Las estructuras conceptuales de representación del conocimiento en Internet

Miguel-Ángel López Alonso
Facultad de Biblioteconomía y Documentación
Universidad de Extremadura

0.1. Resumen

Se aborda el estudio de las estructuras para la representación de la información capaces de extraer conocimiento a partir de su análisis conceptual, como manifestaciones de las ideas, el conocimiento o el razonamiento humano. El enfoque parte del establecimiento de las distintas técnicas lineales y no lineales para la representación del conocimiento. Se continúa con el estudio de los modelos conceptuales más avanzados de organización y recuperación de la información, ontologías y taxonomías virtuales. Se concluye con la propuesta de un "marco de implantación" mediante la interrelación de: un espacio conceptual de contenidos que integre una ontología de conocimientos, y un espacio documental que aporte las bases documentales de un área del conocimiento. (Autor)

Palabras clave: Representación del conocimiento. Técnicas lineales y no lineales. Modelos conceptuales. Espacio conceptual de contenidos. Espacio documental.

0.2. Abstract

The structures for information representation that allow the extraction knowledge —such as demonstrations and reasoning— are studied. First, the different linear and no linear techniques for knowledge representation are considered. Second, the more advanced cognitive models for organizing and retrieving information are analysed, that is, ontologies and virtual taxonomies. Finally, an integrated "installation framework" is proposed which consists of two layers: a Cognitive Contents Space organized by a knowledge ontology, and a Documentary Space with full-text databases on a specific knowledge area. (Author)

Keywords: Knowledge representation. Linear and no linear techniques. Cognitive models. Cognitive contents space. Documentary space.

1. Introducción

Soergel (1997), en su manifiesto con motivo de la entrega de la medalla de honor anual de ASIS en diciembre de 1997, dejó claramente acotados los dos problemas claves, convergentes y dependientes entre sí, de la Ciencia de la Información en la actualidad:

1) El diseño de modelos conceptuales capaces de dar respuesta a las demandas de conocimientos de los usuarios, y …
2) La búsqueda de las estructuras de representación de la información que permitan extraer conocimientos ("making sense") a partir de su análisis.

Desde el punto de vista del usuario, el mejor sistema para la recuperación de la información es el que soporta cualquier tipo de búsqueda con facilidad, dentro de un espacio multidimensional (bibliográfico o hipermedia) de la información. La operación básica de la recuperación de la información es siempre la misma, a partir de un objeto (descriptor o multimedia) el usuario sigue los enlaces proporcionados por el sistema para localizar uno o varios objetos pertinentes que puedan ser consultados.

La investigación futura deberá encaminarse a la búsqueda de interfaces "inteligentes" que tengan a los usuarios como destinatarios principales, dada la necesidad de desarrollos avanzados para interactuar con la información. Para evitar que se pierda la oportunidad de diseñar sistemas en perfecta armonía con las preferencias de los usuarios, sus interfaces deberán tener como mínimo las siguientes características:

a) Ser "verdaderamente distribuidos", de modo que, independientemente de su localización, sean capaces de modificar (en tiempo real) los diferentes caminos de acceso a los datos suministrados, por cualquier usuario, en cualquier lugar y en cualquier tipo de formato (Belkin, 1987).
b) Aceptar la "manipulación directa" durante el rastreo sofisticado de los documentos, mediante la realización de las siguientes tareas "inteligentes":
   — Una continua representación de los objetos de interés,
   — la facilitación de las acciones físicas por medio del uso de distintas herramientas: teclado, ratón, joystick, pantalla sensible, etc.,
   — operaciones rápidas, incrementales y reversibles cuyo impacto en el objeto de interés sea inmediatamente visible, y…
   — un aprendizaje progresivo, por etapas, que pueda ser utilizado con conocimientos mínimos previos (Shneiderman, 1986).

c) Proporcionar "ayuda continuada", a través de un diálogo interactivo con estrategias alternativas de búsqueda, mediante las siguientes operaciones concatenadas:

— El análisis de las preguntas del usuario,

— la identificación de sus necesidades de información y su fácil conversión en preguntas,

— la entrega al usuario de los resultados de forma clara y concisa, y

— la sugerencia de alternativas para búsquedas más ajustadas (Turtle, 1992).

d) Integrar diferentes mecanismos de precisión: bases de conocimientos, tesauros, etc., aplicaciones y medios de difusión, con la finalidad de que los usuarios se comuniquen entre sí o accedan a diferentes bases de datos. Y utilizar diferentes servicios globales en línea, mediante nuevas formas más naturales, intuitivas y flexibles, en el marco general de un Modelo Conceptual Integrado de la Información (Ingwersen, 1996).

e) Describir en sus interfaces las estructuras estratégicas de control del grado de implicación del usuario en las búsquedas, y el tipo de capacidades exigibles al sistema (Bates, 1990).

Blair (1990) coincide con Soergel en que el problema clave de la recuperación de la información pasa por la búsqueda de los procedimientos teóricos para su representación. Para ambos, se trata de un problema de uso del lenguaje que, siguiendo la Lingüística del Texto, se acomete desde el análisis de contenido de los textos. A los problemas teóricos de sintaxis o semántica, se añade la dificultad de representar el conocimiento obtenido durante el procesamiento del lenguaje natural.

La cuestión se complica con la masificación de la información digitalizada, que como "sustituto" del documento original utiliza soportes multimedia: audio, vídeo, informático, etc. Su procesamiento implica escaneado de imágenes, comprensión de sonido, compilación de programas, etc., para obtener algún tipo de "conocimiento derivado" representable en cualquier soporte digital.

Los problemas de representar el del conocimiento (textual o multimedia) de forma codificada preocupan por igual a los especialistas en psicología, lingüística, documentación e inteligencia artificial, todos ellos del ámbito de la Teoría Cognitiva. La clave es: ¿cómo incorporar las entidades abstractas de la teoría sintáctica y semántica, y sus relaciones, al complejo mundo de los sistemas informáticos, de manera que participen en el proceso informático-matemático de codificación de la información analizada (indización o resumen) y su posterior decodificación para recuperar el conocimiento demandado por los usuarios?.

En este marco surgen las estructuras de representación del conocimiento como manifestaciones de las ideas, conocimiento, o raciocinio humanos. Sin embargo, en un sistema de recuperación de la información debe contenerse a la vez diferentes estructuras:

- Las cognitivas de los diseñadores, mediante estructuras específicas de bases de datos o algún algoritmo de comparación.
- Las de los contenidos de los textos o de las imágenes, mediante estructuras conceptuales que se comuniquen con las del sistema.
- Las del estado cognitivo de los usuarios durante la formulación de sus preguntas al sistema, mediante estructuras conceptuales más o menos transformables mediante la manipulación de algún tipo de interfaz interactivo.

En resumen, lo que nos interesa es el establecimiento de técnicas y modelos de representación del conocimiento que soporten teóricamente tales necesidades de información variables y mal definidas *a priori* por los usuarios.

Puesto que la memoria humana es asociativa, crea estructuras complejas del conocimiento que permiten relacionar diferentes contenidos o piezas informativas. Cuando un autor escribe un documento, convierte el conocimiento de su memoria (en forma de compleja estructura neuronal) en una representación externa. Como los soportes físicos los textos en papel o las cintas de vídeo solo nos permiten representar la información de forma lineal, y esta no es la forma natural de representación de nuestra mente, el autor se ve obligado a proporcionar información adicional (por ejemplo, los índices) para ayudar al lector a entender la codificación completa de su información.

El proceso de lectura transforma la información externa en una representación interna del conocimiento, y la integra en las estructuras del conocimiento existentes en el cerebro. Durante este proceso de descodificación el lector fragmenta la información externa en pequeños segmentos textuales (Paice, 1991) y los vuelve a reunir en base a sus necesidades.

Las técnicas hipermedia imitan parcialmente los procesos de escritura (codificación) y lectura (descodificación) humanos como si se realizasen dentro del cerebro, utilizando enlaces informáticos. Con ellas podemos crear estructuras de información no lineales que asocien "segmentos textuales" mediante vínculos hipertextuales (binarios o múltiples). Además, podemos utilizar una combinación de objetos multimedia textos, imágenes, vídeos, sonidos y animaciones que enriquezcan la representación de la información.

Esto evita al autor el proceso de codificación lineal de sus estructuras del conocimiento durante la escritura, además de permitir al lector el acceso directo a la estructuras complejas de conocimiento del autor. Todo ello permite que el

lector cree, a partir del entorno intensivo del texto que el autor transmite (descodificación), sus propias representaciones del conocimiento, y las integre de forma natural con las estructuras de su memoria permanente.

La información debe obtenerse de las complejas interacciones entre las fuentes de conocimiento disponibles y las representaciones de las necesidades del usuario (Cortez, 1995). Éste puede ayudarse, para la enunciación de sus preguntas, con la información adicional de una base de conocimientos. La recuperación es considerada como un proceso de razonamiento, en el que las fuentes de conocimientos sobre el contenido de la pregunta y de los documentos se combinan, para estimar la probabilidad de que un documento se ajuste a una pregunta.

2. Técnicas de representación del conocimiento

Existen técnicas para representar el conocimiento de una forma lógica bastante elementa lineales o jerárquicos que vienen utilizándose tradicionalmente para la indización documental. Y otros más complejo redes o grafos que comenzaron a utilizarse para la estructuración de los registros en las base de datos, los programas informáticos, la estructura lógica y semántica de lenguaje natural, la representación del conocimiento y los modelos de memoria artificial:

a) Las técnicas lineales se usan de diferentes maneras en los sistemas hipermedia, habitualmente para retener la estructura secuencial del documento original en papel. Es costumbre mantener la estructura lineal del documento original, ordenada por el autor original de manera que añada conocimiento intensivo. El lector de este documento en un sistema electrónico debería tener acceso a este orden de la misma manera que el lector de la versión en soporte papel.

b) Las técnicas jerárquicas se usan para mantener la estructura original no secuencial de la información contenida en las bases de datos hipermedia. La estructura jerárquica de los libros, divididos en capítulos con secciones y subsecciones, se suele imitar en las bases de datos hipermedia mediante el uso de los vínculos jerárquicos. Un lector puede entrar en la tabla de contenidos o índice de un sistema hipermedia, y seleccionar un lugar para leer dentro del espacio de la información, de la misma forma que lo haría en el soporte papel.

c) El tercer tipo de técnicas disponibles son los grafos y las redes, compuestas de vínculos asociativos por naturaleza, semánticos (entorno intensivo) o pragmáticos (entorno extensivo), verdaderamente no secuenciales y capaces de unir conceptos similares o que precisan relacionarse en el espacio conceptual de la información. La habilidad para rastrear este espacio de conceptos (navegación hipertextual) de forma

progresiva, no secuencial, es una de las mayores ventajas de los sistemas hipermedia.

Los sistemas tradicionales de recuperación de la información no operan directamente sobre el significado de los conceptos. Son estáticos y sólo proporcionan una colección de herramientas genéricas para manipular redes de objetos de una forma inteligente, nunca su contenido. Los sistemas del futuro deberán ser capaces de procesar el conocimiento de manera dinámica (*sense making*), mediante una rica semántica de objetos y sus asociaciones, distinguiendo los objetos por "aquello que representan sus conceptos", e interpretando el conocimiento de las ecuaciones de búsqueda de los usuarios (Carlson, 1990).

Los modelos de representación del conocimiento han evolucionado desde las reglas de producción, los paneles (Minsky, 1975), las redes (Quillian, 1968), etc. y se han desarrollado en distintos lenguajes especializados: KRL, FRL, Lisp, etc.). Su terminología ha cambiado desde su primera utilización en las bases de datos relacionales, pasando ahora a denominarse a los nodos (representan conceptos) como objetos y a las relaciones (representan relaciones de dependencia entre conceptos) como asociaciones, en las bases de conocimientos de los sistemas de recuperación de la información.

Los paneles son estructuras de datos que caracterizan una colección de conceptos relacionados, similares a los registros de los lenguajes tradicionales. A los diferentes campos de los paneles se les denomina guiones (*scripts*), y están formados por términos que contienen valores o atributos (*slots*). Cada guión se subdivide a su vez en facetas, dotadas de un valor predeterminado que sirve para identificar otros guiones con los que se relaciona el primero.

Las relaciones conectan pares de paneles a través de punteros (vínculos) y encuentran sus relaciones específicas almacenadas en los guiones. La herencia de relaciones, junto con sus procedimientos y datos asociados, proporciona a un panel "descendiente" los valores heredados de sus paneles "antecesores".

Las redes neuronales se componen de gran cantidad de paneles conectados. Cada panel realiza simples operaciones aritméticas, al recibir diverso grado de activación (según las entradas que recibe de sus vecinos) y procesar un valor de salida que es enviado a su vez a los demás vecinos. Las conexiones entre los paneles son estimulantes o inhibitorias; las estimulantes incrementan las entradas que activan los paneles próximos, mientras que las inhibitorias tienden a disminuir su activación.

La efectividad con que una señal es transportada, a través de los paneles relacionados, puede variar y se expresa mediante valores ponderados. Los procesamientos de cada panel son muy simples, y la activación final es el resultado de la de los múltiples paneles conectados con el primero y de los pesos atribuidos a sus

interrelaciones ("expansión activada"); lo que da lugar a que la "inteligencia" de las redes neuronales sea una propiedad que resulta de las complejas correspondencias entre sus unidades de procesamiento o paneles vecinos (Cohen, 1987).

Soergel (1993) y Gray (1992) defienden que los modelos de datos de los sistemas expertos, de las bases de datos, y de los sistemas recuperación de la información tienen un alto grado de similitud matemática y una "homogeneidad conceptual" complementaria, y que con una "estructura unificada" se puede mejorar su diseño, dado que se apoyan en estructuras de objetos genéricos (con atributos o asociaciones) que forman una red estructural mediante referencias internas entidad-relación (Barlow, 1989).

Además de asumir que la navegación hipermedia y las preguntas documentales comparten el mismo concepto, proponen diferentes aproximaciones para las búsquedas y sus procedimientos inferenciales, del tipo de la "expansión activada" (con búsquedas booleanas o ponderadas en casos especiales), la herencia jerárquica y la equivalencia de estructuras. Consideran la indización documental o la hipermedia como técnicas para establecer la estructura, dado que para entonces los hipertextos y los documentos se diferencian únicamente por su mayor o menor facilidad en alterar su estructura física.

La modelos o técnicas de representación del conocimiento de este "espacio integrado de la información" estarían formados por objetos (o nodos) y sus asociaciones (o relaciones) que les definen conceptualmente; además de preguntas, series, guiones, etc. Para ampliar la inferencia de las búsquedas, incluirían secciones (grupos de objetos con sus asociaciones, también llamados regiones, grupos o nodos compuestos virtuales), algún tipo de vecinos antecesores o descendientes, y conexiones (también llamadas cadenas de asociaciones).

Los objetos o nodos son los componentes básicos de la información contenida en los documentos, películas o registros sonoros (por ejemplo, párrafos, fotogramas o secuencias musicales). Mientras que, para el usuario, los nodos y sus asociaciones contienen conceptos organizados en estructuras del conocimiento, para el sistema, tanto los nodos como sus asociaciones son una misma cosa: objetos para almacenar, recuperar, visualizar, interconectar, etc.

Una asociación se puede definir como un grupo de objetos (o vecinos reunidos) que tienen vínculos comunes entre ellos. Una pregunta especifica una búsqueda y por tanto una asociación, cuya especificación se torna dinámica por contener objetos que reúnen el criterio de búsqueda en el momento que se realiza la pregunta. Una serie es un tipo especial de objeto (o nodo) que define una asociación de otros objetos y especifica una secuencia de éstos, que también puede verse como un tipo especial de guión. Un guión es un objeto que contiene instrucciones para organizar la visualización de otros objetos. Tanto las series como los

guiones guían al usuario a través de una base de informaciones de cualquier tipo, eligiendo los tipos de objetos y de relaciones habituales adecuadamente.

Para que una información sea útil, deberá realizar "declaraciones de los objetos que abarca". Estas declaraciones estarán formadas por asociaciones con uno o más argumentos, y sus espacios estarán ocupados con un valor (o atributo) para cada objeto específico. Numerosas asociaciones tienen dos argumentos, son binarias, considerándoselas como un vínculo (o relación binaria) entre dos argumentos. Mientras que los sistemas expertos y/o las bases de datos utilizan "asociaciones de grado superior" que son aplicables a diferentes tipos de objetos (además de numerosas asociaciones binarias), los sistemas hipertextuales usan solamente asociaciones binarias, a las que denominaremos "vínculos", siendo aplicables a un determinado tipo de objeto.

Los tipos de objetos y sus asociaciones especifican conjuntamente que tipo de datos se pueden localizar en un sistema de bases de datos/hipermedia, es decir, el esquema conceptual. En su nivel interno, la estructura unificada de uno de estos sistemas integrados combina adecuadamente el poder inferencial selectivo de un sistema experto/base de datos, con el poder rastreador y visualizador de un sistema hipermedia, permitiendo potentes asociaciones que serían mucho más engorrosas en sistemas diversificados.

Un ejemplo de la debilidad de la estructura de los sistemas de bases de datos lo tenemos en el caso de los lenguajes controlados, que se ven como un subconjunto restringido de su correspondiente lenguaje natural del que toman prestado su terminología básica y su estructura conceptual. El significado dado a los términos puede no corresponder exactamente con el que tienen en su emplazamiento "natural", o no utilizan todos los significados de los lenguajes naturales para expresar sus asociaciones estructurales.

Esto provoca que algunos sistemas documentales no logran representar sus asociaciones de manera adecuada, y deben recurrir a métodos inferenciales para la identificación de las relaciones, sus participantes y sus papeles.

Para paliar esta debilidad se han venido utilizando diversas estrategias estadísticas de ponderación (por ejemplo la búsqueda por proximidad, la relación histórica entre términos afines, el uso de indicadores de significado, etc.), que pueden evitarse con la estructura integrada de la aproximación semántica propuesta por Soergel. Sin embargo, estos métodos conceptuales obligan a construir y mantener una base de conocimientos terminológicos que incluya información sobre el significado de los términos y las relaciones semánticas entre ellos (Warner, 1994).

Dado que una asociación es cualquier grupo de objetos junto con las relaciones existentes entre ellos, los miembros se seleccionan habitualmente en base a su asociación con uno o más objetos distintos. En el caso más simple, la asociación se forma por todos los objetos que pueden alcanzarse desde un objeto inicial, mediante vínculos con sus vecinos de un tipo determinado. Por ejemplo, en un sistema documental la asociación puede estar formada por todos los documentos que critican un determinado documento, el cual puede alcanzar a los demás documentos siguiendo vínculos binarios del tipo "criticado por".

Existe una relación directa entre asociación y pregunta. Una pregunta conduce a una asociación, y cada asociación corresponde a una pregunta. Halasz (1988) habla de los "nodos compuestos virtuales" que resultan de la formulación de una pregunta y aclara que: "esta puede permitir que el lenguaje usado para descripciones estructurales virtuales sea el mismo que el lenguaje de la pregunta usado para las búsquedas y para los filtros del internase".

Para acabar de definir los elementos o técnicas básicas de la estructura unificada de la información propuesta por Soergel, debemos añadir que se denomina serie a un tipo de asociación en la que relaciones del tipo "continua por" inducen un orden lineal de todos los objetos contenidos en ella. A menudo un lector se satisface más con la lectura de una secuencia organizada de objetos multimedia que saltando de un nodo a otro. Una serie de almacenamiento preparada por un editor permite esto, dado que se corresponde con un artículo o libro tradicional. Un objeto de una serie puede también ser una serie, dado que un libro puede representarse como una serie de capítulos, cada capítulo como una serie de secciones, y una sección como una serie de párrafos y figuras.

Un guión es un objeto que contiene instrucciones para organizar la visualización de otros objetos, de manera que se conduzca al usuario a través de una información básica. Por ejemplo un guión puede organizar una secuencia de instrucciones programadas que establezca una base hipermedia. Cuando se ejecuta el guión, se presenta la información, se le formulan preguntas al usuario, y el nuevo tema de la información que se presenta se obtiene a base de las respuestas anteriores.

Otro guión puede tener instrucciones para reunir un documento a partir de sus objetos textuales, recuperando datos y haciendo que un programa los represente gráficamente, o recuperando otros datos y aplicando un generador de lenguaje natural, etc. A este tipo de guión se le llama documento virtual.

3. Modelos conceptuales de organización y recuperación de la información

Algunos modelos conceptuales de representación de la información nos son familiares desde su utilización para la indización y recuperación de la información conceptual, por ejemplo, tesauros facetados, matrices conceptuales de indización, etc. Otros se han desarrollado específicamente para la indización de la información hipermedia, por ejemplo, los hiperíndices semánticos, etc., o han derivado de los formalismos de representación de la Inteligencia Artificial, por ejemplo, las redes inferenciales, las redes semánticas, etc.

Recientemente, con el desarrollo masivo de los recursos en las bases de conocimientos de la Web, se habla de ontologías o taxonomías virtuales como modelo conceptual de organización y representación del conocimiento de tipo cooperativo, capaces de servir de clasificaciones virtuales, de representar conocimientos y de razonar con ellos mediante reglas de inferencia.

El tesauro, como estructura de indización más utilizada en los sistemas convencionales de recuperación de la información, está formado por una serie limitada de conceptos y de asociaciones entre ellos. Solo se representan tres tipos de asociaciones relacionadas: las jerárquicas (conceptos generales o específicos), las de equivalencia (conceptos preferentes o equivalentes) y las asociativas (conceptos afines conceptualmente).

Partiendo del análisis facetado, como técnica de indización en la que los conceptos se clasifican en estructuras jerárquicas independientes, y en la que cada una captura un punto de vista o propiedad diferente de los documentos; un tesauro facetado se forma por un determinado número de tesauros diferentes, entre los que cada uno de ellos se utiliza para indizar los documentos con una perspectiva diferente del conocimiento. Su ventaja sobre los tesauros documentales es que la estructura facetada permite mayor exhaustividad y precisión en el proceso de indización hipermedia, dado que los documentos pueden indizarse teniendo en cuenta todos los aspectos que se consideran relevantes.

Las matrices conceptuales de indización (Wille, 1992) son (matemáticamente hablando) una poderosa extensión de la estructura del tesauro, al componerse de una serie ordenada de conceptos en la que cada pareja tiene un único concepto común. La estructura de indización resultante es similar a la del tesauro pero pluridimensional, debido a que cualquier concepto puede tener otros generales o específicos, no necesariamente en un nivel más alto o más bajo en la jerarquía de conceptos, ya que nos movemos en una matriz n-dimensional.

Los hiperíndices (Bruza, 1990) son una técnica de indización desarrollada para los sistemas de información hipermedia, en la que el contenido de los documentos se representa por una "serie de indización" formada a partir del título del

documento con los términos de indización vinculados entre sí. A partir de la primera serie de indización se puede derivar otra "serie reforzada", que forme una matriz de "series de indización" y sea usada como una estructura de indización hipertextual. Cada vértice de la matriz puede considerarse como una pregunta predefinida lanzada al espacio documental, que puede ampliarse (convertirse en menos específica) o reducirse (convertirse en más específica) moviéndose hacia el vértice de su antecesor o descendiente, respectivamente.

La potencialidad de la indización con los hiperíndices descansa en que la matriz de hiperíndices puede ser generada automáticamente a partir de los conceptos que caracterizan el contenido de los nodos, aunque no se tenga en cuenta la manera en que se relacionan semánticamente estos conceptos. Para superar esta limitación se han desarrollado los "hiperíndices semánticos" (más "inteligentes"), que permiten el uso de asociaciones entre los conceptos pertenecientes a diferentes dominios. Su uso permite validar los conceptos de la matriz de hiperíndices, excluir algunas de las asociaciones de conceptos generadas automáticamente e incluir otras que el indizador considera imprescindibles para el dominio del conocimiento al que pertenecen o para un tipo determinado de tareas de usuarios específicos.

Las redes inferenciales (Croft, 1993) están formadas por dos series de redes: una red documental que representa los documentos de la colección, y una red de preguntas que representa la información buscada por los usuarios. Las dos series de redes se unen mediante vínculos entre los conceptos documentales y los conceptos de las preguntas, de manera que durante el proceso de búsqueda los conceptos de las ecuaciones de búsqueda se confrontan con los conceptos documentales, mediante cálculos probabilísticos.

En la red semántica (Woods, 1975) los nodos representan los conceptos y los vínculos representan las relaciones entre ellos. Comparada con los tesauros, los supera en una más rica organización interna de relaciones que pueden soportan diferentes mecanismos de razonamiento. Su estructura nodo-vínculo-nodo es conceptualmente muy próxima a la estructura de una red hipertextual, y soporta por ello la navegación de manera natural.

El uso de las redes semánticas como esquema de representación del conocimiento tiene una larga historia en la Inteligencia Artificial. Quillian trató de desarrollar con ellas un modelo que imitara la memoria humana permanente, y utilizar sus principios teóricos para informatizar las complejas tareas de la memoria, principalmente, las relacionadas con el almacenamiento de los significados de las palabras más usuales del lenguaje.

En un sistema documental, Dewèze formaliza la representación de las relaciones semánticas, con la adopción de una "teoría conceptual extralexical" que

sitúa a un nivel superior al lenguaje natural. En esta teoría, el significado es definido como "un conjunto de semas a los que se pueden atribuir relaciones lexicales...". Una materia se representa por un grafo, los términos son las cumbres y las relaciones son los arcos. Los parámetros de demanda se representan mediante un grafo, cuyos arcos y cumbres representan distintas ponderaciones.

Sus propiedades son:

1) La independencia respecto de las relaciones lexicales,
2) la facilidad de extensión o de encaminamiento hacia otros conceptos, a partir de cualquier nodo de la red,
3) la posibilidad de multilingüismo, al entrar en la red semántica por el léxico de una lengua cualquiera, mediante la fijación de relaciones lexicales en varios idiomas, y…
4) el tiempo de acceso reducido, sea cual sea el punto de entrada en la red, debido a la facilidad de seguimiento de las referencias cruzadas.

En las búsquedas documentales un programa compara el grafo de la demanda con los grafos de los documentos registrados en memoria, y retiene aquellos términos con estructura más parecida. Las redes semánticas proporcionan un mapa de las variables introducidas en la búsqueda y de las obtenidas en la recuperación, y son capaces de almacenar información de forma distribuida, aprendiendo de los problemas cotidianos tras un período de validación que actualice las respuestas.

En el contexto de la reutilización de la información, una ontología (Gruber, 1994) es una descripción de los conceptos (asimilable a la especificación formal de un programa informático) y de sus asociaciones (por ejemplo, la clasificación de Yahoo), existentes en un agente (o *softbot*) o federación de agentes de software de la Web. Lo que significa que una ontología proporciona una estructura y unos contenidos que son independientes del fin y del dominio de la aplicación en la que se reutilizarán sus definiciones.

De su definición se deduce una cierta analogía con otras estructuras conceptuales de organización de la información del tipo de los tesauros o las clasificaciones conceptuales, que igualmente establecen asociaciones entre sus conceptos. En cambio, existen diferencias fundamentales en cuanto a la representación del conocimiento, pues las ontologías permiten representar axiomas (conocimientos ciertos e inmutables) y razonar con ellos mediante las reglas de inferencia, definidas en el núcleo de conocimientos de los agentes que filtran la información de las redes distribuidas de Internet.

Una ontología para una base de conocimientos debe abarcar los diferentes tipos de documentos, descripciones conceptuales, sus relaciones, diferentes pro-

blemas científicos. Además debe incluir índices, descripciones bibliográficas, tesauros, códigos clasificatorios, formalizaciones de validez, información terminológica, etc. Su aplicación proporciona una metavisión de la estructura y de la terminología del dominio que mejora la precisión de las recuperaciones documentales. Se las emplea para unificar vocabularios y definir estructuras comunes, entre diferentes aplicaciones federadas que tengan como objetivo la representación del conocimiento. Sus relaciones son tanto estructuras lógicas como sociales, cuya estabilidad depende de la madurez del conocimiento abarcado (tanto en un sentido científico como práctico).

Hay tres factores que influyen profundamente en el grado de madurez de una ontología: su dinámica (número de nuevos conceptos y asociaciones entre ellas que deben considerarse como parte integrante de la ontología, en una unidad de tiempo), su complejidad (número de conceptos y asociaciones, en total), y su dispersión social (conceptos abarcados que comparten los miembros del dominio). Las tecnologías de la información en redes tipo Internet pueden tener un importante impacto como factor de dispersión, e influir en el desarrollo de las ontologías del conocimiento.

4. Marco de implantación y propuestas.

Como resultado de la necesidad del procesamiento de la información perteneciente a diferentes sistemas, su interconexión ha aumentado en los últimos años mediante el uso de Internet. Sin embargo, una conexión puramente técnica es insuficiente para aquellos usuarios que necesitan encontrar el camino de su información específica. Las arquitecturas federadas mantienen la autonomía de los esquemas de sus bases de datos, y a la vez permiten compartir su información mediante esquemas comunes de exportación e importación.

Horizontalment,e se distinguen servicios a nivel del sistema y a nivel del usuario. Los primeros proporcionan una funcionalidad básica, a saber, acceso a bases de datos o preparación de páginas HTML. Los segundos combinan distintos servicios a nivel del sistema dentro de los servicios a alto nivel accesibles al usuario.

Verticalmente, se distingue entre: información para ayudar al usuario a encontrar servicios e informes (espacio de indización, a saber, metadatos que incorporen uno o varios hipertesauros), y bases de datos con distintos tipos de información (espacio documental).

Se propone el diseño de un modelo para el procesamiento de la información que integre una macroestructura de conocimientos o espacio conceptual de contenidos formado por dos espacios conceptuales relacionados (Frisse, 1989).

1) Un espacio de indización, definido como*una ontología de conocimientos* formada por la terminología usada por los sistemas de gestión, su sistema de clasificación, y sus bases de conocimientos.

El diseño de ontologías especializadas (a partir de las bases de datos con texto completo de las diversas áreas del conocimiento) deberá permitir enlazar con las preguntas específicas del área tratada.

La solución propuesta pasa por la integración de múltiples tesauros conceptuales de manera federada, mediante el marcado de las vías de acceso, el filtrado de los conceptos redundantes, la creación de "vistas" de las bases de datos relacionales, etc. Para la descripción de los conceptos y de la estructura de representación del conocimiento se utilizan diferentes ontologías existentes para cada dominio. Las relaciones entre sus conceptos se expresan mediante funciones de conversión que permiten una serie de operaciones algebraicas para la gestión de dichas ontologías.

Dado que las fuentes de información en Internet son muy diversas, cualquier sistema virtual de clasificaciones u ontologías entre sedes Web debería ser muy general y polivalente en el primer nivel de la jerarquía:

a) Se integraría en una superontología dinámica o catálogo de catálogos, en continua evolución a partir de otras subontologías que se adapten y sobrevivan en sus propias áreas de trabajo habitual (Buckland, 1995). A continuación se utilizarían aplicaciones informáticas de la Inteligencia Artificial, que permitieran la automatización de este proceso de autogeneración de tesauros conceptuales de abajo-arriba.

b) Partiría de las bases de conocimientos terminológicos y mediante análisis semántico: indización semántica latente, etc. (Caid, 1995) se irían extrayendo conceptos ordenados por facetas siguiendo un criterio relacional previamente definido. Desde estos conceptos se desarrollarían diferentes tesauros conceptuales, unidos en una red semántica de estructuras neuronales en la que cada nodo contenga una serie de descriptores asociados con un único concepto semántico que pueda ser igualmente identificado en la red hipertextual de "pequeñas piezas de información" textual del espacio documental, llamadas "entidades de argumentación" por Sillince (1992) o "extractos tesaurizados" por Paice (1991). Una de las recomendaciones es la conexión de las diferentes ontologías externas, aprovechando los recientes avances en tecnologías de interoperatibilidad abierta: WWW, CORBA y Java (Kramer, 1997), de manera que cualquier usuario de Internet se pueda beneficiar al mismo tiempo de varias de ellas que actúen en cooperativa (red de hipertesauros (West, 1997).

2) El espacio documental está integrado por las *bases documentales utilizadas* en cada área del conocimiento.

Ambos espacios deben relacionarse dinámicamente, preferentemente a través de un sistema de redes neuronales que permita un aprendizaje continuado mediante entrenamiento interactivo con ciclos de búsqueda y subsiguiente afinado de los mismos (inferencia).

La implantación de un modelo conceptual integrado de recuperación de la Información en dicho espacio conceptual, deberá desarrollarse mediante aplicaciones WWW, dado que sus rastreadores (programados en KIF o KQML) están disponibles en la mayoría de los sistemas de recuperación, lo que hace que las aplicaciones WWW sean accesibles fácilmente desde cualquier plataforma o sistema operativo. Además, se observa una convergencia total en las tecnologías abiertas (tipo OSI) para la integración de las estructuras básicas de Internet, mediante la federación de arquitecturas de los diferentes sistemas distribuidos de bases de datos, bajo un modelo común (esquema común de importación y exportación) en el que cada participante mantenga su esquema individual de bases de datos o sistemas de información autónomo.

Se deberá priorizar el desarrollo de asociaciones de redes neuronales que conviertan en dinámicos los nodos hipertextuales, permitiendo que los conceptos situados en cada uno de ellos puedan ser consultados simultáneamente (no sólo secuencialmente) de acuerdo con las necesidades de cada usuario. Para ello serán precisos neurordenadores con procesamiento "masivamente paralelo", que trabajen con los múltiples nodos hipertextuales al estilo que el cerebro humano trabaja con múltiples neuronas (Moya, 1998). Para plasmar espacialmente dichos contenidos, se vienen experimentando con mejor o peor fortuna los interfaces interactivos tridimensionales (3D) (Cugini, 1997), cuyo análisis escapa de las pretensiones de este trabajo.

Se reconoce que la precisión es muy pobre en las búsquedas en la Red como "espacio documental" hipertextual distribuido y que, por tanto, la exhaustividad no es el problema más importante. De aquí la importancia que cobra en la actualidad el estudio de las llamadas "Data Mining", "Knowledge discovery" o tecnologías para la selección y recuperación de información especializada, a partir de bases de datos no previamente estructuradas. Éstas deberán permitir un tipo de estructuración rápida que coloque automáticamente la información en bases de conocimientos especializadas. De estas últimas podrá ser recuperada como conocimiento adecuado a las necesidades de los usuarios: conceptos, relaciones, clasificaciones, reglas de decisión, etc..

La extracción del conocimiento es una aproximación pluridisciplinar que integra estadísticas, búsquedas y recuperaciones en bases de datos, ingeniería del

conocimiento y sistemas expertos, máquinas inteligentes, redes neuronales y visualizaciones de datos. Todas estas materias contribuyen a la localización de nuevo conocimiento, cuyo proceso principal se encuentra en el "minado de datos", que consiste en una multitud de etapas que comienzan con: el establecimiento de los objetivos que permitan evaluar los resultados, y con la recuperación de datos que permitan reformular los objetivos según los resultados. El objetivo principal es la extracción, a partir de grandes series de datos, de estructuras o modelos desconocidos previamente.

5.Referencias

Barlow, J. et al. (1989). Expertext: hypertext, expert system theory, synergy and potential applications. // Research and Development in Expert Systems VI. N. Shadbolt (ed.), Cambridge: University Press, 1989, 116-127.

Bates, M. J. (1990). Where should the person stop and the information search interface start?. // Information Processing Management, 26 : 5, (1990), 577.

Belkin, N.J. et al. (1987). Distributed Expert-Based Information Systems: An Interdisciplinary Approach. // Information Processing Management, 23 : 5 (1987) 407.

Blair, D.C. (1990). Language and representation in information retrieval. Amsterdam: Elsevier, 1990, 122.

Bruza, P.D. (1990). Hyperindices: A Novel Aid for Searching in Hypermedia. // Hypertext: Concepts, Systems and Applications. Rizk, A. et al. (eds.), Cambridge: University Press, 1990, 109-122.

Buckland et al. (1995). Partnerships in navigation: an information retrieval research agenda. // ASIS Annual Meeting, Chicago, octubre 1995.

Caid, W.R. et al. (1995). Learned vector-space models for document retrieval. // Information Processing & Management, 31 : 3 (1995) 419-429.

Carlson, D.A.; Ram, S. (1990). HyperIntelligence: The Next Frontier. // Communications of ACM, 33 : 3 (1990) 311-321.

Cohen, P. R.; Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. // Information Processing and Management, 23 : 4 (1987) 257.

Cortez, E. M. et al. (1995). The hybrid application of an inductive learning method and a neural network for intelligent information retrieval. // Information Processing and Management, 31 : 6 (1995) 790.

Croft, W.B.; Turtle, H.R. (1993). Retrieval Strategies for Hypertext. // Information Processing & Management, 29 : 3 (1993) 313-324.

Cugini, J. et al. (1997). Interactive 3D Visualization for Document Retrieval. // NIST, 1997, 1-8. http://zing.ncsl.nist.gov/~cugini/vicd/viz.html

Frisse, M.E.; Cousins, S.B. (1989). Information Retrieval from Hypertext: Update on the Dynamic Medical Handbook Project. // Proceedings Hypertext '89, New York: ACM, 11, 1989, 199-212.

Gray, G.L. (1992). Combining expert systems and hypertext: opportunities and obstacles. // Intelligent Systems in Accounting, Finance and Management, 1 : 1 (1992) 21-28.

Gruber, T.R. (1994). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. // Formal Ontology in Conceptual Analysis and Knowledge Representation, Guarino and Poli (eds.). Kluwer Academic Press, 1994.

Halasz, F.G. (1988). Reflections on Notecards: Seven Issues for the Next Generation of Hypermedia Systems. // Communications of the ACM, 31 : 7 (1988) 836-852.

Ingwersen, P. (1992). Cognitive Perspectives of Information Retrieval Interactions: elements of a cognitive IR Theory. // Journal of Documentation, 52 : 1 (1996) 34.

Kramer, R. et al. (1997). Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies. // International Journal Digital Libraries, 1 (1997) 122-131.

Minsky, M. (1975). A framework for representing knowledge. // The psychology of computer vision, New York : McGraw-Hill, 1975, 211-277.

Moya, F. de; Herrero, V.; Guerrero, V. (1998). La aplicación de Redes Neuronales Artificiales a la recuperación de la información. // Anuario SOCADI 1998, 147-164.

Paice, C. D. (1991). A Thesaural Model of Information Retrieval. // Information Processing Management, 27 : 5 (1991) 435.

Quillian, M.R. (1968). Semantic memory. // Semantic Information Processing, Cambridge, MA: MIT Press, 1968, 216-270.

Shneiderman, B. (1986). Designing menu selection systems. // Journal of the ASIS, 37 : 2 (1986) 57.

Sillince, J.A. (1992). Argumentation-based indexing for information retrieval from learned articles. // Journal of Documentation, 48 : 4 (1992) 387-401.

Soergel, D. (1993). Information Structure Management: A Unified Framework for Indexing and Searching in Database, Expert, Information-Retrieval, and Hypermedia Systems. // Services technical report, University of Maryland: College of Library and Information, 1993.

Soergel, D. (1997). An Information Science Manifesto. // ASIS Annual Meeting Coverage, dec. 1997.

Turtle, H. R. ; Croft, W. B. (1992). A comparison of text retrieval models. // The Computer Journal, 35 : 3 (1992) 279-290.

Warner, A.M. (1994). The role of linguistic analysis in full-text retrieval. // Challenges in Indexing Electronic Text and Images, Medford, NJ: ASIS, 1994, 265-275.

West, L.; Murray-Rust, P. (1997). Steps towards the global linking of knowledge. // Managing Information. (May 1997) 36-39.

Wille, R. (1992). Concept Lattices and Conceptual Knowledge. // Systems, Computers & Mathematics with Applications, 23 : 6-9 (1992) 493-515.

Woods, W.A. (1975). What's in a Link: Foundations for Semantic Networks. // Representation and Understanding: Studies in Cognitive Science. Bobrow and Collins (eds.), New York : Academic Press, 1975, 35-82.