Propuesta para la utilización de estructuras verbales aplicadas a la recuperación y representación de la información

Miguel Angel Marzal García-Quismondo
Pilar Beltrán Orenes
Jorge Luis Morato Lara
Juan Llorens Morillo
José Antonio Moreiro González
Sonia Sánchez Cuadrado
Universidad Carlos III Madrid

0.1. Resumen

La inclusión de categorías morfológicas distintas a los sustantivos en los tesauros es actualmente la línea de investigación más desarrollada para mejorar las deficiencias de este tipo de herramientas en la recuperación de información. Esta nueva concepción del tesauro está generando estructuras de tesauros más ricas, pero al mismo tiempo más complejas, lo que conduce a la necesidad de automatización de las mismas. Se hace una revisión de los beneficios de la inclusión de verbos como descriptores, y de las ventajas que presenta la construcción de tesauros de verbos. Entre estas ventajas destacamos que este tipo de tesauros de verbos facilita también la construcción automática de tesauros documentales clásicos.

Palabras clave: Recuperación de información. Tesauros. Verbos.

0.2. Abstract

The lack of efficiency in the application of thesauri in some retrieval environments is mainly due to some limitations in the thesaurus structure. Recently, a new standard, the ISO/IEC 13250:1999 about topic-maps, has been approved to overcome these restrictions. The estandard proposed to include verbs as descriptors and to identify relationships. One advantage of this approach consist in the possibility to employ verb descriptors as a tool that merges with classical ones improve them in the indexing tasks. In this document a bibliographic review is presented. This estate of the art is focused in the approach's advantages and disadvantages. Later, a methodology to develop a thesaurus of verbs is shown. In

our method, this verbal thesaurus will be employed to implement the standard about topic-maps automatically.

Keywords: Information retrieval. Thesauri. Verbs.

1. Introducción

Los tesauros son un tipo de lenguaje documental que tiene, como tal, dos funciones principales: representar un área de conocimiento y servir de herramienta de recuperación de los documentos de dciha área. Sin embargo, si bien es cierto que cumplen estas dos tareas como cualquier otro vocabulario controlado documental, tienen una consideración especial como herramientas de identificación documental (recuperación) debido a su alta coherencia en la indización, y a la posibilidad de expansión del vocabulario de las consultas empleando las relaciones entre los términos. En suma, los tesauros son, de entre los lenguajes documentales, los más aptos para la misión de la recuperación de los documentos.

La estructura de un tesauro contempla un conjunto de relaciones entre los términos, definidos ambos, relaciones y términos, *a priori* (Slype, 1991). Por lo tanto constituyen sistemas fijos o estáticos de descriptores que deben sufrir constantes actualizaciones, tanto del vocabulario como de las relaciones entre términos, para que su coherencia quede garantizada. Son precisamente estas actualizaciones constantes a las que se debe someter un tesauro la principal razón de su cuestionamiento como herramientas eficientes de cara a la recuperación. Esto es, aunque el índice de coherencia en la indización y establecimiento de relaciones de un tesauro sea muy alto, la constante evolución del lenguaje en el que se expresan los documentos que un tesauro describe conduce a la rápida obsolescencia de su definición inicial, provocada, a su vez, por la falta de pertinencia entre su vocabulario y relaciones respecto a los nuevos documentos.

2. Dificultades en la creación de tesauros

Tal como señalábamos más arriba, la estructura que define un tesauro: vocabulario y relaciones, definidas *a priori*, dan lugar a dos graves inconvenientes. El primero es común a todos ellos —es por tanto un problema histórico—, y es el relacionado con el elevado coste que conlleva, por un lado, su creación —debido a la necesidad de especialistas documentales y del área concreta a la que se extiende el tesauro, y el laborioso proceso de toma de decisiones para representar óptimamente el área—, y, por otro, su mantenimiento para evitar la obsolescencia, que sólo se podría sortear con la automatización o, al menos, con la semiautomatización. El segundo está relacionado con la dificultad de aplicación de una estructura como la de un tesauro a la descripción de colecciones documentales procedentes de campos con un nivel alto de abstracción (como el de la Informática),

o a documentos cuya traducción a descriptores conlleva necesariamente perdidas muy importantes de información (por ejemplo, la descripción de imágenes en movimiento) o a aquellos corpus que no tienen ni una estructura ni una temática bien definida, que son todos aquellos textos de libre formato y que cada día son más gracias a Internet.

3. Posibles mejoras en la creación y eficacia en la recuperación de los tesauros

El primer problema, como ya se comentó, podría ser solucionado mediante la construcción y mantenimiento automático (o semiautomático) de tesauros. Las distintas aproximaciones asumidas se pueden resumir en dos grupos:

1. *Estadística*: Entre los desarrollos basados en algoritmos estadísticos se encuentran los de agrupación en clases (*k-means*, por ejemplo), los de coocurrencia terminológica o las redes neuronales, como la de Kohonen (Díaz, 1998), pero sólo resultan eficaces en ámbitos muy concretos (Chen, 1995). Este intento de solución representa un gran avance en el camino hacia la automatización, pero obvia la realidad de los documentos al presentar una descripción algorítmica y, consecuentemente, muy simplificada de los mismos al despreciar el conocimiento lingüístico y contextual en el análisis.
2. *Lingüística*: Se concreta en la creación de sistemas de representación del conocimiento basados en el Procesamiento del Lenguaje Natural. En estos sistemas se lleva a cabo el análisis lingüístico (sintáctico-morfológico) de los documentos, y se representan gráficamente los esquemas que aparecen en las proposiciones que en él se recogen. Este campo se encuentra en constante expansión. Existe una mejora en aquellos estudios que han empleado recursos lingüísticos preexistentes, tipo WordNet (Harabagiu, 2000), utilizándolos como semilla para expandir las relaciones con el análisis de un corpus concreto (Iwanska, 2000). En la práctica, los sistemas lingüísticos tienen un componente estadístico elevado.

El segundo problema pasa por replantear la estructura clásica de un tesauro (ISO 2788), y readaptarlo a aquellas áreas o corpus documentales más problemáticos.

1. *Mejora del estándar clásico sobre tesauros*: Existe una abundante bibliografía en la que se propone la inclusión de más subtipos de relaciones en los tesauros de cara a la mejora de la recuperación (Tudhope, 2001). Es de destacar en este punto el estándar ISO:ICE 13250:1999 sobre *topic maps*. Este estándar surgió de los trabajos del grupo de investigación alemán liderado por H. Holger y S. Pepper (Holger, 1999). Los mapas conceptuales quedan definidos mediante un grupo de documentos

interrelacionados en un espacio multidimensional en el que los conceptos se interpretan como las localizaciones, y la medida de distancia entre dos conceptos la da el número de conceptos por los que es necesario viajar para llegar de uno a otro. En esta propuesta queda abierta la posibilidad de incluir los verbos para la recuperación, haciendo una interpretación conceptual de ellos, es decir, admitiendo que sobre ellos también recae parte de la información conceptual de la proposición.

2. *La aproximación documental*. Desde la que se complementa la estructura clásica con otros aspectos como la inclusión en él de "vistas" diferentes (facetas o metadatos) que suponían la ampliación del tesauro (Maniez, 1993). Este intento de automatización se encuentra con dos inconvenientes: los *corpora* analizados pueden no encontrarse estructurados, y en la fase de recuperación desaparece el elemento intuitivo del que hacen uso especialmente los usuarios no expertos o no familiarizados con la herramienta. Actualmente, la estructuración documental y la inclusión de metadatos han tenido un gran impulso gracias a lenguajes como XML y RDF.
3. *La aproximación informática*: Se trasladan los metamodelos propios de la ingeniería del software a las estructuras de recuperación mediante tesauros (Rumbaugh, 1998). La ventaja que supone este enfoque es un alto nivel de abstracción para describir determinados dominios complejos. Tal como se apunta desde esta última perspectiva, parece que la solución no puede provenir de una sola de estas soluciones. Lo más adecuado parece entonces la integración de ideas de todos ellos o una adecuación entre determinado dominio y una solución concreta.

4. Propuesta

Tras un análisis de las distintas aproximaciones para crear tesauros automáticamente, se propone una metodología que, con un enfoque principalmente lingüístico, permita una representación compatible con cualquiera de los modelos de representación, bien sea facetados, metamodelos de software o *topic maps*. Lo que se propone es:

1. *Aumentar las clases de relaciones de un tesauro clásico*: El aumento de estas relaciones deberá ir guiado por el dominio concreto, pudiendo haber, eventualmente, tantas como verbos diferentes aparezcan en el corpus documental generador del tesauro. Denominaremos el producto como "tesauro basado en verbos".
2. *Crear un recurso lingüístico consistente en un "tesauro de verbos"*: Este recurso se diferencia de una clasificación de verbos (Levin, 1993) en su

objetivo, ya que este recurso deberá indicar que estructuras verbales se podrían relacionar con determinada relación de un tesauro; por ejemplo, "procede de" o "venir de", o por otro lado "se encuentra en" o "se localiza en" definen una relación de tipo asociativa de procedencia. Mediante las jerarquías del "tesauro de verbos" se podrán agrupar los verbos de las relaciones del "tesauro basado en verbos" bajo un mismo hiperónimo, reduciendo de este modo el número de tipos de relaciones posibles. El tesauro verbal permitirá navegar y recuperar entre descriptores de tipo verbal, pero su principal aplicación será la creación de tesauros con relaciones verbales más diversificadas. Análogamente, los verbos que definan una misma relación a determinado nivel pueden estar agrupadas en asociaciones análogas a los "synsets" de WordNet (Miller, 1995).

3. Mejorar la representación de los tesauros al estereotipar las relaciones entre descriptores mediante el análisis de los verbos con los que concurren dichos descriptores en determinado corpus documental.

Véase el siguiente ejemplo:

1. Texto del documento: "La patata vino de América"
2. Etiquetado: <vegetal:patata><venir de><lugar:América>
3. Consulta al "tesauro de verbos": <venir de>, <proceder de> + lugar = Asociación de procedencia de lugar.
4. Modelado en el "tesauro basado en verbos": Patata-Relación asociativa de procedencia de lugar-América.

En definitiva, el objetivo es que mediante el análisis léxico-sintáctico y de frecuencias de un corpus documental se pueda crear una representación del dominio documental. Aunque las funciones de los verbos en el tesauro pueden ser muchas, tal como ya hemos señalado, la principal es la identificación del rol de una asociación mediante un verbo. Lo cual multiplica el número de relaciones posibles, ampliando así la adaptabilidad a dominios concretos. Esta versatilidad conduciría a la indización automática flexible y, consecuentemente, mejoraría las posibilidades de representar en dominios concretos y aumentaría la precisión y eficacia en la recuperación.

4. Beneficios

1. *Coste y actualización*: A nuestro juicio, la creación de un tesauro de verbos que complemente a los tesauros estáticos puede ser la clave para mejorar su eficiencia actual, y sortear los problemas del coste de creación y actualización y la obsolescencia que, tal como señalábamos antes, son comunes a todos los tesauros al uso.

2. *Disminución de la ambigüedad y del ruido en la recuperación*: La consecuencia de tipificar más específicamente la relación entre dos descriptores conlleva dos resultados: una mayor definición del contexto de la relación definida por los descriptores, y, por tanto, una disminución de la ambigüedad, y, consecuentemente, del ruido.

5. Empleo de verbos en los lenguajes controlados: estudios previos

Salvo escasos ejemplos, como Tharp (1973) o Nakamura (1994), los verbos han sido excluidos tradicionalmente como descriptores de los vocabularios controlados. Verbos, conjunciones, artículos y adverbios, por ejemplo, eran sacrificados en beneficio de los sustantivos para, así, reducir los costes almacenamiento y simplificar el mantenimiento.

Por otra parte, las inexistencia de aplicaciones informáticas que posibilitaran la creación automática de tesauros, y la escasa capacidad de procesamiento y almacenamiento de los ordenadores hasta hace muy pocos años, hacía que no se plantease la automatización de los mismos de una forma efectiva. Existen antecedentes de estos intentos en los años 60 y 70 como son los indicadores de rol, mediante los cuales se facetaban y unían los descriptores a través de los verbos; el sistema ECJ (Engineers Joint Council), en el que en cada una de las facetas (causa, efecto, material utilizado, …) se unen los descriptores mediante verbos (Lancaster, 1998); o el Proyecto Cadena Lingüística, herramienta que se basa en el análisis del lenguaje natural y cuyo fin era la proyección de un texto en una estructura regular (Grisham, 1991). Pero ninguno de ellos tuvo mayor repercusión debido a la carestía, ininteligibilidad o falta de hardware adecuado para concretarlos. Sin embargo, y a pesar de que todos los intentos anteriores fracasaron, la inclusión de tesauros de verbos para su uso en la descripción de los documentos presenta una serie de ventajas y atractivos que los justifican plenamente en la actualidad. Por otra parte, la implementación del tipo herramientas que los incluyan se encuentra hoy en día con muchas menos trabas técnicas que hace unos años.

6. Aplicaciones

En principio, las aplicaciones que pueda tener el logro del presente proyecto parecen ser numerosas, pero cabe destacar:

- *Identificación automática del género documental*: Los trabajos de Swales (1990) sobre el análisis de género a principios de la pasada década mostraron como determinados verbos y flexiones verbales están fuertemente relacionadas con determinados géneros. Análogamente, Losee (1996) identificó pautas similares para las estructuras documentales.

- *Aplicación a la descripción de materiales especiales*: Entre las ventajas más destacables se encuentran la versatilidad para indizar materiales especiales como pueden ser las imágenes, cada día más presentes en los corpora (medios de comunicación audiovisuales, Internet, …), que se ven mejor representados por un verbo (acción) que por un sustantivo, que no deja de ser un concepto estático. Además, al tener mayor riqueza descriptiva se favorece la representación de documentos de campos con un nivel de abstracción alto.
- *Mejora de tesauros "de sustantivos" existentes*: La indización doble de un tesauro estático de corte clásico, esto es, con tesauros de sustantivos y el tesauro de verbos, permitirá previsiblemente la reformulación de las relaciones entre sustantivos del primero con los verbos del segundo.

7. Desarrollos futuros

Se están estudiando los siguientes desarrollos para el proyecto:

- *Inclusión de otras categorías morfológicas*: Actualmente se está barajando la posibilidad de incluir categorías morfológicas diferentes a los verbos y sustantivos, como puedan ser adjetivos o adverbios.
- *Identificación de complementos circunstanciales*: Frecuentemente, la adecuada representación de los conceptos de un tesauro no se puede inferir de los verbos y sustantivos que aparecen en determinada frase, sino de la identificación de aquellas partículas como las preposiciones que nos permiten caracterizar adecuadamente los argumentos de la oración.
- *Identificación automática de géneros y de estructuras documentales*: Como ya se indicó en el apartado de aplicaciones, la identificación del género de una frase podría ayudar significativamente a mejorar la recuperación documental.

Agradecimientos

La presente investigación ha sido financiada dentro del proyecto TIC2000-0383 dentro del Plan Nacional de I+D+I (2000-2003).

Bibliografía

Chen, H.; Yim, T.: Fye, D. ; Schatz, B. (1995). Automatic Thesaurus Generation for an Electronic Community System. // Journal of the American Society for Information Science. 46:3 (1995) 348-369.

Díaz, I.; Velasco, M.; Lloréns, J.; Martinez, V. (1998). Semi-Automatic Construction of Thesaurus Applying Domain Analysis Techniques. // International Forum on Information and Documentation. 23:2 (Mayo-98) 11-19.

Grishman, R. (1986). Computational Linguistics, Cambridge University Press, Cambridge, 1986. Versión castellana de Antonio Moreno Sandoval: Introducción a la lingüística computacional. Visor Distribuciones Madrid, 1991. (Colección Lingüística y Conocimiento).

Harabagiu, Sanda M. ; Moldovan, Dan I. (2000). Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text. // Natural Language Processing and Knowledge Representation. Menlo Park (CA) : AAAI Press ; MIT Press, 2000. 301-333.

Holger Rath, Hans ; Pepper, Steve (1999). Topic Maps at work. // The XML Handbook. 2nd Edition. New Jersey : Prentice Hall, 1999.

ISO /IEC JTC 1/SC34. Information Technology: Document Description and Processing Languages. URL: http://www.topicmaps.com/content/resources\iso13250/iso13250-1999-fcd.htm.

ISO 2788. Guidelines for the establishment and development of monolingual thesauri: international standard ISO 2788, ISO. 2nd ed. [Geneve]: ISO, 1986.11-15.

Lancaster F. W. (1998). Indexing and abstracting in theory and practice. 2nd ed. Library London : Association Publishing, 1998. 182-189.

Levin, B. (1993). English verb classes and Alternations: a preliminary investigation. Chicago (Ill.): University of Chicago Press, 1993.

Loose, R.M. (1996). Text windows and phrases differing by discipline, location in document, and syntactic structure. // Inform. Processing & Manag. 32:6 (1996) 747-767.

Maniez, Jacques (1993). Los lenguajes documentales y de clasificación: concepción y utilización en los sistemas documentales. Madrid, Salamanca: Fundación Sánchez Ruipérez, Pirámide, 1993

Miller, G. A. (1995). WordNet: A lexical database. // Communication of the ACM. 38:11 (1995) 39-41

Nakamura, Yukio (1994). A Language for Knowledge Representation. // Advances in Knowledge Organization. 4 (1994) 127-133.

Rumbaugh, James (1998). Modelado y diseño orientado a objetos: Metodología OMT. Madrid: Prentice Hall, 1998.

Slype, Georges Van (1991). Los lenguajes de indización: concepción, construcción y utilización de los sistemas documentales. Madrid: Fundación Germán Sánchez Ruipérez, 1991.

Swales, J. M. (1990). Genre analysis: English in academic and research settings. Cambridge [UK]: Cambridge University Press. 1990.

Tharp, Alan L. (1973). Using verbs to automatically determine text descriptors. // Inform. Stor. Retr. 9 (1973) 243-248.

Tudhope, Douglas; Alani, Harith and Jones, Christopher (2001) Augmenting Thesaurus Relationships: Possibilities for Retrieval. // Journal of Digital Information. 1:8 (2001). URL: http://jodi.ucs.soton.ac.uk.