Un prototipo de motor de búsqueda para los diarios de sesiones del Parlamento de Andalucía basado en modelos gráficos probabilísticos

Luis M. de Campos, Juan M. Fernández-Luna. Juan F. Huete, Alfonso E. Romero
Departamento de Ciencias de la Computación e Inteligencia Artificial
E.T.S.I. Informática, Universidad de Granada, C.P. 18071, Granada
{lci, jmfluna, jhg}@decsai.ugr.es, aromero@correo.ugr.es

Resumen

En este trabajo se presenta el prototipo de sistema de recuperación de información diseñado para permitir la recuperación de los diarios de sesiones del Parlamento de Andalucía. Este software está basado en el *Context-based Influence Diagram Model*, modelo de recuperación cuya base formal son las redes bayesianas y los diagramas de influencia, modelos gráficos probabilísticos ampliamente utilizados en el tratamiento de la incertidumbre.

1. Motivación e Introducción

El Parlamento de Andalucía, desde su primera legislatura, que comenzó en 1982, transcribe todas las sesiones que se llevan a cabo, ya sean plenos, comisiones o diputaciones permanentes, dando lugar a una gran colección de documentos que plasman literalmente los temas tratados en esas asambleas. Estos documentos siguen fielmente la estructura establecida de las diferentes sesiones.

Actualmente, la Web del Parlamento andaluz [10] tiene accesible todos los documentos para su consulta por parte del público en general, en formato Portable Document Format (PDF). El acceso a los documentos se hace mediante unos formularios de consulta a bases de datos, que permiten búsquedas en los diarios almacenados sólo por fechas de celebración o rangos de éstas, legislatura o números de plenos, sin que se posibilite efectuar una consulta de texto libre. Son varios los problemas con que los usuarios se pueden encontrar con este tipo de búsqueda. Por ejemplo, el caso en el que el usuario tiene más o menos claro lo que quiere buscar, pero no conoce ni la fecha ni el número de diario en el que localizarlo. Además, en caso de encontrar las publicaciones adecuadas, éste debe revisarlas todas hasta ubicar lo que realmente le interesa.

En este sentido, la aplicación de técnicas de Recuperación de Información (R.I.) [5] a la búsqueda de diarios de sesiones claramente mejoraría el acceso a la información buscada. En este caso, tras efectuar una consulta en lenguaje natural donde se expresa la necesidad de información, el sistema de recuperación de información (S.R.I.) devuelve el conjunto de diarios relevantes (aquellas publicaciones que tratan, en más o menos grado, sobre el tema de la consulta), ordenados de manera decreciente según un valor de relevancia. Así, el usuario no necesitará conocer información previa y adicional sobre los documentos que busca. Esto también posibilita el hecho de que el usuario, en el caso en que no tenga muy claro lo que busca, pueda refinar su consulta de manera sucesiva, conforme vaya inspeccionando los documentos obtenidos.

Además, las técnicas de R.I. son capaces de aprovechar la estructura de los diarios de sesiones y determinar qué partes de éstos son las verdaderamente relevantes. Así, la salida sería un conjunto de *partes*, de diferente granularidad (desde el diario íntegro hasta un párrafo, por ejemplo), también ordenadas según su relevancia. De esta manera, sólo se suministra

al usuario las partes más relevantes, ahorrando al usuario la búsqueda del material que le interesa dentro de los diarios. A este tipo de colecciones documentales donde el material sigue una estructura bien definida se les conoce como colecciones estructuradas, y a las técnicas que se aplican, recuperación de información estructurada [6].

Las investigaciones realizadas dentro del grupo de investigación *Tratamiento de la Incertidumbre en Sistemas Inteligentes* de la Universidad de Granada en el campo de la R.I. han originado varios modelos de recuperación fundados en técnicas de inteligencia artificial, concretamente en modelos gráficos probabilísticos [7], como son las redes bayesianas [2] y los diagramas de influencia [2, 3].

Así, en este trabajo presentamos la de aplicación del modelo *Context-based Influence Diagram Model for Structured Documents Retrieval (CID model)* [3] a los diarios de sesiones del Parlamento de Andalucía, dando lugar al desarrollo de un prototipo de S.R.I. para documentos estructurados. La idea básica subyacente es transferir la investigación efectuada en este área, creando un software en explotación real que integre los logros alcanzados desde el punto de vista de la investigación, y adaptado al dominio concreto del parlamento autonómico. Con objeto de describir las características de esta aplicación, este trabajo queda estructurado como sigue: en la siguiente sección se expondrán brevemente los principales conceptos de R.I. estructurada, así como una breve introducción a los modelos gráficos probabilísticos relevantes al trabajo. La sección 3 presenta las peculiaridades de los diarios de sesiones y cómo han debido ser convertidos a XML [11] para facilitar su procesamiento; una descripción del modelo CID y cómo ha sido adaptado a la colección documental se expone en la sección 4. En la siguiente, la 5, se describe la arquitectura y funcionalidad del prototipo. Finalmente, la sección 6 cierra este trabajo con las conclusiones del mismo y algunos comentarios sobre las líneas de trabajo futuras.

2. Preliminares

2.1. Recuperación de información

Los S.R.I. son herramientas informáticas potentes y efectivas para localizar información por contenido. Un usuario especifica ese contenido mediante una consulta que a menudo se formula utilizando una expresión en lenguaje natural. Los documentos que se estiman relevantes respecto de la consulta se presentan al usuario mediante un interfaz. Los nuevos estándares en representación de documentos multimedia están impulsando el diseño y la implementación de nuevos modelos y de R.I. para indexar, recuperar y presentar documentos en consonancia con la estructura de los mismos (R.I. estructurada) [6]. De hecho, mientras que los modelos de R.I. clásicos consideran los documentos como entidades individuales e indivisibles, los métodos más modernos necesitan trabajar con representaciones más elaboradas, como por ejemplo documentos escritos en SGML, HTML, XML o MPEG-7. Estos formalismos de representación de documentos permiten representar y describir documentos *estructurados*, es decir documentos cuyo contenido está organizado de acuerdo a una estructura bien definida. Otros ejemplos de tales documentos son libros, artículos científicos, manuales técnicos, vídeos educativos, etc. Esto significa que los documentos no deberían considerarse como entidades indivisibles, sino como agrupaciones de objetos interrelacionados que necesitan indexarse, recuperarse y presentarse tanto de forma separada como en grupos, dependiendo de las necesidades del usuario. Así pues, dada una consulta, un sistema de R.I. debería poder recuperar los subconjuntos de componentes de los documentos que son más relevantes para esa consulta, y no sólo los documentos completos. Estos componentes se denominarán unidades estructurales.

La inclusión de la estructura de un documento en los procesos de indexación y recuperación afecta al diseño e implementación de un S.R.I. de varias formas. En primer lugar, el proceso de indexación debe tener en cuenta la estructura de una forma apropiada, de modo que los usuarios puedan buscar informa-

ción tanto por contenido como por estructura. En segundo lugar, el proceso de recuperación debería de usar tanto el contenido como la estructura para estimar la relevancia de los documentos y de sus componentes. Finalmente, el interfaz entre el sistema y el usuario debe permitir a éste hacer uso de la estructura del documento. De hecho, consultar por contenido y estructura sólo puede conseguirse si el usuario puede especificar en la consulta *qué* está buscando y *dónde* podría localizarse dentro de los documentos lo que busca. El qué implica el contenido, mientras que el dónde está relacionado con la estructura de los documentos.

De los diferentes tipos de modelos de R.I. existentes, los que más nos interesan son los modelos probabilísticos [4], que emplean la *Teoría de la Probabilidad* para manejar la incertidumbre que aparece en muchos de los procesos de R.I. Estos modelos calculan la probabilidad de relevancia de los documentos para una consulta. Algunos de los modelos de este tipo más recientes utilizan redes bayesianas como formalismo de representación del conocimiento, como es el caso del modelo subyacente al prototipo que presentamos en este trabajo.

2.2. Redes bayesianas y diagramas de influencia

Una red bayesiana consta de dos componentes [8]: (a) uno *cualitativo*, representado por un grafo dirigido y acíclico $G = (V, E)$, donde los nodos del conjunto finito V se corresponden con las variables aleatorias del problema a resolver, y los arcos de E indican causalidad, relevancia o relaciones de dependencia entre variables, y (b) otro *cuantitativo*, codificado mediante distribuciones de probabilidad condicionadas. En una red bayesiana, la ausencia de arcos entre pares de variables representa la existencia de relaciones de independencia condicional entre esas variables que pueden obtenerse. Las distribuciones de probabilidad permiten cuantificar nuestra incertidumbre sobre (la fuerza de) las relaciones existentes entre las variables del problema. Si $V = \{X_1, X_2, \ldots, X_n\}$, cada variable $X_i \in V$ tendrá asociado un *conjunto de padres*, $Pa(X_i)$, que son aquellas otras variables de V desde las que existe un arco dirigido hacia X_i, $Pa(X_i) = \{X_j \in V \mid X_j \rightarrow X_i \in E\}$. Para cada variable X_i se almacena una familia de distribuciones de probabilidad $p(X_i|pa(X_i))$, una para cada *configuración* $pa(X_i)$ de $Pa(X_i)$, es decir, para cada asignación de valores a todas las variables del conjunto de padres de X_i. Si X_i no tiene padres, $Pa(X_i) = \emptyset$, entonces $p(X_i|pa(X_i))$ es igual que $p(X_i)$. La inferencia consiste en el cálculo de las probabilidades a posteriori de todas o algunas de las variables, dada alguna evidencia conocimiento relativo a los valores que han tomado determinadas variables (las cuales se dice que han sido *instanciadas*).

Los diagramas de influencia [7] constituyen una generalización de las redes bayesianas para estudiar y resolver problemas de decisión con incertidumbre. Además de los nodos que representan variables aleatorias (*nodos de azar*, dibujados mediante círculos), el grafo que modeliza un diagrama de influencia contiene también *nodos de decisión* (representados mediante rectángulos) y *nodos de utilidad o de valor* (con forma de rombo). Cada nodo de decisión modeliza precisamente una decisión que hay que tomar, representada por una variable cuyos valores son las diferentes alternativas a disposición del decisor. Los nodos de valor representan el beneficio obtenido (utilidad) expresado numéricamente) de las consecuencias que se deriven de las decisiones adoptadas. Los arcos entre nodos de azar representan dependencias probabilísticas. Los arcos de un nodo de decisión a un nodo de azar o a un nodo de utilidad indican que la decisión tomada influirá en el valor del nodo de azar o en el beneficio que se obtenga, respectivamente. Los arcos de un nodo de azar a uno de decisión únicamente indican que el valor del nodo de azar será conocido en el momento de tomar la decisión. Finalmente, los arcos de un nodo de azar a uno de utilidad representan que el beneficio obtenido dependerá del valor que tome dicho nodo de azar. El objetivo final de un diagrama de influencia es calcular la utilidad esperada de cada una de las decisiones, con objeto de tomar aquella(s) decisión(es) que la(s) maximice.

3. Los diarios de sesiones del Parlamento de Andalucía

En esta sección describiremos brevemente la colección sobre la que actúa el prototipo de S.R.I. desarrollado. Esta base de datos documental está compuesta por aproximadamente dos mil documentos pertenecientes a las siete legislaturas de vida del Parlamento de Andalucía. Concretamente, diarios de sesiones de los plenos del Parlamento, de las diferentes comisiones existentes y de la diputación permanente. Estos diarios son transcripciones literales de lo debatido en cada una de las sesiones celebradas más información adicional, como es el caso del orden del día o un breve sumario.

Una de las principales peculiaridades de esta colección es su carácter dinámico, en el sentido de que el número de documentos que la forman crece continuamente cada año. Esto requiere que el S.R.I. que trabaje con ella tenga la potencialidad de incorporar a la colección documental nuevos documentos de una manera sencilla y eficiente.

Básicamente, todos los tipos de diarios tienen la misma estructura general:

A) Información de identificación de la sesión: número, fecha, presidente, legislatura, ...

B) Orden del día, compuesto por la relación de los puntos que lo integran. A su vez, cada punto queda formado por varias iniciativas (por ejemplo, en el punto titulado "Proposiciones de Ley-podrán existir varias iniciativas de este tipo, cada una de ellas con un identificador único, un tema y quién la propone).

C) Sumario. Siguiendo el orden del día, esta sección presenta información adicional para cada una de las iniciativas tratadas. Ofrece la información básica sobre la iniciativa y añade los ponentes del debate y el posible resultado de la votación.

D) Desarrollo de la sesión. La unidad básica de esta sección es de nuevo la iniciativa, que queda compuesta por las diferentes intervenciones de los diputados autonómicos. Cada intervención está compuesta por el nombre del ponente y del texto de la intervención. A su vez, se puede dividir en un conjunto de párrafos, siendo éstos unidades atómicas.

El Parlamento de Andalucía publica en su web los diarios en formato Portable Document Format (PDF), aunque actualmente el estándar de facto para representar documentos estructurados es XML (eXtensible Markup Language). Así, para que el S.R.I. pueda procesar correctamente los diarios y hacerlos disponibles para la búsqueda, los ficheros se deben someter a un proceso de conversión de formatos: de PDF a XML. Consiguientemente, se diseñó, en colaboración con personal del Parlamento de Andalucía un DTD (Document Type Declaration), es decir, se formalizó la estructura de los diarios de sesiones por medio de una gramática que debe ser cumplida por los documentos generados en XML. De esta manera se ha desarrollado una aplicación que forma parte del S.R.I. para hacer la conversión previa a la indexación. A continuación, y como ejemplo, se muestra un extracto del DTD, referente al desarrollo de una sesión plenaria:

```
<!ELEMENT pleno (numero_pleno, presidente,
numero_sesion_plenaria, fecha_celebracion,
pagina_inicio, pagina_fin, (orden_del_dia|
unica_iniciativa), sumario, desarrollo)>
...
<!ELEMENT desarrollo (intervencion*,
          debate_iniciativa+)>
<!ELEMENT debate_iniciativa
          (descripcion_iniciativa,
          intervencion+)>
<!ELEMENT descripcion_iniciativa (#PCDATA)>
<!ELEMENT intervencion (ponente,discurso)>
<!ELEMENT discurso (parrafo|acotacion)+>
<!ELEMENT acotacion (#PCDATA)>
<!ELEMENT parrafo (#PCDATA)>
```

4. El motor de búsqueda de diarios de sesiones: El modelo Context-based Influence Diagram

El modelo sobre el que se implementa el prototipo objeto de estudio es la evolución natural del *Bayesian Network Retrieval Model* [1], el cual trabaja con documentos planos, sin estructura. Éste ha sido extendido para manejar documentos estructurados y dotado con capacidad de decisión. Aunque una descripción muy detallada del modelo CID completo (topología, distribuciones de probabilidad, infe-

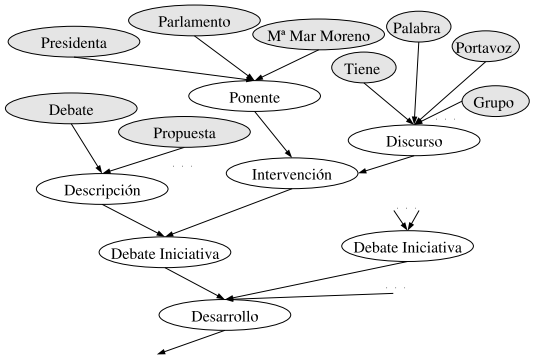

Figura 1: Red bayesiana representando una porción de un diario de sesiones.

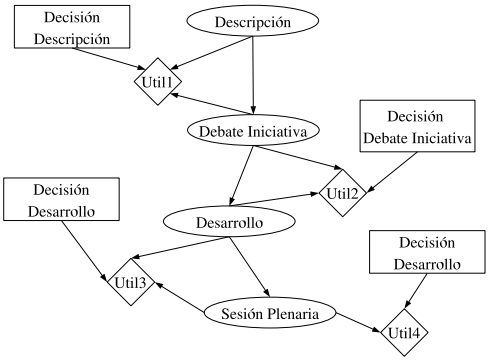

Figura 2: Parte del diagrama de influencia de un diario de sesiones.

rencia) se pueden encontrar en [2, 3], en esta sección presentaremos las características básicas del mismo. Desde el punto de vista de la descripción que vamos a hacer del modelo, al ser un diagrama de influencia un modelo gráfico probabilístico en el que se encuentra inmersa una red bayesiana, pasaremos a describirlo en dos etapas: en primer lugar la red bayesiana, y seguidamente las características especiales del diagrama de influencia.

En la figura 1 se muestra la topología de la red bayesiana que representa una parte del desarrollo de un diario de sesiones, concretamente la reflejada en la porción de DTD mostrada anteriormente. Cada unidad estructural del documento se representa como un nodo de azar en la red (elipses). Por ejemplo, *Desarrollo*. Estos nodos podrán tomar los valores *{no relevante, relevante}*. Los nodos de estructura se enlazan entre ellos mediante arcos, indicando que una unidad está contenida en otra (*Desarrollo* está incluido en *Sesión Plenaria*). Además, existe un segundo tipo de nodos de azar (sombreados en la figura 1) que representan a los términos de indexación (por ejemplo *Presidenta*), y que igualmente pueden tomar dos valores: *{no relevante,relevante}*. Estos nodos, sombreados en el gráfico, se enlazarán a los nodos de unidades donde estas palabras clave aparecen (así, la palabra *Presidenta* aparece en la unidad *Ponente*), mediante arcos que se originan en los nodos término y apuntan a los nodos unidades. La red bayesiana subyacente al diagrama de influencia queda totalmente especificada indicando que cada nodo de azar almacenará un conjunto de distribuciones de probabilidad que establecen el componente cualitativo de la red, y miden la fuerza de las relaciones expresadas en ella. Se obtienen a partir de la frecuencia de aparición de los términos en las unidades.

Además de los nodos de azar, en la sección de diagrama de influencia de la figura 2 se aprecian dos tipos más. Por un lado, los nodos de decisión. Dibujados como rectángulos, existe uno por cada nodo representando una unidad estructural (por ejemplo, *Decisión Desarrollo*), e indican la decisión de mostrar o no el contenido de la correspondiente unidad estructural al usuario. Toman los valores *{no recuperar, recuperar}* la correspondiente unidad. Por último, los nodos de utilidad (dibujados como rombos), que también están asociados a cada unidad estructural existente en la colección (como es el caso del nodo utilidad etiquetado como *Util1*, que está asociado con el nodo de azar *Descripción*).

Los arcos que involucran estos dos últimos tipos de nodos son los siguientes: Arcos desde cada nodo de azar representando a una unidad y su nodo de decisión asociado hacia el correspondiente nodo de utilidad, los cuales indican que la función de utilidad que almacenan éstos nodos depende obviamente de la decisión que se tome y del valor de relevancia de la unidad estructural considerada. Finalmente, para

cada unidad, desde la unidad donde está contenida surge también un arco apuntando al nodo utilidad. Esto implica que la utilidad de la decisión de presentar o no una unidad también depende del valor de relevancia de la unidad que la contiene. De esta manera se tiene en cuenta el contexto, lo que permitirá poder devolver, por ejemplo, la discusión completa de una iniciativa en lugar de una intervención completa cuando en dicha discusión se trate también el tema de la consulta, a pesar de que la intervención sea relevante. La función de utilidad asignará un valor para cada posible configuración de valores que una unidad dada, la que la contiene y el nodo decisión pueden tomar. Así también se pueden tener en cuenta las preferencias del usuario, el cual determinará si le pueden interesar unidades más grandes o más específicas.

Una vez que el diagrama de influencia se ha creado, desde el punto de vista cualitativo como cuantitativo, a partir de la colección de diarios de sesiones, el usuario puede efectuar una consulta. En su forma más simple, ésta será formulada en lenguaje natural. Los términos que la componen servirán de evidencias. En una primera fase, se calculará la probabilidad a posteriori de cada unidad dada la consulta mediante un método de propagación muy eficiente y específicamente diseñado para la topología de la red bayesiana subyacente. Una vez calculadas estas probabilidades, se utilizarán para obtener la utilidad esperada de cada unidad, de nuevo mediante una evaluación del diagrama muy eficiente. Con estos valores, el sistema podrá ofrecer una ordenación de todos los componentes según su grado de relevancia, o podrá discernir qué unidades será las que entregue al usuario.

5. Descripción del prototipo

El sistema sobre el que se implementa el modelo anterior se ha confeccionado a propósito para tal efecto, ya que aunque existen utilidades para indexación y recuperación de texto estructurado, su uso implica realizar el desarrollo de una manera muy rígida, estando encorsetado por la arquitectura y el diseño de éstos. En este sentido, el nuevo sistema desarrollado es flexible y nos permite totalmente adaptarlo a las necesidades de cada momento. También añadir, siguiendo esta línea, que el sistema ha sido implementado en un lenguaje compilado (C++), siguiendo una arquitectura orientada a objetos. De esta forma, es posible realizar una extensión o adaptación del sistema fácilmente.

Fundamentalmente, y de manera general, el prototipo es capaz de trabajar con colecciones de documentos estructurados basados en lenguaje de marcas XML, haciendo uso de la metainformación proporcionada por los posibles DTD (por lo que son necesarios si queremos trabajar con una colección).

A grandes rasgos, el sistema puede dividirse en dos módulos principales: el subsistema de indexación y el de consulta. El primero será el encargado de construir una serie de estructuras de datos persistentes (llamadas *índices*) sobre la colección, para facilitar el proceso de búsqueda. El segundo por su parte gestiona el procesamiento de las consultas que formulan los usuarios, recuperando las unidades relevantes según nuestro modelo y mostrándolas al usuario clasificadas según relevancia o utilidad esperada.

5.1. Indexación

El proceso de *indexación* se suele realizar una vez antes de poder usar el sistema y, en nuestro caso, implica la construcción de una estructura conocida como *índice invertido* [5]. Esta estructura es similar a la de índice terminológico que aparece en cualquier libro: dado un término, se proporciona la lista de identificadores de los documentos en los que aparece. Puesto que en nuestro caso estamos tratando con documentos estructurados, dicha lista tendrá que ser la de las unidades estructurales en las que aparece el término. Así, para una consulta que incluya varios términos, es fácil obtener la lista de las unidades comunes para dichos términos, simplemente haciendo una intersección de las listas individuales obtenidas mediante el índice invertido.

Al tratarse de documentos estructurados, la indexación como tal no termina aquí, pues-

to que lo dicho anteriormente sólo es válido para el texto. Para la información estructural (qué unidades son contenedoras, qué unidades están contenidas en una concreta,...) utilizamos otras estructuras de datos más específicas, que nos proporcionan las relaciones necesarias para procesar una consulta de forma eficiente tanto en espacio como en tiempo. Este índice podría considerarse de tipo *estructural*, mientras que el anterior corresponde a un índice *textual* (por la naturaleza de los datos que almacenan).

5.2. Motor de búsqueda

Aunque el tamaño de estas estructuras anteriormente descritas puede ser bastante grande y no suelen caber en memoria, se mantienen en ésta "partes críticas" de los índices, aprovechando la información obtenida tras un cierto tiempo de uso del sistema. Con esto pretendemos ajustar el tiempo de respuesta del sistema en función de las necesidades de información de los usuarios, manteniendo en una caché parte de las estructuras más referenciadas en las consultas. También, para ahorrar espacio (tanto en memoria como en disco), utilizamos técnicas de compresión de datos [9].

Obviamente, el tiempo de respuesta al usuario a la hora de procesar una consulta formulada por aquél es vital en un sistema de este tipo, y por tanto, todas las estructuras construidas durante el proceso de indexación van orientadas a acelerar la espera en una consulta. De esta forma, junto con el índice textual y el estructural, se almacenan convenientemente colocados valores precalculados dependientes del modelo propuesto (véase apartado 2.2), generalmente orientados a facilitar el cálculo de las probabilidades a posteriori de una unidad dada una consulta.

Las redes bayesianas y los diagramas de influencia que sustentan el prototipo no se construyen explícitamente, sino que se utilizan estas estructuras que albergan los índices estructurales y textuales. Los mecanismos de inferencia parten de los términos de cada consulta, considerados como evidencias, y efectúan un proceso de propagación y evaluación implementado de una manera muy eficiente, lo que redunda en un tiempo de espera para el usuario muy bajo.

En R.I. estructurada, son dos los tipos de consultas que se pueden efectuar sobre una colección. La primera es la consulta que expresa una necesidad de información basándose exclusivamente en contenido. Por ejemplo, "reforma de los estatutos de Andalucía". En este sentido, la recuperación se hace de acuerdo al proceso expuesto al final de la sección 4. Por otro lado, el segundo tipo de consulta es el que se conoce como "contenido + estructura". En este caso, se establecen restricciones estructurales que indican las unidades recuperables y/o las unidades en dónde buscar. Un ejemplo de este segundo tipo puede ser: "Intervenciones sobre la igualdad de género cuyo ponente sea la presidenta del parlamento en iniciativas relacionadas con proyectos no de ley".

En una primera aproximación a este tipo de consultas, y teniendo en cuenta que se debería hacer un análisis de las mismas desde un punto de vista del lenguaje natural con objeto de detectar unidades estructurales, se ha optado por una descripción de las mismas basadas en un formulario, en el que aparece un campo de texto por tipo de unidad estructural existente. En ese campo se introduce la parte de consulta por contenido relacionada con la unidad estructural. Además se añade un componente visual, concretamente una lista, donde el usuario marca exclusivamente las unidades en las que está interesado.

La inferencia en este tipo de consultas se hace de manera guiada. El texto asociado a las restricciones estructurales que indican dónde buscar hace que los términos que se instancien sólo influyan en las unidades estructurales que han sido especificadas en la consulta. Por otro lado, las restricciones que especifican el tipo de unidad devuelta, hacen que la propagación sea guiada, calculando la probabilidad a posteriori, y posteriormente la utilidad esperada exclusivamente de éstas unidades, y de aquellas que están incluidas, directa o indirectamente, en las objetivo. De esta forma, este tipo de consultas se responden de una manera rápida, pues únicamente se trabaja con aquello que se necesitará.

6. Conclusiones y trabajos futuros

En este trabajo se ha presentado un prototipo de S.R.I. para recuperar los diarios de sesiones del Parlamento de Andalucía, el cual se fundamenta en un modelo de recuperación cuya base son las redes bayesianas y los diagramas de influencia. Para que esta aplicación pueda operar sobre los diarios, los documentos que esta cámara autonómica genera en formato PDF han debido de convertirse a XML, tras realizar el diseño previo del DTD que define su estructura. La aplicación permite a un usuario interrogar a la colección documental en busca del material relevante, expresando tanto consultas por contenido, como por consultas de estructura más contenido, y obtener los resultados de una manera rápida y cómoda.

En cuanto a los trabajos futuros, una vez que éste se haya instalado para su explotación final en la web del Parlamento, se pretende realizar un estudio con usuarios reales a partir del cuál se obtendrá información relevante sobre la calidad general del mismo. Esta validación real examinará tanto la capacidad de recuperación de material relevante, como la forma en que el usuario interactúa con él, pudiendo modificar el sistema según los resultados obtenidos con objeto de mejorarlo.

Además, todos las nuevos resultados obtenidos de la investigación en el área, experimentalmente comprobados, se añadirán al mismo, originando un sistema totalmente actualizado y que incorpore las últimas novedades de investigación.

Agradecimientos: Este trabajo ha sido financiado por el Fondo de Investigación Sanitaria (FIS) bajo el proyecto PI021147.

Referencias

[1] L.M. de Campos, J. M. Fernández-Luna, J.F. Huete. *The BNR model: foundations and performance of a Bayesian network-based retrieval model.* International Journal of Approximate Reasoning, 34:265–285, 2003.

[2] 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 & Management, 40(5), 829–847, 2004.

[3] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete. *Improving the Context-based Influence Diagram Model for Structured Document Retrieval: Removing Topological Restrictions and Adding New Evaluation Methods*, Lecture Notes in Computer Science, 3408, 216–230, 2005.

[4] F. Crestani, M. Lalmas, C.J. van Rijsbergen, L. Campbell. *Is this document relevant?... probably. A survey of probabilistic models in information retrieval.* ACM Computing Survey, 30(4):528–552, 1998.

[5] R. Baeza-Yates y B. Ribeiro-Nieto. *Modern Information Retrieval*, Addison-Wesley, 1999.

[6] Y. Chiaramella. *Information retrieval and structured documents*, Lectures Notes in Computer Science, 1980:291–314, 2001.

[7] F.V. Jensen. *Bayesian Networks and Decision Graphs.* Springer Verlag, 2001.

[8] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[9] I.H. Witten, A. Moffat, T.C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images.* Morgan Kaufmann, 1999.

[10] Página web del Parlamento de Andalucía: *www.parlamento-and.es.*

[11] Especificación del Extensible Markup Language. http://www.w3.org/XML/