tecnim@p Sevilla 2006
30 mayo a 2 junio
060.es
24x7x365

Comunicación

393

TECNOLOGÍAS LINGÜÍSTICAS Y ACCESO A LA INFORMACIÓN: LA EXPERIENCIA DE BITEXT EN LA ADMINISTRACIÓN PÚBLICA

Antonio S. Valderrábanos
Director General
Bitext.com

Josu Gómez Pérez
Director Técnico
Bitext.com

Antonio S. Valderrábanos
Josu Gómez Pérez

Palabras clave

Gestión de documentos, gestión de contenidos, gestión documental, recuperación de la información, búsquedas, buscadores, tecnología lingüística, stemming, truncamiento, acceso a la información, bibliotecas virtuales.

Resumen de su Comunicación

Las instituciones y entidades que se ven obligadas a manejar grandes cantidades de datos escritos (bases de datos documentales, repositorios de ficheros, etc.) se encuentran comúnmente con el problema de conseguir un acceso a esa información rápido y seguro, en el momento en que lo necesitan.

La tecnología lingüística está siendo reconocida como la solución más apropiada para los entornos públicos, en los que el problema del acceso a la información es cada vez más relevante por el actual proceso de digitalización e informatización integral de todos los procesos documentales.

Estas soluciones, implementadas ya en entornos como el INAP, Ministerio de la Presidencia o RENFE, contribuyen a hacer asequible a los usuarios y ciudadanos las tecnologías de búsqueda más avanzadas, poniendo a su alcance toda la información y documentación necesaria con rapidez y seguridad.

Antonio S. Valderrábanos
Josu Gómez Pérez

TECNOLOGÍAS LINGÜÍSTICAS Y ACCESO A LA INFORMACIÓN: LA EXPERIENCIA DE BITEXT EN LA ADMINISTRACIÓN PÚBLICA

1. El problema del acceso a la información

Las instituciones y entidades que se ven obligadas a manejar grandes cantidades de datos escritos (bases de datos documentales, repositorios de ficheros, etc.) se encuentran comúnmente con el problema de conseguir un acceso a esa información rápido y seguro, en el momento en que lo necesitan.

En los ámbitos de aplicación profesional (derecho, farmacia, ingeniería, medicina, etc.), hay dos hechos que hacen preciso el uso de una herramienta de búsqueda que garantice la exhaustividad de las búsquedas:

- la proliferación de bases de datos textuales (de jurisprudencia, legislación, fármacos...) en los más distintos niveles (local, autonómico, nacional, supranacional), y a menudo en diferentes formatos e idiomas.

- la imposibilidad de conocer en detalle todas y cada una de las posibles fuentes de conocimiento para un determinado caso.

La Administración Pública se enfrenta diariamente con este problema. Las tomas de decisiones requieren de la seguridad de contar, en cada momento, con los documentos más relevantes para determinado asunto, de no haberse dejado nada atrás, y de no tener que escoger entre miles de posibles resultados de los cuales sólo unos pocos serán realmente pertinentes para la cuestión tratada.

Las distintas líneas de software dedicado a la recuperación de información intenta ofrecer soluciones para esta problemática, pero se encuentra con una limitación de base: los programas no son capaces de analizar el texto como tal, esto es, como construcciones lingüísticas que expresan un contenido. En esta ponencia trataremos los problemas prácticos que surgen de esta limitación, y la solución que ha dado Bitext en varios proyectos llevados a cabo en los últimos años en la Administración Pública.

2. Soluciones basadas en palabras claves

Los sistemas de gestión de bases de datos documentales más utilizados en la actualidad funcionan de la siguiente manera:

- **Paso 1.** Un usuario introduce una determinada consulta que refleja el asunto sobre el que precisa información. Por ejemplo: “penas condonadas”.

- **Paso 2.** El sistema de gestión documental buscará en la base de datos correspondiente la aparición de las palabras “penas” y “condonadas”.

- **Paso 3.** El sistema de gestión documental proporcionará al usuario los documentos encontrados.

Este procedimiento, sencillo y de fácil implantación, está generalizado ya para una gran mayoría de los sistemas de búsqueda. Pero sus limitaciones son grandes, presentando dos importantes problemas que reducen considerablemente su fiabilidad: el “silencio” y el “ruido”.

3. El problema del "silencio"

Si existen documentos en los que aparecen las palabras "pena" o "condonada", el sistema de gestión documental no los localizará y, por tanto, no los mostrará al usuario. Sin embargo, estos documentos sí son relevantes para el usuario. Este problema se conoce generalmente como "silencio" o cobertura incompleta.

Dado que ésta es una situación frecuente, los distintos sistemas de gestión documental que se comercializan para profesionales han puesto en marcha distintas alternativas. Podemos describir el funcionamiento de estas alternativas de la siguiente manera.

Dada nuestra consulta anterior, "penas condonadas", el sistema de gestión documental suprimirá un número variable de caracteres en cada palabra. Así, la consulta del usuario se reformulará de la siguiente manera.

penas condonadas > PEN CONDONAD

A continuación, el sistema de gestión documental buscará palabras que contengan la consulta en su nueva forma (PEN CONDONAD) y buscará la cadenas PEN y CONDONAD seguidas de cualquier carácter (a, as, s, o, etc).

Efectivamente, esta reformulación de la consulta permitirá al sistema de gestión documental localizar documentos en los que aparecen las palabras

PENa
CONDONADa
PENas
CONDONADas

Así, podría parecer que el problema del silencio queda resuelto satisfactoriamente. Sin embargo, esta solución genera un segundo problema: el ruido.

4. El problema del "ruido"

El problema del ruido se produce porque, con la consulta reformulada (PEN CONDONAD), el sistema de gestión documental encontrará no sólo documentos en los que aparecen las palabras anteriores:

PENa
CONDONADa
PENas
CONDONADas

sino también documentos en los que aparecen las cadenas

PENalidad
PENacho
PENalti
PENal
PENe

Dado que no existe relación alguna de contenido entre, por ejemplo, PENa y PENacho, el usuario obtendrá como resultado de su búsqueda documentos en los que no está interesado. Éste es el problema que se conoce generalmente como "ruido".

5. Las causas del problema

La causa de ambos problemas, tanto del silencio como del ruido, está en que los sistemas de gestión documental tratan la consulta del usuario de una manera mecanicista, atendiendo únicamente a su forma e ignorando su contenido. Como consecuencia, establecen relaciones de contenido entre palabras que simplemente comparten una serie de caracteres.

Esta situación proviene del hecho de haber sido creado en entornos angloparlantes, donde, por las características del idioma más utilizado (inglés), presenta un grado de fiabilidad aceptable. Esta situación, sin embargo, no es trasladable a otros entornos donde el idioma mayoritario de uso es otro, como, por ejemplo, el español. En estos entornos, la aplicación de este sistema de "truncamiento" presenta más problemas que beneficios, y hace patente la necesidad de afrontar esta problemática desde presupuestos adaptados a la realidad española, sólo asequibles desde un profundo conocimiento de las peculiaridades del idioma.

6. La solución: tecnología lingüística (1-DataLexica)

El principal campo de aplicación para DataLexica es la gestión de grandes bases de datos documentales y, más específicamente, la búsqueda y recuperación de la información en este tipo de bases de datos.
La principal ventaja que DataLexica aporta a los sistemas existentes de gestión documental es que garantiza la exhaustividad y exactitud de la búsquedas, es decir, garantiza que el usuario recuperará todos y cada uno de los documentos relacionados con el asunto sobre el que busca.

DataLexica trata las palabras de acuerdo con su contenido, no con su forma, de la misma manera que lo hace un usuario. DataLexica es una base de conocimiento lingüístico que contiene información acerca de más de 3.000.000 palabras y sus relaciones.

Así, DataLexica puede establecer relaciones entre palabras como

"condonó", "condonada", "condonaron"
"ser", "era", "siendo"
"régimen", "regímenes"

y, de la misma manera, puede determinar que no existe relación alguna entre

"pena", "penacho" y "penalidad"
"condonado" y "condenado"
"caso" y "casa"

De esta forma, DataLexica garantiza que el sistema de gestión documental recuperará todos los documentos que son relevantes para el usuario y sólo éstos, excluyendo los no relevantes. Así, se recuperarán documentos en los que aparezcan frases como:

"penas condonadas"
"se condonó la pena"
"condonaron la pena"
etc.

En otras palabras: DataLexica resuelve tanto el problema del ruido como el del silencio porque expande la consulta del usuario de acuerdo con su contenido.

7. La solución: tecnología lingüística (2-DataGrammar)

DataGrammar es un componente que, junto con DataLexica, permite el análisis inteligente del texto.

DataGrammar permite que los usuarios puedan interactuar con las aplicaciones que gestionan texto (típicamente los buscadores) en lenguaje natural, convirtiendo al buscador en un documentalista. El análisis inteligente del texto identifica las expresiones relevantes de las consultas y las expande convenientemente, evitando al usuario el uso de comodines u operadores.

El usuario, con DataGrammar, dispone de un interfaz en lenguaje natural, al que puede interrogar con consultas como:

- Necesito documentación sobre planes de pensiones
- Dame los textos que hablen de jubilaciones anticipadas
- Quiero la información que haya sobre parques temáticos pero que no sean de Disney

8. Casos prácticos: Bitext y la Administración Pública

La tecnología lingüística está siendo reconocida como la solución más apropiada para los entornos públicos, en los que el problema del acceso a la información es cada vez más relevante por el actual proceso de digitalización e informatización integral de todos los procesos documentales.

Una de las instituciones que viven de forma más acuciante esta problemática es el Instituto Nacional de Administraciones Públicas (INAP). Para cumplir sus objetivos de promover y facilitar la investigación en materias relacionadas con la Administración, y favorecer la formación de los funcionarios, le es indispensable disponer de un acceso diáfano y seguro a una gran cantidad de fuentes de información diversas y heterogéneas.

Las tecnologías de Bitext, integradas en una solución de "Biblioteca Virtual", facilitan a todos los usuarios (externos e internos al Instituto) la consulta, recuperación y estudio de todas las fuentes de documentación digitalizadas, tanto internas (publicaciones periódicas, monografías y catálogos) como externas (bases de datos o páginas web). La expansión lingüística asegura que la recuperación sea completa, a la vez que sencilla.

El Ministerio de la Presidencia, en concreto la Dirección General de Relaciones con las Cortes, es otro de los ámbitos donde se han implementado soluciones de búsqueda con tecnología lingüística. Por este medio, múltiples instancias de la Administración se han beneficiado de instrumentos de apoyo a la toma de decisión que han permitido llevar a cabo decisiones informadas, evitando así los riesgos de "vacíos informativos". Los usuarios de este sistema han podido mantenerse al día de las fuentes de información más relevantes, de forma sencilla, automática, y adaptada a sus necesidades.

Otros proyectos, como el interfaz de lenguaje natural desarrollado por primera vez en RENFE y presentado actualmente al Boletín Oficial del Estado, amplían las posibilidades de acceso a la información para una amplia gama de usuarios, colocan las potencialidades de las búsquedas avanzadas al alcance de personas que no desean aprender los intrincados mecanismos que suelen ser necesarios para utilizarlas, y, en definitiva, contribuyen a acercar a ciudadanos y usuarios la documentación que precisan en el ámbito de la Administración.