitido mejorar aún más la calidad de la búsqueda mediante instrumentos como las sugerencias de consultas y documentos y listas de “mejores resultados”, que evitan la consulta de grandes listas documentales.

Con todo, los análisis de los registros de los usuarios siguen evidenciando que más de la mitad de las consultas formuladas a las bases de datos legales, no obtienen buenos resultados.

Los nuevos desafíos que esa evidencia plantean, permiten pensar en funcionalidades aun por desarrollar, pero cuyas mejores posibilidades parecen orientarse a los diálogos abiertos

con los usuarios para concretar el alcance de la búsqueda intentada y en agentes de búsqueda, capaces de sugerir búsquedas alternativas a la formulada o de construir búsquedas enteramente nuevas, basadas en la identificación de la necesidad de información del usuario y de su intención al construir la búsqueda.

Agradecimientos

Los autores quieren agradecer a Raquel Fernández Cestero, responsable de diseño de Wolters Kluwer España, su colaboración en la preparación de los gráficos e imágenes incluidos en este trabajo.

Referencias

Anderson, Chris (2004). The Long Tail. // Wired. (October 2004).

Brin, S.; Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf

Broder, A. (2002). A taxonomy of web search. http://www.sigir.org/forum/F2002/broder.pdf

Jackson, P.; Al-Kofahi, K. (2010). Human expertise and artificial intelligence in legal search. http://labs.thomsonreuters.com/Files/HomepageFeatures/AI%20in%20legal%20search.pdf

Morville, P.; Callender, J. (2010). Search Patterns. Sebastopol (CA): O'Reilly.

Morville, P.; Rosenfeld, L. (2006). Information Architecture for the World Wide Web. 3.ª edición. Sebastopol: O'Reilly.

Sancho Ferrer, Á.; Mateo Rivero, J. M. (2008 b), La búsqueda de información jurídica: problemas habituales y nuevas tecnologías para mejorar la calidad de los resultados, Diario La Ley, 7038 (21 de octubre de 2008), 20-24.

Sancho Ferrer, Á.; Fernández Hernández, C.; Mateo Rivero, J. M. (2011). From Thesaurus Towards Ontologies in Large Legal Databases. // Sartor, G.; et al (eds.). Approaches to Legal Ontologies: Theories, Domains, Methodologies. Heildelberg: Springer, 2011. 179-200.

Sancho Ferrer, Á.; Mateo Rivero, J. M.; Mesas García, A. (2008 a), Improvements in Recall and Precision in Wolters Kluwer Spain Legal Search Engine. // Casanovas, P.; et al. (eds.). Computable Models of the Law: Languages, Dialogues, Games, Ontologies. Heidelberg: Springer, 2008. 130-145.

Wilson, J. (2010). On WestlawNext, State of the Art & Steve Jobs: A Conversation With Peter Jackson, Chief Scientist for Thomson Reuters. http://www.slaw.ca/2010/06/24/a-conversation-with-peter-jackson/

Enviado: 2012-04-10.
Aceptado: 2012-07-05.