Seda: un motor de búsqueda para las colecciones (XML) del Parlamento de Andalucía

Luis M. de Campos[1], Alberto Ching[1], Juan M. Fernández-Luna[1], Juan F. Huete[1], Antonio Tagua-Jiménez[2], Carmen Tur-Vigil[2], and Eduardo Vicente-López[1]

[1]Departamento de Ciencias de la Computación e Inteligencia Artificial, ETSIIT, CITIC-UGR, Universidad de Granada, 18071, Granada, España
[2]Servicio de Publicaciones Oficiales del Parlamento de Andalucía, Sevilla, España
[1]{lci,alch,jmfluna,jhg,evicente}@decsai.ugr.es, [2]{aj.tagua,mc.tur}@parlamento-and.es

Keywords: Recuperación de Información Estructurada, XML, Colecciones Parlamentarias.

Resumen En este trabajo presentamos *Seda*, una aplicación de búsqueda para las colecciones documentales del Parlamento de Andalucía. Está basada en un campo de la Recuperación de Información denominado Recuperación de Información Estructurada, el cual permite tratar los documentos, no como unidades atómicas e indivisibles, sino como una composición de partes que pueden ser susceptibles de ser recuperadas separadamente. La colección documental del Parlamento de Andalucía está formada principalmente por Diarios de Sesiones y Boletines Oficiales. Estos documentos tienen una organización interna claramente definida y son adecuados para poder operar con ellos desde esta perspectiva. *Seda* permite el acceso a dichos documentos de una forma eficaz y amigable, con funcionalidades novedosas aprovechando estas peculiaridades.

1. Motivación e introducción

Ya en el siglo XXI, el avance tecnológico que se está produciendo en todos los campos de nuestra sociedad hace que el uso de los ordenadores esté integrado en nuestra vida diaria. Sin casi darnos cuenta los empleamos para cualquier tarea. Y no sólo utilizamos frecuentemente los ordenadores, digamos de sobremesa o portátiles, sino que los dispositivos móviles están teniendo un crecimiento casi exponencial, llegando a emplearlos más incluso que los primeros. Podríamos decir que los tenemos integrados en nuestro día a día para tareas tan variadas y diferentes como comunicarnos, en sus diferentes formas, acceder a nuestros bancos o divertirnos, entre otras muchas posibilidades.

Las empresas y organizaciones también están cambiando radicalmente su forma, tanto de relacionarse con el exterior, como de gestionar sus propios procesos internos, y se están volcando en las TIC para mejorar su productividad y ofrecer servicios de más calidad, e inéditos hasta hace poco tiempo.

Las administraciones públicas no pueden quedarse atrás en esta evolución, a pesar de los tiempos que corren, y deben permitir que el ciudadano acceda a sus servicios de forma telemática, acercando así la Administración al contribuyente, simplificar trámites que hasta ahora han sido largos y tediosos, y facilitar el acceso a la información. Todo esto se desarrolla en el horizonte contemplado por la Ley de acceso electrónico de los ciudadanos a los servicios públicos.

Un caso particular lo tenemos en los parlamentos nacionales y autonómicos, en donde sus sesiones parlamentarias quedan transcritas para posterior consulta general, tanto de los propios diputados, como de cualquier persona que desee saber qué se ha discutido en ellas. Esta información, por tanto, debe estar disponible para el ciudadano. Inicialmente dichos diarios, que así se denominan a los documentos que contienen esas transcripciones, se imprimían en papel y se distribuían a varios organismos para su consulta. Con la expansión de la web, los diarios pasan a publicarse también electrónicamente en los sitios oficiales de los parlamentos, llegando el momento en que las copias impresas desaparecen.

Así, es necesario ofrecer servicios que permitan, no sólo obtener electrónicamente un diario de sesiones por su fecha o por su número, sino también teniendo en cuenta su contenido. Y es aquí donde entran en juego los motores de búsqueda, ya que estos permiten que el usuario establezca una serie de términos de búsqueda, que reflejan su necesidad de información, y el buscador le devuelva aquellos diarios que hablen del tema en el que el usuario está interesado.

Actualmente, todas las cámaras españolas ofrecen estos servicios, de una u otra forma, a los ciudadanos que estén interesados en acceder a sus colecciones documentales. Pero si consideramos que dichos diarios de sesiones tienen una estructura interna muy definida, es decir, su contenido está muy bien organizado, tendría sentido pensar que si un usuario está interesado en lo que han hablado los parlamentarios sobre el tema concreto de la Ley Orgánica para la Mejora de la Calidad Educativa, por ejemplo, el sistema de recuperación de información pudiera devolverle no un documento completo, sino la parte del mismo que trata sobre la discusión de la LOMCE. Este documento podría contener otras partes que pueden tratar sobre la Ley Orgánica relativa a la Justicia Universal, o sobre las preguntas al Presidente versando sobre la Política Agrícola Común de la UE, por citar otras temáticas totalmente diferentes, y que podrían ser discutidas en un mismo pleno, pero estas partes no serían del interés del usuario. De esta forma, se le presenta al usuario sólo el material relevante, ahorrándole el tiempo de buscar lo que verdaderamente le interesa en el documento completo devuelto.

Y así surge *Seda*, un motor de búsqueda para las colecciones documentales del Parlamento de Andalucía, basado en el concepto de recuperación centrada en la parte del documento que es más relevante (concepto conocido como recuperación estructurada ó XML [11]). Nace como un ejemplo de transferencia tecnológica de las técnicas de Recuperación de Información (RI) estructurada que ha desarrollado el grupo *Tratamiento de la Incertidumbre en Inteligencia Artificial (TIC103)*, bajo el amparo de dos proyectos de investigación de la Junta de Andalucía, puestos en práctica en un problema real, con la estrecha colaboración del personal del Servicio de Publicaciones Oficiales de esta cámara autonómica.

Este trabajo pretende ofrecer una visión general de *Seda* y mostrar sus principales prestaciones y funcionalidades, que lo hacen diferente de la mayoría de los sistemas de RI implantados en los distintos parlamentos nacionales, y por qué no decirlo, en muchos otros fuera de nuestras fronteras. Y para tal fin, lo estructuramos de la siguiente forma: la sección 2 tiene como objeto presentar la colección documental con la que *Seda* trabaja, tanto en tipos de documentos, como a nivel de estructura interna. La sección 3 muestra las principales funcionalidades del buscador, ofreciendo una idea general de los tipos de búsqueda y opciones disponibles para el usuario. Un breve repaso de los modelos de recuperación subyacentes se hace en la sección 4. El trabajo finaliza con una revisión de algunas propuestas para el acceso a colecciones parlamentarias (sección 5) y con una exposición de las conclusiones y trabajos futuros (sección 6).

2. La colección documental del Parlamento de Andalucía

Antes de comenzar por la descripción de la colección documental, que será la fuente del sistema de RI, debemos introducir el concepto principal sobre el que gira la vida parlamentaria: la iniciativa. Básicamente es una propuesta realizada por un diputado o un grupo parlamentario que implica una acción parlamentaria como consecuencia. Por ejemplo, la discusión de una ley o solicitud de comparecencia del presidente o un consejero en un pleno, entre otros tipos.

Como ya hemos indicado anteriormente, el Parlamento de Andalucía publica dos tipos de documentos oficiales: por un lado, el Diario de Sesiones, que es una transcripción literal de los debates que se producen en cada sesión parlamentaria, pero organizado en torno al concepto de orden del día, es decir, los puntos que se van a tratar en la sesión, o mejor dicho, el conjunto de iniciativas sobre las que se discutirá en ella. Este documento puede ser de dos tipos: el que se refiere a una sesión plenaria, con todos los diputados participando en ella, comúnmente llamado pleno, o el asociado a una comisión, centrada en un ámbito concreto (agricultura, economía, etc.), en la que sólo participan los diputados adscritos a dicha comisión. Por otro lado, el Boletín Oficial del Parlamento de Andalucía es el segundo tipo de documento de la colección, registrándose el ciclo de vida completo que sigue una iniciativa desde que se presenta.

Los diarios de sesiones, tanto de plenos como de comisiones, tienen una estructura interna común: una primera parte en donde aparecen datos sobre el propio diario, como el número, la fecha, el tipo (ordinario, extraordinario,...), y una segunda, que contiene el desarrollo en sí. En este caso, y a grandes rasgos, éste está conformado en torno a las iniciativas, las cuales quedan compuestas por una descripción (extracto) y su correspondiente identificador (número de expediente), una serie de materias que han sido asignadas por documentalistas y que clasifican el contenido de la iniciativa según el tesauro EUROVOC [7], y seguidamente una sucesión de intervenciones, cada una compuesta por el nombre del diputado que la protagoniza y el texto de la transcripción de su discurso (agrupado en párrafos). Por otro lado, los boletines oficiales están compuestos por una primera parte con datos de identificación y un conjunto de secciones que agrupan iniciativas, que, a su vez, se componen de número de expediente y extracto, el estado en que se encuentra y su desarrollo o fase de tramitación.

Esta estructura que acabamos de comentar es muy adecuada para que un sistema de RI, que indexe los documentos y los ponga a disposición de los usuarios a partir de una consulta, pueda ofrecer partes de documentos en lugar de documentos completos. ¿Por qué devolver un diario de sesiones completo si lo que realmente le interesa a un usuario es una intervención de un diputado sobre un tema concreto? ¿O un boletín entero cuando lo relevante sólo es una iniciativa?

El formato de publicación de las colecciones documentales del Parlamento de Andalucía es el PDF, proveniente de un documento en formato Microsoft Word y posteriormente editado en InDesign. Aunque los indexadores puedan extraer el texto plano contenido en los ficheros PDF, estos no son los más apropiados para recuperar partes bien definidas, ya que la estructura interna del documento está implícita en el propio texto, totalmente transparente, pues estos módulos sólo ven texto, y no están dotados de elementos que establezcan explícitamente las diferentes porciones que los componen. Así, es necesario que los documentos estén marcados en XML, haciéndose evidente su estructura y permitiendo ser manipulados por el indexador convenientemente, extrayendo cada una de sus partes y haciendo posible su indexación separada, dejando todo listo para su posterior recuperación.

En el caso que nos ocupa, los documentos de la colección del Parlamento se transforman de forma automática a partir de los ficheros en Adobe InDesign o PDF a XML bajo unos DTD[1], los cuales definen de forma detallada y completa todas las partes de que constan los diarios y los boletines, habiendo sido definidos conjuntamente con el personal del Servicio de Publicaciones del Parlamento. Cada elemento XML que compone cada documento, denominado unidad, es conceptualmente susceptible de ser recuperado si es relevante a una consulta.

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	Total
Documentos	82	138	146	232	522	494	671	748	512	3.545
Iniciativas	285	980	1.309	1.254	3.510	3.506	4.754	5.275	12.397	33.270
Intervenciones	11.142	18.840	22.338	18.363	48.507	44.445	53.893	58.709	19.825	296.062
Párrafos	37.203	77.196	97.149	110.329	115.557	76.616	134.669	136.209	42.974	827.902

Cuadro 1. Número de documentos y unidades principales por legislatura.

En la tabla 2 mostramos el número de documentos que hay hasta la fecha de escritura de este trabajo, organizados en las nueve legislaturas de vida de esta comunidad autónoma. Como se puede observar, aunque la colección en términos absolutos no es muy grande, el número de partes principales (iniciativas, intervenciones y párrafos) que debe gestionar es muy alto.

3. Funcionalidad de Seda

Las prestaciones generales que ofrece *Seda* al usuario interesado en consultar la colección documental del Parlamento de Andalucía son las siguientes:

- Búsqueda por texto libre sobre un conjunto heterogéneo de documentos, donde el propio motor de búsqueda decide qué partes de los documentos son las

[1] Document Type Definition – ficheros que contienen las reglas que establecen las partes que componen la estructura interna de los documentos XML.

idóneas para ser devueltas al usuario según la consulta o, alternativamente, el usuario puede establecer un tipo preferido.
- Filtrado de resultados por legislatura, tipos de documentos, rango de documentos por su número y fechas de publicación.
- Opciones de presentación de resultados, indicando la forma de mostrar los resultados aisladamente o en el contexto de los documentos donde están incluidos y número de resultado máximo, entre otras.
- Visualización de los documentos XML obtenidos como resultado de una consulta, permitiendo navegar por las distintas partes relevantes de un documento, así como consulta del correspondiente documento PDF.
- Búsqueda avanzada de material relevante mediante consultas estructuradas, en dónde el usuario puede decidir dónde buscar, qué buscar y qué partes del documento quiere obtener como respuesta a su necesidad de información.
- Uso en la búsqueda avanzada de las materias por las cuales las iniciativas han sido indexadas y sugerencias de éstas a tenor de los términos de búsqueda empleados para formular consultas.
- Ayuda que permite en cada momento saber cómo emplear el buscador.

En general, podemos decir que *Seda* ofrece un interfaz de usuario amigable y usable, que permite al usuario manejar de forma intuitiva el buscador sin necesidad de conocimientos previos. Además, ha sido diseñado también para permitir una visualización e interacción adecuada en dispositivos móviles. *Seda* ha sido evaluado positivamente, tanto por personal del Parlamento de Andalucía, como por usuarios no especializados en estudios de usuario. Su uso se encuentra abierto a cualquier persona en la dirección `http://irutai2.ugr.es/SEDA`. En las secciones siguientes vamos a describir más detalladamente las funcionalidades de *Seda* esbozadas anteriormente.

3.1. Búsqueda básica

Denominamos búsqueda básica a aquella que el usuario realiza mediante la introducción de una consulta de texto en el campo de búsqueda. En este caso, la salida del buscador podrá ser cualquier unidad del documento de las consideradas como principales (iniciativas, intervenciones o párrafos). La interfaz de usuario para tal fin se muestra en la Figura 1. En ella, además del campo de texto para incluir la consulta, podemos observar varios filtros que permiten al usuario restringir la salida según sus criterios. Así las cosas, el usuario podrá requerir documentos que fueran creados en una o varias legislaturas, desde la I a la IX, ésta última en la que nos encontramos actualmente; el tipo de documento que desea (puede elegir de forma múltiple entre diarios de sesiones de plenos, de comisiones, diputaciones permanentes[2] y boletines oficiales. En caso de que los diarios de comisiones hayan sido seleccionados, el sistema le da la posibilidad al usuario de elegir también una o varias comisiones. Finalmente, las búsquedas podrán quedar focalizadas en rangos de fechas o números de diarios o boletines.

[2] Una diputación permanente es el órgano de la cámara autonómica que se reúne en momentos en los que se encuentra cerrado el periodo de sesiones del Parlamento. Su correspondiente diario tiene la misma estructura que el de plenos y comisiones.

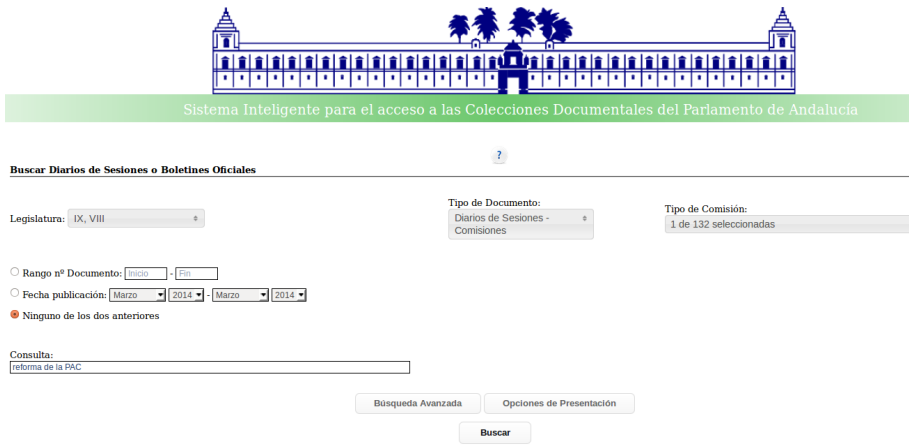

Figura 1. Interfaz para realizar la consulta "Reforma de la PAC", con resultados restringidos a la VIII y IX legislaturas pero sólo de comisiones de agricultura.

Una vez que el usuario ha expresado su consulta, el sistema de RI calcula el grado de relevancia de cada unidad, que compone un documento incluido en el índice del motor de búsqueda. El resultado de dicho proceso se muestra al usuario tal y como aparece en la Figura 2.

En ella, podemos observar que las unidades recuperadas, en este caso 200, número que puede establecerse a voluntad del usuario, se distribuyen en páginas con 10 resultados cada una, por las que se puede navegar. En cada una de ellas se puede observar, agrupados por documentos, las unidades recuperadas. Así, en el ejemplo de la figura, se devuelve, en primer lugar, el debate agrupado de tres iniciativas (con números de expediente 8-08/APC-000024, 8-08/APC-000036 y 8-08/APC-000093) perteneciente al diario de sesiones de comisiones nº 29 de 21 de mayo de 2008. Sin embargo, los siguientes resultados pertenecen todos ellos al diario de sesión de la comisión nº 120 de 20 de marzo de 2013. Esta forma de presentar las unidades relevantes agrupadas según el documento al que pertenecen se conoce técnicamente, según la terminología de INEX[3], como *all in context*, la cual procesa la lista de unidades relevantes agrupándolas por documentos, creando finalmente una lista de éstos, en cuyo "interior" aparecen las unidades relevantes. De esta forma, al usuario se le muestran las unidades dentro del contexto del documento en el que aparecen y no aisladamente. En *Seda*, además de esta última forma denominada "Todos los resultados agrupados por documento", existen otros dos estilos de presentar las unidades recuperadas: "Todos los resultados" (*focused retrieval* en nomenclatura de INEX), que muestra la lista de unidades relevantes tal cual la devuelve el sistema de RI, eliminando el solapamiento que exista entre ellas, y "Sólo un resultado por documento" (best in context), que presenta lo que se conoce como el mejor punto de entrada

[3] *Initiative for Evaluation of XML* [9] – Iniciativa internacional que tiene como objetivo el desarrollo y evaluación de sistemas de recuperación XML.

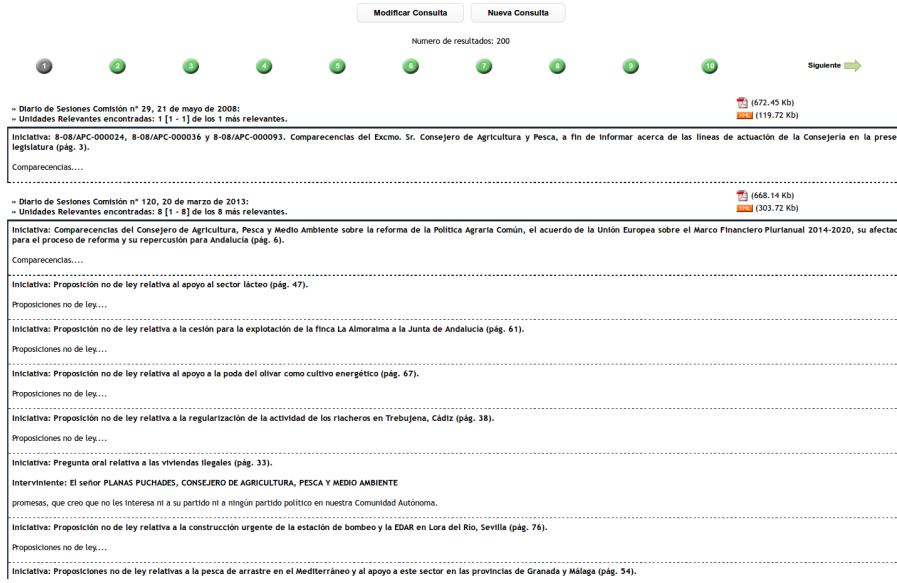

Figura 2. Resultados obtenidos tras efectuar una consulta simple.

al documento, es decir, una unidad relevante a partir de la cual se recomienda la lectura del documento. El usuario podrá elegir qué técnica de presentación desea mediante la selección de la opción correspondiente desplegada tras pulsar el botón "Opciones de Presentación" de la página principal.

Para acceder al contenido del documento, tendríamos dos alternativas: hacer clic en el icono del documento PDF situado en la parte superior del recuadro que agrupa los resultados relevantes de un documento, en cuyo caso se abriría el documento oficial en este formato, o hacer lo propio sobre el pequeño rectángulo naranja etiquetado con XML, momento en el que se abre una nueva ventana dividida en dos partes (véase Figura 3): la parte izquierda presenta los extractos de las unidades relevantes del documento; la derecha, el documento completo en su versión XML, aunque formateado para su correcta visualización (en verde cada una de las unidades relevantes). Al hacer clic en cualquier unidad relevante de la izquierda, en la zona principal se mostrará el comienzo de dicha unidad. De esta manera se presenta un método sencillo de navegación a través de las unidades relevantes de un documento, permitiendo al usuario que pueda moverse por él para encontrar el contexto que necesite de forma rápida y sencilla.

Volviendo de nuevo a las opciones de la búsqueda que se encuentran al presionar el botón de "Opciones de Presentación", comentar, por último, que existe la posibilidad de establecer si se desea, que el buscador exija la aparición de todos los términos de la consulta en los resultados que va a devolver al usuario (al estilo de un *AND* estricto), o por el contrario, que pueda relajarse en dicha restricción, permitiendo que algún término pueda no estar (salvando las distancias, y con objeto de que se entienda, que pudiera parecerse más a una consulta tipo *OR*).

Esto equivale, tal y como está expresado en el interfaz de usuario, a un nivel de exigencia de aparición de todos los términos de búsqueda en los resultados muy alta o mínima, respectivamente (con dos niveles intermedios adicionales).

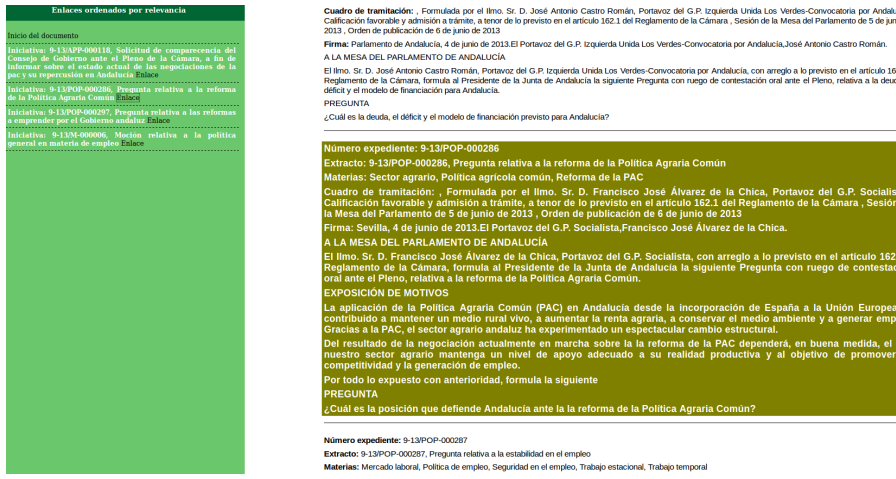

Figura 3. Visualización de las unidades relevantes en el documento al que pertenecen.

3.2. Búsqueda avanzada

La búsqueda avanzada de *Seda* está basada en el concepto de consulta estructurada, consulta CAS (Content and Structure) según INEX, la cual aprovechando la estructura interna explícita de los documentos, permite establecer restricciones estructurales sobre qué se quiere recuperar y dónde buscar, además del contenido textual en sí (véase [3] para conocer la tecnología de resolución de consultas CAS que implementa *Seda*). La formulación de la consulta CAS se apoya internamente en el lenguaje de consulta NEXI [14], el cual establece su sintaxis y semántica, que a groso modo, son los siguientes:

//A[about(.,términos de consulta 1)//B[about(.,términos de consulta 2)]

En esta consulta se desea recuperar unidades de tipo B (objetivo de la recuperación) que traten sobre los términos de la consulta 2, dentro de unidades de tipo A (contexto) que versen sobre los términos de la consulta 1.

Es evidente que a un usuario no especializado no se le puede requerir que conozca, ni la sintaxis de NEXI ni la estructura interna de los documentos a consultar para poder formular consultas CAS en *Seda*, por lo que hay que buscar una forma alternativa para tal fin. Una de estas formas sería el uso del lenguaje natural, pero no es una solución fácil de implementar. Por tanto, la opción más sencilla es establecer un mecanismo gráfico intuitivo, por el que el usuario pueda

expresar una consulta CAS simplemente rellenando campos de texto o seleccionando alternativas, y así expresar qué partes de los diarios y BOPA quiere obtener y en qué unidades se debe buscar.

Así, bajo el botón "Búsqueda Avanzada" se despliegan varios componentes gráficos que permiten esta tarea. Concretamente se observan dos partes diferenciadas: la primera, denominada "Objetivo", tiene como fin seleccionar la unidad deseada como objetivo de la recuperación del desplegable "Dónde buscar" y establecer la consulta textual sobre ella en el campo de texto "Texto a buscar"; la segunda, etiquetada como "Dar preferencia a los elementos en el contexto", establece con los mismos métodos las restricciones de contexto. El usuario puede introducir tantos elementos del contexto como desee, que serían cláusulas *about* en la parte de la consulta relacionada con el contexto (ver Figura 4).

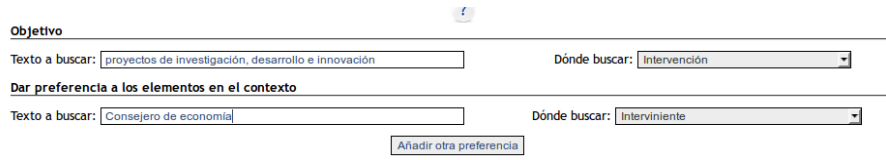

Figura 4. Ejemplo gráfico de una consulta CAS donde se buscan aquellas intervenciones que traten sobre proyectos de investigación, desarrollo e innovación (objetivo) cuyo interviniente sea el Consejero de Economía (contexto).

3.3. Incorporación de materias en el buscador

Una opción interesante de la búsqueda avanzada es el hecho de poder emplear las materias del tesauro EUROVOC, que han asignado manualmente los documentalistas del Parlamento de Andalucía a cada iniciativa. Así, por ejemplo, podríamos buscar párrafos de intervenciones que traten sobre cursos de formación ocupacional sólo en iniciativas que hayan sido indexadas bajo la materia "`acceso al empleo`". El problema evidente de esta aproximación es el hecho de que el usuario no tiene por qué conocer las materias del tesauro que han sido asignadas a las iniciativas ni tampoco las que puede emplear. Así, se ha incluido un módulo que, a petición del usuario, le sugiere materias a partir del texto de la consulta del objetivo y que las añade como una preferencia adicional a la parte del contexto de la consulta CAS. La tecnología subyacente de este módulo es un clasificador de materias que, de forma automática, sugiere las más adecuadas según los términos de la consulta, ordenándolas según su grado de relevancia y las dispone en un cuadro de selección para que se elijan las oportunas [4].

3.4. Diferencias con el buscador oficial del Parlamento de Andalucía

En esta sección vamos a establecer las diferencias existentes entre el buscador oficial del Parlamento de Andalucía, al que denominaremos BPA, y *Seda*, con objeto de mostrar las mejoras y avances tecnológicos que incorpora *Seda*, y que lo convierten en un buscador muy potente y versátil, para las colecciones documentales del Parlamento de Andalucía:

- Sobre la formulación de la consulta:
  - En el buscador del Parlamento se tiene que elegir en qué tipo de documentos se va a realizar la búsqueda, mientras que *Seda* puede manejar y buscar simultáneamente en las diferentes subcolecciones documentales (Diarios de Sesiones, Comisiones, BOPAs...), o en una concreta. Dentro de las Comisiones, se puede especificar también una combinación en concreto o todas.
  - El BPA sólo tiene la capacidad de buscar en una única legislatura, mientras que *Seda* puede hacerlo en varias a la vez especificadas previamente por el usuario.
- Sobre la búsqueda:
  - *Seda* hace búsquedas aproximadas, de forma muy eficiente, basadas en tecnologías de Inteligencia Artificial. El motor del Parlamento lleva a cabo la recuperación mediante un modelo booleano, con su correspondiente emparejamiento exacto.
  - *Seda* permite efectuar consultas por contenido y estructura (CAS), mientras que la aplicación BPA sólo por contenido.
- Sobre los resultados de la búsqueda:
  - *Seda* devuelve los resultados de la búsqueda ordenados por orden de relevancia, al contrario que el BPA, que devuelve y muestra todos los documentos que emparejan con la consulta, sin priorizar.
  - El motor BPA devuelve los documentos completos, y el usuario tiene después que buscar en ellos las partes relevantes. Sin embargo, *Seda* devuelve, según se desee, aquellas partes de los documentos que son relevantes, o los documentos completos, indicando cuáles son las partes relevantes de éstos.
  - *Seda* muestra documentos en formato PDF y XML, resaltando en este último las partes relevantes devueltas por el sistema. Por el contrario, el buscador oficial de la cámara autonómica sólo ofrece PDF completos.

4. La tecnología subyacente: redes bayesianas y diagramas de influencia

Seda es un interfaz de usuario que permite interactuar con el verdadero motor de búsqueda, con el que se comunica, encargándose éste de llevar a cabo la recuperación de las unidades relevantes propiamente dichas. Este software recibe el nombre de *Garnata* [2] e implementa un modelo de RI, el modelo CID [1], basado en modelos gráficos probabilísticos, más concretamente, en redes bayesianas y diagramas de influencia [10]. Este modelo obtiene el grado de relevancia de cada parte del documento mediante la combinación de dos dimensiones diferentes: la especificidad y la exhaustividad de una unidad con respecto a la consulta. El primer concepto se basa en el hecho de que cuantos más términos de la unidad aparezcan en la consulta, más relevante será el elemento (tratará de forma más clara sobre los términos de la consulta). Esta dimensión se calcula mediante razonamiento probabilístico, en una red bayesiana que representa los documentos

(y sus partes) de la colección, a partir de la consulta y que obtiene la probabilidad de relevancia de cada unidad. La segunda gira en torno al concepto de que cuantos más términos de la consulta emparejan con los de una unidad, más relevante es ésta (es una especie de cobertura, en la que cuando se da esa situación, más cubre la unidad el tema de la consulta). En este caso, la red bayesiana se transforma en diagrama de influencia para obtener la utilidad de recuperar cada elemento de la colección.

Garnata ha sido evaluado en varias campañas de la iniciativa INEX (véase, por ejemplo, [5]), ofreciendo garantías para considerar que su operación tiene un rendimiento bastante aceptable, en cuanto a la calidad de la recuperación se refiere, en un entorno real con usuarios.

5. Otros sistemas de acceso a colecciones parlamentarias

En la literatura se pueden encontrar unos pocos ejemplos, en primer lugar, de cómo se emplea XML para representar documentos originados en parlamentos, y en segundo lugar, de cómo llevan a cabo la recuperación sobre ellos. En [12] se describe, dentro de un software más general para gestión de flujos de trabajo, la aplicación para convertir automáticamente los documentos del Parlamento de Portugal a XML, para posteriormente realizar su recuperación, pero desde un punto de vista de BBDD nativas de XML. Maarten Marx ha realizado una extensa labor de conversión a XML de los debates del Parlamento de Holanda (véase, por ejemplo, [8]). La recuperación sí la hacen, en este caso, con técnicas de RI estructurada [13]. Por otro lado, en la web `www.theyworkforyou.com`, se presenta información parlamentaria, más orientada a los diputados y a sus acciones en la cámara británica, toda ella marcada en este metalenguaje, aunque empleando una aproximación clásica desde el punto de vista de la recuperación. Por último, mencionar la iniciativa "Africa i-Parliaments" (`www.parliaments.info`), que promueve entre otras cosas, el marcado en XML de los documentos parlamentarios de África.

6. Conclusiones y trabajos futuros

En este trabajo se ha presentado *Seda*, un buscador de los documentos oficiales del Parlamento de Andalucía. El primer hecho diferenciador con la gran mayoría de los motores de búsqueda existentes, es su capacidad para determinar el grado de relevancia de las partes que componen los documentos, y devolver éstas en lugar de sólo los documentos completos, como harían los buscadores de texto plano. Esta característica ofrece un gran ahorro de tiempo al usuario, sobre todo cuando interactúa con documentos grandes, pues permite focalizarlo exclusivamente en el material relevante, y facilitar así el proceso de localización de lo que le interesa dentro del documento completo. *Seda* es un claro ejemplo de cómo se pueden transferir resultados de investigación en el campo de la RI a entornos reales.

En cuanto a trabajos futuros, además de un estudio de usuario que permita conocer de primera mano la opinión de los usuarios de *Seda* sobre su usabilidad y utilidad, se está investigando actualmente en métodos que permitan personalizar las búsquedas, considerando las peculiaridades de los usuarios (véase [6] como ejemplo de métodos propuestos). En este sentido, asumiendo la premisa de salvaguardar el universo de intereses de cada usuario, se ha abandonado la posibilidad de realizar perfiles de usuario individuales, por lo que nuestra investigación se está orientando a crear perfiles genéricos, a los que el usuario pueda adherirse cuando lo vea oportuno. Otras líneas de investigación futura podrían estar orientadas a dotar a *Seda* de la capacidad de reconocer entidades, y su aprovechamiento en un proceso de búsqueda, así como poder realizar una recuperación de datos analizando criterios de temporalidad.

Agradecimientos: Este trabajo ha sido financiado por la Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía (P09-TIC-4526) y el Ministerio de Educación y Ciencia (TIN2011-28538-CO2-02).

Referencias

1. L.M. de Campos, J.M. Fernández-Luna & J.F. Huete. Using context information in structured document retrieval: An approach using influence diagrams. Information Processing and Management 40(5):829–847, 2004.
2. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete & A.E. Romero. Garnata: An information retrieval system for structured documents based on probabilistic graphical models. Proceedings of the IMPU'06 conference, 1024–1031, 2006.
3. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, C. Martín-Dancausa Managing structured queries in probabilistic XML retrieval systems Information processing & management 46 (5), 514-532, 2010
4. L.M. de Campos, A.E. Romero. Bayesian network models for hierarchical text classification from a thesaurus. International Journal of Approximate Reasoning 50(7):932-944, 2009
5. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, C. Martín-Dancausa, and A.E. Romero. New utility models for the Garnata information retrieval system at INEX'08. Lecture Notes in Computer Science, 5631:39-45, 2009.
6. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, and E. Vicente-López. Using personalization to improve XML retrieval. IEEE Transactions on Knowledge and Data Engineering, to appear. DOI: 10.1109/TKDE.2013.75.
7. http://eurovoc.europa.eu/drupal/?q=es
8. T. Gielissen, M. Marx. Exemelification of Parliamentary Debates. Procs. of DIR conference, 19–25, 2009.
9. `https://inex.mmci.uni-saarland.de/about.html`
10. F.V. Jensen. Bayesian Networks and Decision Graphs. *Springer Verlag*, 2007.
11. M. Lalmas. XML Retrieval. Morgan & Claypool Publishers, 2009.
12. R. Nacimento, J. Martins, J. Pinto, P. Almeida. The Portuguese Parliament Workflow Process Automation based on FLoWPASS. Proc. ICWI, 1027–1030, 2003.
13. M. Marx, N. Aders. From documents to data: linked data at the Dutch Parliament. Proc. of the Online Information Conference, 17 – 22, 2010.
14. Trotman, A. & Sigurbjörnsson, B. (2005). Narrowed Extended XPath I (NEXI). LNCS., 3493, 16–40.