Técnicas de recuperación de información aplicadas a la construcción de tesauros

Information retrieval techniques applied to the development of a thesaurus

Blanca GIL URDICIAIN[1]
Rodrigo Sánchez JIMÉNEZ[1]

Resumen

El artículo propone la aplicación de un conjunto de técnicas propias del ámbito de la Recuperación de Información a la elaboración de Tesauros. Las propuestas que se presentan se aplicaron en la selección de la terminología, en la categorización de términos mediante *clusters*, y en el establecimiento de relaciones semánticas entre los términos, por procedimientos de similitud, que dieron como resultado un Tesauro de Comercio Exterior, de 7.790 términos. De tales resultados se puede concluir que las técnicas utilizadas simplifican de forma considerable las tareas para la recopilación de la terminología, y pueden suponer una mejora de la calidad del Tesauro resultante, en tanto que permiten el análisis de las condiciones de la colección para la que se utilizará el Tesauro, así como aportar información extra a los expertos que es difícilmente obtenible de forma manual.

Palabras clave: Construcción de tesauros. *Clustering*. Modelo de espacio vectorial. Modelo generalizado de espacio vectorial. Semántica latente.

Abstract

The aim of the article was to propose the application of a set of techniques used in Information Retrieval for the development of a Thesaurus. The proposed ideas have been applied in the selection of the terminology; categorization of terms by creating clusters; and establishment of semantic relationships between terms through semantic similarity, which resulted in a Foreign Trade Thesaurus of 7,790 terms. From these results, we concluded that the techniques used significantly simplified the tasks of obtaining the terminology, and they can improve the quality of the final thesaurus. In addition, the techniques enabled the analysis of the conditions of the collection for which the thesaurus is used and provide extra information that would be hard to obtain manually.

Keywords: *Thesaurus development. Clustering. Vector space model. Generalized vector space model. Latent semantic indexing model.*

Introducción

Este trabajo propone la aplicación de técnicas del ámbito de la Recuperación de Información a la elaboración de un Tesauro sobre Comercio Exterior. Durante su desarrollo se aplicaron las técnicas existentes y se perfeccionaron con objeto de profundizar en la investigación de nuevas metodologías para la elaboración de Tesauros. De las fases en las que se desarrolla un Tesauro, las medidas propuestas benefician concretamente a la selección previa del léxico, su depuración y posterior agrupación en campos

[1] Universidad Complutense de Madrid, Facultad de Ciencias de la Documentación. Calle Santísima Trinidad, 37, 28010, Madrid, España. Correspondencia a nombre de/*Correspondence to*: B. GIL URDICIAIN. *E-mail*: <mbgil@pdi.ucm.es>.

Recebido el día 18/2/2013, re-presentado el 24/9/2013 y aceptado para su publicación el 4/7/2013.

semánticos, y al establecimiento de relaciones semánticas entre los términos. También afectan a la elaboración de un protocolo de mantenimiento y actualización de dicho lenguaje documental.

Existen dos líneas de investigación para la generación automática de Tesauros, una de ellas considera los Tesauros como herramientas lingüísticas, y se desarrolla en el ámbito del procesamiento del lenguaje natural. Esta línea está representada en los trabajos de Grefenstette (1994) y Curran (2001). La investigación se enfoca hacia la utilización de los Tesauros como herramientas lingüísticas, prestando especial atención a las relaciones existentes entre los términos y el contexto. La segunda línea de investigación tiene por objeto la utilización de los Tesauros como herramientas de recuperación de información siendo, por tanto, la de mayor interés desde el punto de vista de la Documentación. Estas aproximaciones se centran en la utilización del modelo de Espacio Vectorial y, al contrario que las investigaciones anteriores, se centran en la relación existente entre los términos y los documentos, tendencia cuyos principales trabajos recoge Pérez Agüera (2005). Uno de los puntos fuertes de la investigación en el área es el desarrollo de nuevas medidas de similitud que permitan mejorar la detección de las relaciones entre términos (Chen *et al*., 1995), así como la utilización de técnicas de *clustering* o agrupación para la localización de los conjuntos de términos semánticamente más próximos (Crouch & Yang, 1992) y (Aitchison *et al*., 2007). Son precisamente estas dos ideas las que se estiman de interés para el presente proyecto. Además de estos focos de investigación, el empleo de diccionarios para el establecimiento de relaciones jerárquicas (Moreiro González *et al*., 1999) y la utilización de conocimiento en forma de Tesauros preexistentes conforman los puntos de interés fundamentales para los investigadores.

Aunque comparte varias de estas técnicas, este trabajo difiere en la intención de llevar a cabo un proceso paralelo supervisado para la creación de un Tesauro. El objetivo no es tanto llegar a elaborar de forma automática los Tesauros, como profundizar en la utilización de sistemas de ayuda a su creación semiautomática, compatible con la metodología tradicional. Para conseguirlo, se establecen procedimientos de compilación de una terminología caracterizada por el habitual empleo de expresiones extranjeras, y se proporcionan recursos para la futura actualización del lenguaje.

Desde un punto de vista aplicado, el proyecto de elaboración de un Tesauro de Comercio Exterior se desarrolla para proporcionar al Instituto Español de Comercio Exterior de España (ICEX) un lenguaje controlado capaz de cubrir las necesidades de normalización del vocabulario de las bases de datos de dicho Instituto y de las 92 oficinas que intercambian información con esta institución en todo el mundo. El Tesauro da cobertura a 20.000 documentos electrónicos, 3.600 documentos textuales y 10.000 fotografías que forman en estos momentos las bases de datos del Instituto. El Tesauro permitirá una indización homogénea de los documentos y facilitará la recuperación de información, permitiendo el intercambio de información.

Metodos

Se combinaron procedimientos automáticos con las fases de trabajo habituales en la construcción de Tesauros, así, se realizó el proceso de selección del léxico, establecimiento de campos temáticos, asignación de descriptores a campos temáticos y establecimiento de relaciones en orden secuencial, ya que el trabajo automatizado debía realizarse con anterioridad a la supervisión de los expertos, y las diferentes técnicas utilizadas crecían sobre los resultados de técnicas utilizadas durante la fase inmediatamente anterior.

Para diseñar un sistema de asistencia a la generación de Tesauros en el que las decisiones finales fueran tomadas por personas, se estimó la posibilidad de aplicar el Modelo Generalizado de Espacio Vectorial, que podía aportar un conjunto de técnicas utilizables de forma rápida para la selección de la terminología. Del análisis de otros modelos más complejos se extrajeron algunas de las ideas que proponen, utilizándose para refinar el sistema y mantener una perspectiva crítica sobre el modelo elegido, que sirvió para mejorarlo.

Del análisis del Modelo Generalizado de Espacio Vectorial también se obtuvo la idea de trasponer la matriz. La aplicación que se le da a este modelo clásico es la de obtener formas de representación de los documentos

que sean utilizables directamente por un ordenador para efectuar los cálculos oportunos. Sin embargo, como se desprende de lo dicho, en esta investigación se mantiene la hipótesis de que también se puede representar una colección como un conjunto de un espacio terminológico, por oposición a un espacio documental. Para llevar esto a cabo se procedió a trasponer la matriz. Este procedimiento se basa en la hipótesis de que si dos términos se utilizan con la suficiente frecuencia en los mismos contextos, se pueden llegar a considerar como semánticamente relacionados; esta condición de trabajo se refuerza mediante la utilización de algoritmos de *Lematización* o *Stemming* que permiten profundizar en el nivel conceptual y obviar algunas de las servidumbres propias del nivel léxico.

Para seleccionar la terminología se combinaron los métodos deductivo e inductivo. Con el primer procedimiento, se indizó automáticamente la colección de documentos que facilitó el ICEX; un indizador automático proporcionaría los términos con mayor poder de resolución y, por tanto, los términos con mayor capacidad de recuperación. Se implementó un indizador automático basado en *Term Frequency - Inverse Document Frequency* (TF-IDF) con un algoritmo de lematización de Porter incluido (Frakes & Baeza-Yates, 1992), lo que dio como resultado una lista de candidatos a descriptor. Sin embargo, el problema de la utilización de esta técnica de indización por unitérminos es que la tasa de precoordinación es obviamente nula, lo que limita la capacidad expresiva del lenguaje resultante. Aparte de esta pega, los pesos asignados por el sistema a cada uno de los descriptores resultantes concordaban bastante bien con lo esperado, de forma que la terminología resultante, una vez revisada manualmente, representaba globalmente los contenidos de la colección.

Este sistema de indización mostraba otros problemas, tales como la inclusión de una gran cantidad de verbos y otras partículas del lenguaje, poco viables como candidatos a descriptor. Esto, evidentemente no se solucionaba ampliando la ya extensa lista de palabras vacías, sino que era necesario tratar la información con otras técnicas propias del Procesamiento del Lenguaje Natural. Como resultado de estos ensayos preliminares conseguimos información acerca del carácter del indizador a utilizar: debía reconocer correctamente sintagmas o n-gramas, así como verbos, sustantivos, adjetivos, etcétera. Se decidió aplicar la fórmula del TF-IDF que parecía idónea desde el punto de vista teórico, y había proporcionado ponderaciones adecuadas de los términos.

De los indizadores automáticos disponibles de forma gratuita para propósitos de investigación, se escogió el indizador *Keywords and Keyphrases* (KEA), desarrollado por la Universidad de Waikato (Witten *et al*., 1999). KEA posee un *stemmer* propio, así como un módulo de sintagmatización con capacidades de reconocimiento de entidades (distinción entre nombres, verbos y otras partículas). Además está basado en TDF-IDF, por lo que cumplía todos los criterios antes mencionados. Así pues, se indizó la colección, con el resultado de un listado de términos preliminares de algo más de 10.000 términos, que proporcionaba información de interés para determinar su posterior aceptación, como sobre el peso asignado por la aplicación y el número de documentos en los que aparecía, por ejemplo:

> 10 El descriptor es "ACCESORIOS DE BAÑO "(Posición en la lista total de términos 10)
>
> Sus raíces son acceso de bañ
>
> La suma de sus pesos es 0.3324
>
> Aparece en 1 documento
>
> 11 El descriptor es "ACCESORIOS Y COMPONENTES "(Posición en la lista total de términos 11)
>
> Sus raíces son acceso y compon
>
> La suma de sus pesos es 1.2494
>
> Aparece en 3 documentos

Sobre este léxico se hicieron dos filtros, el primero utilizando técnicas de *Dimensionality Reduction* desarrolladas por Yang y Pedersen (1997), que en nuestro caso no tenían como objetivo reducir la carga computacional sino eliminar los candidatos a descriptor con menos capacidad para establecer relaciones, así como los que simplemente aparecían con poca frecuencia en la colección. Para ello se utilizó la técnica de reducción de la dimensionalidad, *Selection of termsbyfrequency of occurrence in documents* (Yang & Pedersen, 1997). Esta técnica rechaza los términos que no aparezcan en un número determinado de

documentos que por lo general se sitúa entre 1 y 5; para este proyecto se eliminaron los términos que aparecían una única vez en la colección, y se constató que el impacto que esto tenía sobre la capacidad de recuperación de los candidatos elegidos era nulo. No se perdía exhaustividad ni precisión con el cambio, y el resto de las técnicas a aplicar sobre la colección tampoco se resentirían, ya que se procedió a excluir aquellos términos sobre los que no se podía establecer comparación positiva alguna. El conjunto de descriptores eliminados, en torno a un 40%, tenían una capacidad de recuperación muy limitada, y por tanto resultaban candidatos a descriptor irrelevantes para el Tesauro.

En un segundo filtro se seleccionó manualmente la terminología, ya que muchos de los términos propuestos parecían adecuados para un sistema de indización automático, pero no se adaptaban a la morfología de los descriptores recomendada por las normas *International Organization for Standardization* (ISO) 2788/1986: *guidelines for the establishment and development of monolingual thesauri* y *American National Standards Institute/National Information Standards Organization* (ANSI/NISO) Z39.19/2005: *guidelines for the construction, format, and management of monolingual controlled vocabularies* (International Organization for Standardization, 1986; American National Standards, 2005). Por ejemplo: Las siglas: *AAPEX, AIW;* términos tomados directamente del inglés: *account, bakeryequipment;* o adjetivos: *alimentaria etc.*

Concluido este proceso, se generó un léxico de 4.500 palabras que se completó con términos extraídos de otros lenguajes documentales, diccionarios, glosarios y de las siguientes fuentes:

- Tabla de Aranceles de la Unión Europea (Taric, 2013).

- Tesauro Instituto de Información y Documentación en Ciencias Sociales y Humanidades (ISOC) de Economía (Centro de Información y Documentación Científica, 1995).

- Tesauro Spines (Centro de Información y Documentación Científica, 1988).

- Tesauro Eurovoc de la Unión Europea (Parlamento Europeo, 1987).

- Tabla de Sectores del ICEX (Material sin publicar).

- Thesaurus del Centre Français du Comerse Exterieur (Material sin publicar).

- Diccionario de Comercio Internacional (Hernández Muñoz, 2002).

- Tesauro de términos de Comercio Internacional, del Centro de Comercio Internacional (Organización Mundial del Comercio, 2004).

Este último repertorio merece mención especial, así como la adaptación del mismo, facilitada por el Instituto Español de Comercio Exterior, por la valiosa información que aportaron para cubrir, de forma particular, las áreas relacionadas con el desarrollo del comercio y comercialización de las exportaciones; así como de los productos y promoción del comercio.

Para simplificar la organización de la terminología en campos semánticos, se utilizaron técnicas de *clustering,* para generalizar características comunes en un conjunto de documentos, a partir de elementos agrupados en torno a su grado de similitud. Por este procedimiento se localizaron los temas fundamentales de la colección y se consiguió información de interés acerca de la estructura semántica de la misma. Al tratarse de una técnica no supervisada, implica que no sólo no depende del conocimiento previo del tema, sino que aporta más datos sobre la estructura de la colección de los que ya se conocían. Además de estas técnicas, se emplearon otras supervisadas de clasificación automática, en las que se manejaron los datos obtenidos durante esta fase.

De la multitud de algoritmos de *clustering* existentes, se utilizó en un principio *clustering* plano frente al *clustering* jerárquico, ya que podía revelar más información acerca de la estructura de la colección. Se estudió posteriormente la posibilidad de trabajar con algoritmos de *clustering* duro o blando, teniendo en cuenta que el *clustering* duro realiza una única asignación para cada objeto, mientras que el *clustering* blando admite la posibilidad de que la pertenencia de un elemento a un *cluster* pueda ser graduada de alguna manera, y es más típico de agrupaciones planas que de agrupaciones jerárquicas. Sin embargo estos dos tipos comparten la noción de que un determinado objeto sólo debería pertenecer a un grupo, aunque el *clustering* blando admite cierto grado de incertidumbre a la hora

de decidir cuál es el *cluster* más adecuado. Por otra parte, el *clustering* blando asume que un mismo objeto puede pertenecer a varios grupos al mismo tiempo. Se optó por el *clustering* duro, ya que los resultados que presenta son más fácilmente interpretables.

Tras varias pruebas preliminares con un Algoritmo Jerárquico Aglomerativo basado en enlace sencillo, se podían considerar cada uno de los elementos a agrupar, es decir, los documentos o los términos, como un *cluster* independiente. Se generó una jerarquía de *clusters* de términos para la colección, sin embargo se planteó un problema a la hora de interpretar los resultados. Los algoritmos de *clustering* jerárquico describen una jerarquía de *clusters* de *t-1 clusters*, siendo *t* el número de términos total de la colección; esto significa que la jerarquía de *clusters* resultante representaba todas las relaciones jerárquicas posibles, tantas como *t-1* términos tenga la colección, lo que la hacía difícil de interpretar. Por este motivo no proporcionó una información útil, aunque es una posible alternativa sobre la que investigar en el futuro.

Algoritmos no jerárquicos

De la diversidad de algoritmos de agrupación planos o no jerárquicos existentes, que describen Rijsbergen (1979), Frakes y Baeza Yates (1992) y, Manning y Schütze (2002), una de las mejores opciones desde el punto de vista de la calidad de las agrupaciones resultantes es el algoritmo de *clustering* mediante *k-vecinos* o *k-medias*. Se trata de un algoritmo de *clustering* duro no jerárquico que sólo requiere de la asignación del número de *clusters* deseado. Al margen de este aspecto, es un algoritmo no supervisado, aunque se pueden mejorar sus resultados aportando cierto grado de información en forma de semillas, o núcleos previos que sirven como base para el desarrollo de las siguientes fases. Las semillas pueden ser elegidas de forma aleatoria y en el número deseado; y cada una de ellas pasa a formar un *cluster* en una primera instancia, cuyo centroide (el vector cuyos valores suponen la media del conjunto de los miembros del *cluster*) es exactamente el vector que representa la semilla.

Se aplicó este procedimiento, llevándose a cabo sucesivas pasadas en las que cada uno de los términos se asignó al *cluster* con el que guardaba mayor grado de similitud, lo que se obtuvo mediante la aplicación de una medida de similitud *Jaccard,* que aportó mejores resultados que los procedimientos del *coseno* o de *Dice*. Después de cada pasada se calculó de nuevo el centroide de los *clusters* resultantes, con objeto de actualizar la representación de las agrupaciones conforme a los nuevos miembros, y se realizaron tantas pasadas como fueron necesarias hasta que el algoritmo convergió, o hasta que se determinó que los cambios en la distribución de agrupaciones entre pasada y pasada no eran ya relevantes, lo que solía ocurrir después de 10 ó 12 vueltas. El resultado de estas operaciones fue un conjunto de agrupaciones distribuidas sobre el espacio terminológico que se identificaron con los conjuntos terminológicos más destacados dentro de la colección. Estas agrupaciones proporcionaron el fundamento de los campos temáticos del Tesauro.

Asignación de descriptores a campos temáticos

Para realizar la asignación de descriptores a campos temáticos se aplicaron técnicas de clasificación automática, que permiten utilizar la información previa sobre la terminología en forma de datos de entrenamiento, para proceder a la asignación automática de los descriptores a los campos temáticos. Se eligió una técnica basada en el algoritmo de *k-vecinos* Yang (1994). Para realizar la clasificación de los términos de la colección era preciso contar, en primer lugar, con las categorías en las que iban a introducirse los descriptores, las cuales eran, evidentemente, los campos temáticos, así como ejemplos representativos de dichos campos.

Con los descriptores de entrenamiento existía ya un conjunto de campos temáticos con descriptores modelo que iban a servir como datos de aprendizaje automático para el algoritmo de *k-vecinos*. A continuación se midió la similitud entre cada descriptor y cada uno de los descriptores de entrenamiento, y se tomaron los *k* descriptores de entrenamiento más similares al documento (en nuestro caso 10). Posteriormente se generó un listado en orden decreciente de similitud al descriptor con los descriptores de entrenamiento, se sumaron las similitudes de los descriptores de entrenamiento que pertenecían a una misma categoría,

y la categoría con mayor similitud agregada se designó como aquella a la que pertenecía el descriptor (en este caso, un campo temático). El proceso fue rápido, y la clasificación mediante *k-vecinos* tuvo un grado de precisión y exhaustividad interpolada de hasta el 96%.

La aplicación de la clasificación automática evitó tener que asignar manualmente un elevado número de descriptores, llevándose a cabo la tarea, además, en pocos minutos, y virtualmente sin fallos.

Establecimiento de relaciones semánticas

Para el establecimiento de relaciones semánticas entre descriptores se adoptó un enfoque en el que el sistema hallaba la similitud entre un término y el resto de los términos de la colección y presentaba los 10 descriptores más similares al investigador.

Al proceder a la aplicación de la fórmula se localizaron relaciones entremezcladas: preferenciales, jerárquicas y asociativas, ya que no es posible, *a priori*, discriminar el tipo de relación resultante mediante este sistema. Lo que sí se consiguió con éxito fue integrar la información que proporcionaba el sistema en la aplicación de gestión de Tesauros TemaTres (Ferreyra, 2009), que se menciona más adelante, de forma que cada descriptor presentaba un conjunto de sugerencias para establecer relaciones de la siguiente forma:

productos químicos*0.24059024818974784

fosfato*0.21161536786283966

sector del calzado*0.1200831737733705

Este listado ordenado por grado de similitud al descriptor, en este caso *caucho*, le proporciona al experto la posibilidad de apreciar relaciones que en un principio no había considerado, anotarlas, y corregir la información para su posterior uso. Como ya se indicaba anteriormente, el carácter de las relaciones que se pueden establecer con los descriptores sugeridos va más allá de las posibilidades del sistema, pero no se descarta perfeccionarlo en el futuro.

Resultados

Como resultado del trabajo, se consiguió una solución específica para la gestión y actualización del Tesauro y su adaptación a las necesidades de compatibilidad con el *software* diseñado para la aplicación de técnicas automáticas. La selección automatizada de terminología, 1) ofreció una idea bastante aproximada del tipo de terminología utilizada en el ámbito, sin necesidad de leer abundante documentación acerca del área en cuestión, y sirviendo de orientación para las fases de recopilación de terminología posteriores; 2) permitió la selección de una buena parte del léxico de forma casi automática, lo que constituyó un considerable ahorro de tiempo. Por otra parte, la consiguiente alta tasa de adaptación a las necesidades reales de terminología para el centro repercutió en una mejor calidad del Tesauro.

La aplicación del *software* de gestión de Tesauros TemaTres, creado por Diego Ferreyra (Ferreyra, 2009) facilitó las tareas de gestión y edición del Tesauro. Esta herramienta, que sigue la norma *ISO 2788-1986* permitió el trabajo conjunto y coordinado del equipo, dado que se trata de una aplicación *Web*. TemaTres utiliza MySQL para almacenar los datos y el lenguaje de script PHP para realizar las consultas y recibir la información de los expertos a partir de formularios *Web*. Por sus características, este programa permitió su instalación en un servidor *Web*, y, mediante la autorización pertinente, la edición de todos los componentes del Tesauro. Además, al ser compatible con *Dublin Core*, SKOS-*Core* y *Zthes*, puesto que se trata de una aplicación distribuida bajo licencia *General Public License* (GPL), fue posible modificar el código fuente para adaptarlo a las necesidades de este proyecto, y generar el grado de compatibilidad necesaria con el sistema del ICEX, de forma que fue posible introducir de forma automática en la aplicación, tanto los candidatos a descriptor procedentes de la fase de extracción terminológica, como las relaciones existentes entre dichos descriptores y sus campos temáticos, así como las relaciones observadas por el sistema en la forma antes descrita. Igualmente se consiguieron modificar la base de datos y los formularios para conectarlos con nuestro sistema, ampliando la información disponible para cada descriptor, de cara al mantenimiento y actualización del Tesauro de acuerdo con la propuesta metodológica de Gil Urdiciain (2004), como se puede observar en la Figura 1.

La utilización de este *software* y la aplicación de las mencionadas técnicas materializaron en un Tesauro compuesto por 7.790 descriptores, provistos de

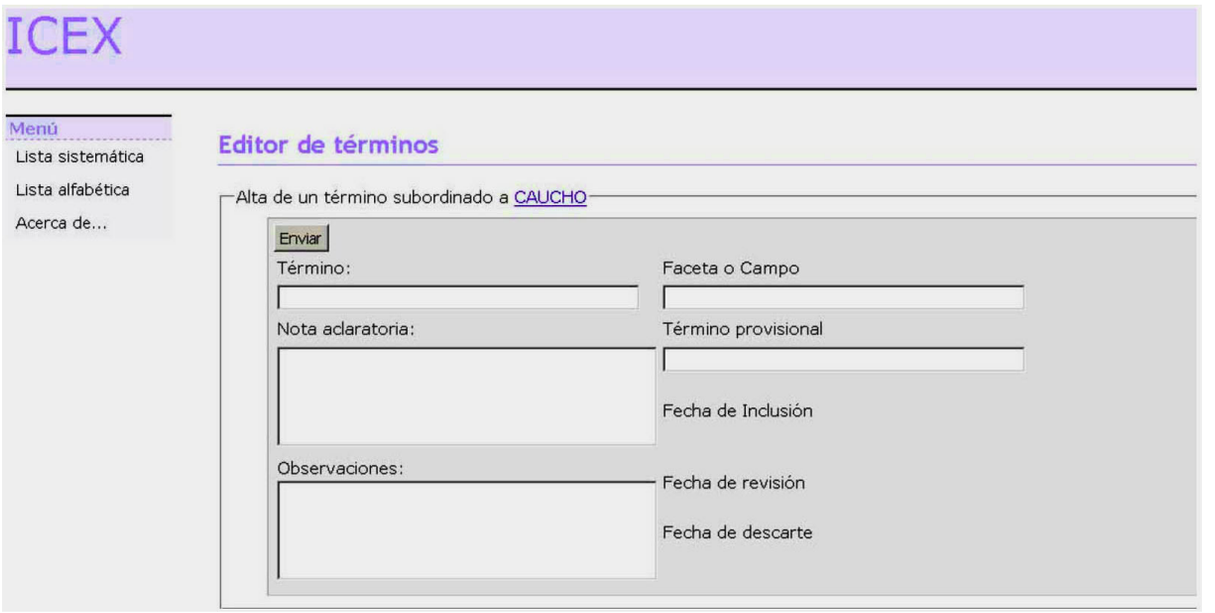

Figura 1. Ficha modificada de los descriptores.

Fuente: Modificación del *software* de Ferreyra (2009) de acuerdo con el sistema de Gil Urdiciain (2004, p.211).

relaciones de equivalencia, jerárquicas y asociativas, que se representan mediante un índice alfabético, un índice sistemático y un buscador, que sustituye al índice permutado. El Tesauro se completa con un índice de topónimos, elaborado teniendo en cuenta el código numérico de tres dígitos propuesto por la *Norma ISO 3166, Code for the representation of Names of Countries*, y en los códigos de dos dígitos utilizados por la *División de Estadística de Naciones Unidas* para la clasificación de los países en grupos económicos o geográficos.

Conclusiones

La indización automática de documentos de la colección, seleccionando los términos de mayor peso mediante la aplicación de la fórmula de Salton y Buckley (1988) TF-IDF, proporciona una terminología unívoca, claramente representativa de los contenidos de la base de datos.

La asociación de términos en *clusters* facilita la agrupación de descriptores en campos temáticos.

Hallar la similitud entre un término y el resto de los términos de la colección supone una gran ayuda para el establecimiento de relaciones.

Las técnicas utilizadas pueden suponer una mejora de la calidad del Tesauro resultante, en tanto que permiten analizar las condiciones de la colección para la que se utilizará dicho Tesauro, así como aportar a los expertos información difícilmente obtenible de forma manual. Esta constatación se sitúa fuera de nuestras ideas originales, en tanto que se esperaba conseguir una mejora en los aspectos relacionados con la eficiencia en el trabajo, mientras que se encontró que la aplicación de una metodología asistida puede ofrecer, además, mejoras en la calidad del Tesauro resultante, gracias a la aplicación de nuevas perspectivas sobre la misma tarea.

La aplicación de procesos automatizados en combinación con los procedimientos tradicionales para la recopilación de la terminología de un Tesauro, así como para la distribución de los descriptores en campos semánticos, reduce de forma considerable las tareas para su elaboración.

Sería muy útil integrar todo el *software* desarrollado en una única aplicación compatible con un sistema de gestión de Tesauros. Esta es, sin embargo, una tarea que implica mucho tiempo y que puede desarrollarse en futuros trabajos.

Referencias

Aitchison, J.; Gilchrist, A.; Bawden, D. *Thesaurus construction and use*: A practical manual. 4th ed. London: Aslib, 2007.

Ansi/Niso Z39.19. *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. Bethesda, Maryland: NISSO Press, 2005. Available from: http://www.niso.org/apps/group_public/download.php/6487/. Cited: Jan. 12, 2013.

Centro de Información y Documentación Cientifica. *Tesauro Isoc de Economía*. Madrid: IEDCYT, 1995. Disponible en: <http://thes.cindoc.csic.es/alfa_esp.php?thes=ECON&letra=A>. Acceso: 7 enero 2013.

Centro de Información y Documentación Cientifica. *Tesauro Spines*. Madrid: ICYT, 1988. Disponible en: <http://thes.cindoc.csic.es/index_SPIN_esp.php>. Acceso en: 7 enero 2013.

Crouch, C.J.; Yang, B. Experiments in automatic statistical thesaurus construction. *In*: International ACM/SIGIR Conference on Research and Development in Information Retrieval, 5., 1992, Copenhagen. *Proceedings*... Copenhagen: 1992. p.77-88.

Curran, J.R. *Automatic thesaurus extraction*. 2001. PhD (Thesis) - Edinburgh University, School of Informatics, 2001.

Chen, H. *et al*. Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science,* v.46, n.3, p.175-193, 1995.

Ferreyra, D. *TemaTres*: aplicación para la gestión de lenguajes documentales (versión 1.033) [Software]. R020.com.ar. 2009. Disponible en: <http://sourceforge.net/projects/tematres/>. Acceso en: 7 enero 2013.

Frakes, W.B.; Baeza-Yates, R. *Information retrieval*: Data structures and algorithms. London: Prentice Hall, 1992.

Gil Urdiciain, B. *Manual de lenguajes documentales*. Gijón: Trea, 2004.

Grefenstette, G. *Explorations in automatic thesaurus discovery*. Boston: Kluwer Academic Publishers, 1994.

Hernández Muñoz, L. *Diccionario de comercio internacional*. Madrid: Instituto Español de Comercio Exterior, 2002.

International Standard Organization. *Documentation 2788-1986*: Guidelines for the establishment and development of monolingual thesauri. Genève: ISO, 1986.

International Standard Organization. ISO *3166-1:2006*: codes for the representation of names of countries and their subdivisions - Part 1: Country codes. Genéve: ISO, 2006.

Manning, C.; Schütze, H. *Foundations of statistical language processing*. 2nd ed. Cambridge: The Mit Press, 2002.

Moreiro Gonzalez, J.A. *et al*. Generación automática de tesauros: propuesta de un método lingüístico-estadístico. *Ciencias de la Información*, v.30, n.4, p.139-147, 1999.

Organización Mundial del Comercio. *Tesauro de términos de comercio internacional*. Ginebra: Centro de Comercio Internacional, 2004.

Parlamento Europeo. *Tesauro Eurovoc*. Comisión de las Comunidades Europeas. Oficina de Publicaciones Oficiales. Luxembourg: Parlamento Europeo, 1987.

Pérez Agüera, J.R. *Generación automática de tesauros documentales*: trabajo para la obtención de Diploma de Estudios Avanzados (DEA) en Informática. Madrid: Universidad Complutense, 2005.

Rijdbergen, K. *Information retrieval*. 2nd ed. London: Butterworths, 1979.

Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* v.24, n.5, p.513-523, 1988.

Taric S.A. *Aranceles de la Unión Europea*: Arancel netTaric. Grupo TARIC, 2013. Disponible en: <http://www.taric.es/services/nettaric/nettaric.asp>. Acceso el: 20 enero 2013.

Yang, Y.; Pedersen, O.J. A comparative study on feature selection in text categorization. In: Internacional Conference on Machine Learnig, 14., 1997, San Francisco. *Proceedings*… San Francisco: Morgan Kaufmann Publishers, 1997. p.412-420.

Yang, Y. Expert network: Effective and efficient learning from human decisions in text categorisation and retrieval. In: ACM International Conference on Research and Development in Information Retrieval, 17., 1994, Dublin, Ireland. *Proceedings*... New York: Springer-Verlag, 1994. p.13-22.

Witten, I.H. *et al. KEA*: Practical automatic keyphrase extraction. Hamilton, New Zealand: University of Waikato, 1999.