30(4):51-60, diciembre, 1999

Generación automática de tesauros. Propuesta de un método lingüístico-estadístico[1]

José A. Moreiro González
Irene Díaz
Juan Lloréns
Jorge Morato
Manuel Velasco

En este trabajo se presenta la investigación realizada durante los dos últimos años dentro del proyecto *Generación Automática de Tesauros Orientada a la Arquitectura de Componentes.* En este trabajo se han desarrollado varios métodos para construir semiautomáticamente un tesauro de descriptores. Para ello se han intentado automatizar, en muchas ocasiones con éxito, todas las fases para esta construcción automática, poniendo especial énfasis en las fases de adquisición y organización del conocimiento.

Introducción

Se presentan en este trabajo los resultados obtenidos a lo largo de las investigaciones realizadas dentro del proyecto Generación Automática de Tesauros Orientada a las Arquitecturas de Componentes (GATOAC), que ha sido financiado por la Comisión Interministerial para la Ciencia y Tecnología (CICYT) española. El desarrollo de esta investigación se inició en 1997 y se ha extendido hasta el momento presente.

El principal objetivo consiste en diseñar y construir una plataforma de gestión de repositorios (tesauro de software) capaz de almacenar, procesar, gestionar y recuperar cualquier tipo de documento, sin importar su presentación, soporte y forma de acceso, creando un sistema autogenerable (que el tesauro crezca por sí mismo desde su nacimiento o desde algún punto de desarrollo).

Procesar y gestionar implica todo el trabajo interno del sistema, el cual permite que la información sea analizada vocablo por vocablo, determinando su naturaleza, el número de incidencias, las relaciones entre sí, etc. Se pretende ayudar de manera eficiente y relevante en la recuperación de la información de acuerdo a las necesidades de los usuarios. Por otra parte, si se logra que el sistema sea autogenerable se podrá analizar toda la cantidad de información que se genera día a día en las "autopistas de la información", facilitando la labor de documentalistas e informáticos. El proyecto busca desarrollar una estructura específica orientada a la recuperación documental automatizada. Dicha estructura de información se define como tesauro autogenerable y está basada en la teoría documental de los tesauros de descriptores en documentación (según la norma ISO 2788).

Como características generales del trabajo podemos definir las siguientes:

- Está basado en un tesauro de descriptores global, es decir, utiliza la

superestructura de los tesauros.

- Es autogenerable lo que permite que el propio sistema se actualice conforme vaya almacenando, mediante la construcción automática de relaciones.
- Es de aplicación multilingüe.
- El proceso de filtrado se realiza mediante analizadores sintácticos, semánticos y morfológicos.
- Está basado en la reutilización de software.
- Se aplica a la indización de todo tipo de documentos (textos, imágenes, etc.).
- Permitirá la gestión y recuperación de la información.

Además de los parámetros anteriores, el proyecto persigue el diseño y la ejecución de herramientas que permitan una gestión documental mucho más ágil:

- – Trabaja en un entorno multiusuario: el proyecto trabaja en un ambiente que permite su acceso a varios usuarios de manera simultánea (sistema de multiproceso).
- – Permite la gestión del repositorio que almacena la información desde fuera por especialistas.
- – Presenta la información de la manera más clara y sencilla.
- – Permite desarrollo y perfeccionamiento por unidades y también facilita su ampliación, cobertura y especialización, debido a su estructura modular.
- – Permite gran amplitud en las relaciones semánticas de los términos: es ilimitada la cantidad de específicos que se puede asignar a cada término, igual sucede con los términos relacionados, tanto los circunstanciales como los permanentes.
- – Contiene un sistema de gestión bibliográfica (registro bibliográfico con información secundaria de los documentos).
- – Se conseguirá un sistema de indización y recuperación automática (búsquedas inteligentes).
- – Supone la integración de diversos tipos de aplicaciones, con la ayuda de protocolos de interconexión de aplicaciones que permiten el traspaso de información entre aplicaciones (OLE-DDE).
- – Realiza clasificación temática de los descriptores.
- – Posee un sistema de importación y exportación de la ficha bibliográfica del documento propiamente dicho.

Se han considerado como fundamentos teóricos de esta investigación el concepto de información y el de los diversos soportes que puede contener dicha información, así como la adopción de la teoría general de tesauros de descriptores en cuanto fundamento para la gestión y recuperación de dicha información.

Tesauro de descriptores

Para afrontar esta labor seguimos los supuestos teóricos planteados por Van Slype por los que se concibe al tesauro de descriptores como una lista controlada de términos de un área del conocimiento estructurada semánticamente. No brinda significado gramatical de la palabra, pero sí establece relaciones semánticas entre los términos. El tesauro de descriptores es pues una herramienta para la “representación de conceptos” expuestos en los documentos y cuya finalidad es la recuperación de la información

contenida en ellos [1].

Un tesauro está constituido básicamente por dos elementos: unidades léxicas y relaciones semánticas. Las unidades léxicas se dividen en cuatro categorías:

1. 1. Campo semántico o grupo de familia de términos.
2. 2. Descriptor o término preferente, que designa un concepto y que sirve para representar el contenido de un documento y realizar consultas.
3. 3. No-descriptor o término no preferente, originados en los sinónimos o cuasi-sinónimos. No sirven para indizar, pero reenvían la indización o la consulta hacia los descriptores. Su misión es servir de inferentes hacia los descriptores.
4. 4. Descriptores auxiliares que por sí solos no aportan ningún concepto, pero sumados a los descriptores forman conceptos o descriptores compuestos.

Los tipos de relaciones semánticas entre los términos pueden ser:

- – Equivalencias interlingüísticas, conceptos iguales o equivalentes en diferentes lenguas o equivalencias semánticas intralinguísticas, dentro de una misma lengua (de un descriptor a un no-descriptor). Dentro de esta relación encontramos sinonimia verdadera, sinonimia por variante ortográfica, siglas, variantes de escritura, extranjerismos, antonimias, lenguaje usual *versus* lenguaje científico.
- – Relaciones de jerarquía basadas en niveles de super o subordinación, en que un término superordenado representa un todo o clase y los términos subordinados corresponden a los miembros o partes del término superior.
- – Relaciones de asociación: se muestra así la necesidad de establecer asociaciones que sugieran desde un concepto otros con los que esté relacionado, sin que entre ellos exista dependencia jerárquica. Como tipos de relaciones asociativas podemos nombrar las que se generan en la causalidad, la instrumentación, la sucesión en el espacio y en el tiempo, la concomitancia, la similaridad, la antonimia, las propiedades de los objetos y hechos, la localización, y los objetos de acciones, procesos o disciplinas.

Construcción semiautomática de un tesauro de descriptores

Para realizar este trabajo, se han llevado a cabo procedimientos orientados a definir las herramientas de semiautomatización utilizadas en un sistema de generación de tesauros. El tesauro de software (TS) se muestra como una variación del tesauro de descriptores orientada a la reutilización, lo que le permite actuar como base sobre la que poder representar un dominio cualquiera. El TS proporciona una riqueza de componentes tal que mejora enormemente el proceso de recuperación de información [2, 3, 4, 5, 6].

Buscando generar automáticamente representaciones del dominio [7, 8, 9, 10] y, por ello, buscando construir también de forma automática el tesauro, se han integrado técnicas de origen tanto informático, como estadístico y de inteligencia artificial, así como otras provenientes de las Ciencias de la Documentación. La adecuada combinación de estas técnicas fundamenta la viabilidad del trabajo [11,12,13].

Las fases de construcción de un tesauro y, por tanto, de la estructura de repositorio que alberga el dominio, se muestran en la figura 1.

(Insertar figura 1, página 54 posible nombre de archivo F990405-01)

Se han logrado avances en varios de los procesos básicos de construcción de dominios y, por tanto, de construcción de tesauros. Principalmente en los relativos a identificación y adquisición de componentes representativos de un dominio y al de filtrado y búsqueda de relaciones entre ellos.

El estado actual permite automatizar partes importantes del proceso, aunque globalmente, aún sigue siendo necesaria la intervención humana en algunas de ellas.

Identificación y adquisición de componentes representativos de un dominio

La identificación y adquisición de componentes representativos de un dominio es el proceso a través del cual se consiguen aquellos componentes considerados como una representación fiel del dominio que se está estudiando. Dentro de este proceso de identificación y adquisición de componentes, como es bien conocido, se encuentran los siguientes subprocesos: análisis léxico, tratamiento de palabras vacías, tratamiento de términos flexionados, tratamiento de palabras compuestas y filtrado de términos.

Se ha conseguido automatizar los procesos de eliminación de palabras vacías, de tratamiento de términos flexionados, así como el tratamiento de términos compuestos, lo que ha supuesto un gran avance para posteriormente seleccionar la raíz de una jerarquía e incluso hacer la primera de las jerarquías del tesauro basándose en términos compuestos [14, 15, 16].

Análisis léxico

El análisis léxico [17] tiene como objetivo transformar una cadena de caracteres en un conjunto de palabras o *tokens*. Estos son grupos de caracteres que presentan un significado colectivo. El análisis léxico siempre es la primera parte dentro del proceso automático de adquisición de componentes. Esta etapa se encarga de proporcionar los términos (posibles descriptores) para que sean posteriormente examinados por otros procesos (filtrados, palabras compuestas, etc.).

Lo primero que se ha previsto al desarrollar un analizador léxico, y teniendo en cuenta que la información de la que se dispone no presenta errores ortográficos, es decidir cuáles son los caracteres o símbolos que no son interesantes y que en muchos casos sirven para delimitar un *token* o palabra.

Eliminación de las palabras vacías

Los procesos de filtrado que se han utilizado, con sus correspondientes cálculos estadísticos, tienen la posibilidad de eliminar previamente los términos vacíos mediante su confrontación con una lista de palabras vacías, construida previamente. También pueden suprimirse a posteriori, eliminándolas si consiguen eludir el proceso de filtrado. Las palabras vacías sólo son descartadas cuando se trate de obtener descriptores simples, ya que pueden formar parte de descriptores compuestos. Existe para cada idioma un conjunto de palabras vacías, comunes a todos los dominios, fácilmente identificable: artículos, preposiciones, conjunciones, etc. En el sistema desarrollado se

toman como antidescriptores. Aunque algunos son considerados partículas de unión: los artículos, conjunciones y adverbios para todo tipo de dominio; y adjetivos y pronombres en determinadas situaciones.

Tratamiento de términos flexionados

Flexionados son aquellos términos relacionados morfológicamente entre sí como, por ejemplo, “león”, “leona”, “leones”, “leonas”,…, y que, en algunos casos, puede considerarse que tienen un significado común. Los flexionados de un término canónico presentan entre ellos variaciones de género, número o tiempo verbal.

El tratamiento de flexionados, que consiste en reducirlos a su término canónico, se utiliza para mejorar la efectividad en la recuperación de información y para reducir el tamaño de los resultados de adquisición de componentes. Esta aplicación resulta también aprovechable para agrupar términos con vistas a los tratamientos estadísticos asociados a la creación automática de la representación del dominio: filtrados, creación de relaciones entre descriptores, etc.

Tratamiento de palabras compuestas

Las técnicas clásicas de adquisición manual de componentes resuelven fácilmente el problema de indización de palabras compuestas porque el experto selecciona directamente aquellos términos compuestos que considera representativos. En el caso automático es necesario diseñar un algoritmo para poder incluir palabras compuestas como componentes del dominio. Para realizar este tratamiento se ha utilizado un autómata de estados finitos [18], que trabaja conjuntamente con el proceso de referenciación de descriptores.

El autómata consta de cuatro estados y en cada uno de ellos se siguen unas reglas específicas para la identificación y adquisición de los términos compuestos. Habrá de tenerse en cuenta que el proceso de identificación y adquisición guarda información referente a las palabras que trató con anterioridad, pero sólo procesa una palabra cada vez. El funcionamiento se basa en la utilización de una pila (estructura en la que se almacenan elementos, de tal modo que en la primera posición está el último que se ha guardado) para el almacenamiento de las palabras pendientes, dos capas para el intercambio de información entre estados y el consiguiente núcleo de identificación y adquisición de componentes.

Con el reconocimiento de los descriptores compuestos se produce un primer acercamiento a la construcción de las relaciones entre descriptores, en este caso con las relaciones jerárquicas.

Filtrado de términos

Además de los tratamientos anteriores es muy interesante realizar filtrados sobre los posibles términos representativos de un dominio, ya que a la hora de buscar relaciones entre los términos es necesario que el número de estos sea reducido, debido a que los métodos estadísticos y de redes neuronales que proporcionan estas relaciones trabajan con un conjunto limitado de elementos.

Las distintas técnicas que se han analizado son capaces de discriminar entre los términos que consideran representativos de un texto y los que consideran sin importancia. En su aplicación se han desarrollado dos algoritmos diferentes.

- – *IDF:* Son las siglas correspondientes a Indización estadística de términos por frecuencias [19, 20]. Este sistema de filtrado está basado en la ley de Zipf [21] que establece que las palabras con mayor frecuencia absoluta son las palabras vacías, mientras que las más infrecuentes reflejan el estilo y riqueza del vocabulario del autor. Aquellas que aparecen en la zona media de la función de distribución de frecuencias son las que mejor representan al documento. La técnica IDF establece un sistema de pesos en función de la frecuencia relativa de cada término en cada documento. En el caso de que un término tenga una frecuencia en un documento mayor que la media fijada en el resto de documentos, se tomará como descriptor. En el momento que se tome como descriptor para un documento será considerado como tal en el resto de documentos, es decir, no es necesario que un término aparezca en todos los documentos a filtrar para que sea descriptor.
  Se aplica primero la ley de Zipf para el cálculo de la zona de transición y después el método IDF para ponderar por documentos. Comentamos ahora la problemática específica de cada método, así como las mejoras introducidas. Se ha modificado la ley de Zipf para aprovechar la información que nos proporciona el tratamiento de palabras compuestas, puesto que no estaba pensada para filtrar por términos compuestos [18, 19, 21].
- – *Método N-grams:* Este algoritmo trabaja con cadenas de caracteres de longitud fija para solucionar el tratamiento de palabras compuestas. Hace un tipo de filtrado parecido a los anteriores de tal forma que la frecuencia se calcula no sobre cada término o palabra compuesta sino sobre cadenas de caracteres de longitud predeterminada y fija. El número *n*, la longitud de la cadena, toma valores entre 3 y 6. En este trabajo se ha tomado el valor 5, para poder tener un carácter central en el *n-gram*
  La construcción del *background* necesario para realizar la comparación de frecuencias con los documentos del *corpus* del dominio no es un paso en absoluto trivial. El filtrado variará en función de la información que componga el *background*.

Para comprobar que el *background* responde a características generales del lenguaje se han utilizado estudios estadísticos propios sobre cómo aparecen las cadenas en cada idioma.

Organización de los conceptos adquiridos

Para poder reutilizar información de un modo óptimo e inteligente es necesario primero clasificarla, de tal modo que se establezcan relaciones entre los componentes que la definen y describen. Las relaciones son muy importantes para poder seleccionar posteriormente, de forma inteligente, la información que contiene un repositorio. Existen numerosos y variados enfoques para realizar este proceso. Se presentan en este trabajo alternativas relativas a campos de investigación muy distintos entre sí en algunos casos. Principalmente se ha trabajado con tres tipos de clasificadores:

- Cienciométricos: Co-wording.

- *Estadísticos:* Max-min, K-vecinos, K-vecinos incremental, Isodata.
- *Neuronales:* Kohonen, Art-1, Art-2.

Clasificadores cienciométricos: Método de Chen

El análisis de coocurrencia de palabras estudia el uso de grupos de palabras que aparecen simultáneamente en varios documentos. Las palabras pueden pertenecer a un lenguaje controlado o a texto libre.

El método de concurrencias capaz de evaluar la relación entre dos descriptores se considera, por tanto, un método de clasificación. Su propósito es establecer un peso a la relación que existe entre dos descriptores. Para aplicar tal método se deben haber identificado los descriptores, y posteriormente se debe proceder a realizar el análisis de coocurrencias para todos los documentos del *corpus* documental. Se calcula un peso para cada término basado en el modelo de espacio vectorial [20] y en una función de semejanza asimétrica [22].

Algoritmos estadísticos de agrupación en clases

La agrupación en clases puede definirse como el proceso de clasificación no supervisada de objetos.

Se dispone de un conjunto de vectores {x1, ..., xp}, que representan a los objetos y a partir de él se desea obtener el conjunto de clases {(1, ..., (n} que los engloban. El problema es que *a priori* no se sabe cómo se distribuyen los vectores en las clases, ni siquiera cuántas clases habrá.

El problema consiste en, a partir del conjunto de vectores de características dado, conseguir realizar agrupaciones de estos vectores en clases de acuerdo con las similaridades encontradas.

Se presentan a continuación a modo de ejemplo, dos de los clasificadores estadísticos que han sido seleccionados para su aplicación en este trabajo:

- – *Algoritmo K-vecinos*. Es un algoritmo rápido y eficaz, si la distancia que utiliza es adecuada para el problema considerado.
  Busca minimizar un índice de rendimiento, basado en la suma de distancias euclídeas cuadráticas de todos los miembros de un cluster a su centroide.
  Exige conocer el número de clústeres *k* en los que se desea clasificar la muestra de vectores de la población. Si el número de clases no se conoce por adelantado, se puede dejar que el algoritmo determine el número de clústeres utilizando parámetros definidos por el usuario.
  El modo de funcionamiento del algoritmo consiste en mover cada vector al clúster cuyo centroide esté más cercano al mismo, y actualizar después los centroides de los clústeres. Su convergencia depende mucho del número de clases.

- – *Algoritmo K-vecinos axial o incremental*. Este algoritmo, como su nombre indica, calcula los clústeres de forma incremental. Pertenece a la familia de algoritmos de clasificación por centros móviles. Es una variante

del algoritmo k-vecinos en su versión adaptativa, y del algoritmo de Forgy, en el caso iterativo.
Dado un patrón de entrada, el algoritmo debe actualizar la representación de los clústeres y devolver el índice del clúster actual al cual pertenece el patrón, sin necesitar tener presentes los demás patrones. De este modo puede tratarse una sucesión arbitrariamente grande de patrones en tiempo real. Los algoritmos de clúster incremental son muy atractivos para el tratamiento de patrones documentales, dado el gran espacio de almacenamiento que requieren dichos patrones.
El algoritmo de clúster euclídeo no converge necesariamente en un conjunto fijo de prototipos: los prototipos pueden variar infinitamente, sin converger en el tiempo. El número de clústeres creados tampoco es necesariamente finito, y depende de las funciones utilizadas en el algoritmo.

– – *Redes neuronales.*Las redes neuronales se utilizan como herramientas o métodos para resolver problemas, fundamentalmente relacionados con el conocimiento humano. Especialmente para el reconocimiento de patrones, reconocimiento del lenguaje hablado, reconocimiento de imágenes, procesos de control adaptativo y en el estudio del comportamiento de ciertos problemas para los que no están muy bien dotados los computadores tradicionales.

En este trabajo se han utilizado varios tipos de redes neuronales: Kohonen, ART1 y ART2.

Obtención de relaciones

Los clasificadores llevan a cabo conjuntamente la tarea más compleja de todo el trabajo presentado en esta investigación: la obtención de relaciones jerárquicas. Aquí se presentan métodos, que integrados, producen resultados que permiten asegurar, con nuevos desarrollos añadidos, la automatización definitiva del proceso.

Se proporcionan también las asociaciones temáticas, pero no la forma de nombrar los grupos temáticos obtenidos. Existen conocidos trabajos sobre obtención de este tipo de relaciones [19].

El método presentado para la obtención de jerarquías y asociaciones temáticas parte de la integración de las distintas técnicas, que trabajan en paralelo, como filosofía de trabajo.

Todos estos clasificadores realizan un proceso de clusterización, que agrupa en clases aquellos descriptores que responden a una serie de características comunes. *A priori* no puede establecerse la ventaja de un método respecto a otro en cuanto a calidad de resultados. La integración de las jerarquías obtenidas nos dará los criterios sobre la bondad de cada clasificador.

Al utilizarse los procesos de clusterización para la construcción automática de tesauros tendrán que tenerse en cuenta factores específicos de esta problemática. El tamaño de cada clúster no debe ser muy dispar, ya que las áreas temáticas suelen tener un número parecido de descriptores, el número de clústeres en los que se divide uno dado tampoco debe ser muy alto, ya que cada nuevo clúster representa un conjunto de términos que

serán globalmente específicos, aunque sólo alguno(s) en un primer nivel de jerarquía. A mayor número de clústeres generado a partir de uno dado, mayor número de específicos de primer nivel. Un número alto de específicos de primer nivel no suele ser común en un tesauro.

La construcción de la representación del dominio se hace mediante aproximaciones *top-down* en la jerarquía. A partir del total de descriptores filtrados se irá formando la jerarquía desde el más general hasta el más específico.

El primer paso consiste en encontrar la raíz o raíces de la jerarquía. Se utilizan técnicas de extracción de componentes principales. Se intenta encontrar el concepto más significativo utilizando diferentes grados de pertenencia al clúster. Se han tenido en cuenta cuatro formas de obtención de raíces que son:

Mediante el cálculo del centroide que representa el centro de masas del clúster o conjunto de descriptores en consideración.

- Seleccionando el descriptor más general del clúster, tomando aquel que tenga mayor número de apariciones en el total de documentos del *corpus*.
- Seleccionando el descriptor más general del cluster, escogiendo aquel que aparezca en un número mayor de documentos.
- Seleccionando el descriptor más general, combinando las dos ideas anteriores.

Una vez seleccionada la raíz o raíces se realiza clusterización o agrupación en clases del resto de los descriptores mediante cada técnica de clasificación en su caso.

Terminado el proceso de clusterización, los distintos clústeres creados pueden considerarse simbólica y globalmente específicos de la raíz o raíces obtenidas en el paso anterior. Cada uno de ellos constará de un número de descriptores no determinado *a priori*.

Este proceso de clusterización proporciona implícitamente el primer nivel de la clasificación temática. Cada clúster representa una aproximación a la formación de nodos del árbol de áreas temáticas, identificándose directamente en muchos casos con un nodo específico.

Repitiendo la extracción de componentes principales en cada uno de los clústeres se obtiene el próximo nivel en la jerarquía del tesauro. Las nuevas raíces son consideradas términos específicos de las raíces de primer nivel.

Los dos pasos anteriores (clusterización + extracción de raíces) se van repitiendo hasta que se cumplan unas determinadas condiciones que paran el proceso.

Para realizar asociaciones temáticas, se toman como primeras áreas aquellas generadas en el primer paso de clusterización. La generación de áreas temáticas [5] comprende valores óptimos de términos por área, en torno a 50 componentes. De esta forma puede decidirse si crear o no nuevas áreas, en función del número de elementos de cada clúster.

Debe tenerse también en cuenta que no debe sobrepasarse un número máximo de

niveles en la jerarquía del tesauro. Este número máximo puede tomarse con un valor alrededor de 4.

Puede efectuarse también una construcción temática a partir de una aproximación *bottom-up* al agrupar las áreas definitivas (descriptores simples o grupos pequeños de descriptores) teniendo también en cuenta el número máximo de niveles en la jerarquía para un tema dado.

Se obtienen solapamientos típicos de las clasificaciones temáticas durante el procedimiento de integración.

Integración de relaciones

A partir del proceso de generación de relaciones semánticas, debe disponerse un proceso de contraste, ya que es muy posible que la ejecución en paralelo de los distintos clasificadores proporcione relaciones distintas para dos descriptores dados. La integración de relaciones en este trabajo se ha realizado de forma manual, siguiendo ciertas pautas, de las cuales las más importantes son:

Disposición de un sistema de pesos que potencie los resultados obtenidos por los mejores clasificadores. A partir de conocimiento previo, puede obtenerse una escala de eficiencia de clasificadores para ser utilizada en posteriores clasificaciones, teniendo en cuenta con mayor consideración aquellos clasificadores que se encuentren en posiciones más altas en la escala.

Establecimiento de una primera escala de relaciones (para los cinco tipos de relaciones existentes) para definir la calidad y riqueza de estas.

Posibilidad de instauracción de una segunda escala, respecto a los distintos tipos de relaciones, en función de la dificultad de encontrar específicamente cada relación.

Selección de un representante del grupo de sinónimos, si dos o más descriptores se consideran sinónimos.

Conclusiones

Las conclusiones a las que este grupo de investigación ha llegado en el transcurso de esta investigación han sido las siguientes:

- – La intervención humana es imprescindible para arrancar el proceso global de construcción de un tesauro, dado que hoy por hoy, no existe, o al menos este grupo de investigación no conoce, ningún mecanismo para delimitar áreas de conocimiento ni para identificarlas.
- – Una vez que el dominio ha sido identificado y la información relativa a él ya está disponible, es posible automatizar casi por completo el proceso de construcción de un tesauro, lo que ahorrará tanto tiempo como esfuerzo y costes a informáticos y documentalistas.
- – El proceso de adquisición de conceptos ha sido automatizado con resultados satisfactorios. Para ello se han utilizado dos técnicas de filtrado de información diferentes, que ha tenido que ser modificadas adecuadamente para mejorar los

resultados. Los resultados de ambas técnicas posteriormente han debido ser integrados, obteniéndose así conceptos que representan de un modo más fiel el dominio de trabajo.

- – El proceso de organización de los componentes también se ha automatizado, utilizando, en algunas ocasiones, técnicas cuyo cometido no estaba orientado a esta tarea. Otra vez, los resultados de cada una de estas técnicas han debido ser integrados, con lo que se ha comprobado que se obtiene una clasificación más rica.
- – El tiempo utilizado para conseguir un tesauro se ha reducido considerablemente, con lo que se consigue ahorrar esfuerzo, tiempo y costes.
- – Este proceso aún se podría mejorar, no tanto en cuanto a la automatización del proceso como a los resultados finales de esta herramienta, mejorando los parámetros de filtrado, tratando de reflejar más semántica en el proceso de organización de componentes, mediante el uso de otras técnicas (lógica difusa, por ejemplo).
- – Se deben desarrollar técnicas de validación del tesauro construido, y verificar la utilidad de las ya desarrolladas.

Referencias

1) Van Slype, G. Les Langages d'Indexation. Conception, Construction et Utilisation dans les Systèmes Documentaires. Paris, Les Editions d'organisation. 1991.
2) Díaz, I., J. Lloréns, V. Martínez, y M. Velasco. Semi-Automatic Construction Of Thesaurus Applying Domain Analysis Techniques. International Forum on Information and Documentation. 23(8): 11-19. 1998.
3) Lloréns, J. Definición de una metodología y una estructura de repositorio orientadas a la reutilización. Tesis doctoral. Universidad Carlos III de Madrid. 1995.
4) Lloréns, J., A. Amescua y M. Velasco. Software Thesaurus. a Tool for Reusing Software Objects. Actas del 4º IEEE Assessment on Software Tools. Toronto, Canada. 1996.
5) Lloréns, J., A. Amescua y M. Velasco. A Software Thesaurus as an Intelligent Tutorial System for Software Specifications. Tercera Conferencia Internacional en Intelligent Tutorial Systems. ITS-96. Montreal. Canada. 1996.
6) Lloréns, J., A. Amescua, M. Velasco, J. A. Moreiro y V. Martínez. Automatic Domain Analysis using Thesaurus Structures. (Aceptado en publication in Journal of the American Society of Information Science, 1998.)
7) Prieto-Díaz, R., P. Freeman. Classifying Software for Reusability. IEEE Software. 1987.
8) Prieto-Díaz, R. Domain Analysis for Reusability. COMPSAC'87, pp. 23-29, Tokyo, 1987.
9) Lloréns, J., M. Velasco y R. Pérez. Modelización Activa. Técnicas y Métodos. INFONOR 97. Chile. 1997.
10) Lloréns, J.; A. Amescua, V. Martínez y M. Velasco. The Reuse Maturity Model. 3RMM. Symposium on Computer and Information Science. ISCISXII, Antalya, Turquía. 1997.
11) Díaz, I; M. Velasco y J. M. Molina. Aplicación de la lógica borrosa al cálculo de distancias entre componentes representativos de un dominio, en un proceso global de análisis de dominios. *Actas del VIII Congreso Español sobre Tecnologías y Lógica Fuzzy*. Pamplona, 1998.
12) Hopfield, J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Actas del National Academy of Science (81):3088-92. National Academy of Science, 1982.

13) Zimmermann, H. J. Fuzzy Set Theory and its Applications. Kluwer Academic Publishers. 1990.
14) Velasco, M., J. A. Moreiro y J. Lloréns. Estado actual del proyecto GDA (gestión documental automatizada). Planteamiento teórico y descripción práctica. ISKO 97. Madrid, Noviembre, 1997.
15) Velasco, M., V. Martínez, J. Lloréns y A. Amescua. Automatic Domain Analysis. Generation of Domain Representations. Information Technologies and Knowledge Systems. IT-KNOWS. International Federation for Information Processing (15th IFIP). Vienna, september, 1998.
16) Frakes, W. B. y R. Baeza-Yates. Information Retrieval. Data Structures & Algorithms. Prentice Hall. 1992.
17) Fernández Herrero, J. A., J. Lloréns, J. y M. Velasco. Indización básica contra el tesauro autogenerable. Universidad Carlos III de Madrid. 1995.
18) Muñoz, A. Redes neuronales para la organización automática de información en bases documentales. Tesis Doctoral. Universidad de Salamanca. 1994.
19) Salton, G. Automatic Text Processing. the Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, cop. 1989.
20) Zipf, G. K. Human Behaviour and the Principle of Least Effort. An Introduction to Human Ecology. Haffner. New York. 1972.
21) Chen, H. y K. J. Lynch. Automatic Construction of Networks of Concepts Characterizing Document Databases. IEEE Transactions on Systems, Man and Cybernetics, (22):885-902, 1992.

Recibido: 1 de octubre de 1999.
Aprobado: 15 de octubre de 1999.

José A. Moreiro González
Departamentos de Biblioteconomía y Documentación.
Universidad Carlos III de Madrid. España.
Correo electrónico: <<jamore@bib.uc3m.es>>.

Nota

[1]Los autores de este trabajo desean expresar su agradecimiento a la Comisión Interministerial para la Ciencia y Tecnología (CICYT) de España su financiación durante estos dos años de trabajo.