DESARROLLO DE UN MÉTODO PARA LA CREACIÓN AUTOMÁTICA DE MAPAS CONCEPTUALES

Moreiro, José; Marzal, Miguel Ángel; Beltrán, Pilar
[jamore, mmarzal, pbeltran]@bib.uc3m.es
Departamento de Biblioteconomía y Documentación

Morato, Jorge; Sánchez-Cuadrado, Sonia; Llorens, Juan
Departamento de Informática, Universidad Carlos III
[jorga, ssanchec, llorens]@ie.inf.uc3m.es
Universidad Carlos III de Madrid – España

Resumen: Se muestra la metodología empleada en el proyecto REID. El proyecto está encaminado a la creación semiautomática de mapas semánticos. Se exponen los distintos problemas encontrados a lo largo del proyecto y las distintas soluciones lingüísticas, informáticas, de representación del conocimiento y documentales que se han estimado más prometedoras. Los mapas conceptuales obtenidos son de aplicación en indización y recuperación de información, estudiándose extensiones para la navegación conceptual y la vigilancia científica.

Abstract: It is shown the methodology used in project REID. The project is directed to the semiautomatic creation of semantic maps. The different problems found throughout the project and the different solutions are exposed: linguistics, from computer science, and from documentary representation of the knowledge and that have been considered more promising. The obtained conceptual maps are of application in indexing and information retrieval, studying extensions for conceptual navigation and the scientific monitoring.

1 INTRODUCCIÓN

Se describe aquí el desarrollo del proyecto REID[1], cuyo objetivo es la representación automática de mapas semánticos. Se analizan, en concreto, los problemas que se han encontrado durante los tres años de duración del proyecto, así como las soluciones adoptadas. Es poco frecuente encontrar este tipo de información en el desarrollo de proyectos por considerarse errores de análisis o de planteamiento. Por el contrario, los autores pensamos que esta información es útil para el planteamiento de extensiones y proyectos similares ya que pueden ayudar a ahorrar recursos y esfuerzos.

El documento esta organizado según las etapas de desarrollo del proyecto, haciendo especial hincapié en el desarrollo de aplicaciones informáticas para el estudio. Tras la

[1] El proyecto REID: *Desarrollo de un tesauro de verbos para entornos de información dinámica. Aplicación del estándar ISO/ICE: 1 32 50:1999*, está siendo financiado por la CICYT. Plan General del Conocimiento, con el número de proyecto TIC 2000-038. Su duración abarca los años naturales de l 2000 al 2003.

descripción de cada herramienta se ha procedido a describir los problemas y soluciones adoptados.

2 METODOLOGÍA

Mediante el estudio de las palabras encontradas en los documentos aportados como entrada al sistema (corpus documental) se obtiene una lista con todos los términos que describen el dominio en cuestión.

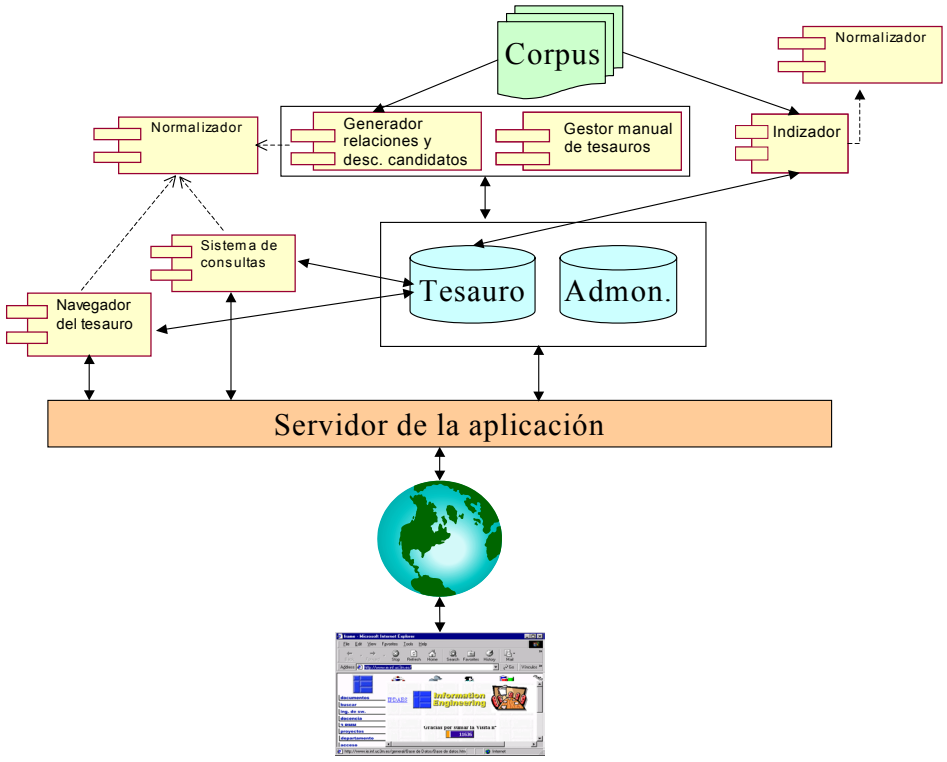

Figura 1. Metodología del proyecto REID

Para ello se deberán cumplir los siguientes pasos:

1. Identificación del vocabulario:
    a. Normalización de cada una de las palabras encontradas
    b. Extracción del tipo correspondiente de cada palabra
    c. Identificación de palabras compuestas
2. Filtrado del vocabulario
3. Identificación de las relaciones entre los términos del vocabulario
4. Filtrado de las relaciones
5. Generación de la primera versión del mapa conceptual que representa el dominio

Para el cada uno de los módulos anteriores se creó un software específico. En el caso, del filtrado se realizó un módulo de ayuda a la eliminación manual.

El almacenamiento de los datos se realizó mediante una base de datos relacional en SQL Server.

2.1 PROBLEMAS PLANTEADOS

En los inicios del proyecto se planteó la hipótesis de si era factible identificar un corpus documental idóneo que permitiera una representación conceptual del área relacionada con el corpus.

Los problemas encontrados en este planteamiento se centran en los siguientes aspectos:

1) Debidos a la hipótesis de partida:
   a) Se presupone que los documentos contienen de forma exhaustiva el vocabulario del área, o al menos aquellos términos más utilizados –preferidos- por los profesionales que trabajan en dicho área.
   b) Las relaciones entre conceptos se pueden obtener analizando la concurrencia entre pares de términos en la misma frase dentro de los documentos. Es decir, partiendo de un término y analizando los otros términos de la frase se pueden recorrer la mayoría de los conceptos presentes en los documentos.
2) Metodológicos:
   a) Problemas asociados a la identificación del corpus documental
   b) Los problemas propios del Procesamiento del Lenguaje Natural: normalización terminológica, anáforas, desambiguación de la categoría gramatical, frases ambiguas, etc.
   c) Problemas originados en la consecución de una representación semántica idónea
   d) Problemas relacionados con las herramientas desarrolladas para el filtrado terminológico y relacional de los elementos más relevantes

2.2 DESCRIPCIÓN DEL PROYECTO

2.2.1 CREACIÓN DEL CORPUS DOCUMENTAL

Un corpus documental apropiado debe tener las siguientes características:

i) El discurso y la tipología documental en todos los documentos seleccionados debe ser similar, para que las variaciones debidas al estilo sean escasas.

ii) Los documentos del corpus deben pertenecer a un área de conocimiento bien delimitada, con un mínimo de solapamiento con áreas próximas. El solapamiento puede ocasionar un cambio en las frecuencias para identificar la terminología preferida o ambigüedad en los términos debido, por ejemplo, a la homonimia.

iii) No debe existir un conocimiento implícito en el área no haya sido expresado en los documentos. Es decir, no existe un conocimiento en el área que se presuponga en el corpus documental.

Resulta evidente que la creación de un corpus documental con las características expuestas es muy complicada. Las soluciones lingüístico-documentales adoptadas han consistido en crear un corpus con un discurso preferentemente científico-técnico, con una tipología de artículo científico en áreas tan concretas como el representado dentro del tema petróleo para el refino.

Para incluir el conocimiento implícito se ha recurrido a dos estrategias:

- Obras de referencia: como diccionarios, libros de texto, etc, en general material que no presupone ningún conocimiento previo. Lamentablemente, esta solución no es siempre factible al no tener no tener en electrónico para el área estas obras.
- Opinión de expertos: mediante un software que permite la inclusión manual de términos no presentes en los documentos

En el proyecto, los corpus se construyeron a partir de documentación procedente de los siguientes dominios:

- Derecho
- Documentación
- Petróleo
- Informática

La selección de estos temas vino motivada principalmente por la disponibilidad de documentos electrónicos en español en dicho campo escritos para profesionales del dominio.

La razón para seleccionar varios dominios reside en estudiar si existen diferencias significativas en los resultados obtenidos mediante la metodología expuesta al variar de campo.

2.2.2 IDENTIFICACIÓN DE LA TERMINOLOGÍA DEL DOMINIO

Durante esta fase se van a identificar aquellos términos que determinan el dominio representado por los documentos de partida. Esta fase puede ser realizada mediante herramientas informáticas que filtren la información electrónica contenida en los documentos de partida y proporcionen un conjunto de términos simples y/o compuestos que describan los conceptos del dominio.

Esta fase se realizará en las siguientes etapas:

1. Extracción automática de las palabras simples incluidas en los documentos.
2. Normalización/Lematización de las palabras en un mismo término canónico.
3. Generación de términos compuestos a raíz de un análisis léxico.
4. Filtrado terminológico de acuerdo con criterios estadísticos, tipográficos y de localización.

Todas estas sub-tareas se describen con mayor profundidad en los siguientes apartados.

Extracción automática de palabras simples

Mediante un sistema informático de tratamiento de la información textual y gráfica producida por las aplicaciones ofimáticas mas comunes, se consigue extraer las palabras simples que forman los documentos del corpus.

Normalización

El objeto de este módulo es conseguir identificar conceptos, bien representados mediante palabras simples, o bien mediante términos compuestos a partir de un documento textual. El problema que se presenta radica en el hecho de que palabras como ‘casa’ y ‘casas’ que realmente son distintas, deberían representar el mismo concepto. Para solucionar este problema, este módulo seguirá los siguientes pasos:

- Generar todas las variantes flexionadas de cada término de entrada en función de un conjunto de reglas de lematización para cada idioma tratado.
- Confrontar cada variante flexionada contra un vocabulario del mismo idioma con vistas a identificar su categoría morfológica real. Tomar decisiones en función del valor de dicha categoría. Así, el valor semántico de un posible sustantivo será mayor que el de una preposición.
- Proceso de normalización: las distintas variantes de términos flexionadas y derivadas son unificadas bajo una forma normalizada, para de esta manera eliminar las distintas variantes debidas al número, género, … La consecución de este paso se articula mediante la aplicación de una serie de reglas que sustituyen las distintas terminaciones con otra normalizada que se estime oportuna.

Por ejemplo, si se procesa la palabra "perforadores" y tuviésemos como únicas reglas:

"es"→[nada]

"s"→[nada]

nos resultarían dos términos, "perforador" y "perforadore". Como la segunda palabra no pertenece al vocabulario controlado la única salida válida del proceso sería "perforador". En caso de no existir la palabra "perforador" se incluiría el término "perforadores" sin normalizar.

Problemas identificados

Cambios en la semántica debido a la normalización. Solución adoptada: Es necesario un etiquetador que identifique la categoría gramatical del término para no aplicar reglas erróneas. La identificación de la categoría gramatical trabaja con una maquina de estados que de forma probabilística asigna la categoría según el elemento precedente.

p.e "casas" [verbo] "casas" ↛ [sustantivo]

Estas transformaciones de la categoría gramatical basadas en la regla de normalización están prohibidas en el sistema

Actualmente, la desambiguación no se produce correctamente en todos los casos, este problema se intenta solucionar con las siguientes estrategias:

- Mediante las frecuencias de aparición en la fase de filtrado
- Creando nuevos tipos de términos problemáticos como poner a los participios como un tipo más junto con verbos y sustantivos. Esto soluciona por ejemplo normalizaciones erróneas de adjetivos.
- Utilizando listas de topónimos, gentilicios y nombres propios de personas, para que estos términos no sean normalizados

Generación de términos compuestos

La utilización de un vocabulario controlado basada en unitérminos supone a menudo un aumento de la ambigüedad terminológica que la creación del tesauro pretende evitar. Para ello este módulo construye términos compuestos que son, por lo tanto, más específicos. Para ello se identifican las categorías gramaticales de los elementos implicados en la construcción del término compuesto y de aquellos elementos susceptibles de "pegar" dichos elementos para formar un término cuyas partes por separado no tienen la misma semántica que el término compuesto.

Por ejemplo, si tenemos las palabras “plataforma” y “perforación” el término compuesto será “Plataforma de perforación”, aunque independientemente “plataforma” o “perforación” puedan tener otra semántica al funcionar de forma aislada e independiente).

Por lo tanto, la principal ventaja de esta fase no es sólo la de poder encontrar conceptos compuestos, sino también la de poder encontrar relaciones automáticamente entre los términos compuestos, y sus términos simples constituyentes (en caso de que los conceptos simples se consideren como relevantes al dominio).

Un primer tipo de relación estructural –no basada en verbos- se obtiene en esta etapa, al identificar como elemento genérico de “plataforma de perforación” el término “plataforma”.

Filtrado

El número de términos obtenidos en la anterior etapa puede resultar excesivo. En un primer momento se intento utilizar a un modelo tipo TF-IDF totalmente automatizado para eliminar automáticamente términos.

Problemas detectados:

Lamentablemente los resultados no fueron esperanzadores por lo que se recurrió a usar una estrategia semimanual que combinase varios filtradores.

Los tres sistemas utilizados se describen a continuación:

- Módulo de filtrado asistido inteligente
- Filtrado terminológico por frecuencias
- Validación manual de los términos restantes por expertos en el área

Módulo de filtrado asistido inteligente

Este módulo permite tratar de forma masiva con términos especialmente problemáticos. Por ejemplo, los términos inferiores a tres caracteres son a menudo errores tipográficos en los documentos. Este módulo muestra cada uno de los términos con n caracteres permitiendo manualmente eliminarlos, editarlos o unificarlos con otro término del vocabulario.

El sistema permite entre otras opciones seleccionar aquellos documentos que solo varían en los términos centrales para optar por su unificación como términos equivalentes. Por ejemplo, muestra permitiendo la unificación de los términos “Motores de gasoil” – “Motores de funcionamiento por gasoil”

También permite el uso de listas planas conteniendo términos con errores frecuentes del sistema.

Filtrado terminológico por frecuencias

A. Filtrado mediante un corpus de comparación

Este apartado se subdivide en las siguientes tareas:

- Cálculo del número de veces que aparece un mismo normalizado en distintos documentos.

Comparación del dato anterior con el que sería esperable en un corpus de comparación. El corpus de comparación está formado por documentos de un dominio general o de varios dominios con un bajo índice de solapamiento con el dominio estudiado pero con un discurso y tipología documental similar. Su función es comparar la frecuencia de los términos en esta colección con la obtenida en el procesamiento del corpus documental. Cuando la frecuencia relativa es similar, se presupone que el término pertenece al vocabulario científico-tecnológico del idioma y no al vocabulario específico del dominio de estudio.

B. Validación mediante frecuencias

También se recurre a revisar aquellos términos con muy escasa frecuencia de aparición en número de documentos o en número total de apariciones en la colección, de forma análoga a TF-IDF.

C. Evaluación mediante formato y localización del término

En este apartado se hicieron numerosos experimentos con la posición del término en la frase, presuponiendo que cuanto antes aparezca en la frase mayor es su importancia. Y con el formato, un término en negrita o formando parte del título puede tener mayor importancia que el resto. Aunque ambas estrategias dieron buenos resultados existe el problema de asignar esos pesos, ya que las variaciones de un autor a otro también parece conllevar cambios en la asignación de pesos.

Validación por expertos

Los expertos del área pueden evaluar los términos restantes siempre que el número no sea excesivo. En el proyecto se hizo una aplicación para validación remota vía web para este fin, mostrando antes aquellos términos que aparecen en un mayor número de frases para su aceptación.

Almacenamiento del filtrado

Una vez que se han encontrado, éstos deberán ser almacenados en la base de datos. Pero, a diferencia del módulo de indización, el resultado del filtrado no se almacena documento por documento, sino que se deberá generar un documento de tipo *Mapa*

Conceptual. Este nuevo artefacto tiene el nombre indicado por el usuario, y contiene una relación de ocurrencia donde colgarán todos los términos que hayan pasado el proceso de filtrado.

2.2.3 Generar relaciones entre los conceptos seleccionados

Mapeado de Pares Sintagma Verbal con Relación Semántica

El objetivo de esta fase pretende ser la identificación automática de relaciones entre los conceptos mediante la aplicación de algoritmos informáticos.

El método propuesto se basa en el análisis de los sintagmas de la frase. Aunque existe una literatura muy diversa que sigue el planteamiento que se explicará a continuación, consideramos nuestro precedente, precisamente porque este proyecto es continuación de su tesis doctoral, el de Díaz (2001).

Básicamente, el método esta compuesto por las siguientes actividades:

- Identificación de verbos que unen el sintagma nominal sujeto y el objeto o complemento.
  - por ejemplo: < un motor de vapor> <es un tipo de> <motor>
    - problemas:
      - solo un subconjunto de las relaciones implícitas las establece el sintagma verbal, p.e “un tipo de motor, el de vapor, es empleado frecuentemente”
      - la normalización terminológica previa puede eliminar relaciones en pasos siguientes: p.e. participios llevados a infinitivos, o supresión de preposiciones en la frase.
    - solución adoptada:
      - modificación del autómata de relaciones para evitar estos problemas en el segundo problema. Ampliación de los objetivos del proyecto en el primero.
- Asociar estos verbos con diferentes tipos de relaciones de tipo tesauro.
  - en el ejemplo: < un motor de vapor> <es un tipo de> <motor>, la estructura <ser un tipo de> puede asociarse a una relación generalización.

    Problemas:
    - las relaciones no son biunívocas, una misma frase en distintos contextos puede estar indicando relaciones distintas dependientes del contexto. Por

supuesto, un mismo sintagma nominal puede estar señalando a varias posibles relaciones.

- Uno de los principales problemas es que no todos los términos del vocabulario validado está relacionado con otro término validado en las sentencias de los documentos, por lo que se deberá crear otro tipo de relación no de tipo tesauro para aquellos términos sin relacionar.
- Las relaciones obtenidas con formas impersonales de verbos parecen ser mejores que con formas personales

Soluciones:

- utilizar criterios de frecuencia para establecer una relación entre conceptos. Por ejemplo, es más fiable si tres estructuras verbales diferentes indican la misma relación

- Actualización de la base de datos para insertar la relación identificada entre los distintos descriptores del tesauro.
- Opcionalmente: Como ya se ha visto más arriba la calidad aumenta si se utilizan herramientas estadísticas, de localización en el documento y tipográficas.

A pesar de que este proceso identifica todas las relaciones que se encuentren entre un sujeto, un verbo y sus complementos, pero teniendo en cuenta que el principal objetivo de esta fase es encontrar en los textos las relaciones que deban aparecer en el tesauro, estos verbos se pueden clasificar en dos grandes grupos de interés para la generación del dominio:

1. Aquellos con una semántica alta: reprogramar, fotografiar, ...
2. Los que tengan una semántica con una semántica similar a las relaciones encontradas en el tesauro. Por ejemplo, "ser un", "estar asociado con", "ser un tipo de", "ser equivalente a" , "ser genérico de", … Es decir, relaciones que pudieran encontrarse en cualquier dominio y cuyo valor principal es relacionar conceptos.

Creación de un Tesauro

Una vez seleccionada la lista de aquellos conceptos que son relevantes para el dominio seleccionado, el siguiente paso para obtener una representación de conocimiento en base a un mapa conceptual es la generación de una serie de relaciones entre dichos conceptos. Las relaciones que se han de generar son las definidas en la norma ISO 2788, es decir:

- Generalización/Especialización
- Asociación

- Sinonimia

Para obtener estas relaciones, se tendrán en cuenta los términos obtenidos en el paso anterior. Para cada uno de ellos, se estudiarán las frases en las que aparezcan conjuntamente dos o más de estos términos. En dichas frases, se hará un estudio de los posibles verbos que unen a los conceptos, de tal forma que se intenten relacionar algunos de los verbos más comunes del castellano con las relaciones de la norma ISO.

Así, por ejemplo, si el sistema detecta la frase "Un coche es un tipo de vehículo que ...", y en el caso de que tanto 'coche' como 'vehículo' hayan pasado el filtro de la generación de conceptos, el sistema será capaz de deducir una relación de tipo 'generalización/especialización', donde 'vehículo' es el concepto genérico y 'coche' represente al concepto específico.

Problemas encontrados

- Los tesauros no parecen ser la solución idónea, ya que otras estructuras se muestran más adecuadas por su mayor riqueza semántica.

Solución aplicada

- Utilizar el estándar *Topic Maps*, ya que a pesar de ser más complejo en estructuras, parece más próximo al vocabulario del corpus.

Creación de *Topic Maps*

Topic Map representa un nuevo estándar de representación de conocimiento. Un *Topic Map* tiene como finalizar normalizar los elementos y la notación utilizada para estructurar la información mediante la construcción de una red de enlaces semánticos que relacionen diferentes recursos informativos. Se caracteriza fundamentalmente por representar un documento en base a los conceptos (*topics*) que aparece en él, pero no de forma aislada, sino desde el punto de vista de las relaciones que aparecen entre los distintos *topics* del documento. Así, la frase 'Lorca nació en Granada', no se caracteriza por los conceptos 'Lorca' y 'Granada' como hacen la mayoría de los sistemas previos, sino que se caracteriza por una relación etiquetada como 'nacer' entre ambos conceptos. Tras *Topic Maps* están distintos objetivos como:

- Crear una estructura que favorezca la fusión de distintas representaciones semánticas
- Crear una herramienta que posibilite la web semántica, primando la navegación conceptual, sobre la navegación documental típica del web
- Combinar conceptos y recursos en un mismo diseño

- No limitar el número de relaciones posibles entre términos, sino dejarlo de forma totalmente abierta

Esta idea potencia enormemente la exactitud en la fase de recuperación documental y, por ello, ha sido tomada como idea a la hora de representar conocimiento en el presente proyecto.

Por lo tanto, el presente proyecto utiliza *Topic Map* como forma almacenar la información contenida en un documento textual, pero además representa un sistema novedoso para la generación automática de *Topic Maps* a partir de documentos textuales.

2.2.4 Mantenimiento y corrección del tesauro

Tras los pasos anteriores es muy posible que dicho tesauro tenga que ser actualizado con el tiempo. Esto puede ser debido a varios problemas:

- El dominio representado en un primer momento no era el adecuado o ha de cambiarse ligeramente por alguna otra razón (si el dominio cambiase por completo, en lugar de modificarlo en base a la herramienta de mantenimiento, sería más eficaz empezar el proceso desde el principio)
- No se ha cambiado el dominio objetivo, sin embargo el dominio original se ve alterado por la inclusión de nueva terminología (tecnicismos, anglicismos, ...)

Por lo tanto, se debe proporcionar una herramienta que permita las siguientes operaciones:

- Inserción y eliminación de conceptos
- Inserción y eliminación de relaciones
- Renombrado de conceptos

Todas estas tareas, que aparentemente son triviales, se complican por el hecho de que hay ciertas reglas que siempre han de mantenerse para que se pueda considerar que el tesauro es consistente. Entre las reglas que han de mantenerse se encuentran, por ejemplo, las siguientes:

- Sólo los conceptos representativos del conjunto de sinónimos podrá estar relacionado con otros conceptos
- Entre cada pareja de conceptos sólo puede haber un tipo de relación
- No se pueden dar ciclos en las relaciones de generalización
- ...

Tras recoger todas estas relaciones, se deberá estudiar cuáles de ellas son verdaderamente significativas para el dominio. Asimismo, se deberá comprobar si entre los distintos documentos aportados se producen relaciones incompatibles entre sí.

Una vez terminado este proceso automático, deberá utilizarse una herramienta que permita la gestión (mantenimiento) tanto de los conceptos como de las relaciones del mapa conceptual.

3 CONCLUSIONES

Se ha expuesto una metodología para creación y representación automática de dominios. Durante el desarrollo del proyecto se han encontrado distintos problemas a los que ha habido que dar respuesta. Estos problemas son frecuentemente olvidados en los documentos que tratan del desarrollo de los proyectos, a pesar de ser una de las fuentes más valiosas de información para futuros investigaciones.

Entre otros se ha intentado dar respuesta a los problemas encontrados en el procesamiento del lenguaje natural, a los filtrados terminológicos y de relaciones y a la validación definitiva de la representación automatizada.

Los experimentos realizados hasta el momento no han permitido llegar a la conclusión de que sea prescindible la supervisión manual en las etapas de creación del corpus, filtrado y validación. Si bien varios de estos pasos pueden ser ayudados por los ordenadores.

4 REFERENCIAS

- Díaz Rodríguez, S. I.- *Esquemas de representación de información basados en relaciones: aplicación a la generación automática de representaciones de dominios*. Tesis doctoral, Director, Juan Lloréns Morillo. Leganés: Universidad Carlos III de Madrid, Departamento de Informática , 2001.
- ISO /IEC JTC 1/SC34 *Information Technology - Document Description and Processing Languages*.

  http://www.topicmaps.com/content/resources\iso13250/iso13250-1999-fcd.htm .
- ISO. *Guidelines for the establishment and development of monolingual thesauri: international standard ISO 2788*. ISO. 2nd ed. 1986-11-15. [Geneve]: ISO, 1986.
- Moreiro, J.A.; Sánchez Cuadrado, S.; Morato, J.- Panorámica y Tendencias sobre *Topic Maps* (en línea), en Rovira, C. y Codina, L. (dir.).- *Documentación digital*. Barcelona: Sección Científica de Ciencias de la Documentación.

Departamento de Ciencias Políticas y Sociales. Universidad Pompeu Fabra, 2002. ISBN 84-88042-39-6: http://www.editaweb.com/docdigitalinfo/

- Holger Rath, H. y Pepper, S.- Topic Maps at work, en *The XML Handbook*. 2nd Edition. New Jersey: Prentice Hall, 1999.
- Moreiro González, José A.; Llorens Morillo, Juan; Marzal García-Quismondo, Miguel Ángel; Morato Lara, Jorge ; Sánchez Cuadrado, Sonia; y Beltrán, Pilar.- Utilización de estructuras verbales en la identificación de relaciones y descriptores en tesauros, en *Ciencias de la Información*, 2002, 33, nº 2: 3-20. http://www.cinfo.cu/cinfo2002/v33n2a2002/index2_2002.htm