RECONOCEDOR-ASIGNADOR PARA SEMANTIZACIÓN EN HIPERTEXTO (RA)

Autores: Mela Bosch
Héctor Thompson
Universidad Nacional de La Plata- Argentina
Facultad de Periodismo y Comunicación Social
sibila@perio.unlp.edu.ar

Resumen: Se propone, utilizando los principios básicos de una Lógica Descriptiva o Terminológica, el diseño de un instrumento interactivo para construir esquemas conceptuales y detectarlos en documentos etiquetados con lenguajes de marcado. Se plantea un modelo de construcción de esquemas conceptuales utilizando estructura de Tesauro y etiquetas XML. Este proyecto presenta dos niveles: uno de análisis, diseño, e implementacíón en el cual se utiliza el paradigma de Orientación a Objetos. El otro nivel es el estudio de las posibilidades de interacción de usuarios expertos con documentos hipertextuales con el objetivo de normalizar la terminología, en este caso, en el ámbito de Periodismo y Comunicación Social

Palabras claves: Análisis y Diseño Orientado a Objetos; documentación; estructuras conceptuales, lógica descriptiva; Periodismo y Comunicación Social; recuperación de información; representación del conocimiento; terminología; tesauros.

Abstract: We intend, using the basic principles of Descriptive or Terminological Logics, to outline an interactive instrument that could be used to build conceptual structures and detect them in documents with markup languages tags. The model for the construction of conceptual structures uses a thesaurus structure and XML tags. This project has two levels: the first is that of analysis, design and implementation using the Object Oriented paradigm. The second is the study of the possibilities of interaction between expert users and hypertext documents in order to standarize the terminology –for this particular project– in the field of Journalism and Social Communication
.

Keywords: Conceptual structures; description logic; documentation; information retrieval; Journalism and Social Communication; knowledge representation; Object Oriented analysis and design; terminology; thesauri.

El problema

La extracción de conocimiento de masas textuales recorrió un camino abarca desde la detección exacta por medio de *pattern matching* hasta el desarrollo de operadores inteligentes, y llega a los actuales motores de minería de datos. En todos estos casos el enfoque está puesto en la masa textual. (Leloup, 1998).

Otra tendencia, basada en los estudios cognitivos, es enfocar el

problema desde la perspectiva de cómo el conocimiento es estructurado por los usuarios para elaborar estrategias de búsqueda. Se trabaja en el otro extremo de la interacción entre personas y documentos: en la manera en que se expresa lo que se desea detectar y qué conjunto de términos y cuáles relaciones entre ellos son las que mejor lo representan.
Observar esta interacción ofrece la posibilidad de comparar los esquemas conceptuales planteados por el usuario y los que aparecen en los documentos. Así es posible crear un ámbito de trabajo valioso para la normalización terminológica y que es la vez útil para la generación de lenguajes controlados. (Bosch, 1998).
Debimos considerar, además, la inserción de nuestro trabajo en el panorama del la representación del conocimiento en los sistemas informáticos.

Marco de este trabajo

Debimos considerar, además, la inserción de nuestro trabajo en el panorama del la representación del conocimiento en los sistemas informáticos. Diferenciamos en este campo, entre la representación declarativa y la representación procedimental del conocimiento. (Hatton, 1986).
En la representación procedimental el conocimiento está integrado en el programa que lo explota. Su especificidad es la clave de su eficacia. En su contra obran la falta de versatilidad y la dificultad de modificarse de acuerdo con los cambios que se producen, ya que debe establecerse *a priori* cada paso del proceso.

La representación declarativa permite un almacenamiento de conocimiento de manera modular independiente de los pasos de la utilización ulterior.

En cada una de estas modalidades la representación del conocimiento en el sistema informático se asocia a una estructura de datos. En las formas procedimentales es de tipo algorítmico, pero en las declarativas se requiere que los datos operen con determinados formalismos: tenemos así las representaciones lógicas. Aquí los conocimientos están representados por medio de fórmulas lógicas construidas con el apoyo de los operadores clásicos de la lógica formal. Son simples y concisas y con fundamento matemático consistente, pero se adaptan con dificultad al modelado del conocimiento de un experto humano. De manera que su rigurosidad es útil sólo en algunos dominios.

Se están desarrollando otros sistemas lógicos más ricos, entre éstos encontramos la lógica descriptiva denominada también lógica terminológica. Es una lógica ideada para representar, organizar y manipular conocimiento de un dominio particular de aplicación, y se expresa por medio de una terminología. (Meghini; Straccia, 1996)

Lo interesante de esta lógica es que permite utilizar los avances de otras formas de representaciones tales como las redes semánticas, surgidas ligadas al tratamiento del lenguaje natural, su ventaja está en la facilidad de comprensión y en que resuelven el problema de la herencia de propiedades por parte de un concepto con varios nodos vinculados. Funcionan bien en dominios acotados.

Justamente para dominios acotados existe otra forma declarativa muy

eficaz que son las reglas de producción. También están muy ligadas a la lógica, son modulares, autónomas, y muy difíciles de abstraer ya que la parcelización del conocimiento en reglas puede conducir a una explosión de ellas. Finalmente tenemos las representaciones estructuradas, no presentan ningún formalismo en particular, sino que reúnen los que consideran utilizables. Estas representaciones abstraen objetos con sus atributos y comportamientos es decir, secuencias de eventos típicos de un dominio.

Para operar pueden utilizar cualquiera de los formalismos citados: redes semánticas, reglas de producción, representaciones lógicas y procedimientos. De manera que encontramos una mezcla entre aspectos procedimentales y declarativos de acuerdo con las necesidades.

En este marco de representaciones estructuradas aparece el paradigma de Orientación a Objetos, en el que cada objeto posee su conocimiento y un conjunto de procedimientos, (métodos) que rigen su comportamiento. Adquieren así el nivel de abstracción y libertad respecto de los pasos de acción de la forma declarativa y encapsulan los procedimientos, lo que los hacen eficientes a la vez que permiten el reuso y la modificación.

Los lenguajes Orientados a Objetos, algunos puros como Smalltalk con su propio ambiente de objetos, hasta variados híbridos, ofrecen interesantes facilidades de implementación ya que unen íntimamente las estructuras de datos con los procedimientos encargados de manipularlos. Esta manipulación además cuenta con recursos de alto valor expresivo y operativo como la herencia, el polimorfismo y el enlace dinámico. (Martin; Odell, 1992).

El proceso más complicado parece ser el modelado en objetos del universo que se desea representar y cómo hacer que estos objetos optimicen sus relaciones. Las diferentes metodologías de análisis y diseño Orientado a Objetos han ido confluyendo en los últimos años en un lenguaje de modelado común cuya función es documentar, visualizar y llegar a un diseño, separando el análisis y el diseño de la implementación, este lenguaje es el Unified Modeling Language, UML. (Fowler, 1999)

Pero el lenguaje de modelización tiene sólo una función expresiva del modelo que un sistema diseñado según el paradigma de Orientación a Objetos realiza de una parte del mundo. El modelo está constituido de objetos individuales que están vinculados por medio de relaciones y agrupados en clases que capturan aspectos comunes a sus instancias. Aquí volvemos a la lógica necesaria para expresar estas relaciones.

Para ello, cuando se trabaja en ámbitos donde el aspecto terminológico es fundamental, puede usarse como apoyo la lógica descriptiva (Description logics)sobre la que hablamos, también conocida como lógica terminológica (Terminological logics) ya que opera en base a la diferencia entre las clases, usualmente llamadas conceptos, y definidas intensionalmente a través de las propiedades que los objetos deben satisfacer para pertenecer a determinado concepto o clase. Estas descripciones usan términos de algún lenguaje natural e incluyen restricciones y relaciones (a menudo llamadas roles) que rigen la conexión entre objetos. Experiencias recientes demuestran que es posible su aplicación a la recuperación de información en Web. (Barreiro; Losada; Ramos, 1999). Aclaramos que en el estado de actual de nuestro trabajo no hemos avanzado en el estudio de esta lógica más allá de sus fundamentos.

Este es, en resumen, el marco de nuestro trabajo: Una representación estructurada de conocimiento utilizando el paradigma de Orientación a Objetos, operando con principios básicos de lógica terminológica. Para representar el modelado nos valemos del Lenguaje de Modelado Unificado (UML) y para la implementación del lenguaje Smalltalk en VisualWorks.

El usuario

El RA está dedicado a un determinado tipo de usuario que hemos denominado **interactuante** y que tiene las siguientes características:
Conocimiento del **dominio temático** de la información. i.e. docentes y estudiantes de una materia específica, comunicadores especializados, investigadores.

Conocimiento de las **estructuras textuales** de los documentos que habitualmente consulta. i.e. bibliografías, curriculas docentes, informes técnicos, artículos científicos, artículos periodísticos.
Hemos hecho esta distinción pues para los usuarios estándar existen muchas herramientas de manejo de XML, p.e. editores de los buscadores.

De manera que el usuario de nuestro sistema requiere un uso particular de las palabras, que es el uso terminológico. Se habla de términos y no de palabras, pues su funcionamiento y objetivos son diferentes. Las palabras corresponden al léxico común, y sus funciones pueden ser muy amplias. Las unidades de un léxico común son consideradas términos cuando tienen un valor referencial, una temática específica, un nivel de conocimiento especializado y una situación comunicativa formalizada. (Cabré, 1993)

Como ya indicamos, el usuario de léxico común posee un gran número de herramientas muy efectivas para la localización de palabras en Internet. Pero para las necesidades terminológicas o bien contamos con instrumentos muy especializados como los Bancos Terminológicos o bien con los buscadores comunes de Internet.

El ámbito de aplicación

La aplicación tiene un objetivo muy acotado: la construcción de esquemas conceptuales, por parte de docentes y estudiantes avanzados de Periodismo y Comunicación Social, y comparación de estos esquemas comparación con los que aparecen en los documentos generados en la actividad académica.

El trabajo se desarrolla como parte de una investigación de la Facultad de Periodismo y Comunicación Social de la Universidad Nacional de La Plata, Argentina, en la cual se estudian los imaginarios profesionales y cómo éstos están representados en la curricula y la documentación que se genera en el ámbito de la Facultad.

A partir de la interacción de los docentes y alumnos avanzados, con los documentos generados, y una explicitación de los esquemas conceptuales, se visualizan los resultados y éstos sirven de base para construir una terminología común de referencia para la producción documental institucional.
Hemos detectado que la inconsistencia terminológica en la ámbito de

Comunicación Social es considerable, pero sabemos que una normalización que no exprese los usos reales aporta poco a la comunicación y calidad del proceso de enseñanza aprendizaje y al acceso al conocimiento en general.

Metodología

Como ya indicamos, los usuarios de terminología especializada requieren una reflexión sobre las características y alcance conceptual en la formulación de una búsqueda y encontramos que es posible expresar ésta en una lógica descriptiva, la cual ve al mundo como un conjunto de conceptos u objetos. Complementariamente los *conceptos* denotan subconjuntos de individuos o instancias y los *roles* denotan relaciones entre las instancias.
Todo ello se expresa con un Lenguaje Conceptual. Los lenguajes conceptuales proveen términos constructores para edificar términos compuestos por variables libres que asocian conceptos y roles para definir nuevos conceptos y roles.

En nuestra aplicación los individuos son los Términos y los roles son: Focal, Genérico, Específico, Relacionado y Equivalente (en todos los casos en el contexto del Dominio que se explicita).
Como se ve, hemos utilizado la estructura conceptual de un tesauro. Por el momento sólo establecemos un limitado número de roles y no hemos planteado aún categorías. Estas irán surgiendo a partir del estudio aplicado. En ese sentido, confiamos en la potencia de expresiva de la estructura conceptual de un tesauro. (López Huertas, 1999).

El presente desarrollo es un modesto intento para lograr la mejor coincidencia entre la expresión del modelo del usuario y la actividad del sistema, ya que consideramos que el dominio temático de un usuario no sólo cubre el compendio del conocimiento sino también, tareas e intencionalidades. (López Huertas, 1997).

El Constructor de Temas o Asignador del RA, explicita el lenguaje conceptual. Permite construir la taxonomía compuesta de conceptos y roles y sus relaciones subsumidas.
Las construcciones previas se mantienen en la Bolsa de Temas que contiene el conjunto de Términos y relaciones entre ellos que dan lugar a una específica concepción del mundo de un determinado interactuante en una sesión de trabajo.

Lo llamamos Asignador, porque desde la perspectiva de la persona que interactúa con el sistema la tarea es asignar nombres, roles y etiquetas.
La parte del Reconocedor es un motor que utiliza lógica terminológica, e interpreta y devuelve los resultados que surgen al operar con los temas asignados y los documentos seleccionados.
Finalmente, nos hemos tomado la libertad de denominar a este proceso semantización para diferenciarlo de la recuperación de información, pues la base está en una tarea de retroalimentación y ajuste entre los esquemas conceptuales de las personas y los que aparecen en los documentos.

Modelado e implementación del RA

Realizamos el modelado del sistema utilizando UML, el cual, según ya mencionamos, es un lenguaje de modelado destinado a documentar, visualizar y llegar a un diseño separando el análisis y el diseño de la implementación, usando específicamente conceptos de Orientación Objetos.
Al ser un lenguaje tiene una función expresiva, no es una forma de proceso de desarrollo, sino que da elementos para representar un modelado
UML es claro y preciso además de ser un estándar aceptado y conocido. La primera versión estándar aparece en el 1997, la versión que usamos es de 1999.
UML presenta tres tipos que diagramas:
1. Diagramas de Casos de Uso
2. Diagramas de Estructura Estática
Que comprenden a su vez:
    Diagramas de Clases
    Diagramas de Paquetes
3. Diagramas de Comportamiento
Que comprenden a su vez:
    Diagramas de Interacción
    Diagramas de Estado
    Diagramas de Implementación

Hacemos notar que cada uno de estos diagramas no incorpora nuevos elementos sino que ofrecen una vista diferente del mismo modelo.
Presentamos el Diagrama de Casos de Uso y el Diagrama de Clases, en el estado actual de desarrollo de nuestro sistema.
En cuanto a la implementación, hemos desarrollado en VisualWorks la parte correspondiente a al Constructor de Temas que opera por asignación de términos roles, etiquetas y restricciones por parte del usuario. El Motor del RA que realiza la parte de Reconocimiento no está aún implementado

Diagrama de Casos de Uso RA

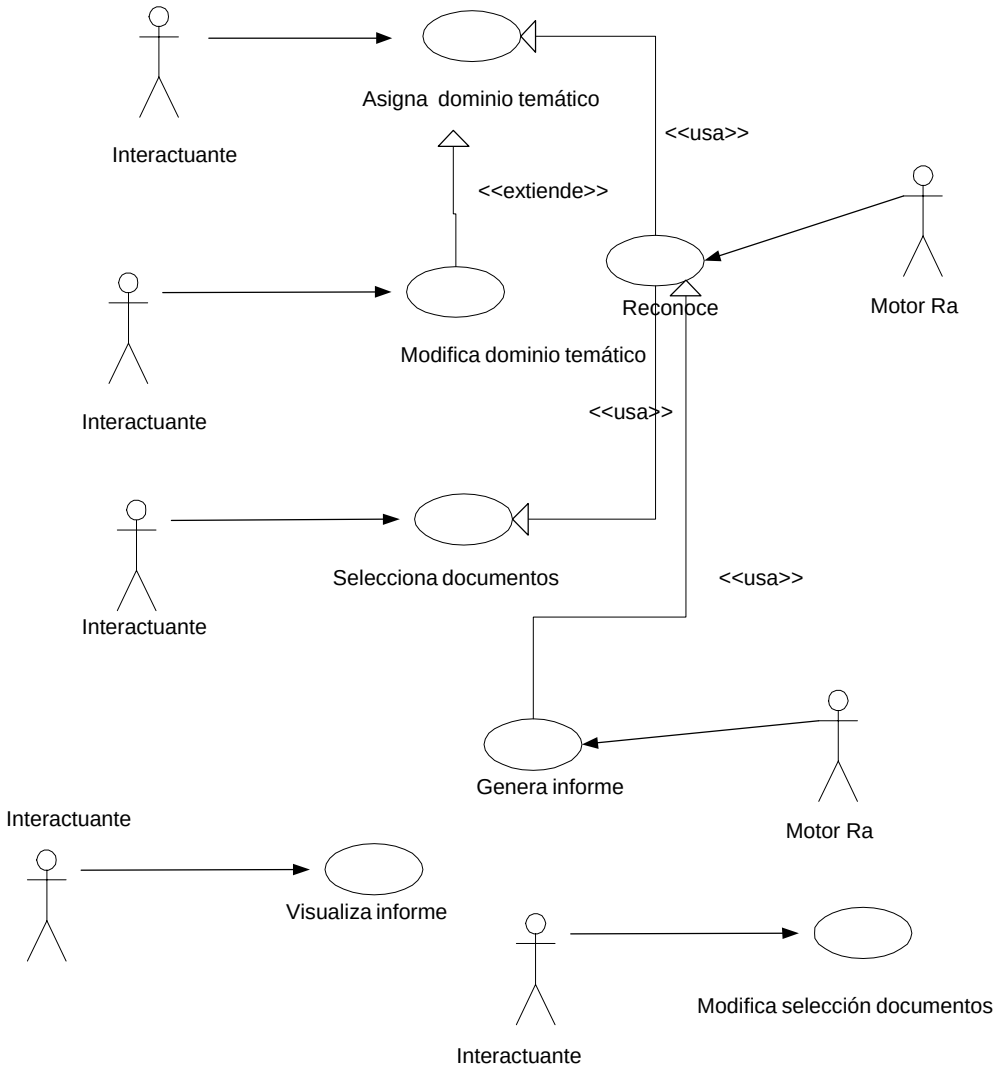
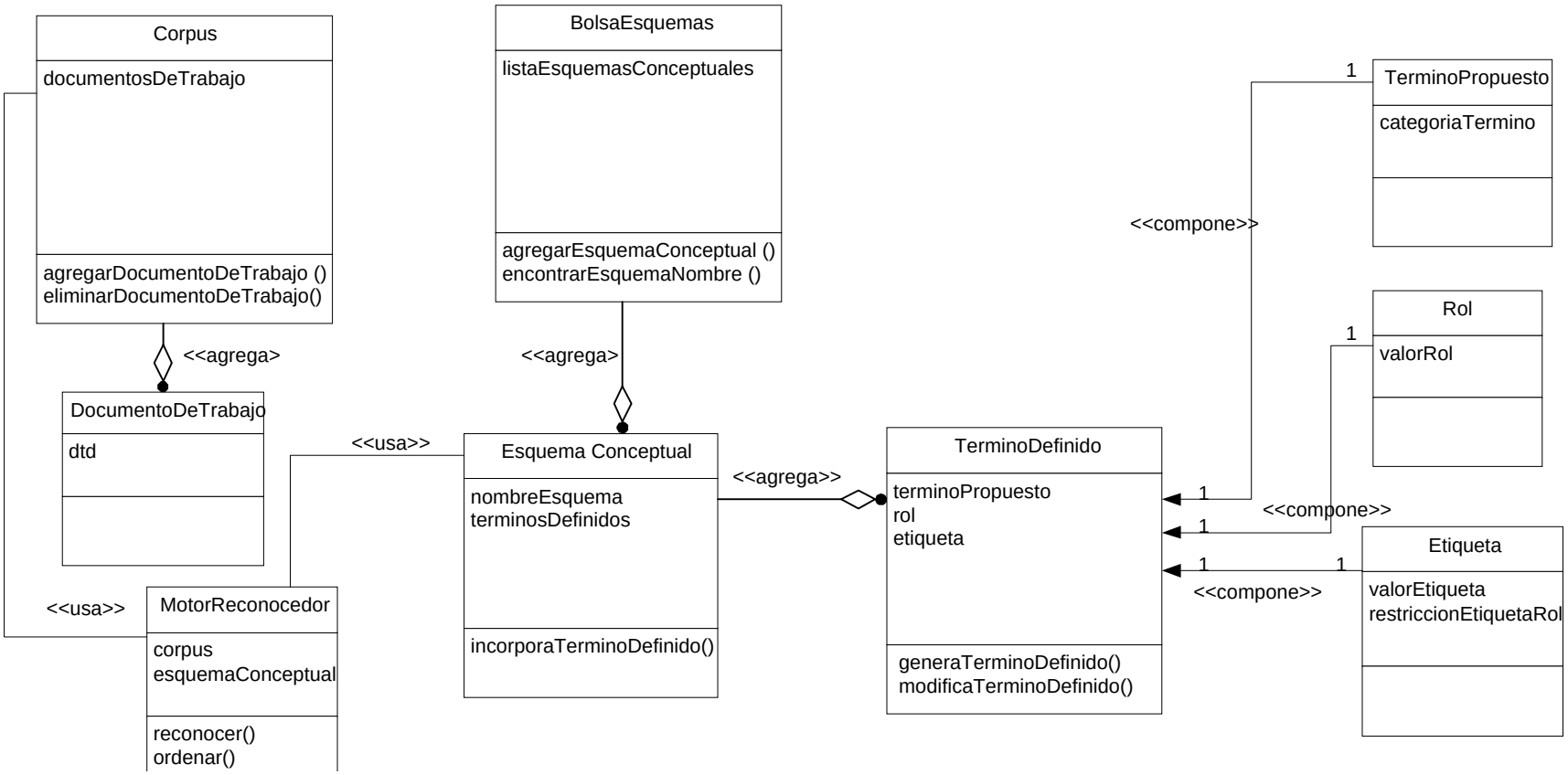


Diagrama de clases del RA



Bibliografía

Barreiro, A. ; Losada, D.; Ramos, R. ; TWebs: *An application of terminological logics in Web Searching*. *In: Actas Congreso 4, International Society for Knowledge Organization,* ISKO, Granada: 1999. p. 253-259.

Bosch, M. *El trabajo terminológico en la Ingeniería de Sistemas Informáticos.* Journal of Translation Studies. In: *Turjuman, Revue de traduction et interprétation.* Tanger: V. 7, n.2. Octubre 1998.

Fowler, M.; Scott, K. UML Gota a gota, México: Addison Wesley, 1999.

Cabré, M. T. La terminología, Teoría, metodología, aplicaciones. Barcelona: Antártida, 1993.

Hatton, J. P. (1986) Inteligence artificielle, panorama des techniques et domains d´applications. In: Le Moigne, J. Intelligence des mécanismes, mécanismes de l´intelligence. Paris: Fayard. p. 57-72.

López Huertas, María J. Potencialidad evolutiva del tesauro, hacia una base de conocimiento experto**.** In: *Actas Congreso 4, International Society for Knowledge Organization, ISKO*, Granada: 1999. p. 133-140.

Leloup, C. Moteurs d´indexation et de recherche. Paris: Eyrolles, 1998.

López Huertas, María J. Thesaurus Structure design: a conceptual approach for improved interaction. In: *Journal of Documentation.* V. 53, n.2. Marzo 1997. p. 139-177.

Martin, J, Odell, J. Object Oriented Analysis and Design, Prentice Hall. 1992.

Meghini, C ; Straccia, U. A Relevance Terminological Logic for Information Retrieval**.** In: *Proceedings of SIGIR-96, 19th International Conference on Research and Development in Information Retrieval*, Zurich: 1996. p.197--205