Hacia la Generación Automática de Tesauros

Manuel Velasco[1], Irene Díaz[1], José Antonio Moreiro[2],Juan Lloréns[1]
Universidad Carlos III de Madrid

Resumen

En este trabajo se presenta una semi-automatización del proceso de generación de tesauros donde el "Tesauro de Software" (TS) se utiliza para representar un dominio cualquiera. Para realizar esta construcción automática de un tesauro, se han utilizado técnicas tanto informáticas, estadísticas, de inteligencia artificial, como extraídas del campo de las ciencias de la documentación. Se han conseguido avances importantes en procesos básicos de construcción de dominios, tales como la identificación y adquisición de componentes representativos, el filtrado y la clasificación de estos componentes.

Palabras clave: Tesauro, análisis de dominios, tesauro de software, descriptor.

Abstract

A semi-automatic approach to generate thesaurus is presented in this paper. Here, software thesaurus is used to represent any domain. Computer science, statistical, artificial intelligence and information science techniques have been used to try to represent a thesaurus automatically. Some important advances have been got in some processes as identification and acquisition of components, filtering and classification of those components.

Keywords: Thesaurus, domain analysis, software thesaurus, descriptor.

1. Introducción

En este trabajo se presenta una descripción de procedimientos en la que se atiende a definir las herramientas de semi-automatización 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. [1], [2], [3], [4], [5].

Para generar automáticamente representaciones del dominio [6], [7], [8], [9] y, por tanto, para 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 [10], [11],[12],[13].

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.

Estas acciones forman parte de la investigación propia del proyecto GATOAC (Generación automática de tesauros orientada a las arquitecturas de componentes), financiado por la CICYT y orientado a la parte textual de la construcción de dominios.

2. El Tesauro de Software

El Tesauro de Software fue diseñado para permitir clasificar y recuperar componentes de software a partir de un conjunto de requerimientos más amplio que los utilizados en los tesauros de aplicación documental. Estos requerimientos incluyen el tratamiento de información textual que posee cada uno de los componentes de software desde la perspectiva de la Documentación.

Los sistemas clásicos de tratamiento de información textual se fundamentan en la idea de que una vez completado un tesauro de descriptores que representen un campo del conocimiento, toda información contenida en algún documento que no se encuentre en el tesauro debe ser desestimada. Sin embargo, al utilizar estas técnicas para la clasificación de componentes de software, resulta fundamental considerar la posibilidad de que cada uno de ellos pueda incorporar nueva información al tesauro, la cual pueda ser recuperada y reutilizada en posteriores consultas (proyectos). Esta nueva posibilidad obliga a la modificación de la estructura clásica del tesauro de descriptores, puesto que se necesita almacenar también la información sobre palabras sin significado (vacías) para que el sistema las elimine como posible información a clasificar.

3. Identificación y Adquisición de Componentes Representativos de un Dominio

Es el proceso por el 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].

3.1 Análisis léxico

El análisis léxico [17] tiene como objetivo transformar una cadena de caracteres en un conjunto de palabras o tokens. Éstos 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.

3. 2 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.

3. 3 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..

3.4 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 aquéllos términos compuestos que considera repreentativos. 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.

3.5 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 éstos 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:

3. 5. 1 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. Aquéllas 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 la ley de Zipf no estaba pensada para filtrar por términos compuestos [18], [17], [21].

3. 5. 2 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.

4. Obtención de Relaciones entre Componentes

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.

4.1. 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 coocurrencias 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].

4. 2 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.

4. 2. 1 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 clusters 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 clusters utilizando parámetros definidos por el usuario.

El modo de funcionamiento del algoritmo consiste en mover cada vector al cluster cuyo centroide esté más cercano al mismo, y actualizar después los centroides de los clusters. Su convergencia depende mucho del número de clases.

4. 2. 2 Algoritmo K-vecinos axial o incremental. Este algoritmo, como su nombre indica, calcula los clusters 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 clusters y devolver el índice del cluster 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 cluster incremental son muy atractivos para el tratamiento de patrones documentales, dado el gran espacio de almacenamiento que requieren dichos patrones.

El algoritmo de cluster euclídeo no converge necesariamente en un conjunto fijo de prototipos: los prototipos pueden variar infinitamente, sin converger en el tiempo. El número de clusters creados tampoco es necesariamente finito, y depende de las funciones utilizadas en el algoritmo.

4.3 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.

El aprendizaje de una red neuronal está relacionado con los pesos de las conexiones entre sus nodos [23]. Cuando se presenta un patrón a la red, ésta produce una respuesta. Si la respuesta o salida de la red no es la supuesta, habrán de hacerse modificaciones para acercar la respuesta obtenida a la esperada. La señal que se recibe en la capa de neuronas de entrada cuando se le presenta el patrón se mueve a través de los enlaces o conexiones entre capas, hacia las neuronas de la capa de salida. Estos enlaces modulan la señal a su paso con los pesos que los caracterizan. Por lo tanto, si se quiere modificar la señal que llega al final a la capa de salida, habrá que actuar sobre dichos pesos.

Las reglas de aprendizaje especifican cómo se irán modificando los pesos de las conexiones a medida que se entrena la red para mejorar el rendimiento de la misma, es decir, que la salida se vaya aproximando cada vez más a la esperada.

Existen dos tipos fundamentales de aprendizaje: Supervisado y No supervisado. Un clasificador es un sistema que va a permitir determinar cuál de las M clases es la más representativa para un patrón de entrada no estático que contiene N elementos.

El clasificador neuronal actúa en dos etapas, contabilizándose en la primera el número de elementos que pertenecen a cada clase y en la segunda se selecciona el máximo. La primera etapa se alimenta con los N elementos del patrón de entrada en paralelo, produciéndose aquí la comparación del patrón de entrada con los prototipos de las distintas clases y pasando los resultados intermedios a la siguiente etapa en paralelo. En la segunda etapa se selecciona el máximo. Habrá salida para todas las clases, pero al acabar la clasificación sólo será apreciable la salida para la clase con mayor probabilidad, y el resto serán valores muy bajos o inapreciables. Se pueden utilizar las salidas como realimentación de la primera etapa adaptando los pesos iniciales según un determinado algoritmo de aprendizaje (principio de realimentación negativa).

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

4. 4 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 drá 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 cluster no debe ser muy dispar, ya que las áreas temáticas suelen tener un número parecido de descriptores, el número de clusters en los que se divide uno dado tampoco debe ser muy alto, ya que cada nuevo cluster representa un conjunto de términos que serán globalmente específicos, aunque sólo alguno(s) en un primer nivel de jerarquía. A mayor número de clusters 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 cluster. Se han tenido en cuenta 4 formas de obtención de raíces que son:

- Mediante el cálculo del centroide que representa el centro de masas del cluster o conjunto de descriptores en consideración.
- Seleccionando el descriptor más general del cluster, tomando aquél que tenga mayor número de apariciones en el total de documentos del corpus.
- Seleccionando el descriptor más general del cluster, escogiendo aquél 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 clusters 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 cluster 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 clusters 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 aquéllas generadas en el primer paso de clusterización. La generación de áreas temáticas [24] [4] 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 cluster.

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 agrupando 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.

4. 5 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 cuáles las más importantes son:

- Se dispone 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.
- Se establece una primera escala de relaciones (para los cinco tipos de relaciones existentes) buscando definir la calidad y riqueza de éstas.
- Puede instaurarse una segunda escala, respecto a los distintos tipos de relaciones, en función de la dificultad de encontrar específicamente cada relación.
- Si dos o más descriptores se consideran sinónimos, uno de ellos tiene que seleccionarse como representante del grupo de sinónimos.

BIBLIOGRAFÍA

[1] Díaz, I.; Lloréns, J.; Martínez, V.; Velasco, M.: Semi- Automatic Construction Of Thesaurus Applying Domain Analysis Techniques. *International Forum on Information and Documentation* . Vol 23, nº 8. 1998

[2] Llorens, J. Definición De Una Metodología Y Una Estructura De Repositorio Orientadas A La Reutilización. *Tesis Doctoral. Universidad Carlos III de Madrid*. 1995

[3] Lloréns, J.; Amescua, A.; Velasco, M.: Software Thesaurus: a Tool for Reusing Software Objects. Actas del *4º IEEE Assessment on Software Tools*. Toronto, Canada. 1996.

[4] Lloréns, J.; Amescua, A; Velasco, M.: A Software Thesaurus as an Intelligent Tutorial System for Software Specifications. *Tercera Conferencia Internacional en Intelligent Tutorial Systems. ITS-96*. Montreal. Canada. 1996.

[5] Lloréns, J.; Velasco, M.; Amescua, A.; Moreiro, J. A.; Martínez, V.: Automatic Domain Analysis using Thesaurus Structures. Aceptado en publication in *Journal of the American Society of Information Science*, 1998.

[6] Prieto-Díaz, R.; Freeman, P.: Classifying Software for Reusability. *IEEE Software*. 1987.

[7] Prieto-Díaz, R.: Domain Analysis for Reusability. Actas del *COMPSAC'87*, pp 23-29, Tokyo, 1987.

[8] Lloréns, J.; Velasco, M.; Pérez, R.: Modelización Activa: Técnicas y Métodos. *INFONOR 97*. Chile. 1997.

[9] Lloréns, J.; Amescua, A.; Martínez, V.; Velasco, M.: The Reuse Maturity Model: 3RMM. *Symposium on Computer and Information Science. ISCISXII*, Antalya, Turquía. 1997.

[10] Díaz, I; Velasco, M.; Molina, J.M.: 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.

[11] Hopfield, J.: Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Actas del National Academy of Science, vol 81, 3088-92. *National Academy of Science*, 1982.

[12] Sanchis, A.; Molina, J. M.; Isasi, P.: Learning Reactive Behavior for Autonomous Robots using Classifier Systems. Spatiotemporal Models in Biological and Artificial Systems. *IOS Press*, 1997.

[13] Zimmermann, H. J.: Fuzzy Set Theory and its Applications. *Kluwer Academic Publishers*. 1990.

[14] Velasco, M.; Moreiro, J. A.; Lloréns, J.: Estado Actual del Proyecto GDA (Gestión Documental Automatizada): Planteamiento Teórico y Descripción Práctica. *ISKO 97*. Madrid, Noviembre, 1997.

[15] Velasco, M.; Martínez, V.; Lloréns, J.; Amescua, A.: Automatic Domain Analysis: Generation of Domain Representations. Aceptado en *IFIP 98*.

[16] Velasco, M.: Generación Automática de Representaciones de Dominios. *Tesis Doctoral. Universidad Politécnica de Madrid*. 1998.

[17] Frakes, W. B. & Baeza-Yates, R.: Information Retrieval: Data Structures & Algorithms . *Prentice Hall*. 1992.

[18] Fernández Herrero, J. A., Lloréns, J. & Velasco, M. : Indización Básica contra el Tesauro Autogenerable *Universidad Carlos III de Madrid*. 1995.

[19] Muñoz, A.: Redes Neuronales para la Organización Automática de Información en Bases Documentales. *Tesis Doctoral. Universidad de Salamanca*. 1994.

[20] Salton, G.: Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer. *Addison-Wesley*, cop. 1989.

[21] 21.Zipf, G. K.: Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. *Haffner*. New York. 1972.

[22] Chen, H.; Lynch, K. J.: Automatic Construction of Networks of Concepts Characterizing Document Databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22, pp 885-902, 1992.

[23] Hebb, D.: Organization of Behaviour . *Wiley & Sons*, New York, 1949.
[24] Van Slype, G.: Les Langages d'Indexation: Conception, Construction et Utilisation dans les Systèmes Documentaires. Paris. *Les Editions d'organisation*. 1991.