Informe Técnico UNED-IA-98-01 *Diciembre 1998*

ESTUDIO DE LOS MÉTODOS DE *EXPLICACIÓN EN REDES BAYESIANAS*

Carmen Lacave Rodero
Dpto. Informática
Universidad de Castilla-La Mancha
Escuela Superior de Informática
Paseo de la Universidad, s/n
13005 - Ciudad Real
Correo-e: clacave@inf-cr.uclm.es
www: http:\\www.inf-Cr.uclm.es\www\clacave

Fco. Javier Díez Vegas
Dpto. Inteligencia Artificial
UNED
Facultad de Ciencias
Avda. Senda del Rey, s/n
28040 - Madrid
Correo-e: fjdiez@dia.uned.es
www: http\\www.dia.uned.es\~fjdiez

ÍNDICE

0. INTRODUCCIÓN

Las redes bayesianas, surgidas en los años 80, proporcionan un modelo para representar el conocimiento de un sistema experto, que utiliza la probabilidad como medida de incertidumbre. Al ser de tan reciente aparición, todavía quedan aspectos por resolver. Uno de los más importantes es el de la explicación, tanto del conocimiento que modelan, como del proceso de razonamiento para la obtención de resultados y de la justificación de los mismos.

Y es precisamente esta carencia, la que ha constituido una de las grandes críticas que han sufrido los sistemas probabilísticos desde sus orígenes. El argumento esgrimido era la poca relación existente entre la forma de razonamiento humano y los métodos de razonamiento probabilístico. Sin embargo, estudios posteriores han demostrado que el problema no radica en la ausencia de relación entre ambos tipos de razonamiento, sino en que la mayoría de la gente no hace interpretaciones correctas de las probabilidades.

Por eso, si se dota a este tipo de sistemas de un mecanismo de explicación que justifique los resultados obtenidos, la aceptación de los mismos crecerá considerablemente.

A pesar de que distintos autores han realizado grandes esfuerzos por dotar a este tipo de modelos de cierta capacidad de explicación, no se ha encontrado todavía el método adecuado, aunque existen algunos que proporcionan explicaciones plausibles y que nos servirán de gran ayuda a la hora de *construir un nuevo método de explicación para redes bayesianas que mejore los defectos y lagunas que existen en los métodos actuales.*

1. OBJETIVOS DEL TRABAJO

Antes de emprender la tarea que nos proponemos, es necesario estudiar a fondo lo que se ha investigado hasta ahora para intentar definir, en la medida de lo posible, nuestras metas. Con este fin se elabora este trabajo, que se plantea como objetivos generales los siguientes:

- determinar las **características básicas** de la explicación para sistemas expertos en general, y para redes bayesianas en particular,

- recopilar información sobre explicación en sistemas expertos,

- recopilar información bibliográfica sobre explicación en redes bayesianas,

- analizar las ventajas e inconvenientes o dificultades de cada método propuesto,

El estudio de la bibliografía recopilada se abordará a partir de su clasificación en tres grandes grupos, del más general al más específico. Así, en el primero, se analizarán las publicaciones sobre explicación en *sistemas expertos en general*; el segundo, abarcará los trabajos sobre explicación en el caso concreto de los *sistemas expertos probabilísticos* y, el último grupo, englobará los realizados en el tema de la explicación en *redes bayesianas*, que son un subconjunto de los anteriores.

Ya que se hace recomendable aplicar el método a un caso real para proceder a su evaluación, nosotros lo pretendemos implementar sobre una red bayesiana para el diagnóstico de enfermedades infecciosas infantiles. En consecuencia, el análisis que realicemos, a pesar de que intentaremos crear un método de explicación que sea genérico para cualquier red bayesiana, lo enfocaremos, de momento, a las aplicaciones en medicina.

Finalmente, el apartado de conclusiones, tratará de sintetizar todo lo estudiado y propondrá las características que deberá poseer el sistema de explicación que intentaremos desarrollar posteriormente.

2. DEFINICIÓN DE EXPLICACIÓN

El término *explicación* tiene múltiples acepciones, dependiendo del momento histórico y del contexto en el que se aplique [30]. Desde los orígenes de la filosofía y de la ciencia, numerosos estudiosos de ambas disciplinas han intentado determinar su significado, vinculándolo con la comprensión y la descripción, pero sin fijar un sentido único. Así, por ejemplo, para algunos autores, la explicación consiste en la *búsqueda de las causas* de un hecho; otros, distinguen entre *explicación y comprensión*; algunos científicos opinan, de manera diferente, que consiste en una *deducción de leyes y condiciones* que describan un proceso.

En el ámbito de la lingüística aparece también esta multiplicidad de significados: "Declaración o exposición de cualquier materia, doctrina o texto con palabras claras o ejemplos, para que se haga más perceptible" | ... | "Manifestación o revelación de la causa o motivo de alguna cosa"[1]. Sin embargo, entre toda esta variedad subyace un objetivo común: el intento de buscar el **cómo** y/o el **porqué.**

En el caso de los sistemas expertos, ocurre lo mismo que en otras disciplinas: no existe unanimidad de criterios a la hora de determinar qué es la explicación y, no menos importante, cómo ofrecerla al usuario. De hecho, podemos encontrar, desde los trabajos que consideran una *explicación* del modo más simple, como una mera descripción de datos, hasta los más sofisticados, en los que constituye un diálogo "inteligente" en lenguaje natural con el usuario del sistema.

A pesar de ello, sí existe coincidencia en el objetivo que se persigue al dotar a los sistemas expertos de una herramienta de explicación, que es el de dar respuestas a preguntas del tipo "¿por qué el sistema ha obtenido este resultado?" o "¿cómo ha sido el proceso que ha llevado a esta conclusión?", con el fin de ayudar al usuario a comprender las conclusiones que le presenta el sistema que, para el caso concreto de las redes bayesianas, se traduce en ayudar al usuario a interpretar las probabilidades a posteriori que el sistema le proporciona.

3. CARACTERÍSTICAS BÁSICAS

Pese a tanta diversidad, el hecho de coincidir en un objetivo común, bien concretado, permite aislar ciertas características básicas, comunes a todos los métodos de explicación y que intentaremos exponer a continuación. Para intentar conseguirlo, nos basaremos en el resumen del seminario sobre explicación organizado por la AAAI en 1988, publicado por **Wick** [65], en el que se muestran las distintas líneas de investigación en el área de explicación en sistemas expertos y las cuestiones básicas que se plantean a la hora de construir este tipo de herramientas, que son:

- el contenido de la explicación,
- la consideración de la explicación como un proceso inicial junto con el de la adquisición del conocimiento,

[1] Acepciones que proporciona el diccionario de la Real Academia Española.

- los distintos tipos de usuario a los que va dirigida,
- los tipos de preguntas a realizar por estos,
- y el modo de presentar la consiguiente explicación.

3. 1. Clases de explicación

Se puede hacer una clasificación general de los modelos de explicación, atendiendo al contenido de las mismas:

- Explicación como **descripción** de los estados que representan el modelo y de los valores resultantes que proporciona el sistema. En este caso, la explicación consistirá en presentar un conjunto de variables o un conjunto de reglas y, posiblemente, ciertos valores numéricos asociados. Por ejemplo, en el caso concreto de las redes bayesianas, la descripción de un estado vendría dada por una colección de variables, los valores que toman y las probabilidades asociadas.

- Explicación como **comprensión**. En este caso, a diferencia del anterior, la explicación consistirá en la presentación de la conclusión mediante un discurso, expresado en lenguaje natural, en modo gráfico, etc. en el que se indiquen las relaciones entre los datos, de qué manera influye cada uno, bien individualmente o en unión con otros, sobre la conclusión y, sobre todo, en el que se identifiquen y justifiquen las causas que producen los resultados obtenidos. Además, se debe ofrecer una *interpretación* de los datos numéricos resultantes.

3. 2. Tipos de Explicación

Entre las facilidades que debe proporcionar una herramienta como la que se intenta construir, está la de ofrecer los siguientes tipos de explicación:

- Explicación ***del modelo*****,** ya que así se ofrece la posibilidad de detectar errores durante el proceso de construcción del sistema, además de ayudar al usuario a afianzar su propio conocimiento sobre el mismo. A este tipo de explicación se le denomina también *estática*.

- Explicación ***del proceso de razonamiento*** seguido por el sistema después de suministrarle cierta evidencia. También se conoce a este tipo de explicación como *dinámica* y debe ser capaz de proporcionar, básicamente, dos tipos de justificaciones, que son:

    - *los resultados obtenidos por el sistema*, tanto para demostrar al usuario que dicho proceso de razonamiento se ha llevado a cabo correctamente, como para que sirva de aprendizaje para aquellos usuarios que deseen adquirir o ampliar sus conocimientos sobre el sistema experto,

    - *las conclusiones esperadas por el usuario pero no obtenidas por el sistema* para, por un lado, comprobar si la representación del conocimiento se ha realizado correctamente y, por otro, mejorar o educar el conocimiento del usuario.

3. 3. Niveles de Explicación

Existiría otra clasificación para los métodos de explicación, que fue propuesta por **Sember** y **Zukerman** [53] para el caso concreto de las redes bayesianas pero que vale para cualquier sistema experto. Dicha clasificación, que se realiza atendiendo a la línea de razonamiento y del nivel de abstracción ofrecido en la explicación, es:

- **Nivel Micro**: la explicación consiste en una justificación detallada de las variaciones que se producen en un nodo particular de la red o en una regla concreta.

- **Nivel Macro**: en este caso, se muestran las principales líneas de razonamiento que llevan a una conclusión, explicando cualitativamente el razonamiento sobre la red completa o sobre un camino concreto de la red.

3. 4. Presentación de la explicación

En cuanto a las formas de presentar la explicación al usuario, podemos distinguir básicamente las siguientes:

- **Verbal**, es decir, mediante la generación de un discurso expresado en un lenguaje natural, que sea coherente y claro para el usuario.

- **Gráfica**, de dos formas:

  - Una consiste en aprovechar que la representación de las redes bayesianas se realiza mediante grafos y colorear los nodos o añadirles texto, números, etc. para mostrar los cambios sufridos en la red al hacer la propagación de la evidencia. Del mismo modo, los enlaces entre las variables del grafo pueden ayudar a entender de qué forma influyen ciertos datos sobre otros, en particular, para las explicaciones a nivel micro.

  - La otra se basa en la utilización de gráficas, diagramas de barras, funciones, etc. para representar la evolución de las probabilidades de las variables durante la propagación de la evidencia, la relación entre ciertas variables, etc.

- **Multimedia**, integrando las anteriores en un entorno de hipertexto, que permitiría generar explicaciones interactivas, y multimedia. Los entornos multimedia ayudarían enormemente a la comprensión de la explicación, al poder incluir en la misma herramientas tan eficaces, desde el punto de vista educativo, como el vídeo y el sonido.

3. 5. Adaptación al usuario

Por último, uno de los factores que se presentan como determinantes para obtener la mejor efectividad en la explicación, es la capacidad que tenga el sistema para adaptar su comportamiento a las necesidades específicas de cada usuario y a sus expectativas. Sin embargo, éste es un problema complejo, pues implica conocer en cada momento cuál es el nivel de conocimiento del usuario sobre el dominio correspondiente, e ir actualizándolo a la vez que éste progresa, no solo durante el funcionamiento del sistema sino incluso después.

Las soluciones propuestas consisten en clasificar el conocimiento del dominio según diferentes tipos de usuario (principiante, avanzado, experto,...), y que sea el usuario, en el momento de la utilización del sistema, el que se autoclasifique según una escala proporcionada. Sin embargo, así como en el resto de cuestiones enumeradas hasta ahora sí existe unanimidad de criterios entre los autores a la hora de realizar las distintas clasificaciones, no se puede decir lo mismo sobre la catalogación de los distintos tipos de usuario. Por este motivo, esta cuestión la iremos analizando con cada método individualmente.

4. MÉTODOS DE EXPLICACIÓN EN SISTEMAS EXPERTOS

El auge que se viene produciendo, durante esta última década, de los entornos dedicados a la intercomunicación persona-computador, sobre todo en los sistemas aplicados al aprendizaje, se deja notar también a la hora de diseñar herramientas de explicación para sistemas expertos. Por esta razón, analizaremos dichos métodos de explicación atendiendo a su clasificación en dos grandes grupos:

- *métodos no intercomunicativos*, que son aquéllos en los que la explicación se genera de modo cerrado. La principal característica de este tipo de métodos es que no existe posibilidad de que el sistema actualice el conocimiento que a priori posee del usuario, durante el funcionamiento del mismo. Esto implica que el método de explicación queda <u>prefijado</u> de antemano, no pudiendo este modificarse durante el proceso de generación de la correspondiente explicación.

- *métodos intercomunicativos*, formados por los métodos que generan la explicación por medio de entornos en los que existe comunicación entre el sistema y el usuario y en los que, en consecuencia, tanto el usuario como el propio sistema, actualizan el conocimiento mutuo durante el proceso de explicación, lo que implica, fundamentalmente, la adaptación del sistema en cada momento a las necesidades y expectativas del usuario. Por esta razón a este tipo de sistemas se les denomina también <u>adaptativos</u>.

4.1. Métodos no intercomunicativos

Como se ha indicado, este tipo de métodos se caracterizan porque la presentación de la explicación correspondiente se genera de modo estático, sin permitir al usuario “conversar” con el sistema. Esto quiere decir que, por ejemplo, tanto los tipos de preguntas que el usuario puede realizar, como el nivel de detalle de la explicación están fijados previamente. Igualmente, y lo que es más significativo, el sistema no puede modificar, durante su funcionamiento, ni el conocimiento sobre la evolución que se puede ir produciendo durante el proceso de explicación del nivel del usuario, ni en general, sobre el contexto en el que se genera y presenta la explicación.

Aunque hay numerosos trabajos realizados en este campo, nos limitaremos simplemente a analizar aquéllos cuyas aportaciones han sido o pueden ser relevantes para la generación automática de explicaciones en redes bayesianas. El estudio lo haremos atendiendo a una clasificación general de los tipos de sistemas expertos. En consecuencia, empezaremos viendo las herramientas de explicación desarrolladas para los sistemas expertos basados en reglas, que han sido y siguen siendo mayoría, para pasar después a estudiar el caso concreto de los sistemas probabilísticos. Dentro de cada grupo, presentaremos los trabajos más relevantes ordenados según su aparición cronológica.

4. 1. 1. Sistemas expertos basados en reglas

❖ En el sistema de Bleich [2], considerado por algunos autores como uno de los primeros sistemas expertos, se utilizaba un mecanismo de explicación bastante rudimentario, conocido como técnica del "texto enlatado" (*canned text*) que consistía en almacenar, junto con el código del sistema, las respuestas de las preguntas que se le podían plantear. El problema que conlleva la generación de este tipo de explicación es hay que prever todas las preguntas posibles, con la dificultad que esto conlleva.

❖ Uno de los grandes avances se produjo con el desarrollo del sistema médico MYCIN [3], y que fue pionero, entre otras cosas, en el campo de la generación de explicaciones. El método que utilizaba consiste en traducir el conjunto de reglas utilizadas durante el proceso de razonamiento y en encadenarlas para producir un discurso comprensivo. El problema surge cuando hay un número elevado de reglas o de antecedentes de algunas de ellas. En este caso, la obtención de una línea de razonamiento coherente entre las mismas puede resultar bastante complicada y nada clarificador.

❖ Posteriormente **Swartout** [59], con su sistema XPLAIN, desarrolla un método consistente en la utilización de un programador automático para crear el sistema experto y determinar el funcionamiento del mismo, lo que incluye también el mecanismo para la generación de explicaciones. El sistema XPLAIN se utilizó como sistema tutor en la administración de fármacos para la estabilización del ritmo cardíaco y consta de varias partes, ilustradas en el esquema global que se describe en la Fig. 1 y que son:

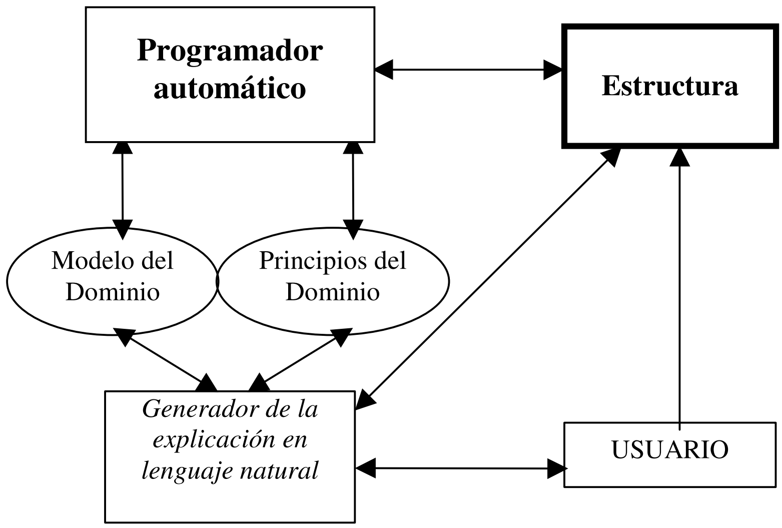


Fig. 1. Esquema del sistema XPLAIN

- El programador automático, encargado de generar el sistema.
- El modelo del dominio, que representa los hechos del conocimiento experto mediante una red causal en la que se plasman las relaciones para la prescripción de la terapia concreta (el *qué*).
- Los principios del dominio, que contiene los métodos y heurísticas del dominio (el *cómo*)
- La estructura para la generación de explicaciones.
- Un generador de explicaciones en lenguaje natural.

La novedad de este trabajo estriba en la utilización del programador automático para generar, durante el proceso de construcción del sistema experto, la información necesaria para producir posteriormente las justificaciones de los resultados proporcionados. Dicha información se almacenará en una estructura que sea capaz de almacenar los mecanismos de razonamiento seguidos por el programador durante la construcción del sistema. En concreto, el método consiste en ir creando, a la vez que el sistema, dicha estructura, que estará representada por un árbol. Como el sistema se va generando mediante un proceso de refinamiento de objetivos generales a específicos, cada nodo del árbol corresponderá a cada uno de los objetivos; por tanto, un subárbol corresponde a una fase de refinamiento del objetivo que está representado en el nodo del nivel superior, es decir, en su nodo *padre*. Las hojas del árbol constituyen las operaciones básicas que el programa debe producir (Fig. 2).

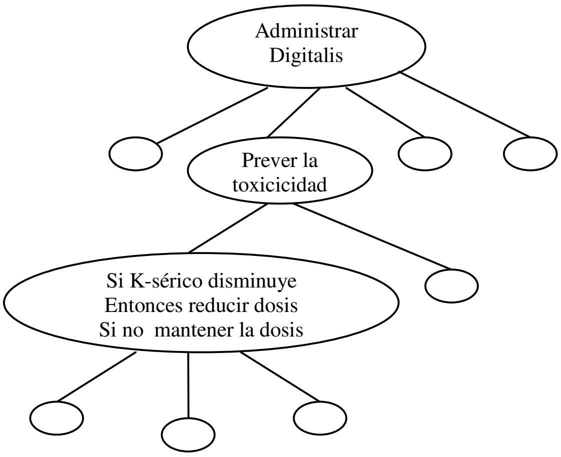


Fig 2. Ejemplo de estructura tomado de [59]

Esta estructura servirá como entrada al módulo que se encarga de generar la explicación en lenguaje natural. Dicho generador de explicaciones consta de dos partes:

- una de bajo nivel, llamada *generador de frases*, que es la que construye las frases directamente de la base de conocimiento de XPLAIN;
- otra, llamada *generador de respuestas*, de alto nivel, que determina qué es lo que se debe decir; su misión es la de seleccionar aquellas partes de la representación del conocimiento que el generador de frases debe traducir a lenguaje natural. Además, otra tarea del generador de respuestas es la de seleccionar el nivel de detalle de la explicación, para lo que necesita conocer en cada momento cuál es el estado de la ejecución del programa, qué es lo que se ha dicho ya y qué es lo que el usuario necesita saber.

❖ En esta línea encontramos también las aportaciones de **Wick** y colaboradores [64, 66], quienes proponen una metodología de explicación reconstructiva, mediante la que se presentan los procesos de razonamiento y de explicación como tareas distintas e independientes, tanto en forma como en contenido. El argumento que utiliza es que cuando un humano proporciona una explicación de cualquier problema, normalmente no realiza una descripción de todos los pasos que le han llevado hasta la solución, sino que hay detalles que se omiten e, incluso, pueden aparecer otros que no formen parte del razonamiento.

Por esta razón, consideran que las técnicas de generación de explicación deben tomar como entrada el proceso de razonamiento, posiblemente incompleto, que se haya llevado a cabo en el sistema para la solución de un determinado problema, y basarse en él para producir la explicación de las conclusiones. Para ello, se incluye un nuevo dominio que contiene todo el conocimiento necesario para la generación de dicha explicación (Fig. 3).

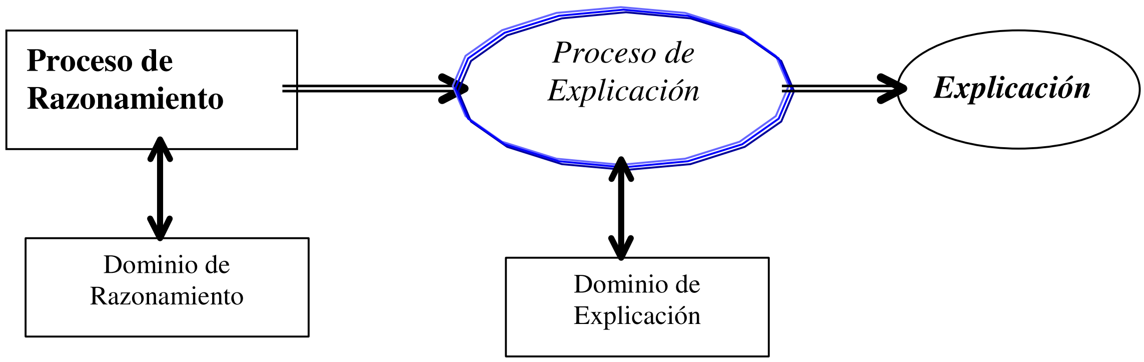


Fig. 3. Proceso de explicación propuesto por Wick.

Todas estas ideas las han plasmado Wick y Thompson en el prototipo REX [66], diseñado para proporcionar explicaciones de cómo un sistema experto obtiene los resultados que presenta. Dicho sistema está basado en un modelo propuesto también por él, para generar explicaciones reconstructivas. El modelo es el siguiente: el sistema experto produce un conjunto de "*pistas*" del proceso de razonamiento que ha seguido; éstas pasan por un filtro que, basándose en las restricciones del problema, se encarga de determinar qué partes del razonamiento se pasarán al sistema generador de la explicación. A continuación, dicha información, filtrada y ajustada al problema concreto que se desee justificar, junto con las restricciones que debe verificar la solución, se proporcionan al módulo encargado de generar la explicación. Este módulo utiliza la información almacenada en el dominio de la explicación, para obtener la línea de la explicación que justifique las pistas de razonamiento proporcionadas. La línea de explicación conseguida se pasa, finalmente, a otro módulo encargado de traducirlo a lenguaje natural y de presentarlo al usuario.

❖ **Di Marco** y colaboradores [17] han desarrollado también un sistema para la generación de lenguaje natural en el ámbito de la medicina, llamado HEALTHDOC, cuyo objetivo es generar informes médicos para los pacientes y materiales sobre educación para la salud, adaptados a las necesidades concretas de cada uno de ellos. El sistema posee las siguientes características:

- Los informes y materiales se generan en forma de documentos escritos, por lo que no se permite interacción entre el sistema y el paciente. Estos documentos se pueden enviar directamente a la impresora o ser generados por un procesador de textos para que el médico lo edite y le haga las modificaciones que estime oportunas.
- La forma de adaptar el texto a las necesidades del paciente está determinada por la condición médica del paciente y por sus características personales y culturales. Para ello, existe una base de datos con toda la información de los pacientes.
- Se permite la generación del texto en varios idiomas.
- El proceso de generación de informes en lenguaje natural está basado en técnicas de planificación de texto.

La producción de cada informe concreto se realiza a partir de un *documento maestro,* específico del tema a tratar en dicho informe. Este documento contendrá toda la información, incluidas las ilustraciones y las anotaciones sobre aquellos datos que se consideren relevantes, que el sistema tenga que incluir en él. En concreto, el documento maestro se representa mediante un conjunto de estructuras y anotaciones en SPL (Sentence-Plan Language). Este lenguaje se utiliza como paso intermedio para la generación de lenguaje natural. Además, cada sentencia se marca con relaciones de coherencia y referencia a otras sentencias. El documento maestro lo edita el médico en cuestión, mediante una herramienta proporcionada por el propio sistema HEALTHDOC, llamada Splat.

El proceso de generación de un informe en lenguaje natural consta de varios pasos, ilustrados en la Fig. 4 y que, básicamente son: a partir del documento maestro, se realiza una selección de su contenido. Esta selección determinará los planes de texto, que permitirán generar el consiguiente discurso en lenguaje natural. Estos planes, que son grupos de proposiciones, pasan a continuación a un módulo encargado de realizar la planificación del discurso, teniendo en cuenta las relaciones retóricas necesarias para la producción del texto. De esta fase se obtiene un conjunto de proposiciones estructuradas, que pasan al proceso encargado de realizar la planificación de las sentencias o frases; en concreto, se encarga de controlar el orden de las mismas, las relaciones entre ellas, el léxico, etc. La salida de este proceso consiste en una secuencia de especificaciones de frases, que son las que se pasan al módulo encargado de convertirlas a lenguaje natural. Como se puede apreciar en la figura, durante todo el proceso de planificación se tienen en cuenta los datos del paciente para así realizar la planificación en cada momento atendiendo a sus características específicas.

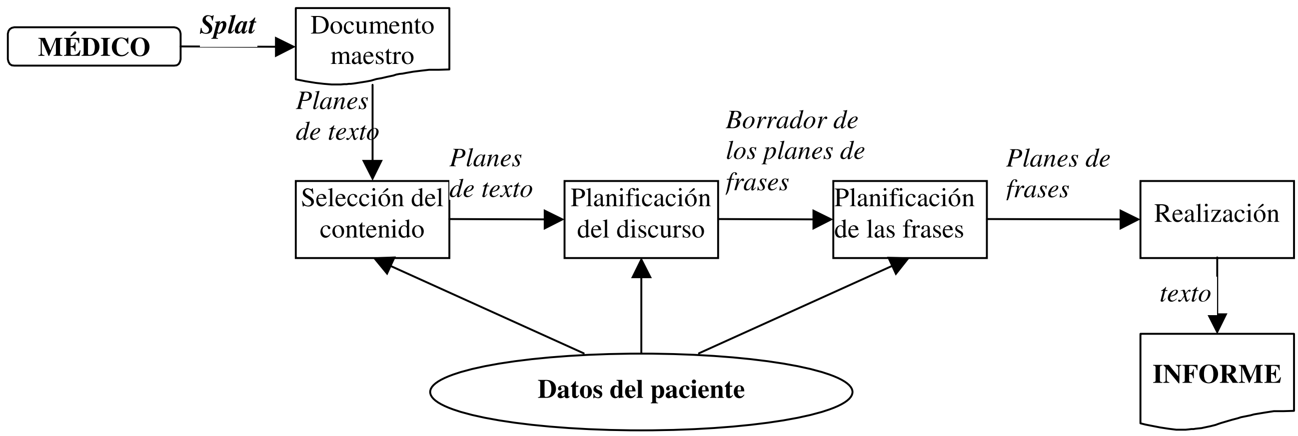


Fig. 4. Modelo de generación de informes en lenguaje natural de HEALTHDOC

❖ Por último, **De Carolis** y colaboradores [14] realizan un trabajo aplicado a la prescripción de fármacos en medicina, implementado en el sistema OPADE, en el que además de hacer un estudio previo de lo que debe contener una explicación, y teniendo en cuenta las necesidades del usuario, proponen un método para generar explicaciones basado en técnicas de *planificación de texto.* Estos autores clasifican en dos los tipos de usuarios: directo e indirecto. El usuario directo es aquél que interacciona con el sistema, mientras que el indirecto es aquél que simplemente recibe un informe (posiblemente escrito) con la explicación de los resultados obtenidos, por ejemplo, aplicado al caso de la medicina, el usuario directo sería el médico y el indirecto el paciente, que es quien recibirá el informe con la explicación del diagnóstico médico. El método se propone exclusivamente para la generación de explicaciones al usuario indirecto, eso sí, teniendo en cuenta las características del usuario directo. Es decir, tratan de generar un informe tal y como lo haría el usuario directo que lo "firmará".

Para ello, se introduce como entrada la historia médica del paciente y el diagnóstico emitido, y a partir de aquí, se genera el texto de la explicación en lenguaje natural de dicha situación. Este proceso se lleva a cabo, fundamentalmente, gracias a las cuatro componentes de las que consta el sistema, que son las siguientes:

- Modelo del usuario, que contiene las características tanto del usuario directo como del indirecto. Del primero representa dos propiedades: su *propensión a hablar* y sus *preferencias sobre los estilos* de los mensajes. Del segundo, su *grado de conocimiento* y sus *intereses* sobre la información a recibir. Ambos tipos de modelos están clasificados de acuerdo con los resultados de estudios psicológicos previos, sobre las explicaciones que proporcionan los médicos y las necesidades de los pacientes a los que va dirigida dicha explicación.

- Planificador de texto, que es el que se encarga de definir la estructura del texto, en función de los tipos concretos de usuario, tanto directo como indirecto, según la clasificación citada anteriormente. También se encarga de determinar los conceptos que aparecerán en el texto de la explicación, su orden de presentación y las relaciones retóricas entre las partes del texto. Los criterios para la generación del texto se describen por medio de una librería de *operadores de planificación*, que son los que establecen cómo conseguir un objetivo cuando se verifican ciertas condiciones y qué efecto produce sobre el usuario. Un operador de planificación es una tupla "(Encabezado, Restricciones, Intenciones, Precondiciones, Efectos, Descomposición, Relación_Retórica)". La acción que realizan dichos operadores consiste en descomponer el objetivo en dos o más subobjetivos, o en una o más acciones que incluir en el texto del discurso. Gracias a estas descomposiciones se va obteniendo un plan de discurso en forma de árbol.

- Módulo de resolución de conflictos, que se encarga de encontrar un equilibrio entre el usuario directo y el indirecto cuando hay discrepancias.

- Generador de texto, basado en redes de transición aumentadas (ATN) anidadas en varios niveles. Dicho generador toma el plan de texto obtenido, en forma de árbol, como entrada, y produce el texto asociado del modo siguiente: la red del primer nivel explora el árbol desde la raíz, examina el tipo de nodo y salta a los nodos de niveles inferiores, las redes de niveles intermedios generan términos lingüísticos de acuerdo con el tipo de usuario (directo e indirecto) y las inferiores generan fragmentos de sentencias, de acuerdo con la acción asociada al nodo hoja. Estas frases se obtienen de la base de conocimiento.

Podemos observar el esquema del sistema en la Fig. 5.

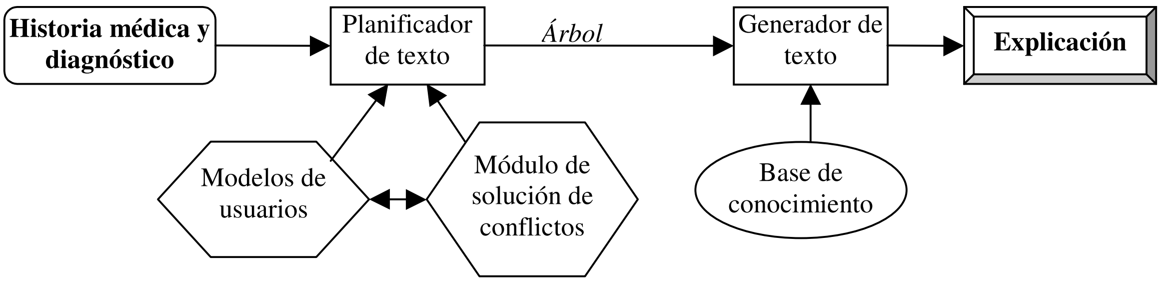


Fig. 5. Estructura del sistema OPADE

4. 1. 2. Sistemas expertos probabilísticos clásicos

Este tipo de sistemas expertos se caracterizan por la utilización de la probabilidad como medida para procesar la incertidumbre. Aunque se disiente sobre si la teoría de Dempster-Shafer puede considerarse una extensión de la inferencia probabilística, coincide con ella en utilizar una técnica puramente numérica basada en la teoría de la probabilidad, por lo que en este apartado incluiremos también los trabajos más relevantes basados en esta teoría. Además, analizaremos también, por ser un caso particular de sistemas probabilísticos, los relacionados con la teoría de la decisión probabilística.

❖ Entre los estudios sobre explicación, relacionados con la teoría de Dempster-Shafer, destaca el de **Strat** y **Lowrance** [55, 56]. La idea principal que subyace en el método que proponen para la generación de las explicaciones es la del *análisis de sensibilidad*, idea que han utilizado también otros autores en otros tipos de sistemas [16, 17, 58]. Este tipo de análisis consiste en realizar distintas variaciones en la evidencia para comprobar la variación en las posibles soluciones. El proceso concreto que desarrollan Strat y Lowrance consta de varios pasos:

1. Análisis del efecto de la evidencia sobre la hipótesis a determinar. Este proceso genera un grafo en el que cada nodo define un estado, que representa una opinión, y cada arco, el mecanismo por el que se cambia de un estado a otro.

2. Representación del flujo de información, mediante el grafo, desde los hallazgos hasta la conclusión, lo cual permitirá generar la explicación. Una explicación para ellos consiste en un subconjunto de hallazgos que forman la evidencia. Para obtenerlos, se realiza el análisis de sensibilidad que consiste en la eliminación de uno o más hallazgos que forman la evidencia para después actualizar el valor de la credibilidad de la hipótesis con arreglo a diversas medidas de especificidad; éstas serán las que determinen la pertenencia de uno o más hallazgos a la explicación. Ya que en estos trabajos se contemplan distintos tipos de preguntas que puede realizar el usuario, las medidas de especificidad serán también distintas dependiendo de la clase de pregunta.

La explicación la presentan en lenguaje natural, y para ello almacenan, junto con cada nodo y arco del grafo, el texto que representa el estado o el argumento para el cambio de estado, respectivamente.

❖ En ocasiones, en los sistemas en los que además de emitir los posibles diagnósticos, tienen que ofrecer una decisión, no basta con dar una explicación del diagnóstico emitido, sino que hay que considerar otras propuestas alternativas que, aunque no corresponden a la justificación de la conclusión, son igualmente *útiles.* Por ejemplo en medicina, además de emitir un diagnóstico hay que decidir el tratamiento a seguir. Entre los métodos de explicación desarrollados para este tipo de sistemas, destacaremos el de **Reggia** y **Perricone** [50], quienes proponen un método para la justificación, a nivel macro, del diagnóstico de ataques neurológicos, basado en clasificación bayesiana.

El proceso consiste en realizar una ordenación de todas las posibles enfermedades que el sistema puede diagnosticar y después ofrecer la justificación de la posición relativa de una enfermedad concreta en la clasificación obtenida. Dicha justificación se realiza teniendo en cuenta un factor de medida asociado a cada dato de evidencia relacionado con la enfermedad en cuestión. Esta medida se obtiene a partir de la comparación de las probabilidades a posteriori de dicho síntoma, suponiendo que se ha producido cada enfermedad.

El método propuesto es interactivo y se inicia pidiendo al usuario que introduzca la evidencia, para después mostrarle las 5 enfermedades más probables, junto con la probabilidad de cada una. Si el usuario lo desea, puede pedir la justificación de una de dichas enfermedades en relación con las demás; en ese caso, se muestran las probabilidades de cada una y las causas que han determinado la obtención de dicha probabilidad.

Para la explicación de por qué una enfermedad D está en una posición determinada en relación con las demás enfermedade[illegible]letectar tanto las causas que hacen que D tenga una probabilidad menor que las que h[illegible]ma de ella como las que hacen que D tenga una probabilidad mayor que las que hay [illegible] de ella en la clasificación.

La mayor limitación del método estriba en que la hipótesis que tienen que verificar los modelos para poder aplicar este algoritmo es demasiado restrictiva: todos los síntomas deben tener causas definidas.

❖ En esta misma línea se enmarca el trabajo de **Langlotz, Shortliffe** y **Fagan** [40]. En él, y basándose en los principios de la teoría de la decisión, proponen una técnica de razonamiento simbólico para generar explicaciones de los resultados del análisis de las posibles decisiones. Al igual que los analistas generan explicaciones basadas en árboles de decisión, la idea es utilizar técnicas de representación del conocimiento para establecer una correspondencia entre los nodos del árbol de decisión y las situaciones médicas que representan. Dicho mecanismo de explicación está formado, fundamentalmente, por dos componentes:

- *Representación del conocimiento mediante marcos*. Cada nodo del árbol se representa mediante un marco, lo que permite definir sus características propias y las relaciones con otros nodos.

- *Razonamiento matemático cualitativo*, consistente a su vez en los cuatro pasos siguientes: cálculo de la utilidad esperada de cada decisión, comparación de las ramas para determinar qué conceptos son candidatos para aparecer en la explicación, selección de los candidatos que aparecerán en la explicación y de qué modo deben ser presentados y, por último, expresión en lenguaje natural utilizando las técnicas de traducción de conocimiento de sistemas basados en reglas. El proceso se ilustra en la Fig. 6.

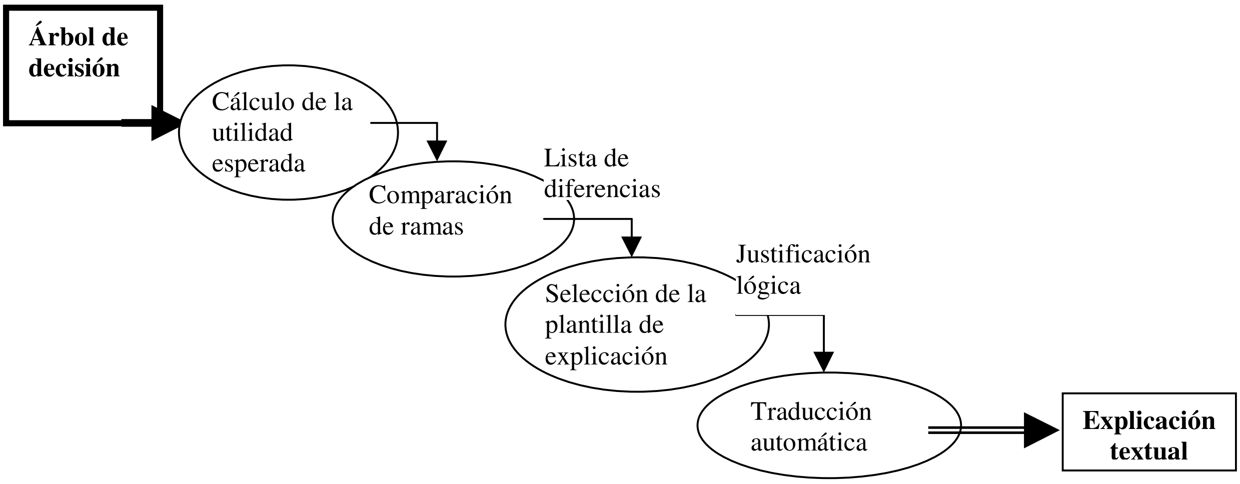


Fig. 6. Procesos para la generación de la explicación propuesto en [40]

La capacidad de decidir el modo más efectivo de organizar y presentar los conceptos se lleva a cabo por medio del análisis de una librería de plantillas y la selección de una de ellas. Este conjunto de plantillas se representa mediante expresiones simbólicas que indican cómo se debe estructurar una explicación y de qué elementos debe constar. La traducción, por tanto, se realiza gracias a la asociación de un fragmento de texto en lenguaje natural con cada plantilla y con cada sentencia lógica que se utiliza para expresar situaciones o diferencias.

En la siguiente tabla encontramos ejemplos de expresiones simbólicas y los textos asociados para la traducción.

Expresión Simbólica	Texto Asociado
P(X)	"la probabilidad de X"
NEFROTOXICIDAD | X	" nefrotoxicidad debido a X"
X > Y	"X es mayor que Y"

❖ Finalmente, aunque no profundizaremos de momento sobre los estudios relacionados con la teoría de la decisión, **Druzdzel**, en el capítulo 9 de su tesis [22], realiza un análisis amplio sobre este tema y podría servirnos de referencia si quisiéramos ampliar el problema de la explicación en redes bayesianas al caso de diagramas de influencia, donde sí aparecen nodos decisión y utilidad. Mi objetivo, como ya he indicado antes, se centrará en proporcionar explicaciones para el caso concreto de las redes bayesianas.

4.2. Métodos Intercomunicativos

Este tipo de métodos surge frente a las limitaciones de las herramientas clásicas de explicación a la hora de realizar explicaciones complejas, debido sobre todo a la falta de comunicación bidireccional entre el sistema y el usuario. Uno de los objetivos que se persiguen al explicar algo es que quien recibe la explicación quede satisfecho. Esto implica que, en determinadas situaciones, el receptor de la explicación necesitará interrumpir el proceso de explicación para formular las preguntas que considere para completar su conocimiento o, por el contrario, para indicarle al emisor que no hace falta que profundice en un determinado tema que ya conoce.

Por tanto, la principal meta de estos métodos es llevar a cabo dicha comunicación, de tal modo que el sistema sea capaz de adaptarse a las nuevas necesidades y/o expectativas del usuario, según varían durante el proceso de generación de la explicación. Por ejemplo, es posible que un usuario con un bajo nivel de conocimiento del modelo representado, sea capaz de mejorarlo durante el proceso de explicación, de tal forma que, al cabo de un cierto tiempo, su nivel haya cambiado, pasando de tener el nivel de un usuario inexperto al de uno superior, por lo que el sistema debe ser capaz de actualizar dicho nivel de conocimiento.

Para conseguir la adaptación entre el usuario y el sistema y que ambos puedan actualizar su conocimiento mutuo de forma dinámica, se debe permitir:

- *al usuario*, solicitar aclaraciones concretas sobre los datos de la explicación en los que necesite profundizar, así como la posibilidad de realizar interrupciones;
- *al sistema*, revisar el nivel del usuario.

La idea sobre la que se fundamentan este tipo de herramientas es la de considerar la tarea de la generación de explicaciones como un problema *independiente del tipo de sistema experto* en sí pero, sin embargo, íntimamente relacionado con el problema de la transmisión de información de cierta complejidad, de la forma más efectiva posible.

❖ Una de las investigadoras que más ha aportado al tema es **J. Moore.** En los trabajos realizados tanto con **Swartout** como con otros autores [4, 44, 45], se describe la explicación como una forma de conversación con el usuario del sistema y propone una técnica de *explicación reactiva*. Es decir, el sistema tiene que ser "consciente" de todo lo que se haya dicho ya y reaccionar a cualquier pregunta del usuario de forma inteligente, identificando cuáles son sus objetivos. Para ello, propone un método basado en la *planificación del texto* que aparecerá como contenido en la explicación y en el *almacenamiento de la explicación* que se va produciendo, para utilizarlo después si fuera necesario. Las técnicas de planificación del texto permiten seleccionar y estructurar dinámicamente la información que se va a presentar en cada caso. Un esquema del sistema que proponen lo podemos ver en la Fig. 7

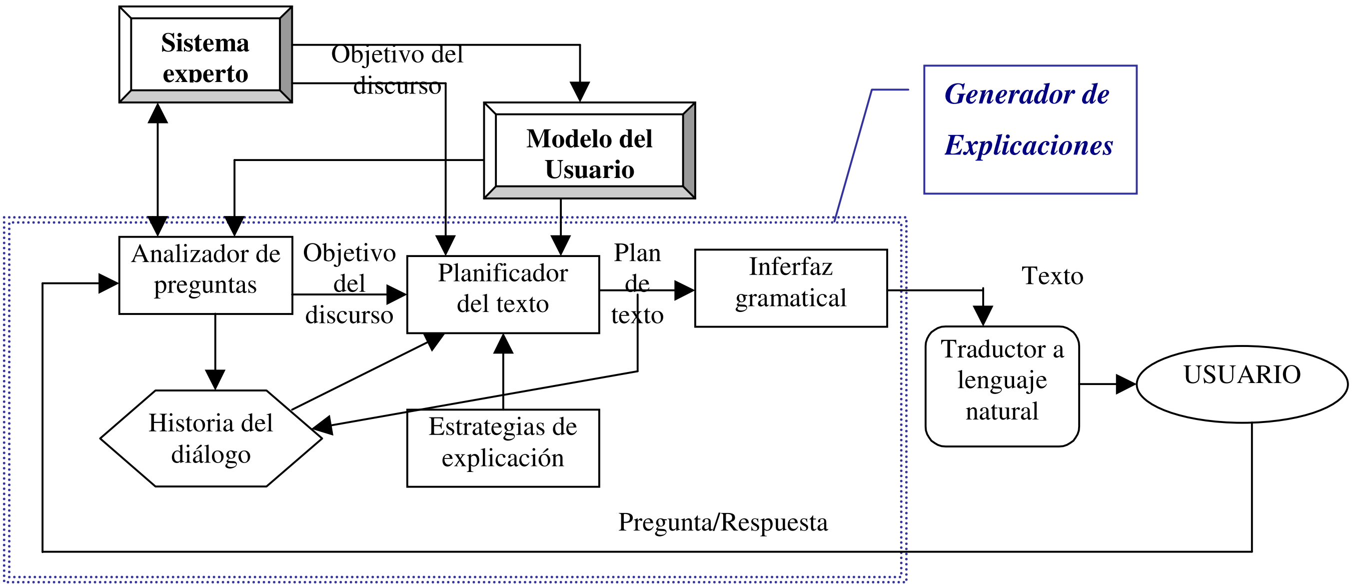


Fig. 7. Esquema del sistema de explicación propuesto por Moore et al. [44]

El proceso, en líneas generales, tiene lugar como sigue: cuando el usuario realiza una pregunta (o una respuesta, pues el sistema le puede solicitar información al usuario), el analizador de preguntas genera un objetivo, que representa la abstracción de la respuesta que se desea producir en el discurso de la explicación. Este objetivo se pasa al planificador de texto, que busca entre las estrategias de explicación que posee almacenadas el sistema, representadas normalmente mediante reglas, aquéllas que verifican el objetivo deseado. La selección de una concreta entre todas ellas se lleva a cabo mediante técnicas de búsqueda heurística, utilizando además el conocimiento del sistema sobre el modelo del usuario, el diálogo mantenido hasta ese momento, etc. Una vez que se tiene el plan del texto de la explicación, representado en forma de árbol, se almacena en la historia del diálogo y se pasa a la interfaz gramatical para que sea transformado, mediante un traductor de lenguaje natural, en el texto de la explicación. Después de presentar dicho texto al usuario, el sistema espera a que éste le responda a dicha explicación: si la ha entendido, si desea concretar algo más, etc. El sistema procesa esta información por de modo similar al caso anterior, y se repite el proceso hasta que el usuario queda satisfecho con la explicación presentada por el sistema.

❖ En una línea de investigación similar, están los estudios de **Alison Cawsey** [6, 7, 8, 9, 10], relacionados con la generación de explicaciones interactivas. El método que ella propone se basa, además de en la *planificación del texto* de las explicaciones, en la *planificación del diálogo* con el usuario. La explicación se genera gracias a la combinación de las reglas que permiten planificar el texto con las de la planificación del diálogo. Dicho método lo han implantado en el sistema EDGE, desarrollado para la generación de tutoriales sobre circuitos electrónicos. Entre las capacidades que ofrece este sistema están: permitir al usuario que interrumpa el diálogo en cualquier momento y obligar al sistema, durante la generación de la explicación, a actualizar, tanto el modelo del usuario como el contenido del resto de la explicación. En este sistema, el proceso de planificación del diálogo se lleva a cabo de modo incremental, como muestra la Fig. 8:

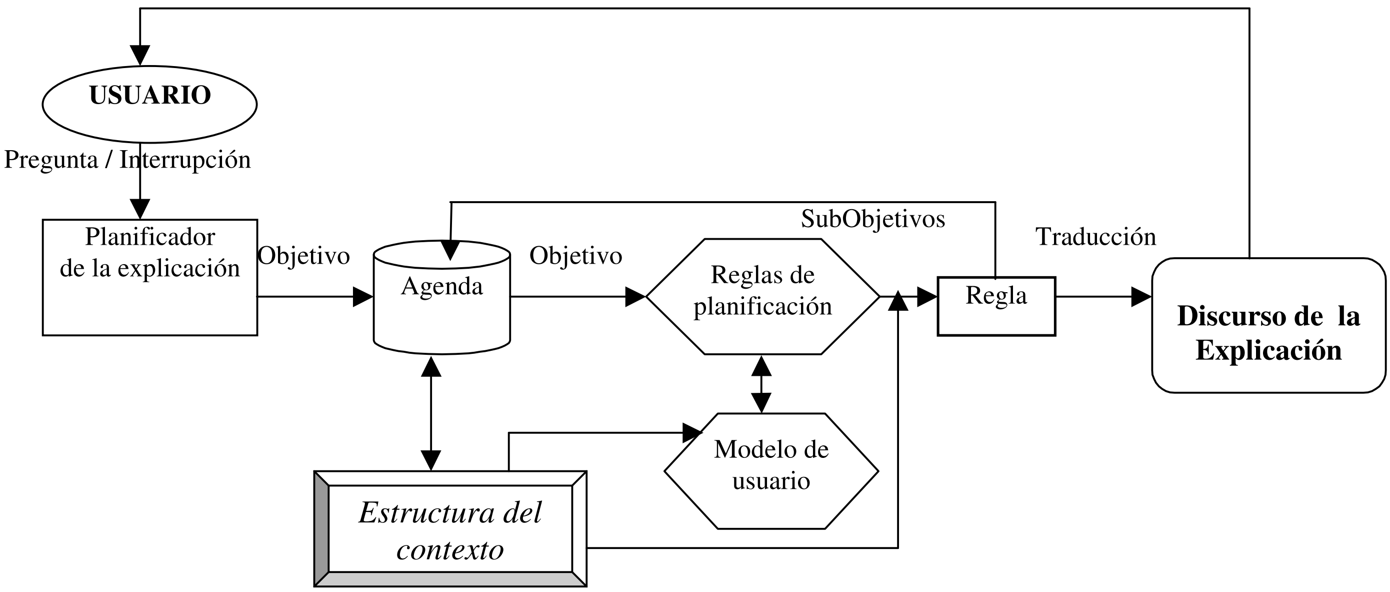


Fig. 8. Proceso de planificación de la explicación.

Los objetivos de la explicación se ponen en una agenda. De ella, se selecciona el objetivo de mayor prioridad. A continuación, se selecciona una regla de planificación, la cual se usa para encontrar nuevos subobjetivos del objetivo inicial, que se colocan a su vez en la agenda. Así, se va creando una estructura jerárquica que servirá para representar el contexto en el que se debe desarrollar la explicación.

Por otra parte, las interrupciones del usuario pueden tener como resultado la inclusión de nuevos objetivos en la agenda y, además, pueden proporcionar información al sistema sobre el conocimiento del usuario para ser actualizado, lo que influirá en el plan del resto de la explicación. Además, después de cualquier intercambio con el usuario, el contexto del discurso se debe actualizar, para que la explicación se genere de acuerdo a las nuevas suposiciones realizadas.

5. *MÉTODOS DE EXPLICACIÓN EN REDES BAYESIANAS*

Una vez realizado el estudio sobre los métodos de explicación para distintos tipos de sistemas expertos, nos centraremos en la revisión de las técnicas desarrolladas para el caso concreto de las redes bayesianas. Debido a la diversidad de trabajos sobre este tema, dicho análisis se llevará a cabo atendiendo a la clasificación inicial de los tipos de explicación que debe proporcionar un sistema experto en general, y bayesiano en particular, y que son:

* *explicación del modelo* del dominio que representa el conocimiento experto, y
* *explicación del razonamiento* que se lleva a cabo en la red al introducirle determinada evidencia.

5. 1. Explicación del modelo

En un sistema experto, y en concreto en redes bayesianas, tan importante como la corrección y eficiencia del método de razonamiento, es que la red refleje exactamente el conocimiento experto que se desea modelar, ya que si éste no se representa de forma apropiada no se puede esperar ningún resultado fiable del sistema.

Sin embargo, aunque se ha investigado bastante sobre los algoritmos de propagación de la evidencia dentro de la red, no ocurre lo mismo con el proceso de construcción de redes bayesianas; de hecho, al no existir una metodología para llevar a cabo esta tarea, está considerada más como un arte que como una técnica. Esta ausencia de metodología implica que, en el caso de que se desee contrastar el modelo con expertos humanos para detectar y corregir posibles errores, queda en manos del ingeniero de conocimiento la tarea de convencer o demostrar a los expertos que dicha representación corresponde con su conocimiento del dominio. Por tanto, la validez del modelo y la aceptación del sistema por los expertos dependerán de algo tan subjetivo y concreto como es la capacidad que tenga el ingeniero para llevar a cabo dicha explicación. Sin duda, esto aumenta más la complejidad del proceso de construcción de la red bayesiana.

Por otra parte, una vez finalizada la construcción del sistema experto, el usuario no tiene posibilidad de obtener ninguna explicación del conocimiento que está representado en dicho sistema. Esto puede constituir una grave deficiencia, sobre todo en los sistemas expertos orientados a la educación.

Por todos estos motivos, se hace necesario proporcionar un método de explicación que sirva por un lado, para convencer, tanto al usuario como a los expertos de los que se extrae el conocimiento, de que la información con la que trabajará el sistema es correcta, y por otro, para comprender el significado de dicha información.

Aunque no se han realizado demasiados trabajos sobre este tema, los estudiaremos atendiendo su clasificación según la forma de presentar dichas explicaciones:

5.1.1. Gráficas

Las presentaciones gráficas se caracterizan fundamentalmente por ser fáciles de entender, aunque en ocasiones pueden resultar confusas. Esto sucede, por ejemplo, en los casos en los que se desee mostrar el grafo que define la red, pero el número de nodos y de enlaces entre ellos es demasiado grande. Aunque tanto los expertos consultados como los ingenieros de conocimiento que diseñan la red están de acuerdo en la importancia de visualizar gráficamente la estructura que representa el conocimiento, sin embargo pocos autores han dedicado esfuerzos a trabajar sobre este tipo de explicación.

- Los únicos autores que hacen referencia a este tipo de explicaciones del modelo son **Durzdzel** [22] y **Suermondt** [58] quienes, en sendas tesis doctorales, proponen un método de explicación que consiste, básicamente, en la visualización del grafo que define la red.
- Respecto a otro tipo de representaciones, como pudieran ser gráficas que indican la evolución de las probabilidades, diagramas de barras para mostrar los valores de las probabilidades de un conjunto de variables, etc., no se ha trabajado prácticamente hasta ahora, pues todos los esfuerzos para la generación de explicaciones se han dedicado para su expresión en lenguaje natural. Sin embargo, sería interesante dotar a las herramientas de explicación con mecanismos de presentación gráfica que sirvan para complementar lo expresado en lenguaje natural.

5.1.2. Verbales

El objetivo fundamental de este tipo de explicaciones es el de adecuarse a la forma en la que se expresan los humanos, además de permitir matizar las relaciones entre los nodos para así tener la posibilidad de distinguir entre varios tipos de enlaces: causa-efecto, factores de riesgo, relaciones ES-UN, TIENE-UN, etc.

❖ En este grupo destaca la propuesta de **Druzdzel** [22], quien para explicar el modelo se centra sólo en las relaciones causales que definen los enlaces entre los nodos.

El método propuesto consiste en describir con una frase cada uno de los enlaces entre los nodos, junto con la probabilidad asociada a cada enlace; para aquéllos que no tienen padres, proporciona la probabilidad a priori de dicho nodo.

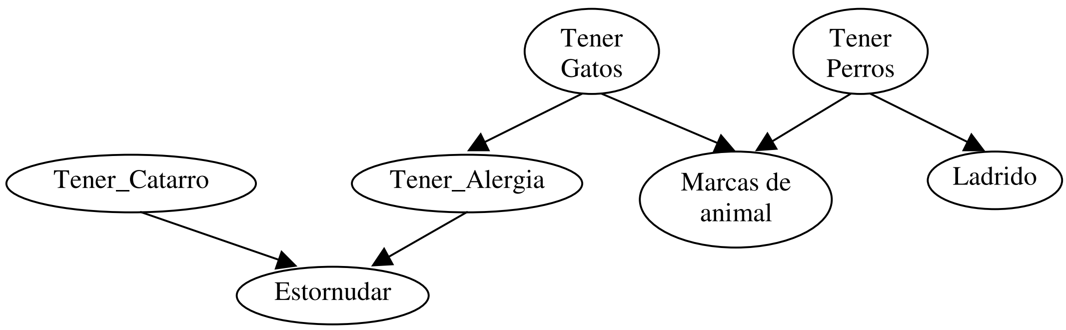


Fig. 9. Ejemplo de red bayesiana propuesto por Druzdzel

Por ejemplo, para la red bayesiana de la Fig. 9, las explicaciones propuestas serían:

La probabilidad de Estornudar depende de Tener_Alergia

La probabilidad de Estornudar depende de Tener_Catarro

...

Tener_Catarro es muy poco probable (0.08)

...

Las influencias entre los nodos se explican junto con la probabilidad así:

Tener_ Gatos frecuentemente (0.8) producen Tener_Alergia

...

Para reflejar las comparaciones entre probabilidades de varias hipótesis, Druzdzel propone la utilización de expresiones de posibilidad relativa, obtenidas de un estudio realizado por **Elsaesser** y **Henrion** [29], en el que realizan un análisis sobre dichas expresiones lingüísticas. De esta forma, se obtendrían explicaciones del tipo:

Tener_Catarro es <u>algo menos probable</u> que Tener_Gatos (0.08 /0.10)

...

En general, Druzdzel propone distinguir varios tipos de enlaces dependiendo de la relación que exista entre los nodos, aunque no realiza ninguna clasificación de los tipos posibles de enlaces que pueden aparecer en una red bayesiana. Sí distingue entre los que pertenecen a una puerta-OR por la ventaja que ésta define: la independencia de los mecanismos causales. Para una puerta-OR propone una técnica particular de explicación:

Tener_Catarro causa muy frecuentemente (0.9) Estornudos

Tener_Alergia causa muy frecuentemente (0.9) Estornudos

Tener_Catarro no afecta a Tener_Alergia para producir Estornudos, y viceversa

No hay otra causa de Estornudos que Tener_Catarro y Tener_Alergia

...

❖ En este grupo destaca también la propuesta posterior de **F.J. Díez**, quien en su tesis doctoral [15], presenta un mecanismo de explicación implementado para el sistema experto bayesiano para ecocardiografía DIAVAL, que sirve tanto para explicación del modelo, como para explicación del razonamiento, y que puede ser un punto de partida para el caso que nos ocupa. Así, distingue varios tipos de enlaces y de nodos.

El mecanismo de explicación utilizado, basado en un sistema de ventanas y menús, se encuadra dentro del nivel micro, pues las explicaciones se refieren, en cada momento, sólo a un nodo o enlace y consisten en ofrecer, para cada uno de ellos, distintas posibilidades de explicación.

- Para cada nodo, se ofrecerán al ususario las siguientes opciones: *prevalencia, probabilidad a posteriori,*lista de *causas o factores,* tabla de *probabilidad condicional* (para los nodos con padres), lista de *efectos, definición, imagen* para los nodos que correspondan a observaciones cualitativas y *pantalla*, para los sistemas expertos que cuenten con interfaz gráfico.

- Para cada parámetro, que es un nodo con un dato numérico asociado, se ofrecerán: *valor, intervalos, significado, fórmula* (para los parámetros calculados a partir de otros), *imagen* y *pantalla.*
- Para los enlaces, clasificados en dos grupos, los que forman parte de una puerta-OR y los que pertenecen al caso general, se presentará: c*ausa, efecto, sensibilidad y especificidad,* texto que explica el *mecanismo causal* asociado y *eficiencia.,* que vendrá dada por las probabilidades condicionales.

La facilidad de explicación propuesta en DIAVAL permite ir navegando por la red y solicitar en cada momento información sobre un nodo o enlace.

❖ Hay que indicar también que algunas aplicaciones que existen actualmente para la construcción de redes bayesianas y, en general de diagramas de influencia, en particular **HUGIN** y **MSBN32**, permiten realizar también una clasificación de los tipos de enlaces y de nodos, lo que puede ayudar a diseñar las herramientas de explicación en los casos en que se utilicen dichos programas para la construcción de redes bayesianas.

* En concreto, en MSBN32 se distinguen varios tipos de nodos según varios criterios.
Atendiendo al papel que desempeñan dentro del problema, las clases de nodos son: *hipótesis, información* y *otro.*
Según su relación con sus padres, se pueden clasificar en: independencia de los mecanismos causales (puerta-OR) o estándar.

5. 2. Explicación del razonamiento

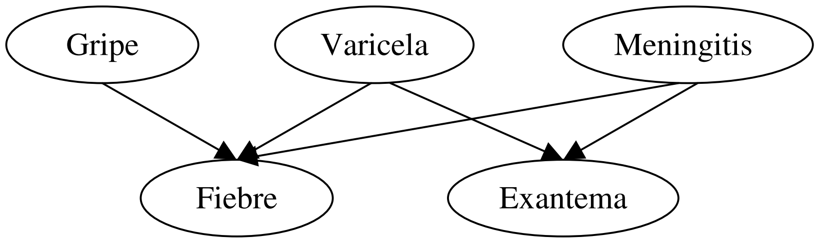
La explicación del razonamiento tiene por objetivo ayudar al usuario a comprender el funcionamiento del sistema y a adquirir nuevo conocimiento o afianzar el que posee sobre el dominio concreto que representa la red bayesiana. Para ello, debe ser capaz de explicar, a partir de las suposiciones implícitas en el modelo y después de suministrarle cierta evidencia, dos tipos de procesos:

i) el que le ha llevado a obtener las conclusiones que presenta.

ii) el que le ha llevado a descartar un resultado esperado por el usuario.

La mayoría de las investigaciones en explicación en redes bayesianas se han realizado para la justificación del primer tipo, por lo que habrá que tenerlo en cuenta con el fin de dotar a la herramienta de explicación que diseñemos de la capacidad de argumentar sobre aquellos resultados suministrados por el usuario y que no formen parte de los proporcionados por el sistema.

Aunque pudieran parecer procesos similares, sin embargo no lo son. Por ejemplo, dada la red bayesiana siguiente:



supongamos que la evidencia proporcionada es *Exantema* y *Fiebre_Moderada* y que el sistema proporciona una explicación del tipo: *La Varicela es la causa de la Fiebre_Moderada y Exantema con una probabilidad del 85%.*. Si el usuario pregunta: "¿por qué existe poca probabilidad de que haya meningitis?", el sistema responderá algo así como: *la Fiebre Moderada reduce la probabilidad de meningitis al 15%,* o bien: *para que la probabilidad de meningitis sea más del 80% debe existir Fiebre Elevada.*

En ambos casos, se presentan las causas que producen la evidencia y la explicación está basada en el proceso de propagación de dicha evidencia sobre la red.

Si la pregunta es, sin embargo, "¿por qué la probabilidad de la gripe es 0,008?" La respuesta será del tipo: *La gripe no produce exantema, lo que descarta la enfermedad.*

Como vemos, el proceso de razonamiento es distinto al anterior, aunque el objetivo es el mismo, ya que en este caso hemos buscado hallazgos que no son efectos de la enfermedad en cuestión. Esto implica que los procesos de justificación de los resultados del sistema y de justificación de los resultados que no forman parte de las conclusiones del sistema pueden ser distintos y, en concreto, este último, llevarse a cabo de diversas formas, mediante la búsqueda de efectos que no son producidos por la evidencia, o en la búsqueda de datos que disminuyen la probabilidad de los datos, etc.

Independientemente de la clase de explicación que se vaya a realizar, se observa una diferencia clara al atender al concepto de explicación *(3.1.)* lo que genera una división del conjunto de trabajos en dos grupos:

- el primero, estaría formado por aquéllos para los que el objetivo de la explicación es realizar una descripción. En este caso, la explicación de una hipótesis, según una evidencia determinada, consiste en mostrar la asignación o asignaciones más probables de valores a las variables que hacen que la probabilidad conjunta de la red sea máxima. Al proceso de búsqueda de esta clase de explicaciones se le denomina *abducción*.

- el segundo grupo, englobaría aquellos estudios de los que consideran una explicación como la forma de presentar mediante un discurso, la línea de razonamiento seguida por la red, ya sea expresado de modo gráfico o lingüístico, que haga comprensible y justificable para el usuario de la misma, la conclusión a la que ha llegado el sistema al partir de la evidencia proporcionada; es decir, explicación como comprensión.

5. 2. 1. Explicación como descripción

❖ Los primeros trabajos relacionados con la explicación en redes bayesianas son los de **Pearl** [47, capítulo 5] quien define una **explicación** como una asignación de valores, consistentes con la evidencia, a todas las variables de la red. Pearl, basándose en los algoritmos de propagación de la evidencia, propone un método para generar varias explicaciones y de ellas seleccionar la más probable. En concreto, su método MPE (*Most Probable Explanation*) para generar explicaciones se basa en la propagación de los mensajes π y λ.

Para entender el método, supongamos que W es el conjunto de variables de la red; la tarea de encontrar una explicación más probable se hace mediante una serie de computaciones locales, para cada variable X del conjunto W, y está basada en el principio de que, para cada valor x de una variable X, existe la mejor explicación de las variables $W\backslash X$. Por tanto, el problema de encontrar la mejor explicación para $X=x$ (proceso basado en la técnica de optimización mediante programación dinámica), se puede descomponer en la búsqueda de la mejor explicación para las variables de $W\backslash X$ y utilizar dicha información para elegir el mejor valor de X.

Dicho método consiste en lo siguiente:

Dado *W* el conjunto de variables de la red, calcular P(w^*| e), siendo W=$\boldsymbol{w}^*$ la explicación más probable para la evidencia *e*.

$$P(w^* \mid e) = \max_{w} P(w \mid e)$$

La tarea de encontrar $\boldsymbol{w}^*$ se realiza localmente calculando para cada variable *X* el valor

$$BEL^*(x) = \max_{w\acute{}_X} P(x, w\acute{}_X \mid e)$$

donde $W\acute{}_X = W \backslash X$. Por tanto, $BEL^*(x)$ representa la probabilidad de la explicación más probable para *e,* consistente con la asignación X=x.

Supongamos, para ilustrar el algoritmo, el caso general de un nodo *X* que tiene *m* hijos $Y_1,...,Y_m$ y *n* padres $U_1,..., U_n,$, como se ilustra en la Fig. 10.

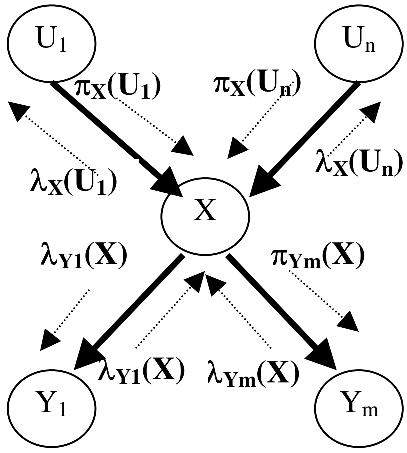


Fig. 10. Ejemplo de fragmento de red

Para calcular $BEL^*(x)$ se procede realizando los cálculos siguientes:

1. Actualización de la creencia, mediante la inspección de los mensajes $\pi^*_{Yj}(x)$ y $\lambda^*_X(u_j)$ de sus sucesores y predecesores, respectivamente, y calculando el producto

$$F(x, u_1 ..., u_n) = \prod_{j=1}^{m} \lambda_{Y_j}^{*}(x) P(x \mid u_1 ..., u_n) \prod_{i=1}^{n} \pi_x^{*}(u_i)$$

Esta función permite a *X* calcular $BEL^*(x)$ e identificar el mejor valor *x* del dominio de X:

$$x^* = \max_x{}^{-1} BEL^*(x)$$

siendo

$$BEL^*(x) = \beta \max_{u_k:1\le k\le n} F(x, u_1, ..., u_n)$$

2. Propagación Ascendente, mediante la actualización de los mensajes λ

$$\lambda_X^*(u_i) = \max_{x,u_k:k\neq i}\left[\ F(x,u_1...,u_n)/\pi_X^*(u_i)\ \right],\quad i = 1,...,n$$

3. Propagación Descendente, mediante la actualización de los mensajes π.

$$\pi_{Y_j}^*(x) = \beta \frac{BEL^*(X)}{\lambda_{Y_J}^*(X)}$$

4. El proceso se repite hasta encontrar un nodo sin sucesores o un nodo que define un hallazgo o bien un nodo sin padres.

En el caso de poliárboles, la complejidad del algoritmo es lineal, pero para el caso de redes con bucles no parece que sea la forma más adecuada, ya que la complejidad del algoritmo puede ser exponencial. Incluso para el primer caso, el proceso de explicación puede complicarse demasiado al tener que incluir en él variables que ni siquiera tienen relevancia sobre la variable que se procesa en cada momento.

No obstante, en su aplicación al diagnóstico en general, y particularmente en el campo médico, existe el inconveniente de que este método solamente es adecuado cuando la diferencia de probabilidades entre la primera y el resto de las explicaciones es apreciable. Además, no es capaz de generar explicaciones alternativas de modo ordenado más allá de la segunda [46].

❖ Una de las mejoras propuestas al método de Pearl, en cuanto a la posibilidad de generar varias explicaciones ordenadas según su probabilidad, la ha desarrollado en su trabajo **Santos** [52], quien propone un método con el que se obtienen varias explicaciones para una evidencia dada. Este método consiste en transformar la red bayesiana en un sistema de restricciones lineales equivalente a la red y resolver dicho sistema.

Un sistema de restricciones *L*(*W*) es una tupla (Γ, *I*, ψ) donde Γ es un conjunto de variables, *I* es un conjunto finito de desigualdades definidas sobre las variables de Γ y ψ es una función de $\Gamma\times$\{verdadero, falso\} a $\Re$.

El proceso de construcción del sistema de restricciones se lleva a cabo en tiempo lineal respecto al número de variables de la red bayesiana, y consiste en añadir restricciones que garanticen las siguientes propiedades: que cada variable sólo tome un valor y que la probabilidad de cualquier asignación de valores a un conjunto de variables se calcule mediante el correspondiente conjunto de probabilidades condicionales.

El método propuesto para encontrar todas las soluciones consiste en ir resolviendo una secuencia de sistemas de restricciones donde cada sistema se deriva del anterior en la secuencia.

El inconveniente que tiene este trabajo es que, al igual que el método de Pearl, asigna valores a todas las variables, incluidas aquellas que no son relevantes. Veremos más adelante cómo hay autores que resuelven este problema sobre el método propuesto por Santos.

❖ Otras mejoras a las deficiencias del método de Pearl y de Santos las desarrolló **Shimony** [54] para intentar resolver el problema de la *abducción parcial* y están basadas en una ligera variación de la definición de **explicación**; dicha modificación consiste en considerar las **explicaciones** como asignaciones parciales (asignaciones de valores a un subconjunto de variables) en las que a las variables irrelevantes no se les asigna ningún valor. Shimony propone tres modos de definir la irrelevancia, a partir de los conceptos de

- *independencia estadística*,
- *δ-independencia*, que es algo menos restrictivo que la anterior, e indica que un hecho es irrelevante si los hechos dados son independientes con una tolerancia de δ, y
- *cuasi-independencia,* que se basa en que el valor de una variable no interesa si no afecta "lo suficiente" a la probabilidad de la explicación. La idea que subyace bajo dicho concepto es que los antecesores sin ningún valor asignado de aquellos nodos *v* con un valor concreto dado, deben permanecer sin valor si no afectan a *v* (y, por tanto, no se pueden utilizar para explicar *v*). Además, considera que todos los nodos que no son antecesores de *v* no se deben utilizar para explicar *v*, pues no son posibles causas de *v*.

A partir de estas definiciones propone un modelo, llamado *MAP (Maximum A-Posteriori) basado en la irrelevancia,* que consiste en la búsqueda de explicaciones cuya probabilidad a posteriori, esto es, después de suministrar determinada evidencia, sea máxima. Con esta teoría pretende solucionar el problema del exceso de especificación que sufre el método de Pearl, al tener que asignar valores a todas variables, aunque sean irrelevantes, para generar una explicación. Para esta teoría propone cuatro casos particulares, dependiendo del criterio de irrelevancia utilizado: *MAPs basados en la independencia estadística*, *MAPs basados en la δ-independencia* y *MAPs basados en la cuasi-independencia.*

Además, demuestra las siguientes propiedades de dicho modelo:

- Las asignaciones parciales basadas en la independencia o en la δ-independencia se pueden reconocer en tiempo lineal.
- La probabilidad de una asignación parcial basada en la independencia se computa en tiempo lineal.
- Existen cotas razonables para computar la probabilidad de una asignación parcial basada en la δ-independencia en tiempo lineal.

Asimismo, propone algoritmos para generar explicaciones según los esquemas propuestos, es decir, un algoritmo MAP basado en la independencia estadística y otro basado en la δ-independencia.

Por último, Shimony propone esquemas alternativos para encontrar explicaciones basadas en la irrelevancia mediante la reducción de la red bayesiana a un sistema lineal de desigualdades, basándose en las ideas de Santos. La mejora consiste, básicamente, en la no asignación de valores a las variables irrelevantes.

❖ Un trabajo más reciente, pero en la misma línea que el anterior, es el trabajo conjunto de **Charniak** y **Shimony** [12]. Dichos autores han desarrollado un algoritmo para obtener la mayor probabilidad a posteriori (MAP) de las variables de una red bayesiana y se puede utilizar dicho método, por tanto, para encontrar las *k* explicaciones más probables en redes bayesianas.

Dicho algoritmo consiste en transformar la red bayesiana, formada únicamente por variables aleatorias discretas, junto con la evidencia proporcionada, en un grafo dirigido acíclico valorado por una función lógica (WBFDAG- *Weighted Boolean Function Directed Acyclic Graph*).

Un WBFDAG es una tupla (G, c, r, ε) donde:

- G es un grafo conexo dirigido acíclico con un conjunto de vértices V y un conjunto de aristas E.
- r es una función de V a F, llamada *etiqueta*. F es el conjunto de todas las funciones con dominio D'^k para algún k y rango D'. D' es el dominio de los valores de los nodos. Si un nodo v tiene k padres, el dominio de $r(v)$ es D'^k.
- c es una función de {V x D'} a los reales no negativos, llamada *función de coste*
- ε, la evidencia, es un par (s, d). s es un nodo sumidero en G y d es el valor de D' asignado a s.

El método para la construcción del consiguiente grafo valorado consiste, fundamentalmente, en crear nuevos nodos, similares a los de la red original pero con una función de coste asociada. Dicha función define, para cada nodo, un valor que depende de los valores de sus nodos padres y mide el coste de asignar un valor determinado a dicho nodo.

Estos autores demuestran que el problema de encontrar la asignación de valores a las variables de la red que genere la máxima probabilidad a posteriori es equivalente a encontrar la asignación de valores de mínimo coste del grafo WBFDAG. Este último problema se resuelve mediante la aplicación de un algoritmo de búsqueda en calidad (*best-first)* para encontrar la asignación de mínimo coste. La aplicación sucesiva de este algoritmo produce la enumeración de las asignaciones de valores a las variables, es decir, de las explicaciones, en orden decreciente de probabilidad.

El inconveniente que tienen estos últimos trabajos al presentar todas las posibles explicaciones, es el de la complejidad tanto computacional como conceptualmente, ya que al presentar todas las explicaciones se podrían producir incompatibilidades entre ellas.

❖ Otro trabajo relacionado con la inferencia causal en modelos basados en grafos dirigidos acíclicos con medidas de probabilidad asociadas, entre los que se incluyen las redes bayesianas, es el de **C. Meek** [43], quien amplia el trabajo realizado por Pearl, Verma y otros, dando repuesta a las siguientes cuestiones:

i) ¿existe una asignación de valores (explicación) consistente con cierto conocimiento previo K y que haga verosímil toda la evidencia observada?

ii) suponiendo que existen varias asignaciones, ¿cuáles son las relaciones causales comunes a todas ellas?

Para Meek, una explicación de un modelo de dependencias *M*, viene dada por un grafo dirigido acíclico G que define las mismas relaciones de independencia condicional que *M*. Un *modelo de dependencias* es una lista de sentencias que designan independencia condicional de la forma $A \perp\!\!\!\perp B | S$ (*A* es independiente de *B* dado *S*), donde *A*, *B* y *S* son subconjuntos disjuntos de V, siendo *V* el conjunto de vértices del grafo que representa el modelo.

Para este tipo de modelos, la representación del conocimiento previo *K* se realiza mediante un par <*F, R*>, siendo *F* el conjunto de arcos dirigidos prohibidos y *R* el conjunto de arcos dirigidos requeridos.

El algoritmo que propone para dar respuesta a la primera pregunta es una modificación de otro desarrollado previamente por Verma y Pearl y consta, a grandes rasgos, de los pasos siguientes:

1. examinar el modelo de dependencias para construir un grafo dirigido acíclico equivalente
2. extender dicho grafo con el conocimiento previo *K* utilizando, para ello, un conjunto de reglas de orientación para la extensión de dicho grafo.
3. generar un grafo dirigido acíclico que sea una extensión consistente del anterior.
4. decidir si este último grafo es una explicación, pues pudiera ser que no exista explicación para la evidencia introducida.

La solución a la segunda pregunta viene dada por un algoritmo que consta únicamente de las dos primeras fases anteriores pues el autor demuestra que las reglas de orientación que se utilizan en la fase 2 generan un grafo que cumple los requisitos buscados, es decir, que consta exclusivamente de las relaciones comunes a todas las explicaciones causales.

❖ Por último, hay que señalar que la búsqueda de la explicación o explicaciones más probables está íntimamente relacionado con el de la *inferencia abductiva* en redes bayesianas, que consiste básicamente en buscar las configuraciones de estados de mayor probabilidad para las variables no observadas, o para un subconjunto de ellas. Como este problema constituye por sí mismo una línea de investigación actual en el campo de las redes causales, conviene destacar aquí el trabajo de **J.A. Gámez** [31] quien, en su tesis doctoral, realiza un estudio detallado sobre el tema de la abducción, así como de los algoritmos desarrollados hasta ahora para llevar a cabo el proceso de la inferencia abductiva en redes bayesianas, con el fin de encontrar la o las explicaciones más probables según la evidencia observada.

Entre las aportaciones más relevantes de su trabajo están:

- el desarrollo de *algoritmos aproximados*, utilizando técnicas de enfriamiento estocástico y de algoritmos genéticos, para resolver el problema de la abducción parcial en los casos en los que los métodos exactos resultan ineficientes, y

- la propuesta de criterios para la *simplificación de explicaciones*, basados en usualidad, es decir, la simplificación consistiría en mostrar únicamente los valores que no coinciden con el valor "usual" a priori de las variables; y basados en independencia y relevancia.

5. 2. 2. Explicación como comprensión

❖ Uno de los primeros trabajos relacionados con la explicación en redes bayesianas, entendida ésta como la justificación de los resultados obtenidos por el sistema de modo que sean comprensibles para el usuario, fue el de **P. Sember** e **I. Zukerman** [53]. En dicho estudio, se centran en la explicación, a nivel micro, de una hipótesis definida sobre una variable dada, llamada *hipótesis focal*, y el cambio de probabilidad producido en el nodo que representa a dicha variable, después de propagar cierta evidencia.

Ya que la probabilidad de un nodo *b* viene dado por la fórmula

$$\text{Probabilidad}\,(b)=\alpha\lambda(b)\pi(b),$$

estos autores consideran la explicación como la justificación del resultado obtenido al combinar los cambios en los valores de los mensajes π y λ.

Además, para que una explicación sea convincente para el usuario, se deben determinar las expectativas del mismo. Dichas expectativas vienen caracterizadas por los cambios que se produzcan en π, pues representa la información causal. Así pues, dependiendo de si $\pi(b)$ aumenta, disminuye o no varía, el usuario esperará que la probabilidad de *b* aumente, disminuya o no varíe, respectivamente.

Por tanto, el algoritmo para la generación de una explicación identifica una expectativa y, si se alcanza, se genera la explicación; si no, se genera la explicación basada en identificar la causa de por qué no se ha obtenido dicha expectativa.

Un ejemplo de explicación producida en el caso de que se alcance una de las expectativas citadas sería:

La probabilidad de [...] ha aumentado debido al incremento del impacto de la evidencia correspondiente a las causas de [...].

Aunque el método parece sencillo e intuitivo, sin embargo tiene dos problemas fundamentales: que sólo sirve para poliárboles y que cuando el número de nodos crece, la complejidad del método también, ya que ésta es exponencial.

❖ El método propuesto por Sember y Zukerman tiene una gran limitación y es que para entender las explicaciones que se generan hay que tener conocimientos sobre teoría de la probabilidad. Para solucionar este defecto, surgieron posteriormente otro tipo de estudios relacionados con la generación de explicaciones mediante discursos en los que la expresión de los datos probabilísticos, en sistemas que utilizaban inferencia bayesiana, se realizaba en términos lingüísticos.

Uno de los autores dedicados a esta tarea es **C. Elsaesser.** En sus trabajos expone el argumento, estudiado por numerosos autores, de que se explica mejor el razonamiento seguido por el sistema mediante expresiones lingüísticas, ya que es la forma habitual de comunicación de las personas, que mediante datos numéricos probabilísticos.

Así pues, dicho autor diseña una técnica para generar explicaciones basada en la conversión de datos numéricos en expresiones lingüísticas (y viceversa), para que el usuario del sistema entienda mejor la explicación proporcionada sin tener conocimientos de probabilidad. En uno de sus artículos [28] proporciona un método consistente en generar explicaciones utilizando una plantilla basada en los modelos inductivos de Polya. Este autor, propuso un conjunto de reglas heurísticas para describir el cambio de credibilidad, llamadas *"patrones de inferencia plausible"*.

Por ejemplo, una de esas reglas, compatible además con la fórmula de Bayes, es:

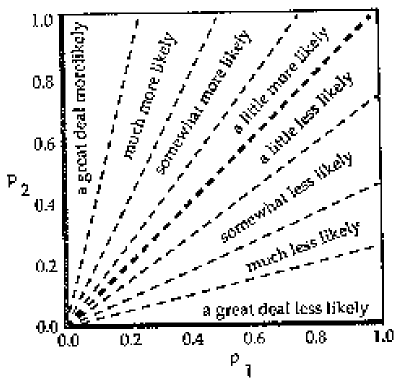
"Si $A \rightarrow B$ y B es verdadero, entonces la existencia de A es más creíble".

La plantilla se rellena con términos expresados en lenguaje natural que representan valores probabilísticos y sus correspondientes variaciones. Dicha plantilla es más o menos así:

> Basándonos sólo en *{información a priori}* es *{Frase-Prob [Priori(H)]}* que {H}
>
> Debido a que *{evidencia E}* es *{Frase-Prob [Priori(E)]}*,
>
> {H} es *{Frase_de_comparación [Posterior(H), Prior(H)]}*,
>
> haciendo que *{Frase-Prob [Posteriori(H)]}* se obtenga {H}

El conjunto de expresiones que denotan probabilidad, reflejadas en la plantilla anterior por *Frase-Prob*, varían entre las que definen probabilidades altas, como *Casi seguro*, para valores de probabilidad comprendidos entre 0,99 y 0,91; *Altamente probable*, para el rango 0,90 a 0,82; hasta las que designan valores de probabilidad pequeños, expresados por expresiones del tipo *Improbable*, para el rango 0,18 a 0,09; *Altamente improbable*, para el rango 0,08 a 0,01.

Las expresiones que definen cambios entre los valores de probabilidad a priori y a posteriori, que en la plantilla se denominan *Frase_de_comparación*, quedan definidas por la tabla siguiente[2], basada en el estudio de la razón entre las probabilidades a priori y a posteriori:



El método propuesto fue aplicado a un sistema de ayuda a la decisión para analizar los resultados de pruebas sobre mutación genética a corto plazo. Los resultados obtenidos fueron aceptables y efectivos para solucionar el problema de la confusión que crea la interpretación de la fórmula de Bayes.

[2] El estudio de las expresiones lingüísticas está hecho para la lengua inglesa, por lo que hemos preferido mostrarla tal y como la expone Elsaesser en su artículo; habría que hacer el correspondiente estudio para el castellano.

Un ejemplo de explicación que se obtiene en dicho sistema es:

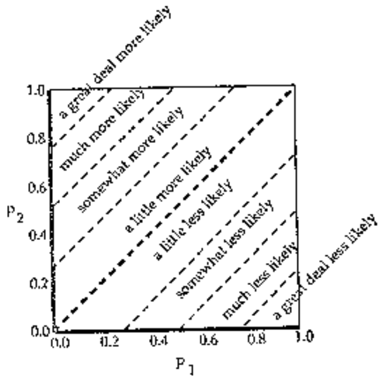
"Basándonos sólo en *su estructura* es *algo improbable (0,35)* que P345-22 sea un gen cancerígeno. Debido a que *un resultado positivo en el cambio de Sister-Chromatid* es *not quite an even chance (0,41)* para un pyrrolizidine, la hipótesis de que P345-22 sea un carcinogen es *un poco más probable (0,35 → 0,68)*, haciendo *algo probable* que P345-22 sea un carcinogen."

❖ En otro artículo [29], **Elsaesser** y **M. Henrion**, amplían el problema anterior y concretan algo más lo que ya había avanzado Elsaesser previamente, sobre la expresión de las variaciones de los valores de las probabilidades mediante términos lingüísticos. En concreto, realizan un análisis de distintos criterios para la selección de la frase que designa dichos cambios. Para ello definen una *función de selección de frases* como aquélla que, dadas dos probabilidades, *p1* y *p2*, devuelve una frase que relacione ambos datos. Esta función genera una partición del cuadrado unidad en regiones que corresponden a frases probabilísticas al igual que en el caso anterior.

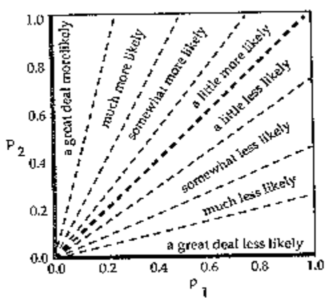
Los autores analizan tres modelos:

a) proporción constante de la probabilidad (*p2=K*p1*),
b) diferencia constante probabilística (*p2=p1+k*), y
c) proporción constante de las razones de probabilidad (*p1*(1-p2)=k*(1-p1)*p2)*),

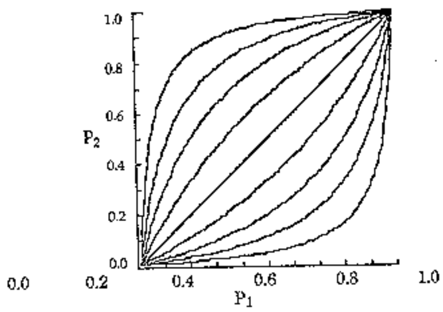
cada uno de los cuales genera una partición distinta del cuadrado unidad, tal y como se ilustra en la figura siguiente:



a) Proporción constante de las probabilidades



b) Diferencia constante de las probabilidades



c) Proporción constante de las razones de probab.

La conclusión obtenida es que el mejor criterio es el de la diferencia constante de las probabilidades. Para llegar a dicha conclusión, estudiaron los resultados obtenidos mediante el siguiente experimento:

Se seleccionaron aleatoriamente del cuadrado unidad pares de probabilidades, excluyendo los valores extremos (0 y 1). A cada una de las 25 personas que participaron en él, se les proporcionó 40 de esos pares; la mitad expresaba variaciones crecientes y la otra mitad, decrecientes. A continuación se les pidió que seleccionaran las frases, de entre 8 que se les ofrecían, que mejor expresaran los cambios entre los valores probabilísticos de cada par. Las 8 frases eran las siguientes (entre paréntesis, su equivalencia en castellano):

Great deal more (Muchísimo más)	*Great deal less* (Muchísimo menos)
Quite a bit more (Mucho más)	*Quite a bit less* (Mucho menos)
Somewhat more (Algo más)	*Somewhat less* (Algo menos)
Little more (Poco más)	*Little less* (Poco menos)

El análisis de los resultados permitió concluir que el 72% de dichos resultados coincidía con una partición similar a la de la diferencia constante de las probabilidades, por lo que se determinó éste como mejor criterio para la selección de frase.

❖ Sin embargo, **Druzdzel** en uno de sus artículos [23], además de realizar una revisión completa de la literatura existente sobre expresiones lingüísticas que expresan incertidumbre, proporciona argumentos suficientes en contra de la utilización en exclusiva (es decir, sin proporcionar a la vez información numérica) de este tipo de expresiones debido, fundamentalmente, a la gran confusión que se puede crear en la comunicación de los resultados, ya que la interpretación de dichas expresiones depende de factores completamente subjetivos.

❖ Este autor ha dedicado después su investigación al tema de la explicación en redes bayesianas; de hecho, es uno de los autores que más publicaciones tiene sobre este tema. Su tesis doctoral [22] es un trabajo muy completo, ya que además de realizar un análisis de los distintos métodos previos de explicación, sugiere distintas propuestas interesantes para el caso concreto de la explicación en redes bayesianas, que estudiaremos en este trabajo.

Además, ha dedicado parte de su trabajo a justificar el uso de redes bayesianas y, en general, diagramas de influencia en la construcción de sistemas expertos basados en la incertidumbre. Estas ideas las ha desarrollado tanto en su tesis como posteriormente en algunos artículos (p. ej. [19, 21, 26]) y en ellos demuestra la relación existente entre la teoría de la probabilidad y el modo en el que los humanos llevan a cabo procesos de razonamiento en situaciones donde existe incertidumbre, basándose en el concepto de causalidad.

Asimismo, justifica la representación de la información que contiene un sistema experto, mediante la descripción de las *relaciones causales* existentes entre los datos que conforman la base de conocimiento, lo que facilita la generación explicaciones en este tipo de modelos y que soluciona una de sus grandes carencias.

❖ Como he indicado antes, Druzdzel ha propuesto distintos métodos para generar explicaciones en redes bayesianas. Una de dichas propuestas es la realizada junto con **Henrion** sobre explicación basada en *escenarios*[3] [22, 24, 26, 38]. Para estos autores, un escenario es una asignación de valores a las variables relevantes[4] para la obtención de la conclusión, ordenadas de tal manera que forman una historia coherente, quizás causal, compatible con la evidencia.

Los trabajos desarrollados en esta línea de investigación sobre explicación basada en escenarios, se fundamentan en los resultados obtenidos a través de diversos estudios psicológicos, en los que se pone de manifiesto que los humanos tienden a interpretar los procesos en los que existe incertidumbre en términos de historias determinísticas y que generan la explicación de dichos procesos después de sopesar las historias más "creíbles" que incluyan la hipótesis a demostrar o sus complementarios.

Entre estos estudios destaca, por haber sido utilizado también como referencia por otros autores [34, 35, 36], el de **Pennington** y **Hastie** [49]. En concreto, su investigación se centra en sistemas basados en la explicación para la toma de decisiones judiciales, mediante la realización de dos experimentos:

- El primero de ellos consistió en obtener de distintos jurados las descripciones detalladas de su representación de la evidencia en un momento concreto. Para ello, partiendo de un

[3] En inglés, *scenarios.* Quizás se puede traducir mejor como "guiones", ya que subyace la idea de historia causal.

caso real sobre el juicio de un asesinato, se dividió el conjunto de jurados en grupos. Cada grupo estaba formado por los jurados que habían dictado el mismo veredicto. A cada grupo, se le presentó una red causal, llamada historia del veredicto, que incluía todos aquellos eventos y relaciones causales que aparecían en el 80% de las historias que cada jurado, por separado, había creado para emitir su veredicto. Además, se les presentó la relación completa de sucesos o pruebas que durante la celebración del juicio aportaron los testigos. Posteriormente, se les presentó un test con frases para que cada sujeto del jurado determinara si la reconocía o no como evidencia para la emisión del veredicto; en caso afirmativo, debían valorarla también respecto a la importancia en la decisión tomada. La conclusión que obtuvieron fue que los individuos evaluaban la evidencia construyendo espontáneamente una representación explicativa de los hechos.

- El segundo experimento consistió en manipular el orden de presentación de la evidencia para que las historias de defensa o acusación fueran más fáciles o difíciles de construir. Para ello, el conjunto de frases se subdividió en dos: uno que contenía las de culpabilidad y otro las de inocencia. Por cada grupo, se generaron dos historias: una que contenía los hechos en su orden temporal y otra cuyo orden era causal. Ambas historias se combinaron para obtener cuatro órdenes diferentes para los mismos hechos. Según las respuestas que proporcionaron los sujetos a las preguntas formuladas sobre los cuatro tipos de historias presentadas, las conclusiones que se obtuvieron fueron que la explicación causal impuesta a la historia es determinante para la elección de la decisión. Además, la coherencia de la historia explicativa y la fuerza de las historias alternativas son los principales factores que determinan la confianza en la decisión tomada.

En la última parte de dicho artículo destacan las principales conclusiones obtenidas, de los trabajos que estos investigadores vienen realizando. La primera es que el modelo de decisión basada en la explicación se confirma como método para la descripción de la decisión del jurado. Además, afirman que la construcción de la representación de la evidencia relevante a la decisión se retiene en la memoria y sirve como dato para el proceso de toma de decisiones. Además, la valoración de la certeza y la incertidumbre se pueden expresar con datos cuantitativos, pero se procesan durante el razonamiento como propiedades semánticas de la evidencia, tales como su completitud, consistencia, plausabilidad y unicidad.

---

[4] Más tarde, en la página 43, especificaremos el significado de relevancia.

❖ Volviendo al caso de la explicación basada en escenarios, el método que proponen Druzdzel y Henrion consiste en generar la explicación, a nivel macro, de una sola hipótesis dada cierta evidencia. Una *hipótesis* es la asignación de un valor a una variable discreta. El proceso para obtener la explicación es sencillo en su origen y consiste, básicamente, en generar todos los posibles escenarios, compatibles con la evidencia, para seleccionar el más probable entre ellos. Pero ya que sobre un modelo de n variables aleatorias binarias se pueden generar 2^n escenarios, se hace necesario restringir el proceso para aquellas variables de la red que son directa o indirectamente relevantes para la obtención de la hipótesis.

La definición del concepto de <u>relevancia</u> depende de la tarea que se vaya a llevar a cabo sobre la red:

- Si lo que se desea es obtener la probabilidad a posteriori de toda la red, los nodos relevantes son todos aquellos cuyas probabilidades dependan de la evidencia observada
- Si el objetivo es únicamente el cálculo de la probabilidad de una variable v, el conjunto de nodos relevantes son aquellos cuya distribución de probabilidad se utiliza para calcular la distribución a posteriori de dicha variable v.

Aun así, cualquier herramienta de explicación debería ser flexible a la hora de determinar el concepto de relevancia, ya que en último extremo debe ser el usuario del sistema el que decida qué datos son relevantes a la hora de incluirlos en la explicación.

Por tanto, para llevar a cabo dicha explicación, primero se seleccionan aquéllos nodos de la red que son relevantes para la hipótesis. Estos nodos se utilizan para generar escenarios mediante la enumeración de todas las combinaciones posibles de sus valores. Los métodos para la generación de escenarios son los citados en el apartado 5.2.1., es decir, todos aquellos dedicados, fundamentalmente, a obtener las k configuraciones más probables para los variables no observadas de la red después de introducir cierta evidencia.

Posteriormente, los escenarios se dividen en dos grupos: aquellos compatibles con la hipótesis (los que incluyen el valor particular de la hipótesis) y los que no lo son. Cada escenario se presenta al usuario con su probabilidad asociada.

La explicación consiste, por tanto, en la presentación de la siguiente información: la evidencia introducida a la red, los escenarios más probables compatibles con la hipótesis, los más probables incompatibles con la hipótesis y una comparación de las diferencias entre las probabilidades de los más importantes/probables pares de escenarios.

Por ejemplo, para la red de la figura 9, un ejemplo de explicación para *Tener-Catarro* después de haber observado la siguiente evidencia *E= {Estornudar, Marcas de animal, Ladrido}* sería:

Estornudar puede haber sido causado por *Tener-Catarro* o *Tener_Alergia.*
Las *Marcas_de_animal* pueden haber sido producidas por *Tener_Gatos* o por *Tener_Perros* u otra causa desconocida.
Los *Ladridos* pueden ser producidos por *Tener_Perros*

Los siguientes escenarios son compatibles con *Tener_Catarro:*

A. *Tener_Catarro* y No *Tener_Gatos* y por tanto No *Tener_Alergia*	0.38
B. *Tener_Catarro* y *Tener_Gatos* que causan *Tener_Alergia.*	0.05
C. Otros escenarios menos probables	0.01
Total probabilidad de *Tener_Catarro*	0.44

Los siguientes escenarios son incompatibles con *Tener_Catarro*

D. No *Tener_Catarro* y *Tener_Gatos* que producen *Tener_Alergia*	0.56

El escenario A es mucho más probable que el escenario B (0.38/ 0.05) porque No *Tener_Gatos* en A es muchísimo más probable que *Tener_Gatos* en B.
El escenario A es algo menos posible que el escenario C (0.38/ 0.56) porque *Tener_Catarro* en A es muchísimo menos probable que No *Tener_Catarro* en C (0.08/ 0.92) aunque No *Tener_Gatos* en A es muchísimo más probable que *Tener_Gatos* en C (0.90/ 0.10)
Por tanto, *Tener_Catarro* es menos posible que No *Tener_Catarro* (p=0.44)"

La última etapa que queda por mejorar sería la de la transformación del escenario a lenguaje natural.

El problema fundamental de este método es el de su complejidad para redes con muchos nodos y enlaces, o variables con un gran número de valores.

❖ Otro método propuesto también por **Druzdzel** y **Henrion**, pero en una línea distinta de las citadas hasta ahora, es la de la transformación de la red bayesiana en una red cualitativa para generar explicaciones sobre ésta.

Las redes probabilísticas cualitativas (QPN - *Qualitative Probabilistic Networks*) fueron introducidas por **Wellman** [61, 62]**.** En este tipo de redes, las relaciones entre los nodos se expresan en términos de influencia positiva, negativa, nula o desconocida. El interés de este tipo de redes radica en que, en general, la gente utiliza expresiones cualitativas para expresar las probabilidades, además de que, a la hora de reconstruir una historia para comprenderla, los datos se agrupan atendiendo a relaciones causales. Otro argumento a favor del uso de este tipo de redes es que un experto tiene que realizar menos esfuerzo para expresar su incertidumbre en términos cualitativos que cuantitativos. Un ejemplo de red cualitativa es la que se ilustra en la figura 11 siguiente[5]:

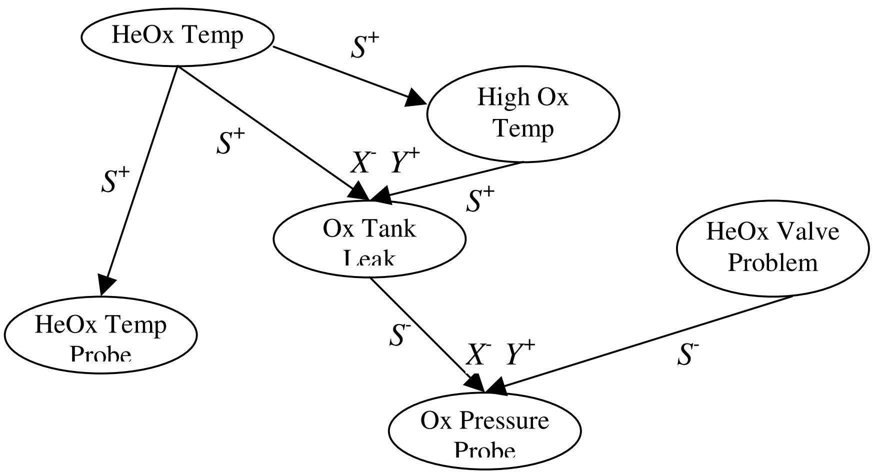


Fig. 11. Ejemplo de red cualitativa.

Los algoritmos de propagación de la evidencia en este tipo de redes han sido diversos: desde los iniciales de reducción del grafo, del propio Wellman, pasando por las mejoras realizadas por Wellman y Henrion [63], hasta las descritas por Druzdzel y Henrion [20, 21, 22, 25, 26, 38]. Sin embargo, lo que nos interesa estudiar aquí, de todos estos trabajos, es la utilización de redes cualitativas para la generación de explicaciones verbales, a nivel macro, en redes bayesianas.

[5] Las etiquetas S de los enlaces indican las influencias cualitativas entre cada par de nodos que une cada enlace. Las etiquetas X e Y describen los signos de las sinergias entre los nodos. El superíndice de cada etiqueta indica si la influencia o las sinergias, respectivamente, son positivas, negativas o cero.

El objetivo en este tipo de redes es la explicación del efecto cualitativo que se produce sobre la distribución de probabilidad de una única variable de interés *t,* al observar el estado de una variable *e,* es decir, al observar la evidencia *e*. Si la evidencia constara de varias variables, el proceso de propagación se aplicaría de forma sucesiva para cada una de ellas. Del mismo modo, si se está interesado en explicar varias variables, se puede generar una explicación para cada una de ellas.

Dicho proceso, en líneas generales, consta de tres fases: primero, a partir de la red bayesiana, de la evidencia seleccionada *e* y de la variable a observar *t*, se construye la red de influencia cualitativa, consistente únicamente en los caminos activos entre *e* y *t*. Es decir, se construye un grafo cuyos nodos serán, únicamente, las variables relevantes para la variable de interés *t*. Esto se hace mediante cualquiera de los métodos existentes para la eliminación de las variables irrelevantes[6] de la red.

A continuación, se realiza la propagación cualitativa de la evidencia sobre dicha red. El algoritmo de propagación cualitativa que proponen Druzdzel y Henrion es una mejora del propuesto por Wellman, basado en la reducción del grafo, y consiste en recorrer la red para marcar todos los nodos con el signo de su impacto en la red, que puede ser ***0***, **+**, **–** ó ***?***. Dicho algoritmo se basa en el paso local de mensajes. Inicialmente, todos los nodos están marcados a 0. El nodo de la evidencia envía el primer mensaje a todos sus vecinos (nodos del *manto de Markov*[7]); el signo de cada mensaje viene determinado por el producto del signo del nodo previo y el signo del enlace que lo atraviesa. Cada nodo sólo se visita una vez. Por otra parte, cada nodo que recibe un mensaje actualiza su signo mediante la suma del signo que tiene y el del mensaje que recibe. Después, comprueba si tiene algún vecino (del manto de Markov) que no ha sido visitado todavía por el mensaje que él envía. Si es así, pasa el mensaje a dicho vecino; si no quedan vecinos por visitar, la propagación del mensaje finaliza.

Finalmente, una vez actualizados todos los nodos con el signo del cambio de creencia, el grafo contendrá toda la información sobre la propagación de la influencia cualitativa entre la evidencia *e* y la variable de interés *t*. Dependiendo del signo final de la variable *t*, se genera la

[6] Druzdzel propone en su tesis [22] seis fases para la eliminación de nodos irrelevantes: propagación de las relaciones deterministas, eliminación de los nodos estadísticamente irrelevantes, eliminación de los nodos computacionalmente irrelevantes en cuanto a estructura, eliminación de los nodos computacionalmente irrelevantes según su distribución, eliminación de los nodos antecedentes pasivos y, por último, eliminación de los nodos relevantes marginalmente. Para cada fase, propone algoritmos para llevar a cabo el proceso de eliminación correspondiente.

[7] El manto de Markov de un nodo *n* lo forman los padres de *n*, los hijos de *n* y los padres de éstos.

explicación a partir de una lista de inferencias cualitativas, junto con un resumen de las conclusiones.

Hay tres tipos de explicaciones elementales correspondientes a los tres tipos de inferencias cualitativas elementales basadas en la propagación cualitativa de la creencia, ilustradas en la fig. 12:

(a) *inferencia predictiva,* en el sentido del arco en la red; la explicación que se genera es del tipo: "*a* puede causar *b*".

(b) *inferencia abductiva,* en el sentido opuesto al del arco; la explicación que se genera es del tipo: "*b* es evidencia para *a* ".

(c) *inferencia intercausal,* que proporciona el impacto cualitativo de la evidencia para una variable *a* sobre otra variable *b* cuando ambas influyen sobre una tercera variable *c*, sobre la que existe evidencia independiente (por ejemplo, *c* ha sido observada); la explicación que se genera en este caso es del tipo: "*a* y *b* pueden causar cada una *c*; como *a* explica *c*, es evidencia para descartar *b*".

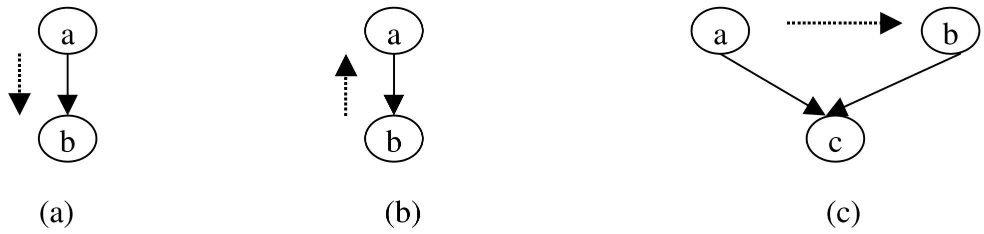


Fig. 12. Los tres tipos de inferencia cualitativa.

En principio, para los tres ejemplos de la figura 12, la generación de la explicación parece intuitiva y sencilla. Las dificultades surgen al intentar generar la explicación sobre un grafo múltiplemente conexo, ya que tanto el razonamiento humano como el lenguaje natural son esencialmente secuenciales. Por tanto, el problema que hay que resolver es el de la secuenciación del grafo de influencia. Druzdzel propone un método para ir generando la explicación por etapas entre los nodos situados entre la evidencia y la variable de interés. En cada momento, se elige un nodo que tenga al menos dos arcos incidentes y, si hay más de uno, se elige el que esté más cerca a la evidencia (en cuanto al número de arcos del camino más corto).

Por ejemplo, para la red causal de la figura 11, tendríamos la siguiente red después de realizar la propagación cualitativa, al observar como evidencia *HeOx Temp Probe*[8]. (En cada nodo, al lado de su identificador, está el signo del impacto de dicho nodo sobre la red)

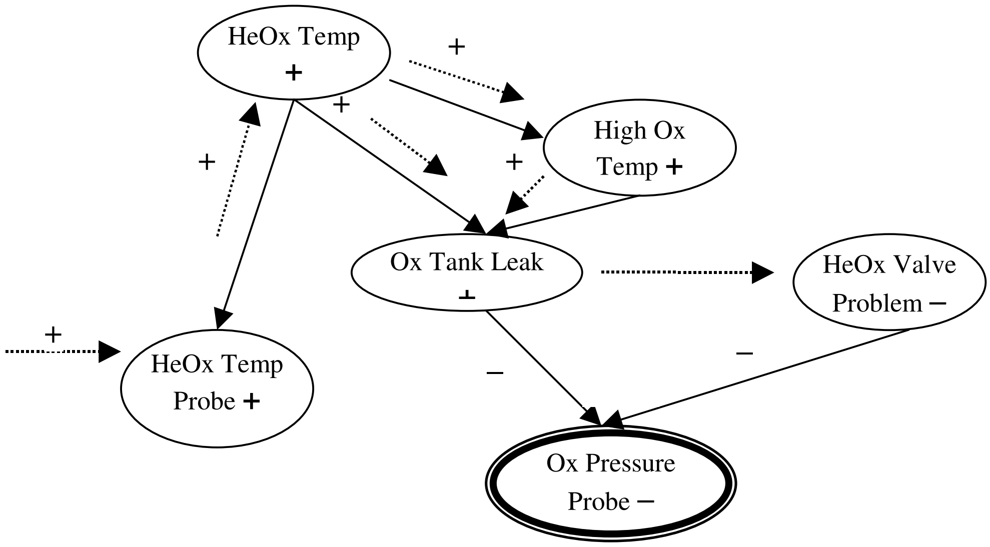


Fig. 13. Ejemplo de red cualitativa

La explicación que se generaría para explicar *HeOx-Valve-Problem* sería:

HeOx-Temp-Probe es evidencia para *HeOx-Temp*.

HeOx-Temp puede causar *High-Ox-Temp* y *Ox-Tank-Leak.*

High-Ox-Temp puede causar *Ox-Tank-Leak*.

Dado que *Ox-Tank-Pressure* es bajo, *Ox-Tank-Leak* es evidencia en contra de *HeOx-Valve-Problem.* Por tanto, la evidencia observada *HeOx-Temp-Probe* descarta *HeOx-Valve-Problem*

❖ Entre los autores que se han dedicado al tema de la explicación en redes bayesianas, uno de los más relevantes es **Suermondt** [57, 58]. La principal idea que aporta es la de que se pueden explicar los resultados de la inferencia realizada en una red bayesiana mediante la determinación de el impacto de los hallazgos sobre una variable determinada. Por tanto, su método INSITE para generar explicaciones, tanto a nivel micro como a nivel macro, consiste básicamente en identificar aquellos hallazgos que influyen en la hipótesis, además de los caminos por los que fluye dicha evidencia.

[8] He preferido dejar los identificadores de los nodos en inglés, tal y como aparecen en el trabajo de Druzdzel, ya que es el ejemplo que él propone en su tesis y en los artículos citados para ilustrar la generación de la explicación mediante la red cualitativa equivalente.

En concreto, el método de explicación pretende dar respuesta a la siguiente pregunta: *¿Por qué la evidencia E afecta a la probabilidad de la variable D?*. Para responderla, estudia los cambios en la distribución de probabilidad de la variable en cuestión mediante el análisis de los valores de la función de coste de *entropía cruzada*[9]:

$$H(P(D \mid E); P(D)) = \sum_i \left[p(d_i \mid E) \log \left[\frac{p(d_i \mid E)}{p(d_i)} \right] \right]$$

El método INSITE desarrollado proporciona explicaciones de dos áreas diferenciadas:

- la de la influencia de la evidencia sobre una variable de interés en el que, mediante un análisis de sensibilidad, trata de determinar cómo influyen uno o varios hallazgos sobre dicha variable, también llamada *hipótesis*. Para llevar a cabo dicho análisis, en el primer caso, se basa en el cálculo del *coste de omisión* de cada hallazgo E_i individualmente, que viene determinado por la fórmula:

$$H^{-}(E_I) = H(P(D \mid E); P(D \mid E \backslash E_I))$$

Así, si $H^{-}(E_i)$ tiene un valor significativo[10], se puede concluir que E_i juega un papel sustancial en la explicación del resultado de la inferencia. Por el contrario, si tiene un valor insignificante, no se puede concluir que no tenga efecto sobre el resultado final.

$$H^{-}(F) = H(P(D \mid E); P(D \mid E \backslash F))$$

Para analizar cómo influyen un subconjunto de hallazgos F sobre la hipótesis se calcula el coste de omisión conjunto mediante la fórmula similar a la anterior:

Además, desarrolla una técnica para la detección de conflictos entre los datos que forman la evidencia, mediante la comparación del efecto de la eliminación de un hallazgo con el efecto obtenido al eliminar todos los hallazgos. De este modo, si un hallazgo E_i no está en conflicto con el resto de los que forman la evidencia E se tendrá que

$$H^{-}(E_I) \leq H^{-}(E)$$

y

[9] Suermondt dedica un capítulo de su tesis a analizar distintas funciones de coste y a justificar la elección de la función de entropía cruzada como mejor función de coste para el método que él propone.

[10] El concepto de "insignificancia" hace referencia a valores menores que una cantidad θ que, o bien la puede determinar el usuario o si no, Suermondt la define por defecto en su sistema INSITE con el valor 0.0001.

$$Dir(P(D \mid E); P(D \mid E \setminus E_i)) = Dir(P(D \mid E); P(D)),$$

siendo $Dir(P(D); P'(D)) = (dir_1, ..., dir_n)$, donde dir_i es el signo de $p'(d_i)-p(d_i)$, es decir, del cambio de probabilidad, y los parámetros d_i son los valores que toma la variable D.

- la explicación de las <u>cadenas de razonamiento</u> entre la evidencia E y la hipótesis D y que están formadas por los nodos que forman parte del camino (o de los caminos) entre E y D y que están computacionalmente relacionados con la hipótesis D dada la evidencia E. Para ello, se siguen los siguientes pasos:
  - Determinar qué nodos de la red son relevantes en la transmisión de la evidencia *E* a la variable *D*.
  - A continuación, eliminar los nodos que no forman parte de ningún camino entre *E* y *D*.
  - Generar todos los caminos entre *E* y *D*.
  - Marcar todos los nodos que son *nudos* en el conjunto de caminos anterior y eliminar las cadenas duplicadas entre ellos. Los nudos son los nodos que separan direccionalmente la evidencia *E* de la hipótesis *D* y que forman parte de los caminos entre *E* y *D*.
  - Eliminar los caminos entre los que la transmisión de evidencia es insignificante.
  - Eliminar arcos estrátegicamente para determinar, del resto de los caminos, cuáles contribuyen al resultado final y cuáles no.
  - Determinar los efectos de la eliminación de cada arco en los caminos que influyen en el resultado final para presentar, paso por paso, las cadenas de razonamiento entre E_i y D.

La aportación más importante del trabajo realizado por Suermondt es que la metodología INSITE desarrollada es independiente del algoritmo de propagación utilizado en la red.

Finalmente, el método lo ha aplicado a la red bayesiana ALARM[11] en el que presenta un interfaz de usuario gráfico combinado con texto que pretende ayudar al usuario a comprender los datos probabilísticos resultantes de las inferencias realizadas en la red.

❖ Basándose en el trabajo de Suermondt, **Haddawy** y colaboradores [34, 35, 36] han diseñado BANTER, una herramienta educativa que presenta la información que contiene una red bayesiana de forma comprensible para el usuario de la misma. Además, está diseñada para ayudar al usuario en la evaluación de hipótesis y en la selección de los diagnósticos óptimos, ya que se pueden realizar varios tipos de preguntas, solicitar las mejores pruebas que pueden diagnosticar o descartar una enfermedad, comprobar su capacidad para seleccionar diagnósticos óptimos y pedir la explicación del razonamiento seguido en la red.

El método utilizado para determinar las mejores pruebas que diagnostican o descartan la enfermedad es el siguiente: se instancian positiva y negativamente cada uno de los nodos que representan dichas pruebas y se calcula la probabilidad a posteriori de la enfermedad, dependiendo del valor de la prueba. La prueba que mejor diagnostica la enfermedad será la que mayor probabilidad a posteriori genere y la que mejor la descarta será la que genere el menor valor.

El algoritmo para la generación de la explicación está basado en el método citado anteriormente de Suermondt, aunque con algunas modificaciones. Al igual que INSITE, consta de dos procedimientos diferenciados:

- el primero, identifica las partes de la evidencia que más influencia tienen en la probabilidad de una hipótesis dada. Para llevarlo a cabo, se eliminan todos los hallazgos de la red y se va instanciando cada uno de ellos individualmente y almacenando la probabilidad a posteriori de la hipótesis.

[11] Beinlich, I.A., Gaba, D.M., The ALARM monitoring system: Intelligent decision making under uncertainty. Anesthesiology, (1989), 71 (3A), A337

- el segundo, para la identificación de los caminos de influencia, selecciona los caminos entre los que se propaga la evidencia basándose en el criterio de separación direccional y de ellos selecciona los 5 caminos más fuertes y menos largos. Para obtener la "fuerza" de un camino se calcula, para cada uno de sus nodos, el valor absoluto de la diferencia entre su probabilidad a priori y su probabilidad condicionada a la evidencia. La menor de estas diferencias determina la fuerza del camino.

❖ **Madigan** y colaboradores [42] proponen un método gráfico para la visualización de las explicaciones generadas por una red bayesiana, que consiste en visualizar cómo se propaga la evidencia a través de la red. Para ello, lo que hacen es medir el impacto de la evidencia sobre cada nodo de la red mediante el "peso de la evidencia" propuesto por Good [33]

$$W(H{:}E) = \log\left(Pr(E|H) \,/\, Pr(E|\neg H)\right)$$

y utilizan la representación gráfica de la red como contexto para describir los flujos de evidencia a través de la misma.

En su sistema, orientado a predecir el resultado del tratamiento ESWL (*Extracorporeal Shock Wave Lithrotipsy*) para la eliminación de cálculos en el riñón, el usuario puede encontrar explicación a dos tipos de preguntas:

- *¿Cuál es la importancia relativa de cada uno de los hallazgos que forman la evidencia sobre la variable de interés?* Esta pregunta la responde la máquina mediante la visualización de los pesos de evidencia que cada hallazgo proporciona a la hipótesis.
- El otro tipo de pregunta que el usuario puede realizar es: *¿Por qué un nodo concreto tiene tanta influencia sobre dicha variable?*. En este caso, la idea es similar a la de los "caminos de evidencia" propuestos por Suermondt, y lo que se hace es mostrar el por el que fluye la evidencia.

La diferencia fundamental frente a otro tipo de métodos radica en que, además de utilizar distintas funciones para medir la influencia de la evidencia sobre los nodos de la red, las explicaciones se generan de modo gráfico. Además, proporcionan distintos tipos de explicación, desde el simple coloreado de un nodo a pantallas más detalladas.

❖ En la **actualidad**, existen distintos grupos de investigación que trabajan en el desarrollo de nuevos métodos de explicación que intenten solucionar los problemas relacionados con la complejidad de los que hemos estudiado, ya que en la mayoría de ellos es exponencial, lo que los hace inadecuados para el caso de las redes con un gran número de nodos y/o enlaces. Entre los proyectos que se están llevando a cabo, merece la pena destacar los siguientes:

- Las investigaciones actuales de **Zukerman** y colaboradores, pues suponen una novedad con respecto a lo que se ha venido haciendo hasta ahora, que radica en el uso de redes bayesianas para la generación de argumentos en lenguaje natural. En sus últimos trabajos, elaborados junto con **Conachy** y **Korb** [41, 67, 68], proponen la utilización de redes bayesianas como base del sistema NAG (*Nice Argument Generator*) para el análisis y generación de razonamientos que sean "suficientemente buenos" para el usuario, es decir, que sean convincentes . Dada una proposición objetivo, introducida en el sistema por el usuario, y el contexto en el que se realiza dicha proposición, junto con el grado de credibilidad que se desee alcanzar, el sistema genera los argumentos que la justifican; además, analiza aquéllos razonamientos que proporciona por el usuario y prepara otros para refutarlos cuando sea necesario.

  La arquitectura básica de dicho sistema se ilustra en la figura 14:

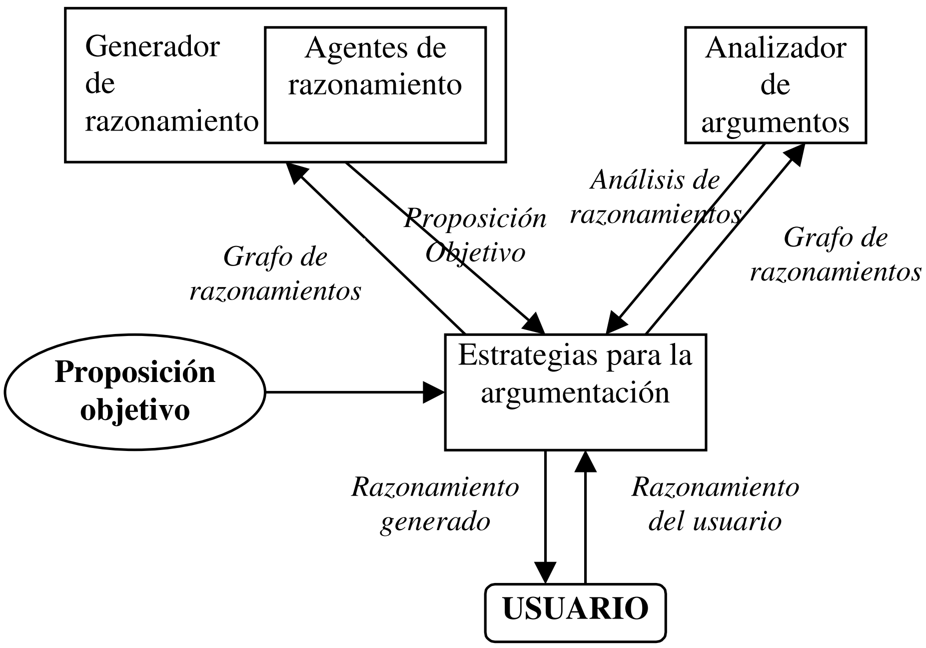


Fig. 14. Arquitectura del sistema NAG propuesto en [67]

Las bases de conocimiento que representan la información en la que se basa el sistema para la construcción y análisis de los razonamientos son: el modelo del usuario, para reflejar sus creencias y sus conocimientos, y el modelo normativo, para generar y evaluar los razonamientos; ambas se representan por medio de redes bayesianas. Los nuevos argumentos que se elaboran, se van añadiendo al modelo normativo. Ya que dichas redes pueden ser demasiado complejas, se utiliza un mecanismo para, en ambos modelos, centrar la atención en las subredes relevantes sobre las que se generará el razonamiento. La intersección de las estructuras de dichas subredes forma el grafo de razonamientos. Para analizar dicho grafo se realiza la propagación sobre las redes bayesianas que representan el modelo del usuario y el modelo normativo, utilizando la información probabilística asociada a cada modelo.

- Otros trabajos que merece la pena citar son:

  - Los que se desarrollan en el Centro Asesor de Agricultura Danesa (Danish Agricultural Advisory Centre), entre los que destacan los dedicados al análisis de herramientas para la generación de explicaciones en redes bayesianas [16] y que han aplicado a la red bayesiana BOBLO para determinar si el parentesco asignado al ganado es el correcto. En concreto, en dicho sistema aplican el *método HUGIN* para la detección de conflictos entre los hallazgos que forman la evidencia, y el *análisis de sensibilidad* para determinar cómo afecta la variación de la evidencia a la conclusión.

  - Proyecto ARPI para la evaluación y explicación de planes representados por medio de redes causales. Este proyecto se desarrolla en el Laboratorio de Palo Alto del Centro Científico Internacional de Rockwell y, como parte del mismo, se intentan están creando nuevos algoritmos para explicación en redes causales [18]

6. CONCLUSIONES

Después del estudio realizado, podemos destacar las siguientes conclusiones:

1. Objetivo de la explicación en redes bayesianas

Las dos grandes críticas realizadas a los sistemas expertos basados en redes bayesianas desde su origen, han sido:

1. la complejidad de los algoritmos que realizan inferencia probabilística es computacionalmente intratable en el peor de los casos, y
2. la teoría de la probabilidad es incomprensible para todos aquellos que no tengan un conocimiento profundo del tema.

Y así como el primero de los problemas se ha solucionado de modo efectivo mediante la definición de nuevos algoritmos de razonamiento, tanto exactos como aproximados, para el segundo, aunque existen soluciones parciales, todavía no se ha creado ningún método definitivo que sea capaz de dotar de significado a los datos probabilísticos y, por tanto, que pueda convencer al usuario de los resultados obtenidos.

Así pues, el objetivo fundamental del desarrollo de un método de explicación para este tipo de sistemas expertos es el de hacer comprender al usuario dichos datos probabilísticos, con el fin de ayudarle, tanto a encontrar errores en el modelo, como a profundizar en el conocimiento del modelo y así facilitarle la tarea de la toma de decisiones, cuando así lo requiera.

2. Definición de explicación

Aunque no existe unanimidad de criterios a la hora de dar un significado concreto del término "explicación", el estudio de sus múltiples acepciones y el análisis de los objetivos planteados, nos permite obtener como conclusión que explicar consiste en *realizar una exposición de algo, de manera que sea* ***comprensible*** *para quien la está recibiendo, es decir, que el receptor mejore su conocimiento sobre el objeto de la explicación, y* ***satisfactoria****, en el sentido de que cubra las expectativas planteadas previamente.*

3. Características de la explicación.

Existen ciertas características comunes a todos los métodos de explicación, con independencia del tipo de sistema experto sobre el que se desee definir. Estas características son:

a) Qué debe proporcionar una herramienta de explicación

Entre las facilidades que debe proporcionar una herramienta de explicación, está la de ser capaz de justificar:

⇒ la representación del modelo, para tener la posibilidad de detectar errores durante el proceso de construcción del propio sistema, además de ayudar al usuario a afianzar su propio conocimiento.

⇒ el proceso de razonamiento seguido, tanto a nivel micro, como a nivel macro.

⇒ la no obtención de una conclusión esperada para, por una parte, poder ratificar si la representación del modelo está bien hecha, y por otra, mejorar o educar el conocimiento del usuario.

b) Modelo del usuario

Aunque la definición de tal clasificación puede involucrar un proceso tanto o más complejo que el que estamos intentando solucionar, varios autores convienen en distinguir, básicamente, dos tipos de usuario:

- **usuario experto,** caracterizado por tener un nivel muy alto de conocimiento del dominio que representa el sistema. Por ejemplo, en el caso de los sistemas expertos para medicina, el usuario experto sería el médico.

- **usuario inexperto,** quien por el contrario, no tiene tanta información sobre la base de conocimiento. Por ejemplo, siguiendo con el ejemplo de la medicina, el usuario inexperto sería el paciente al que hay que diagnosticar una enfermedad, o el A.T.S., a quien hay que darle instrucciones sobre el tratamiento y la medicación a dicho paciente, etc.

c) Nivel de detalle proporcionado

Dependiendo de la clase de explicación que se desee generar, el nivel de detalle de la misma variará considerablemente, ya que no será igual la explicación proporcionada al médico que utiliza un sistema experto para diagnóstico, por ejemplo, en el que la terminología sea muy específica, que la emitida al paciente, al que le interesa que sea lo más parecido al lenguaje natural en el que habitualmente se comunica y no contenga demasiados tecnicismos.

Básicamente, dicho nivel de detalle dependerá del tipo de usuario a quien vaya dirigida dicha explicación por lo que esta cuestión no tiene una solución sencilla, ya que sería casi imposible tener en cuenta los grados de conocimiento de los distintos tipos de usuarios a los que puede ir dirigida una explicación. Sin embargo, creo que una solución podría consistir en establecer por defecto un nivel de detalle medio, y ofrecer al usuario la posibilidad de que, durante el proceso de generación de dicha explicación, él mismo decida si desea una explicación más detallada o menos.

d) Método Intercomunicativo

Ya que el objetivo de las herramientas de explicación es el de convencer al usuario y que éste comprenda los resultados obtenidos, se hace necesario ofrecer la capacidad de comunicación entre el sistema y el usuario y viceversa.

Una herramienta intercomunicativa es útil para el usuario porque le puede ayudar tanto en la utilización del sistema experto como para llevar a cabo su propia formación. Y además, es útil para el sistema, pues puede ir actualizando el conocimiento del usuario durante el proceso de funcionamiento del sistema y de generación de la propia explicación.

Por tanto, sería deseable que los nuevos métodos desarrollados tengan la capacidad de <u>adaptación</u> y permitan, en cada momento, mejorar y actualizar el conocimiento que poseen del usuario. Esto permitiría, además, ajustar el nivel de detalle de la explicación según las exigencias concretas que, en cada momento, requiera el receptor de dicha explicación.

La forma de llevar a cabo esta comunicación será mediante distintas facilidades que la herramienta debe proporcionar, que son: hacer preguntas tanto el usuario al sistema, como a la inversa; la posibilidad de realizar interrupciones en cualquier momento; solicitar explicaciones más detalladas, etc.

e) "Utilidad" de las conclusiones

Como ya hemos indicado en varias ocasiones en este trabajo, existen conclusiones poco probables pero muy útiles para el usuario, por el beneficio que pueden proporcionar. Por ejemplo, piénsese en el efecto psicológico que supone descartar una enfermedad aunque sea muy poco probable (sida, cáncer, ...). Sin embargo, las redes bayesianas no contemplan ni los nodos de utilidad ni los de decisión. En principio, nuestro trabajo se centrará exclusivamente en el caso concreto de las redes bayesianas para, en el futuro, ampliar el método para diagramas de influencia.

3. Explicación del modelo.

Además, de las propiedades generales del método de explicación a desarrollar, la explicación del modelo posee características propias relacionadas, sobre todo, con la forma de presentar dicha explicación y que detallo a continuación:

Explicación verbal: Como hemos visto que en determinadas ocasiones se entienden mejor los datos expresados de forma cualitativa que cuantitativamente, se debe facilitar que el usuario elija, entre ambas, qué tipo de explicación prefiere. En cualquier caso, existen dos posibilidades: generar la red cualitativa y explicarla, tal y como propone Druzdzel [22] o bien sería deseable tener realizada una clasificación de los datos que intervienen en la red, como sugiere Díez [15].
Asimismo, se debería igualmente permitirse generar explicaciones a nivel micro, como plantea Druzdzel [22].

Explicación gráfica: Consistiría, básicamente, en mostrar el grafo que define la red. Si fuera demasiado grande, dicha presentación se podría hacer por grupos de nodos. Por ejemplo, si fuera una red sobre algún caso médico, se pueden agrupar por síntomas, enfermedades, pruebas, etc. También se puede dotar a la herramienta de la facilidad de representación de gráficas que muestren las relaciones entre dos variables introducidas por el usuario, etc.

4. Explicación del razonamiento

Del mismo modo, el proceso de explicación del razonamiento tiene también características específicas que vamos a analizar a continuación:

a) Número de hipótesis

Respecto al número de hipótesis presentadas al usuario, los métodos que proponen solamente la muestra de los diagnósticos más probables, como por ejemplo, el de Reggia y Perricone [50] o el de Haddawy, Jacobson y Kahn [34, 35, 36], nos parece que no son adecuados por la razón de que para que el usuario pueda emitir un diagnóstico, tiene que sopesar personalmente (ya que hemos dicho que no nos vamos a ocupar del caso de diagramas de influencia) otros factores, como el del posible beneficio que se obtiene al descartar un diagnóstico. Piénsese en el caso de los sistemas expertos médicos en los que a lo mejor hay enfermedades poco probables, pero cuya detección, con pruebas sencillas, supone salvar una vida humana. Por otra parte, como el proceso de generación de todas las hipótesis posibles puede ser enormemente complejo, debería ser el usuario el que introduzca su propio *factor de confianza* o *umbral de probabilidad*, es decir, la probabilidad mínima de las hipótesis que desea conocer. Además, en el caso de la abducción, antes de llevar a cabo la consiguiente propagación, se le debe proporcionar el tiempo estimado para obtener todo el conjunto de hipótesis que busca, para que en caso de que sea demasiado, pueda modificar el factor de confianza.

b) Lenguaje utilizado

El lenguaje utilizado para la explicación debe ser una combinación de lenguaje natural y datos numéricos, ya que estoy de acuerdo con las ideas expuestas en los artículos de Druzdzel [22, 23], Elsaesser[28] y Elsaesser y Henrion [29] de que los humanos cometen demasiados fallos al interpretar datos numéricos probabilísticos y de que entienden mejor expresiones lingüísticas. Sin embargo, no nos olvidemos de que, por muchos estudios que se hagan para intentar relacionar dichas expresiones lingüísticas subjetivas con datos numéricos objetivos, no se podrá llegar a obtener una correspondencia exacta entre ambas, ya que la interpretación de datos numéricos siempre depende del contexto en el que se estén evaluando y de las apreciaciones subjetivas del usuario del sistema. Por tanto, aunque la línea de investigación me parece interesante, sin duda no me parece la más indicada a seguir por la imposibilidad de establecer criterios objetivos, que creo que es de lo que se trata. Para trabajar con términos "difusos" ya existen otras disciplinas. Estaríamos más por la propuesta de Druzdzel de que, en cualquier caso, aparte de dar una medida "subjetiva" que exprese los cambios en la probabilidad sufridos por determinadas variables, habría que mostrar al lado los valores numéricos correspondientes. Por tanto, dejaría de tener sentido gastar esfuerzos en intentar calificar dichos valores.

c) Nivel de explicación

Creo que no hay que limitar la herramienta a desarrollar a un único nivel de explicación, sino que a petición del usuario, debe ser capaz de generar explicaciones tanto a nivel micro como a nivel macro.

d) Presentación de los datos

En cuanto a la presentación de los datos, el método de explicación debe permitir tanto mostrar solamente de los datos más relevantes que influyen en el diagnóstico, como también, en un momento dado, sobre todo a la hora de generar explicaciones a nivel micro, presentar todos aquellos factores que influyen en la hipótesis. Además, en ambos casos sería deseable que se ofrecieran tanto los datos que tienen *influencia positiva* sobre la hipótesis, es decir, los que contribuyen a obtener el diagnóstico final, como los que la tienen *negativa*, es decir, los que harían descartar o rechazar el resultado final. En ambos casos, se debería mostrar también en qué medida influyen sobre el diagnóstico.

e) Explicación de los diagnósticos descartados

Además, la herramienta de explicación también debe poder responder a preguntas del tipo: *¿Por qué la evidencia E* no *produce el diagnóstico D?* Este tipo de preguntas ayudará al usuario directo a convencerse del diagnóstico que propone el sistema, aparte de que serán útiles también a la hora de detectar errores en el modelo o de realizar posibles mejoras del mismo.

f) Datos con valor desconocido.

Asimismo, la herramienta de explicación debe incluir algunas de las mejoras que propone Suermondt para su método INSITE [58]:

- ha de ser capaz de estudiar el caso de los datos con valor desconocido y ver de qué modo puede afectar la asignación de un valor a estos datos a la hipótesis proporcionada; es decir, que fuera permitiera realizar razonamiento hipotético para así poder predecir el valor de un hallazgo.
- además, debería ofrecer la posibilidad de recomendar la mejor prueba entre todas las que existan representadas en la red y que sirvan para diagnosticar o descartar un valor de la hipótesis.

g) Fases en la explicación

Los métodos de Suermondt y Druzdzel son bastante interesantes, aunque tienen la limitación del tamaño de la red, ya que la complejidad computacional es exponencial, por lo que si se trata de diseñar un método que sea independiente del número de nodos y de enlaces, habría que estudiar la forma de mejorarlos para que fueran eficaces también para el caso de redes "grandes". Para ello, existen dos posibilidades:

- La primera, consistiría en realizar un proceso de abducción (total o parcial) mediante el que se obtienen los valores de las variables que justifican o descartan la hipótesis a considerar, simplificar el resultado y traducirlo posteriormente en un discurso expresado de cualquiera de las formas vistas en el apartado 3.4. La explicación generada, en este caso, correspondería al nivel macro.

- La segunda, consistiría en estudiar la propagación de la evidencia, tanto a nivel micro como a nivel macro.

En cualquier caso, y para finalizar, el método de explicación debe poseer las características que he citado en este último apartado. Por tanto, el trabajo a realizar ahora consistirá en intentar estudiar de qué forma se pueden mejorar los métodos de Suermondt, Druzdzel y Zukerman para poder diseñar una herramienta que posea lo mejor de cada uno de ellos y que, además, reúna todas las propiedades deseables para un método de explicación en redes bayesianas, independientemente de su tamaño.

7. BIBLIOGRAFÍA

[1] Binsted, K., Cawsey, A., y Jones, R., "Generating personalized patient information using the medical record", *AI in Patient Education Workshop (AIPE),* Glasgow, RU, Agosto, 1995.

[2] Bleich, H.L., "Computer-based consultation: Electrolyte and acid-base disorders", *American Journal of Medicine*, **53** (1972) 285-291.

[3] Buchanan, B.G., y Shortliffe, E.H. (Eds.), *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project,* Addison-Wesley, Reading, MA, 1984.

[4] Carenini, G., Mittal, V., y Moore, J., "Generating patient-specific interactive natural language explanations", *Proceedings of 18th Symposium on Computer Applications in Medical Care (SCAMC 94)*, McGraw-Hill, 1994.

[5] Castillo, E., Gutiérrez, J.M., y Hadi, A.S, *Sistemas Expertos y Modelos de Redes Probabilísticas,* Academia de Ingeniería, Santander, 1996.

[6] Cawsey, A., Galliers, J., Logan, B., Reece, S., y Sparck Jones, K., "Revising beliefs and intentions: A unified framework for agent interaction", *Proceedings of the Conference of the Society for Artificial Intelligence and Simulation of Behaviour (AISB' 93)*, Birmingham, RU, 1993.

[7] Cawsey, A., "Generating interactive explanations", *Proceedings of the 9th National Conference on Artificial Intelligence (AAAI-91)*, 1991, 86-91.

[8] Cawsey, A., "Developing an explanation component for a knowledge-based system: Discussion", *Expert Systems with Applications* ,1994.

[9] Cawsey, A., "User-modelling in interactive explanations", *Journal of User Modelling and User Adapted Interaction*, 1993.

[10] Cawsey, A., *Explanation and Interaction*, MIT Press, Cambridge, MA, 1995.

[11] Chandrasekaran, B., Tanner, M.D., y Josephson, J.R., "Explaining control estrategies in problem solving", *IEEE Expert*, **4** (1989) 9-24.

[12] Charniak, E., y Shimony, S.E., "Cost-based abduction and MAP explanation", *Artificial Intelligence* **66**, 2 (1994) 345-374.

[13] Conati, C., Gertner, A., VanLehn, K., y Druzdzel, M., "On-line student modeling for coached problem solving using Bayesian networks", *Proceedings of the 6th International Conference on User Modeling (UM97),* 1997.

[14] De Carolis, B., De Rosis, F., Grasso, F., Rossiello, A., Berry. D., y Gillie, T., “Generating recipient-centered explanations about drug prescription”, *Artificial Intelligence in Medicine*, **8** (1996) 123-145.

[15] Díez Vegas, F.J., *Sistema Experto Bayesiano para Ecocardiografía.* Tesis Doctoral, Departamento de Informática y Automática, U.N.E.D., 1994.

[16] Dittmer, S.L., Jensen, F.V., “Tools for explanation in Bayesian networks with application to an agricultural problem”, *Proceedings of the 1st European Conference for Information Technology in Agriculture*, Copenhagen, Dinamarca, Junio 1997.

[17] DiMarco, C., Hirst, G., Wanner, L., y Wilkinson, J., “HealthDoc: customizing patient information and health education by medical condition and personal characterisitcs”, *First International Workshop on Artificial Intelligence in Patient Education*, Glasgow, RU, Agosto, 1995.

[18] Draper, D., “Plan Explanation”, Interim Report of the Rockwell Palo Alto Laboratory, 1997, *URL: www.rpal.rockwell.com/~draper/ARPI/expl.html.*

[19] Druzdzel, M., “Explanation in probabilistic systems: Is it feasible? Will it work?”, *Proceedings of the 5th Workshop on Intelligent Information Systems*, Deblin, Polonia, Junio 1996.

[20] Druzdzel, M., “Qualitative verbal explanations in bayesian belief networks”, *Artificial Intelligence and Simulation of Behaviour Quarterly*, **94** (1996), 43-54.

[21] Druzdzel, M., “Some useful properties of probabilistic knowledge representations from the point of view of intelligent systems”, *Proceedings of the 3th Workshop on Intelligent Information Systems*, Wigry, Polonia, Junio, 1994.

[22] Druzdzel, M., *Probabilistic Reasoning in Decision Support Systems: From Computation to Common Sense* PhD Thesis. Department of Engineering and Public Policy, Universidad de Carnegie Mellon, Pittsburgh, PA, 1993.

[23] Druzdzel, M., “Verbal uncertainty expressions: Literature review”, Technical Report CMU-EPP-1990-03-02, Department of Engineering and Public Policy, Universidad de Carnegie Mellon, Pittsburgh, PA, 1989.

[24] Druzdzel, M., y Henrion, M., “Using scenarios to explain probabilistic inference”, *Working Notes of the AAAI-90 Workshop on Explanation*, Boston, MA, Julio 1990, 133-141.

[25] Druzdzel, M., y Henrion, M., “Belief propagation in qualitative probabilistic networks”, *Qualitative Reasoning and Decision Technologies*, CIMNE, Carrete, N.P., y Singh, M.G., (Eds.), Barcelona, 1993, 451-460.

[26] Druzdzel, M., y Henrion, M., “Efficient reasoning in qualitative probabilistic networks”, *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-93)*, Washington DC, Julio 1993, 548-553.

[27] Druzdzel, M., y Suermondt, H. J., “Relevance in probabilistic models: "Backyards" in a "small world"”, *Working notes of the AAAI-1994 Fall Symposium Series on Relevance*, New Orleans, LA, Noviembre, 1994, 60-63.

[28] Elsaesser, C., “Explanation of probabilistic inference”, Kanal, L.N., Levitt, T.S., y Lemmer, J.F., (Eds.), *Uncertainty in Artificial Intelligence 3*, Elsevier, 1989.

[29] Elsaesser, C., y Henrion, M., “Verbal expressions for probability updates How much more probable is "much more probable"?”, *Uncertainty in Artificial Intelligence 5,* Elsevier, 1990.

[30] Ferrater Mora, J., *Diccionario de Filosofía Abreviado*, Editorial Sudamericana, Buenos Aires, 1970.

[31] Gámez Martín, J.A., *Inferencia Abductiva en Redes Causales*, Tesis doctoral. Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, 1998.

[32] Gámez Martín, J.A., Puerta Callejón, J.M. (Eds.), *Sistemas Expertos Probabilísticos*, Cuenca, Ediciones de la UCLM, 1998.

[33] Good, I.J. “Explicativity: a mathematical theory of explanation with statistical applications”, *Proceedings of the Royal Statistical Society* , **354** (1977) 303-330.

[34] Haddawy, P., Jacobson, J. y Kahn, C. “Generating explanations and tutorial problems from Bayesian networks”, *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, 1994, 770-774.

[35] Haddawy, P., Jacobson, J., y Kahn, C., “BANTER: A Bayesian Network Tutoring Shell”, *Artificial Intelligence in Medicine* **10**, (1997), 177-200.

[36] Haddawy, P., Jacobson, J. y Kahn, C., “An educational tool for high-level interaction with bayesian networks”, *Proceedings of the 6th IEEE International Conference on Tools with Artificial Intelligence*, 1994.

[37] Heckerman, D., *Probabilistic Similarity Networks,* MIT Press, Cambridge, MA, 1990.

[38] Henrion, M., y Druzdzel, M., “Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning”, *Proceedings of the 6th conference on Uncertainty in Artificial Intelligence*, July 1990, 17-32

[39] Horvitz, E.J., Breese, J.S., y Henrion, Max. “Decision theory in expert systems and artificial intelligence”, *International Journal of Approximate Reasoning*, **2** (1988) 246-302.

[40] Langlotz, C.P., Shortliffe, E.H., y Fagan, L.M., "A methodology for generating computer-based explanations of decision-theoretic advice", *Medical Decision Making,* **8**, 4 (1988), 290-303.

[41] McConachy, R., Korb, K., y Zukerman, I., "Deciding what not to say: an attentional-probabilistic approach to argument presentation". *Proceedings of the 20th Annual Meeting of the Cognitive Science Society* , Madison, WI, 1998, 669-674.

[42] Madigan, D., Mosurski, K., y Almond, R., "Graphical explanations in belief networks", *Journal of Computational and Graphic Statistics*, **6**, 2 (1997), 160-181.

[43] Meek, C., "Causal inference and causal explanation with background knowledge", *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, 1995, 403-410

[44] Moore, J., y Swartout, W.R., "Pointing: A way toward explanation dialogue", *Proceedings of the 8th Natural Conference on Artificial Intelligence (AAAI-90)*, Boston, MA, 1990, 457-464.

[45] Moore, J., *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context*, MIT Press, Cambridge, MA, 1994.

[46] Neapolitan, R.E., *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*, Wiley-Interscience, New York, NY, 1990.

[47] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco, CA, 1988, 2nd Printing, 1991.

[48] Pearl, J., "Causal inference from indirect experiments", *AI and Statistics List*, **1**, 5 (1994).

[49] Pennington, N., y Hastie, R, "Explanation-based decision making: effects of memory structure on judgment", *Journal of Experimental Psychology: Learning, Memory and Cognition*, **14,** 3 (1988), 521-533.

[50] Reggia, J.A., y Perricone, B.T., "Answer justification in medical decision support systems based on Bayesian classification", *Computational and Biological Medicine*, **15,** 4 (1985), 161-167.

[51] Salmon, W.C., "Statistical explanation", Goldony, R., (Ed.), *The Nature and Function of Scientific Theories*, University of Pittsburgh Press, Pittsburgh, 1970, 173-201.

[52] Santos, E., "On the generation of alternative explanations with implications for belief revision", *Proceedings of the 7th conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, CA, Julio 1991, 339-347.

[53] Sember, P., y Zukerman, I., "Strategies for generating micro explanations for Bayesian belief networks", *5th Conference on Uncertainty in AI,* 295-302

[54] Shimony, S.E., *A probabilistic framework for explanation*, PhD Dissertation, Technical Report CS-91-57, Department of Computer Science, Universidad de Brown, 1991.

[55] Strat, T.M., "The generation of explanations within evidential reasoning systems", *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87),* Milan, Italia, 1987, 1097-1104.

[56] Strat, T.M., y Lowrance, J.D., "Explaining evidential analysis", *International Journal of Approximate Reasoning,* **3** (1989) 299-353.

[57] Suermondt, H.J., y Cooper, G., "An evaluation of explanations of probabilistic inference", *Computers and Biomedical Research*, **26** (1993) 242-254.

[58] Suermondt, H.J., *Explanation in Bayesian Belief Networks,* PhD Thesis. Stanford University. Marzo 1992.

[59] Swartout, W., "XPLAIN: A system for creating and explaining expert consulting programs", *Artificial Intelligence,* **21** (1983) 285-325.

[60] Wallsten, T.S., y Budescu, D.V., "A review of human linguistic probability processing: General principles and empirical evidence", *The Knowledge Engineering Review*, **10**, 1 (1995) 43-62.

[61] Wellman, M., "Fundamental concepts of qualitative probabilistic networks", *Artificial Intelligence,* **44** (1990) 257-303.

[62] Wellman, M., "Graphical inference in qualitative probabilistic networks", *Networks,* **20** (1990) 687-701.

[63] Wellman, M., y Henrion, M., "Explaining "Explaining Away"", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **15**, 3 (1993).

[64] Wick, M., y Slagle, J.R., "An explanation facility for today's expert systems", *IEEE Expert*, **4** (1989) 24-36.

[65] Wick, M.R, "The 1988 AAAI Workshop on Explanation"**,** *AI Magazine,* **10**, 3 (1989) 22-26.

[66] Wick, M.R., y Thompson, W.B., "Reconstructive expert system explanation", *Artifical Intelligence* **54** (1992) 33 – 70.

[67] Zukerman, I., McConachy, R., y Korb, K., "Bayesian reasoning in an abductive mechanism for argument generation and analysis", *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI98)*, Madison, WI, 1997, 833-838

[68] Zukerman, I., Korb, K., y McConachy, R., "Perambulations on the way to an architecture for a nice argument generator", *Notes of the ECAI-96 Workshop on Gaps and Bridges: New Directions in Planning and Natural Language Generation*, Budapest, Hungría, 1998, 31-36.