

TEORÍA PARA UN MODELO CONCEPTUAL DE RECUPERACIÓN DE OBJETOS MULTIMEDIA.

José A. Moreiro González
jamore@bib.uc3m.es

Miguel-A. López Alonso
cuevaslo@bib.uc3m.es

Departamento de Biblioteconomía y Documentación
Universidad Carlos III de Madrid.

Resumen:

Se propone el diseño de un modelo para el procesamiento de la información que integre un espacio de indización ontológico junto con un espacio documental objetivo, de manera que se eviten las limitaciones de las recuperaciones con hipertexto, además de su reutilización para la autogeneración de Tesoros Conceptuales a partir de las bases de conocimientos de los Agentes Inteligentes. Al mismo tiempo, se propone la integración de la tecnología hipertextual con las redes neuronales para la conversión en dinámicos de los nodos hipertextuales. Se termina sugiriendo la forma de establecer interfases interactivos en las bibliotecas digitales.

PALABRAS CLAVE:

BASES DOCUMENTALES; TECNOLOGÍA HIPERTEXTUAL; ONTOLOGÍAS;
TESAURUS CONCEPTUALES; REDES NEURONALES.

Summary:

It's proposed a design of specific ontologies to avoid static retrieval limitations in hypertext systems, then, reusing it to self-generate Conceptual Thesauri from the knowledge databases of Intelligent Agents. Also, the hypertext technology integration with semantic nets, to relate concepts before showing users' solutions on graphic man-machine interfaces.

Keywords:

FULL-TEXT DATABASES; HYPERTEXT TECHNOLOGY; ONTOLOGIES;
CONCEPTUAL THESAURI; NEURAL NETS.

MARCO DE APLICACIÓN.

Aceptando la Teoría Cognitiva como materia interdisciplinar, se propone el diseño de un modelo para el procesamiento de la información que integre una macroestructura de conocimientos formada por dos espacios conceptuales relacionados:

- 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, y
- 2) un espacio documental, integrado por LAS BASES DOCUMENTALES utilizadas en cada área del conocimiento.

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

Deberá modelar temporal y espacialmente las percepciones del usuario en situaciones de evolución continua de sus conocimientos, así como los contenidos de las bases de conocimientos y también las fuentes de información. Actuará tanto frente a la realidad externa (entorno) cuanto a la interna (sus convicciones), así como en el control del acceso a la información en su trabajo profesional.

Se reconoce que la precisión es habitualmente 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. En el caso de la construcción de Intranets, además de los agentes inteligentes de Internet para lograr mayor exhaustividad, se deberá combinar el uso de las herramientas de búsqueda relacionales para mejorar la precisión.

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..

REQUISITOS METODOLÓGICOS.

- a) Deberá tenerse en cuenta que nos movemos dentro de los objetivos de trabajo de cualquier profesional, utilizando sus conocimientos heurísticos¹ y fusionándolos con las estructuras del conocimiento empleadas en las tareas de búsqueda, tanto en el caso de los usuarios novicios como en el de los profesionales y agentes inteligentes;
- b) que el resultado debe permitir la movilidad entre los vocabularios controlados (tesauros, clasificaciones, encabezamientos, etc.) de distintos sistemas de información, en el proceso de navegación de las múltiples bases de conocimientos heterogéneas contenidas en el entorno de las redes descentralizadas de Internet.

En el diseño de este modelo conceptual se plantean como premisas:

1) La construcción de estructuras generales del conocimiento.

El usuario necesita sistemas de recuperación de la información basados en potentes estructuras del conocimiento, capaces de analizar su pregunta, compararla con los documentos de la base de datos, e identificar los fragmentos relevantes para su búsqueda.

Para mejorar la precisión de las recuperaciones documentales, se propone el diseño de ontologías especializadas (a partir de las bases de datos con texto completo de las diversas áreas del conocimiento) que enlacen con las preguntas específicas del área tratada, como modelo de arquitectura para el almacenamiento estructurado del conocimiento. Deberán ser reutilizadas en las bases de conocimientos de los Agentes Expertos de la Inteligencia Artificial, para la autogeneración de tesauros conceptuales internos que permitan distinguir los sinónimos, suprimir los homónimos e inducir relaciones asociativas entre los descriptores.

Una ontología para una base de conocimientos de la IA debe abarcar los diferentes tipos de documentos, descripciones conceptuales, sus relaciones, los diferentes problemas científicos; además de índices, descripciones bibliográficas, tesauros, códigos clasificatorios, formalizaciones de validez, información terminológica, etc. Su aplicación debe proporcionar una metavisión de la estructura y de la terminología del dominio que facilite recuperaciones altamente relevantes.

Diversos experimentos con ontologías, en sistemas de categorización textual, han demostrado que las herramientas basadas en el conocimiento proporcionan una precisión

mayor que los métodos estadísticos o lingüísticos². Sin embargo, con algunos tipos de categorías los métodos estadísticos aportan mayor precisión que los del conocimiento, pero, indisolublemente la combinación de ambos aporta los mejores resultados globales³.

2) La definición de mecanismos hipertextuales interactivos.

Que deben ser capaces de relacionar, en un espacio n-dimensional neuronal, las ontologías de conocimientos con la navegación dinámica e interactiva por los diferentes espacios documentales.

La tecnología hipertextual es la mejor herramienta para estructurar enlaces estáticos entre los documentos, y para relacionar los descriptores de los tesauros y los códigos clasificatorios. La organización hipertextual permite moverse a través de una red conceptual de macroestructuras, definida a partir de las asociaciones de ideas del usuario propias de su nivel de conocimientos en una disciplina y de su filosofía personal.

Mientras la información de los trabajos impresos está dispuesta secuencialmente (en capítulos y párrafos con una variedad de índices), la contenida en las bases de datos digitales con texto completo precisa de cierta estructuración que permita la recuperación a partir de las referencias cruzadas entre sus documentos. Algunos investigadores proponen la aplicación de esta organización hipertextual en los agentes expertos, como apoyo directo en la resolución de problemas concretos profesionales, debido a la complejidad de la tarea de creación de las bases de conocimientos de soporte (representación de axiomas y aplicación a ellos de las reglas de inferencia).

Dado que las bases de datos documentales no se diseñan para la resolución directa de problemas, la organización de su conocimiento requiere el análisis de cada documento y el resumen de sus notas más destacadas, mediante dos fases consecutivas: a) la extracción previa de los datos menos relevantes (clasificación), y b) el posterior examen especializado para estructurar el conocimiento referido a temas específicos (indización o resumen).

Las aproximaciones que se apoyan en la Teoría Cognitiva de Recuperación de la Información, sugieren que el mejor diseño para las bases de conocimientos es el de estructura simple en lenguaje natural, que incluya punteros a los textos completos de los documentos. La base de conocimientos contendrá estructuras semánticas capaces de inducir relaciones inferenciales e incluir además dichos punteros. Dicha estructura facilitará las búsquedas de los usuarios para que no sean demasiado rígidas y, además será un medio de relación entre las diferentes bases de conocimientos de los sistemas descentralizados de Internet.

Se propone una estructura de redes semánticas asociativas, en la que las unidades conceptuales representadas por nodos sean meras entidades semánticas y las relaciones representadas por enlaces sean asociaciones de entidades⁴. En tales sistemas híbridos, las bases de conocimientos serán tratadas como si fueran textos cuyas colecciones se enlacen entre sí para soportar "viajes en el espacio informativo", mediante una indización lingüística apoyada en el espacio de indización ontológico propuesto anteriormente⁵.

a) Una de las limitaciones de estos sistemas es el "desbordamiento cognitivo" o desorientación de los usuarios durante su navegación, a través de los textos y las clasificaciones hipertextuales⁶. La unión de la información de numerosas fuentes con un único nodo conduce a una sobresaturación de enlaces en dicho nodo. La introducción de muchos de éstos dificulta la navegación, reduce el deseo de los usuarios de explorar su información e incrementa las posibilidades de perderse. Se trata de utilizar sólo enlaces significativos e incluir en cada nodo una idea conceptual completa que sintetice el conocimiento contenido en las diferentes fuentes enlazadas.

Los sistemas que aplican el Modelo Cognitivo a como las personas reflexionan delante de problemas complejos, se organizan en esencia como redes semánticas de conceptos unidas mediante relaciones no lineales. Su organización permite la movilidad a través de una red

neuronal de macroestructuras de conceptos, definidos a partir de las asociaciones de ideas del usuario, derivadas de su nivel de conocimientos y de su filosofía personal.

Esta aproximación permite que el emisor utilice un modelo mental propio, aplicando metáforas autogeneradas (nuevas estructuras cognitivas) dentro del contexto de cada aplicación, mientras que la puesta en marcha misma puede basarse en el modelo semántico general. Tal aproximación facilita la navegación y el análisis de las bases de datos en estudio, independientemente de la aplicación específica y de los modelos mentales de cada usuario⁷.

b) Otra de las limitaciones de estos sistemas se deriva de la necesidad de controlar manualmente los enlaces creados durante el trabajo automático, dado que éstos no presentan información sobre el contenido de los documentos relacionados "a priori". Al no poder procesar dinámicamente la información existente en sus fuentes, dichos sistemas recuperan solamente los documentos que tienen enlaces predefinidos con un nodo principal o mantienen relaciones con otros. Se trata de proporcionar al usuario herramientas globales de ayuda a la navegación: mapas cognitivos, historia semántica de los nodos visitados, opción de retorno mediante el marcado de nodos, etc.⁸

Se ha defendido que, en los sistemas de hipermedios, la visualización espacial aporta importantes beneficios en la recuperación de la información. Entre ellos destacan la utilización de los sentidos de distancia y dirección de los usuarios, la facilidad de localización de la información conocida y la asignación a la información desconocida de posiciones fáciles de recordar, además de lo intuitivo de una representación espacial bien diseñada. También sirve al usuario como medio de encontrar conceptos y sus relaciones en el espacio de indización utilizado para la formulación de las búsquedas⁹.

CONCLUSIONES Y RECOMENDACIONES.

El procesamiento de la información se desplaza desde el contenido de la información hacia el estudio de los procedimientos utilizados para su tratamiento, buscando ofrecer un acceso plural que someta su funcionamiento a reglas éticas y se evalúe plenamente la calidad¹⁰.

1ª) Como la relación entre los ordenadores y los usuarios es el principal problema para la integración de los atributos de la inteligencia humana y la tecnología, se estima que la recuperación de la información necesita de una estrecha colaboración entre el usuario final y los programas informáticos de la IA, utilizados en los Sistemas Integrados de Recuperación de la Información. Esto deberá lograrse mediante el desarrollo de potentes interfases, capaces de compartir la experiencia y la alta fidelidad de los canales de comunicación del usuario con el poder de procesamiento y la alta velocidad de los canales de comunicación de la teleinformática¹¹.

Para la obtención de un diseño centrado en la filosofía del usuario, se deberán analizar y describir sus necesidades de manera que permitan especificar las tareas que realizará el sistema:

a) La interpretación del Lenguaje Natural de las ecuaciones de búsqueda del usuario, que puede contener términos lingüísticos poco precisos o ambiguos, y difíciles de traducir a un lenguaje documental,

b) la localización de los documentos en la base de datos, que puede seguir la llamada "manipulación directa" con diversos modos de interacción manual, y la utilización de la "inferencia plausible" basada en reglas declarativas sobre la especificación del modo de visualización, de manipulación y de aplicación de los datos,

c) la presentación de los resultados al usuario en la forma que un consultor humano actúa al resolver problemas: a partir del reconocimiento de las Estructuras de Representación del Discurso y del procesamiento del Lenguaje Natural a distintos niveles.

2ª) 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 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 habitual¹². A continuación se utilizarían aplicaciones informáticas de la Inteligencia Artificial, que permitieran la automatización de este proceso de autogeneración de Tesoros Conceptuales de abajo-arriba.

b) Se partiría de las bases de conocimientos Terminológicos y mediante análisis semántico (indización semántica latente, etc.¹³) se irían extrayendo conceptos ordenados por facetas siguiendo un criterio relacional previamente definido. Desde estos conceptos se desarrollarían diferentes Tesoros 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 Sillince¹⁴ o “extractos tesauroizados” por Paice¹⁵-.

3ª) Como en la red hipertextual de documentos de las grandes bases de datos abarcadas por Internet, es imposible alcanzar “un contenido conceptual diferenciado” con las búsquedas hipertextuales, se provoca “ansiedad en la recuperación de la información”, que conlleva la pérdida del deseo de exploración de los usuarios.

a) Una de las recomendaciones es la conexión de las diferentes ontologías externas, aprovechando los recientes avances en tecnologías de interoperabilidad abierta (WWW, CORBA y Java¹⁶), de manera que cualquier usuario de Internet se pueda beneficiar al mismo tiempo de varias de ellas que actúen en cooperativa (red de hipertesauros¹⁷).

b) Otra solución que se apunta es la utilización del modelo entidad-relación de las bases de datos Relacionales en el núcleo del Agente Experto que impulsa dichos robots rastreadores, de manera que cuando se detecte una “relación global” entre la pregunta del usuario y un documento de la base de datos, se analicen sus oraciones principales buscando las “pequeñas piezas de información” citadas antes, que puedan ser reutilizadas en la enunciación de nuevas ecuaciones de búsqueda, y lleven a un filtrado de pequeñas piezas de información en el texto completo que al combinarse aumenten la precisión en la extracción de documentos por su contenido contextual. Esto permitirá acomodar las aproximaciones de las diferentes perspectivas de varios usuarios al mismo tipo de información (Teoría Cognitiva aplicada a la recuperación de la información¹⁸).

c) Finalmente, para ayudar a alcanzar seguridad a los usuarios, se recomienda la implantación en los interfases de búsqueda de los Sistemas Integrados de Recuperación de la Información de un seguimiento visual de la estructura conceptual de la red neuronal, de manera que abarque el sistema de clasificación utilizado, los conceptos de los tesauros encadenados, y los vínculos dinámicos hipertextuales que se vayan creando con los documentos almacenados de su interés.

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.¹⁹ Para plasmar espacialmente dichos contenidos se vienen experimentando, con mejor o peor fortuna, los interfases interactivos tridimensionales (3D)²⁰, cuyo análisis escapa de las pretensiones de este trabajo.

4ª) La evolución del papel de las bibliotecas como consecuencia de la manipulación masiva de información digitalizada multimedia (imagen, vídeo y sonido), y del cambio de su papel de archivo de información por el de directorio de distribución de fuentes, ha motivado que la naturaleza del acceso a la información se haya redefinido como “un acceso intelectual o cognitivo”²¹, según lo que el usuario demanda en cada momento y no conforme a lo poseído por cada biblioteca.

Para evitar que se pierda la oportunidad de diseñar sistemas en perfecta armonía con las preferencias de los usuarios, sus interfases deberán tener las siguientes características

mínimas:

- a) Aceptar la "manipulación directa" durante el rastreo sofisticado de los documentos²².
- b) Proporcionar "ayuda continuada", a través de un diálogo interactivo con estrategias alternativas de búsqueda²³.
- c) Integrar diferentes mecanismos de precisión: bases de conocimientos, tesauros, etc., aplicaciones y medios de difusión²⁴.
- d) Describir en sus interfases las estructuras estratégicas de control del grado de implicación del usuario en las búsquedas, y el tipo de capacidades exigibles al sistema²⁵.
- e) 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 formato²⁶.

¹ HARTER, S. P. "Psychological Relevance and Information Science". *Journal of the ASIS*, 1992, 43 (9), pp. 602-615.

² HAYES, P.J.; WEINSTEIN, S.P. "CONSTRUE/TIS: a system for content-based indexing of a database of news stories". *Proceedings Second Annual Conference on Innovative Applications of AI*, 1991, pp. 49-64.

³ JACOBS, P.S. "Joining statistics with NLP for text categorization". *Proceedings Third Conference on Applied Natural Language Processing*, 1992.

⁴ DEWÈZE, A. *Réseaux sémantiques: essai de modélisation; application à l'indexation et à la recherche documentaire*. Lyon: Universidad Claude Bernard. Tesis doctoral, 1981, p. 68.

⁵ SPARCK JONES, K. "Fashionable trends and feasible strategies in information management". *Information Processing Management*, 1988, 24 (6), pp. 703-711.

⁶ ROVIRA, C. "Entornos hipertextuales de aprendizaje". *Anuario SOCADI*, 1997, pp. 121-126.

⁷ CARLSON, P.; GONZALES, G. "The knowledge board: Using hypertext as an intelligent workspace for writing issues-based prose". *Journal of educational multimedia and hypermedia*, 1993, 2 (14), pp. 417-431.

⁸ CODINA, L. *El llibre digital*. Barcelona: Centre d'investigació de la Comunicació, 1996.

⁹ SHUM, S. "Real and Virtual Spaces: Mapping from Spatial Cognition to Hypertext". *Hypermedia*, 1990, 2 (2), pp. 133-158.

¹⁰ SOERGEL, D. "An Information Science Manifesto". *ASIS Annual Meeting Coverage*, dec. 1997.

¹¹ MARCHIONINI, G. "Interfaces for End-User Information Seeking". *Journal of the ASIS*, 1992, 43 (2), p. 156.

¹² BUCKLAND et al. "Partnerships in navigation: an information retrieval research agenda". *ASIS Annual Meeting*, Chicago, octubre 1995.

¹³ CAID, W.R. et al. "Learned vector-space models for document retrieval". *Information Processing & Management*, 1995, 31(3), pp. 419-429.

¹⁴ SILLINCE, J.A. "Argumentation-based indexing for information retrieval from learned articles". *Journal of Documentation*, 1992, 48(4), pp. 387-401.

¹⁵ PAICE, C.D. "A Thesaural Model of Information Retrieval". *Information Processing & Management*, 1991, 27(5), pp. 433-447.

¹⁶ KRAMER, R. Et al. "Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies". *International Journal Digital Libraries*, 1997, 1, pp. 122-131.

¹⁷ WEST, L. y MURRAY-RUST, P. "Steps towards the global linking of knowledge". *Managing Information*, mayo 1997, pp. 36-39.

¹⁸ LÓPEZ ALONSO, Miguel-A. "El modelo conceptual en los sistemas digitales integrados de recuperación de la información". *Revista SCIRE*, 1998, pp. (pendiente de publicación).

¹⁹ MOYA, F.de, HERRERO, V. y GUERRERO, V. "La aplicación de Redes Neuronales Artificiales a la recuperación de la información". *Anuario SOCADI* 1998, pp. 147-164.

²⁰ CUGINI, J. et al. "Interactive 3D Visualization for Document Retrieval". *NIST*, 1997, pp. 1-8.
<http://zing.ncsl.nist.gov/~cugini/vicd/viz.html>

²¹ GRIFFIN, S.M. "Taking the Initiative for Digital Libraries". *The Electronic Library*, 1998, 16 (1), pp.24-27.

²² SHNEIDERMAN, B. "Designing menu selection systems". *Journal of the ASIS*, 1986, 37 (2), p. 57.

²³ TURTLE, H. R. y CROFT, W. B. "A comparison of text retrieval models". *The Computer Journal*, 1992, 35 (3), pp. 279-290.

²⁴ INGWERSEN, P. "Cognitive Perspectives of Information Retrieval Interactions: elements of a cognitive IR Theory". *Journal of Documentation*, 52 (1), 1996, p. 34.

²⁵ BATES, M. J. "Where should the person stop and the information search interface start?". *Information Processing Management*, 1990, 26 (5), p. 577.

²⁶ BELKIN, N. J. et al. "Distributed Expert-Based Information Systems: An Interdisciplinary Approach". *Information Processing Management*, 1987, 23 (5), p. 407.