Revista Interamericana de Bibliotecología Medellín Vol. 21 No. 1. Enero-Junio de 1998

LA GESTION AUTOMATIZADA DE TESAUROS: ESTADO DE LA CUESTION

Miguel Angel López Alonso*

RESUMEN

Visión integradora del estado del arte de la gestión automatizada de tesaurus que describe no sólo los gestores tradicionales, sino los incorporados en programas de gestión de bases de datos, en sistemas automatizados de gestión, los interactivos y los sistemas expertos, amén de algunas investigaciones recientes como la cadena de confección automática de tesaurus multilingües (CAT) y el proyecto gestor de tesaurus de la Universidad Carlos III de Madrid (España).

INTRODUCCION

La visión que, se presenta en este artículo de revisión del tema de la gestión automatizada de tesaurus, intenta ser integradora, y abarca no sólo los gestores independientes de cualquier tipo de software, sino también los módulos más empleados en los sistemas integrados de almacenamiento y recuperación de la información o de automatización de centros de documentación.

Se mencionan algunas de las herramientas semánticas contenidas en los sistemas de recuperación de la información con texto completo, que utilizan técnicas de gestión de los términos próximas al NLP (natural language processing) de la inteligencia artificial. Éstos sistemas recopilan los términos candidatos para su posterior revisión por los compiladores humanos que realizan la fase intelectual de compila-

* Miguel Angel López Alonso, Dr. en Documentación. Prof. de la Universidad Carlos III de Madrid, España.
Vallehermoso, 55 28015 Madrid
Email: cuevaslo@bib.uc3m.es

ción de todo tesauro conceptual, y la validación de las relaciones asociativas de tipo inductivo o inferencial.

No es necesario argumentar que se trata de una tarea incompleta, ya que es imposible conocer la evolución de todos los sistemas existentes; además, la información varía de unos países a otros muy rápidamente y da lugar a que las características de los desarrollos, válidas en el momento de redactar este trabajo, se encuentren obsoletas cuando se requiera su consulta.

Gestores tradicionales

Este tipo de gestores de tesauros actúa sobre los tesauros generales precoordinados, diseñados para la mejora de la composición con procesadores de textos, que pueden ampliarse progresivamente mediante la adición de descriptores del campo de trabajo específico del usuario. Todos ellos tratan exclusivamente la superficie de las palabras, pero:

Algunos programas como ASTUTE[1], desarrollado por el Centro de Cálculo de la Comisión de las Comunidades Europeas y que funciona sobre IBM 370 o Siemens 7700, gestionan tesauros con estructuras jerárquicas predefinidas de hasta siete niveles arborescentes que deben ser respetadas y que son rigurosamente controladas[2]. Se ha modificado recientemente para la reedición de tesauros clásicos, como el de la IRRD (International Road Research Documentation, 1991) en un entorno multilingüe, de manera que gestiona hasta cinco idiomas en un mismo tesauro, y puede realizar la impresión en tres lenguas simultáneamente[3].

Otros como MINISIS[4], desarrollado por el Centre de Recherches pour le Dévelopment International (CRDI) en Canadá[5] y que funciona sobre HP 3000, utilizan su módulo

1. Automated System for Thesaurus Updating, Testing and Editing (Luxemburgo).
2. Siguiendo el sistema clasificatorio del «Top-down», que establece una Estructura Clasificatoria general y requiere descriptores que representen conceptos para rellenar dicha estructura.
3. FRETIN, M. y Van SLYPE, G. «Etude de faisabilité de la representation graphique de thesaurus gerés par ASTUTE». Paris: Bureau Marcel Van Dijk, 1980. 39 p.
4. Microcomputer Integrated Set of Information System (Unesco).
5. Minisis. CRDI, apartado postal 8500, Ottawa (Ontario), Canadá, K1G 3H9.

de gestión de tesauros THES para ponderar los descriptores en la indización y en la búsqueda documental[6]. Recientemente ha incorporado una aplicación terminológica con un interfase en cuatro idiomas y un tratamiento de bases de datos terminológicas que incluye: alfabetos latinos y no latinos, intercambio de información de acuerdo con las normas internacionales, ayuda contextual en línea, una red de ayuda internacional y el funcionamiento con plataformas múltiples (DOS, RL/DOS, MPE/IX, Unix, Pathworks en VAX/VMS, etc.).

Gestores incorporados en programas de gestión de Bases de Datos

Se trata de gestores incorporados dentro de los grandes sistemas de recuperación de información más ampliamente difundidos en diversas instituciones españolas[7].

El Gestor de Bases de Datos MISTRAL[8], desarrollado por el Centro de Investigación Informática de Honeywell Bull, que en su versión V funciona en ordenadores HBull y se utiliza en grandes bases de datos jurídicas: el Boletín Oficial del Estado, el Instituto Nacional de las Administraciones Públicas, el Instituto Nacional de Seguridad e Higiene en el Trabajo, el Gobierno Vasco, etc.

La gestión y la puesta al día del fichero tesauro se hace por el método tradicional de tratamiento por lotes. La visualización es convencional, sin posibilidades de navegación por el árbol jerárquico o relacional, no gestionando las relaciones descriptor-no descriptor ni el multilingüismo. Gestiona tesauros precoordinados o admite la incorporación automática en el tesauro de términos nuevos que no figuren en el diccionario de palabras vacías. Permite la impresión completa o de sólo una parte del tesauro, como edición alfabética o como diccionario conceptual; aunque

---

6. Siguiendo el sistema clasificatorio de abajo-arriba («Bottom-up»), utilizado en los vocabularios postcontrolados, que proporcionan una simple lista de términos con referencias cruzadas entre ellos, sin ningún tipo de estructura clasificatoria previa.

   HOPKINSON, A. «The Mini-Micro CDS/ISIS software package». Information Development, 1989, 5 (3), pp. 135-137.

7. PAEZ MAÑA, J. Bases de Datos Jurídicos. Madrid: CINDOC, 1994, pp. 211-231.

8. Mémorisation d'Information, Sélection, Traitement et Recherche Automatique (France).

éste último da la jerarquía completa en la que está situado este término y hace menos manejable la edición[9].

Otro de los programas incorporados en grandes sistemas es BASISPlus[10], desarrollado por el Battelle Institute de Londres y distribuido por Information Dimensions Inc. de Dublin: Ohio (USA), que funciona en ordenadores IBM 4341 y se utiliza en extensas bases de datos jurídicas de la Generalitat y del Parlament de Catalunya, en las bases de datos bibliográficas del CSIC, etc.

Sus funciones son muy completas y evoluciona continuamente, incluso se ha incorporado recientemente un módulo para trabajar como servidor dentro de Internet. Acepta una gestión integral automatizada: catalogación, indización e intercambio de documentos primarios electrónicamente, que utiliza un tesauro tanto en la indización como en la interrogación.

El módulo de gestión y control del tesauro define hasta trece tipos de relaciones: recíprocas, jerárquicas, sinonímicas, etc., y permite verificar los términos nuevos propuestos para validación, a partir del tesauro existente. Incluye varios programas de utilidades para el mantenimiento del tesauro:

- BROWSE, diseñada para la recuperación, que proporciona una lista con sólo términos principales y relaciones subordinadas,

- TFPRINT para impresión del tesauro en distintos formatos seleccionables, con despliegue de un nivel jerárquico de relaciones, y

- TFSCAN para examinar la ficha completa de cada término del tesauro (estructura interna y organización del tesauro)[11].

---

9. ROHOU, C. "La gestion automatisée des thésaurus. Étude comparative de logiciels". Documentaliste, 1987, 24 (3), p. 105.

10. Battelless Automatic Search Information System. Londres: U. K. http://www.idi.oclc.org/html/basisv8.htm

11. A Clarification of Claims made by Verity Corp. vis-a-vis the Capabilities of BASIS plus. La Jolle, CA: Information Dimensions Inc., mayo de 1990.

Se ha citado también a BRS/Search por su gran potencia de procesamiento, en cualquiera de sus versiones, la desarrollada en Ensamblador o la más moderna en C, que funcionan en todo tipo de ordenadores y se utilizan en el Tribunal Constitucional, la Radiotelevisión Valenciana y, más recientemente, en bibliotecas universitarias y en los puntos PIC del Ministerio de Cultura, etc.

La versión en C dispone de un completo módulo para la creación y mantenimiento de tesauros, guiado por menús, que posibilita el tratamiento de las relaciones mediante el procesamiento por lotes, y se encarga de la generación automática de las relaciones recíprocas[12].

Igualmente, se debe mencionar el programa específico UNIDAS, que ha desarrollado Univac para sus ordenadores Sperry-Unisys, y es utilizado en la Comunidad Autónoma de Madrid, el Ministerio de Educación y Ciencia, etc.

El gestor de tesauros ordena alfabéticamente los descriptores usados, y construye un glosario postcoordinado que puede ser estructurado después en forma de tesauro.

Gestores incorporados en Sistemas Automatizados de Gestión

Teniendo en cuenta los diferentes modelos, estadístico, probabilístico, semántico, procedimental, etc., utilizados en los sistemas de indización automatizada, revisaremos el software de gestión de tesauros según los sistemas de indización en que se integran. Los criterios de valoración de los gestores de tesauros deben tener en cuenta preferentemente:

- La facilidad con la que el tesauro generado puede integrarse en el sistema de gestión de la información, y
- la calidad de las funciones de validación del vocabulario que se va incorporando al tesauro.

12. ANON. «BRS-Search full text retrieval software packages». Information Today, 1990, 7 (4), p. 62. (http://www.dataware.com/site/prodserv/online.htm

Los incorporados en Sistemas Estadísticos

Se destaca entre ellos el gestor de tesauros PASSAT, desarrollado por Siemens e incorporado en el Sistema Golem, que utiliza Volkswagen para el mantenimiento de su banco de datos terminológico.

Su desarrollo se fundamenta en el estudio estadístico de los términos del documento, según el conocido principio de que las palabras clave son las que aparecen más de una vez en el texto analizado, pero, no tan a menudo como las palabras más frecuentes (siguiendo la curva de Zipf-Mandelbrot). Al analizar el texto, asocia a cada palabra en lenguaje natural uno o varios descriptores del tesauro con los que crea una «matriz de asociación». Calcula las ocurrencias de estos descriptores asociados, convierte su frecuencia en peso, y retiene como descriptores nuevos aquellos que tienen un peso superior al umbral fijado previamente en la aplicación.

Utiliza el léxico completo de un lenguaje natural: alemán, inglés u holandés, además de listas de sufijos en función del género, número, declinación y conjugación. Cada usuario debe elaborar su propio fichero vocabulario y unirlo con el fichero de desinencias y enlaces[13].

Los incorporados en Sistemas Probabilísticos

Es el caso del veterano CITE, desarrollado por Medlars (Medical Literature Automatic Retrieval Systems) para su integración en el sistema de recuperación en línea CATLINE, cuando todavía el uso del vocabulario controlado del MeSH (Medical Subject Headings) de la NLM (National Library of Medicine) era imprescindible, ya que menos del 50% de sus registros contenían los resúmenes en lenguaje natural.

Se concibió para que las preguntas en lenguaje natural permitiesen localizar los documentos indizados en la base de artículos biomédicos de la NLM, utilizando tesauros especializados. La dificultad de relacionar dicho lenguaje y el vocabulario controlado del tesauro MeSH se resolvió en varias fases sucesivas:

---

13. HOFFMAN, D. et al. PASSAT. Programme danalyse automatique de texte. Paris: Siemens, 1971, 23p.

- Inicialmente, se utilizaba la relevancia de una serie de documentos encontrados en la primera búsqueda por los usuarios, sin estudio de frecuencias ni tamaño de la información, añadiendo CITE a la nueva ecuación de búsqueda y aquellos descriptores que el vocabulario MeSH asignaba a cada uno de estos documentos.

- Posteriormente se añadirían: el estudio de los radicales y la identificación de sus variantes en el vocabulario biomédico, una clasificación de los descriptores preferentes usados en el sistema (divididos en categorías y subcategorías para cada uno de los 82 encabezamientos de materias del MeSH) y un análisis de las frecuencias de los registros que el usuario consideraba más relevantes. Estos elementos permiten que el sistema despliegue una «lista electrónica» de materias relacionadas, como términos complementarios para una nueva selección por el usuario.

- Finalmente, los resultados pueden mejorarse con la incorporación de nuevas rutinas de búsqueda que hayan sido probadas por otros usuarios anteriores[14].

Los incorporados en Sistemas Lingüísticos

Son los más utilizados en el ámbito de la CEE por tener un desarrollo práctico suficientemente avanzado[15] y, entre ellos, debemos resaltar ALETH-DOC que ha sido utilizado por el programa europeo Impact para el desarrollo de su Maqueta de Interrogación Multilingüe. Desarrollado por GSI-Erli (París, Francia) para su uso, con el módulo de indización ALEXDOC, en el Sistema de Indización y Recuperación Documental ALEXIS.

Tras una primera identificación de los términos de la pregunta, efectúa tres niveles de análisis:

- Comparación de los términos seleccionados con los descriptores del tesauro,

14. DOSZKOCS, T. E. CITE NLM: Natural-language searching in an online catalog. Information Technology and Libraries, 1983, 2 (4), pp. 364-380.

15. Apoyados en estudios teóricos con distintos niveles de profundidad analítica: desde el más simple nivel morfológico, pasando por los niveles lexical y sintáctico, hasta el más complejo nivel semántico.

- eliminación de los unitérminos que forman parte de descriptores multitérminos, y
- agregación de los términos genéricos, para formar la ecuación booleana de búsqueda en lenguaje documental[16].

A continuación establece las relaciones entre los distintos términos del tesauro y remite a los documentos indizados que los incluyen.

En la gestión del tesauro, se deben señalar las diferencias existentes entre la fase de indización, en la que el sistema sugiere descriptores (a partir de los textos) que pueden ser validados o no por el indizador, y la fase de consulta, en la que la ecuación de búsqueda ofrecida por el sistema (a partir de la pregunta del usuario) puede ser modificada de manera interactiva.

Como base de conocimientos, incluye un diccionario lingüístico con aspectos sintácticos de los verbos, adjetivos o adverbios. En el plano semántico utiliza dos facetas la de «acción» y la de «humano», para clasificar los términos del tesauro[17].

Gestores interactivos

En estos sistemas se tratan, además de los documentos (en la entrada), las preguntas como texto a indizar (en la salida). La mayoría de ellos poseen módulos analizadores de las preguntas y pueden responder en el lenguaje natural del usuario.

Entre otros, se destacan los dos siguientes:

- El Gestor de Tesauros STRIDE, desarrollado por BNF Metals Technology Centre (Oxfordshire, U. K.) e incorporado en el sistema STATUS\IQ, que posee un analizador del lenguaje natural de las ecuaciones de búsqueda del usuario.

---

16. Mediante análisis morfológico: relaciones de sinonimia y de afinidad, eliminación de homografías, reconocimiento de multitérminos y eliminación de palabras vacías, y, también, mediante análisis sintáctico. http://www.erli.fr/erli/AlethTR/FrTR.htm

17. HILDRETH, C.R. "Intelligent interfaces and retrieval methods for subject searching in bibliographic retrieval systems". Washington, D.C.: Cataloguing Distribution Service, Library of Congress, 1989, p.80.

Genera una lista de documentos con una detallada información estadística de los conceptos hallados en ellos, basada en la frecuencia de la ocurrencia de los términos buscados y en su grado de proximidad, y proporciona la representación y recuperación de fragmentos de dichos documentos[18].

Valida los términos clave, soporta cualquier tipo de relaciones terminológicas definidas previamente (incluidas AND, BTG/NTG Y BTP/NTP del último estándar ANSI para tesauros[19]), así como la definición de menús, campos, etc. por parte del usuario[20].

- El Gestor de Tesauros CLARIT, desarrollado por el Centro de Lingüística de la Universidad Carnegie Mellon (Pittsburgh, USA), que es, a la vez, un sistema automático de indización y un sistema de construcción de tesauros postcoordinados de primer orden[21].

Se parte de un tesauro de terminología contrastada, previamente organizada jerárquicamente mediante relaciones lexicales de principal (BT) o derivado (NT), que se autogenera a partir de la literatura recopilada para una aplicación particular del dominio. Si encuentra sintagmas nominales, los convierte en candidatos para la indización, tras un tratamiento morfológico en el que se comparan con el tesauro y se clasifican como términos exactos, generales o nuevos.

Usa técnicas de procesamiento del lenguaje natural para identificar las estructuras sintácticas que genera el tesauro, aunque sin definir los tipos de relaciones semánticas entre términos individuales. Extiende los descriptores de los tesauros documentales hacia frases ponderadas, a partir de «situaciones específicas» condensadas de los textos analizados[22].

---

18. Sistema basado en el Principio de «Probability Ranking» de Sparck Jones/Robertson, en el que las series de párrafos o secciones puedan ser manipuladas para crear estructuras neuronales conceptuales: expandibles en diferentes direcciones dentro del espacio de la información, fusionadas o simplemente ignoradas algunas de ellas. INGWERSEN, P. «Cognitive Perspectives of Information Retrieval Interactions: elements of a cognitive IR Theory». Journal of Documentation, 52 (1), 1996, p. 34.

19 KROOKS, D. A. y LANCASTER, F. W. «Evolution of guidelines for thesaurus construction». Libri, 1993, 43 (4), p. 328.

20. PAPE, D. L. y JONES, R. L. «STATUS with IQ: Escaping from the Boolean straightjacket». Program, 22 (1), pp. 32-43. http://www.questans.co.uk/p10012.html

21. EVANS, D. A. et al. «A Summary of the CLARIT Project». Pittsburgh, PA: Carnegie Mellon U. Department of Philosophy, Laboratory for Computational Linguistics (Report No. CMU-LCL-91-2).

22. PAICE, C. D. «A Thesaural Model of Information Retrieval». Information Processing Management, 1991, 27(5), p. 435.

Sistemas Expertos

Los sistemas comerciales de recuperación de la información más ampliamente difundidos: Dialog, Medline o Pascal, han ido evolucionando en su forma de acceso a la información, desde los ficheros inversos y la lógica booleana, a la utilización de perfeccionados tesauros conceptuales, diseñados específicamente para la recuperación del conocimiento contenido en la información documental[23].

A pesar de que estos enormes Bancos de Datos no pueden cambiar radicalmente, desde hace años están dando pasos hacia una indización cada vez menos dependiente del factor humano y de la realización de resúmenes:

- DIALOG ha venido utilizando el programa cliente KR ProBase para guiar a los usuarios en las búsquedas con menús inteligentes, corrección de errores y propuesta de nuevos términos. También, investiga algoritmos de ponderación de los descriptores y métodos de medición de la relevancia de las referencias, para mejorar la eficacia de las recuperaciones (ej. convenio con los propietarios del software Personal Librarian[24], desarrollo de Target,[25] etc.),

- MEDLINE ha desarrollado el proyecto MedIndEx System para la indización automatizada a partir de sistemas expertos, con la finalidad de proporcionar ayuda interactiva a los indizadores humanos, mediante bases de conocimientos terminológicas que mejoren la oferta de conceptos de los tesauros documentales[26], y

  El INIST experimenta, en su base de datos PASCAL, el sistema conceptual de indización (Lexinet[27]) con estructuras de análisis[28], para ayudar al indizador

23. LOPEZ-HUERTAS, M. J. Thesaurus structure design: a conceptual approach for improved interaction. Journal of Documentation, 1997, 53 (2), pp. 139-177.
24. Personal Librarian de Personal Library Software, Rockville, MD (USA)
25. TENOPIR, C. y CAHN, P. «Target & Freestyle: Dialog and Mead join the relevance ranks». ONLINE, 1994, 18 (3), pp. 31-47.
26. HUMPHREY, S. M. «Medical Indexing Expert System». Information Processing & Management, 1989, 25 (1), pp. 73-88.
27. CHARTRON, G. «Lexicon management tools for large textual databases: the Lexinet system». Journal of Information Science, 1989, 15 , pp. 339-344.
28. Como conjunto de términos organizados y estructurados dentro de cada Campo del Conocimiento.

manual en la identificación de los conceptos y su transcripción al lenguaje documental, que es capaz de encontrar nuevos términos en los documentos sin utilizar tesauros documentales preestablecidos (Genelex).

Serán, por el contrario, los sistemas pequeños con implantación restringida, los que experimenten con los programas más avanzados para la gestión automatizada de tesauros.

Gestores tradicionales incluidos en Sistemas Expertos

Se ha comparado el Gestor de Tesauros TCS, que es capaz de reconocer conceptos y clasificar textos en el sistema CONSTRUE/TCS de Reuter, con THESYS que desarrolla un análisis de textos contextual avanzado en el sistema TINA de Siemens.

- El Text Categorisation Shell de Liu Palmer, California (USA)[29], se utiliza por la agencia Reuters incorporado en el sistema CONSTRUE/TCS, con dos procedimientos propios de un Sistema Experto durante el proceso de indización, a saber:

  - El reconocimiento de conceptos a través de su definición, como «conjunto de palabras y frases indicativas predefinidas en una regla base», y

  - la asignación de un texto frente a sus términos de indización, a partir de las reglas de programación «If-then», y con gran flexibilidad en la codificación de aplicaciones particulares.

  Durante el proceso de confrontación de los términos de indización del texto y los descriptores del tesauro, se marcan los términos sobre los que existe divergencia para una revisión manual que enriquezca la indización de los documentos más pobremente descritos.

  La autogeneración del tesauro es bastante rápida. Sigue el sistema de relaciones facetadas y obtiene altos niveles de exactitud en las recuperaciones. Cuenta, en

29. http://www.liu-palmer.com

su versión profesional, con el despliegue de validaciones de una amplia gama de relaciones: la eliminación de duplicados, la eliminación de conflictos entre términos emparejados y la prevención de conflictos de relaciones de jerarquías circulares[30].

- A su vez, THESYS del proyecto TINA (Text-Inhalts-Analyse) de Siemens, ha sido concebido como desarrollo de varias herramientas de recuperación correlacionadas: analizador sintáctico, indizador automático de términos, gestor de tesauros, etc. No es un Sistema de Recuperación basado en las técnicas de la Inteligencia Artificial, sino en el perfeccionamiento de los métodos estadísticos y combinatorios lexicométricos.

  Por cada término elegido de la Base de Datos, las relaciones del modificador principal (experto) son generadas nuevamente de manera que puedan ser reinterpretadas de acuerdo con las relaciones terminológicas de los descriptores principales, secundarios o relacionados. También, se ha desarrollado un proyecto de tesauro multilingüe alemán-inglés que efectúa la búsqueda automática de textos en inglés con ecuaciones de búsqueda en alemán.

  El modelo teórico crea estructuras independientes, con el objeto de construir campos semánticos que puedan ser comparados con los términos de un corpus lingüístico, de manera que si ambos especifican palabras idénticas obtengan la misma referencia semántica. En el momento de la recuperación, no solo compara los términos de la pregunta con los de la indización sino que, también, busca coincidencias en las relaciones de los campos semánticos[31].

Gestores originados en el campo de la Inteligencia Artificial

Se diferencian de los gestores de tesauros incluidos en los primitivos Sistemas Expertos en que, además de la gestión, tienen capacidad de conocimiento para analizar la información recibida en las ecuaciones de búsqueda de los usuarios y reutilizarla como alimentación de nuevas búsquedas[32]. La inteligencia artificial

30. HAYES, P. J. y WEINSTEIN, S. P. «CONSTRUE/TCS: a system for content-based indexing of a database of news stories». En: Proceedings of the 2nd Annual Conference on Innovative Applications of Artificial Intelligence. Menlo Park, CA: AAAI Press, 1990, pp. 49-64.

31. SCHWARZ, C. «Automatic syntactic analysis of free text». Journal of ASIS, 41 (6), 1990, pp. 408-417.

32. Proporcionando respuestas a preguntas concretas, de una materia determinada.

hace uso para ello de los útiles de una de sus ramas, los sistemas informáticos, para procesar la información en lenguaje natural, razonar las conclusiones e indicar los caminos a tomar para resolver los problemas planteados.

- Entre los sistemas de indización automatizada pioneros en esta tecnología, destaca el prototipo interactivo MEDINDEX de la National Library of Medicine[33], que utiliza el vocabulario controlado Mesh y fue diseñado para ayudar al experto manual en la indización de literatura médica[34].

  Selecciona los términos de acuerdo con las reglas del *Manual de Indización* de MEDLARS, a saber:

  - conformidad con los descriptores del Mesh,
  - consistencia con las reglas de indización de Medlars,
  - especificidad en la asignación de títulos a los resúmenes, y
  - multiplicidad en la indización, asignando a cada artículo tantas subentradas como el indizador considere útiles, además de la específica del tema.

  Su aportación principal es el desarrollo y experimentación del software para la construcción de la base de conocimientos terminológicos del Sistema Experto que utiliza una estructura de clases en paneles (con diferentes facetas) para la representación de los conceptos (entidades indizables), y los relaciona entre sí mediante redes neuronales (tipo evolucionado de red semántica). Cada concepto citado por una estructura se une con otros[35], para formar la citada base de conocimientos terminológicos.

---

33. http://www.nlm.nih.gov
34. HUMPHREY, S. M. «Medical Indexing Expert System». Information Processing & Management, 1989, 25 (1), pp. 73-88. http://muscat.gdb.org/repos/medl/
35. Junto con sus procedimientos y datos asociados, mediante relaciones específicas del tipo de la leyes de la herencia en la Programación Orientada a Objetos.

- El paso siguiente ha sido el desarrollo de *agentes inteligentes* que buscan directamente los conceptos de indización asignados por los indizadores manuales, y los confrontan con los descriptores de un tesauro conceptual.

- Uno de los primeros sistemas en seguir este procedimiento fue el del servicio de indización de la API-CAIS (American Petroleum Institute), cuyo desarrollo data de 1982 y funciona eficazmente desde febrero de 1985, que ha sido considerado como el primer sistema automático de indización que modela las necesidades de los indizadores manuales.

  Selecciona automáticamente conceptos de indización en los resúmenes de los artículos de bibliografía técnica, realizados por indizadores humanos, y los integra junto con los conceptos encontrados por los indizadores manuales del CAIS (Central Abstracting & Indexing Service), para verificarlos con el tesauro API (American Petroleum Institute) tomado como base de conceptos validados[36].

- Otro de estos sistemas es el MAI (Machine Aided Indexing) de la NASA, que funciona con éxito desde 1983. Se utiliza como tabla de conversión de términos candidatos, para ser validados por el indizador manual, antes de su inclusión en el tesauro de términos (NASA Thesaurus), enfrentado al diccionario NLDB (Natural Language Data Base).

  Sus mejoras con respecto a un sistema manual de indización son:

  - Ahorro de tiempo en la indización manual,

  - ampliación de los puntos de acceso en dicha indización, y

  - utilización permitida de otros sistemas de indización (ej.: el intercambio con el tesauro DTIC de términos sustantivos).

---

36. MARTINEZ, C., LUCEY, J. y LINDER, E. «An Expert System for Machine-Aided Indexing». J. Chem. Inf. Comput. Sci., 1987, 27 (4), pp. 158-162. http://www.api.org/news/1booth.htm

Su evolución a partir de 1984 permitió la selección automática de frases, expandió la red conceptual (especie de Esquema de Representación del Conocimiento del tipo del MEDINDEX de Medlars) y sustituyó el diccionario NLDB por el más condensado LD (Lexical Dictionary) que, ahora, se obtiene a partir del análisis semántico en vez del morfológico, y efectúa los procesamientos en tiempo real.

En la actualidad, este sistema se apoya en la terminología de los Dominios Específicos[37] que forman una «lista autorizada» de palabras y frases del texto, generalmente sinónimas de las del tesauro de términos de la NASA, relacionadas con dicho tesauro en la Base de Conocimientos Terminológicos. Esta lista realiza un control de sinonimias que abarca desde las palabras del texto a los términos específicos de indización y no se limita a las frases nominales, con el fin de ampliar los conceptos de indización[38]. Reduce la lista de términos vacíos a solo 250 estadísticamente comprobados que, junto con los signos de puntuación, dividen el texto en una estructura de conceptos en «paneles».

El Sistema MAI es una herramienta para la validación automática de términos del nuevo tesauro de la NASA, partiendo de registros indizados con anterioridad a su creación. Proporciona una utilidad para verificar el deletreo de los términos nuevos, ya aceptados, y encuentra referencias cruzadas entre las voluminosas definiciones del tesauro[39].

- En el ámbito comercial, se han desarrollado algunos programas basados en la recuperación conceptual de la información que pueden integrarse en cualquier tipo de sistema de recuperación para ordenadores compatibles, estaciones de trabajo, etc. Destaca como uno de los más potentes, TOPIC de Verity Inc. (Mountain View: California, USA), que no es un sistema de conocimientos basado en las técnicas de la inteligencia artificial, pero, mejora la recuperación

37. Partes de la oración no ambiguas semánticamente e indizables del tipo de las «situaciones específicas» propuestas por Paice. PAICE, Chris D. «A Thesaural Model of Information Retrieval». Ibid. Cit. n¦ 22.

38. Pudiendo contener verbos, adverbios e incluso combinaciones de palabras no relacionadas gramaticalmente.

39. SILVESTER, J. P., GENUARDI, M. T. y KLINGBIEL, P. H. «Machine-aided indexing at NASA». Information Processing Management, 1994, 30, pp. 631-645. http://saire.ivv.nasa.gov/saire.html

por conceptos sin precisar previamente de un tesauro o de una base de conocimientos terminológicos[40].

Algunas investigaciones recientes

La cadena de confección automática de Tesauros Multilingües (CAT)

En este apartado, se comparan dos Gestores de Tesauros, uno desarrollado en el ámbito del CINDOC de Madrid: el CAT que ha sido programado en dBase IV[41], y el otro promovido en el ámbito del INFOTERM de Viena: el MTM, pensado para su integración con la base de datos CDS/MINISIS de la UNESCO y programado en Isis-Pascal.

- En el primero, los bloques diseñados para el tratamiento terminológico se conforman en dos cadenas fundamentales, denominadas como Microtesauros y Tesauros por el objetivo en ellas planteadas.

  El CAT ha sido diseñado para facilitar la compilación de tesauros, con una toma de datos inicial a partir del tesauro existente en otro idioma[42]. Construye un glosario de equivalencias idiomáticas con el que posteriormente traduce y obtiene el nuevo tesauro, mediante la aplicación del adecuado soporte de software.

  Se parte de los términos componentes del microtesauro, al que podríamos asimilar con un tesauro reducido a sus términos de cabecera, los de mayor jerarquía, con sus relaciones de equivalencia (Use), alternativas (UF), jerárquicas (BT, NT), afinidad (RT) y nota explicativa en un solo sentido.

---

40. BALLOU, M. C. «Verity system to serve as cornerstone of NAS text retrieval». Digital Review, 8 (20), pp. 1-2.

41. CAT. Cadena Informatizada para la Confección Automática de Tesauros. Madrid: CINDOC (Csic), 1991, 23 p.

42. Inicialmente se desarrolló la «Metodología de rotación del lenguaje pivote», para la elaboración del tesauro SPINES trilingüe de la Unesco, en Madrid (CINDOC).
VALLE BRACERO, A. et al. Confección Automática de Tesauros. Rev. Esp. Doc. Cien., 12 (2),1989, pp.129-140.

Ultimado y validado el microtesauro, mono o multilingüe, se puede pasar a verter el fichero a un nuevo idioma y disponer de las prestaciones previstas en su tratamiento: confección partiendo del microtesauro, conversión de formato, rotaciones, ordenación alfabética, eliminación de redundancias, tratamiento de equivalencias idiomáticas, índice permutado, ediciones impresas, actualización, fusión y ordenación alfabética, clasificación temática, incorporación de nuevos idiomas y sustitución del «lenguaje maestro» por otro de los incluidos en el tesauro multilingüe.

Se llama «lenguaje maestro» al lenguaje fuente en que está estructurado el tesauro primitivo. Se llama «fichero maestro» al tesauro multilingüe considerado, una vez transcrito a soporte legible por el ordenador. Se trata de conseguir la reversibilidad, mediante la obtención de un nuevo fichero en el que aparezcan intercambiados, el contenido de los campos que se refieren al lenguaje maestro y al nuevo idioma que se introduce.

- El MTM ha sido diseñado para la confección automática de tesauros multilingües, a partir de la compilación de los términos en un idioma obligatorio y su simultánea reproducción en los demás lenguajes secundarios[43]. La elección de la lengua de trabajo es independiente de la elección de las lenguas del tesauro multilingüe, se puede por ej. añadir un término en español al tesauro mientras que se trabaja en inglés o se puede usar el software MTM totalmente en francés para realizar un tesauro monolingüe en inglés[44].

  El tesauro se almacena en una base de datos del sistema CDS/ISIS con una estructura específica de cinco tipos de ficheros: dos principales (descriptores y no descriptores) que siempre deben estar presentes, una variante del fichero de descriptores o registro de descriptores locales, y dos tipos de registros optativos: el de palabras vacías y el de instituciones.

  Los ficheros deben contener los descriptores traducidos a las lenguas obligatorias, pero no es preciso que contengan la traducción a las otras lenguas optativas.

---

43. Se partió de una adaptación al Macrotesauro de la OCDE (en inglés, francés y español). «Lista básica de términos relativos al desarrollo económico y social». Santiago de Chile: CLADES (ed. preliminar), 1973. ONU/CEPAL

44. Multilingual Thesaurus Management System. User Documentation v. 3.0. Wien: Infoterm, 1992. 130 p.

Pueden incluir notas de alcance en una o más lenguas, y términos asociados unidos a los descriptores por cualquiera de los tipos de relaciones habituales (UF, BT, NT, o RT). Los ficheros de no descriptores no son multilingües, se deben crear registros separados para cada uno de ellos en cada una de las lenguas del tesauro.

Los descriptores locales son similares a los descriptores, pero, se diferencian en su utilización restringida a una institución, y son reemplazados por los descriptores normales del tesauro cuando se intercambian tesauros o registros de aplicaciones diferentes. El fichero de palabras vacías es necesario para evitar puntos de acceso inútiles en la obtención de los índices permutados que no se crean por el usuario sino mediante manipulación del subsistema KWIC («keyword in context»), y se colocan directamente en la base de datos del tesauro.

Ambos gestores llevan varios años funcionando en los que han probado sus bondades, primeramente en miniordenadores y más recientemente en ordenadores personales. El CAT se desarrolló más rápidamente y ha sufrido la ralentización propia de cubrir ampliamente sus objetivos iniciales. Mientras que el MTM ha sido fomentado para su implantación en los países en vías de desarrollo, apoyado en la base de datos MINISIS de la UNESCO.

Ambos sufren de las virtudes y defectos de sus respectivos lenguajes de programación:

- El CAT, programado en dBase IV, tiene formato de campos fijo (48 caracteres por término), requiere un sistema con el gestor de dBase IV cargado y utiliza ficheros de texto en MS-DOS y ficheros para los datos y resultados en dBase. Su diseño como único fichero multilingüe, con un formato que incluye tanto la información común como la específica de cada lengua, y la potencia de su algoritmo de procesamiento le convierte en idóneo para el tratamiento y modificación de grandes masas de terminología dispersa. Otra de sus posibilidades es la aplicación de un módulo externo que le permite compartir los ficheros generados en ambos sistemas de gestión de tesauros: CAT y MTM, mediante su conversión previa al formato ISO 2709-1981[45].

45. ISO 2709-1981. Format for bibliographic information interchange on magnetic tape.

- En cuanto al MTM, programado en Pascal, necesita utilizar el gestor de base de datos CDS/ISIS. Tiene formato de campos variable y pueden coexistir en cada registro hasta nueve términos de distintos idiomas, pero, su estructura distribuye los identificadores de campo reservando un tramo para cada idioma. Recupera las cadenas identificadoras de términos de cada idioma por separado o todas a la vez, y cuando se crea un descriptor en una lengua obligatoria, automáticamente lo hace en las demás lenguas del tesauro. Similarmente, cuando se borra un descriptor, se pueden borrar todos los términos relacionados con él y en todas las lenguas.

El proyecto Gestor de Tesauros de la Universidad Carlos III de Madrid (GDA)

El Gestor de Tesauros Autogenerables (GDA) desarrolla una estructura orientada a la recuperación documental automatizada mediante la autogeneración de tesauros[46]. En este trabajo se le ha establecido un «hipotético paralelismo» con el sistema Lexinet, desarrollado en el Departamento de Investigación Lingüística del INIST de Francia para la recuperación en su base de datos PASCAL[47].

- El GDA estará formado por una red global de nodos monotemáticos interconectados que recibe el nombre de Tesauro Global (TG). Para ello, será necesario procesar tanto la información genérica del lenguaje como la específica de la propia área de conocimientos.

  El Tesauro Autogenerable (AT) toma como comparación un tesauro de descriptores preestablecido, que se relaciona con un conjunto de estructuras informáticas de apoyo, a través de una clasificación tabular por temas y, dentro de cada tema, por facetas.

  La estructura temática es jerárquicamente simple (árboles) y precisa de la creación de una clasificación previa, el Arbol de Areas Temáticas (AAT), sobre la que configurar la estructura temática que utilizará el tesauro para almacenar

46. Memoria de Investigación curso académico 1993-94. Leganés: Universidad Carlos III, Grupo de Tecnologías de la Información, Departamento de Ingeniería, 1995.

47. CHARTRON, G. «Lexicon management tools for large textual databases: the Lexinet system». Journal of Information Science, 1989, 15, pp. 339-344.

los conceptos. Dicho AAT se almacena en el TA mediante una estructura tabular simple y se crea una tabla auxiliar a la AAT para el caso de términos pertenecientes a diversas categorías o inclasificados.

La red de nodos TG está compuesta por cuatro zonas: la de diccionarios, la de inclasificados, la de índices y la de codificación de tablas de información y almacenamiento de documentos internos al sistema.

- El sistema LEXINET está formado por un conjunto de algoritmos que detectan y extraen los conceptos significativos de un conjunto de documentos pertenecientes a un campo de conocimientos concreto. Se trata de asegurar el procesamiento del mayor número posible de documentos, actuar en diferentes dominios científicos y trabajar en varias lenguas.

Los módulos desarrollados en Lexinet son: el analizador de textos, el lematizador de términos, los eliminadores de palabras vacías y de escasa distribución, el de sinónimos, el gestor de multitérminos, el gestor de unitérminos y el indizador de documentos. Tres unidades terminológicas se actualizan automáticamente: el fichero de descriptores (lexicon), el fichero de sinónimos (synonyms) y el fichero de no descriptores (antilexicon).

Ambos gestores están en proceso de desarrollo, aunque Lexinet ha ido evolucionando continuamente desde 1988 para su aplicación en Pascal y el GDA es un proyecto de investigación, sin concluir su prototipo. Ambos analizan los textos y extraen los términos para que un indizador manual los verifique antes de incluirlos en el Tesauro operativo. Ambos utilizan métodos estadísticos para la selección automática de los términos, dejando a los indizadores manuales la decisión final de corregir los resultados, antes de validar los nuevos descriptores.

Como características diferenciadas, destacan:

- Lexinet comenzó trabajando únicamente con unitérminos, el GDA acomete desde el principio los multitérminos,

- Lexinet trabaja con textos técnicos y científicos condensados (resúmenes de patentes y artículos científicos), y la idea del GDA es trabajar directamente con textos completos,

- Lexinet posibilita que los términos seleccionados puedan ser reutilizados para la indización de posteriores documentos en Pascal, mientras que el GDA perfecciona una fase de investigación en la que verifica los términos extraídos de los textos con los términos de los tesauros previamente cargados, y

- Lexinet utiliza técnicas de fácil aplicación a los Sistemas Expertos, con las que se puede seleccionar automáticamente nuevos términos sin recurrir a su comparación con los términos del tesauro y obtiene una gestión más rápida que la del GDA que utiliza tablas indizadas, en su nivel actual de desarrollo.

Consideraciones Finales

Este artículo no ha sido exhaustivo, dado que sólo se han revisado algunos de los logiciales más conocidos y/o avanzados para la gestión automatizada de tesauros. A veces, se ha descrito someramente la técnica de indización usada por el Sistema de Gestión en el que se encuentran integrados. Otras veces, cuando los sistemas procesan el Lenguaje Natural, se ha procedido a indicar como generan los conceptos para incorporarlos en sus tesauros postcontrolados.

A pesar de que los tesauros conceptuales existentes son poco utilizados por las tecnologías de la lingüística computacional, la inteligencia artificial o la ingeniería del conocimiento, en la búsqueda de nuevas soluciones para el procesamiento del lenguaje natural, se ha detectado un resurgimiento en las ideas desarrolladas por éstos que, como herramienta conceptual bien conocida y establecida que son, deben evolucionar para incorporar nuevos tipos de relaciones asociativas que faciliten la adaptación a sus nuevos usuarios, los Sistemas Expertos y sus nuevas técnicas de Inteligencia Artificial, desde los criterios hasta ahora dominantes en los expertos humanos.