Lógica difusa aplicada a la web semántica

Juan Bernabé Moreno
Universidad de Granada, Departamento de Ciencias de la Computación e Inteligencia Artificial

Abstract— **Este trabajo se centra en demostrar el papel de la lógica difusa en las nuevas tecnologías que emergen en el contexto de la web semántica y más en concreto en el ámbito de la recuperación de información usando ontologías.**

***Index Terms*— Lógica difusa, web semántica, ontologías, recuperación de la información**

I. Introducción

UNo de los motivos que justifican que el Homo Sapient haya conseguido imponer su dominio en este planeta, es precisamente su capacidad de analizar y controlar basadas en técnicas difusas. Para ilustrar esta idea utilizaré un ejemplo de la vida diaria: supongamos que estamos en la autovía que une Granada con Sevilla un día 24 de diciembre, el límite de velocidad es 120 km/h. Para evitar accidentes o colapsos, es recomendable conducir "como los demás", pero cómo definir exactamente –con instrucciones unívocas y precisas- lo que es conducir "como los demás"? Los humanos lo hacen a diario y lo hacen francamente bien aplicando lo que se conoce como lógica difusa: reciben una serie de informaciones "difusas" y en sus cerebros empiezan a procesarlas, asignándoles pesos, aplicando medias, filtrados, etc para llegar a una decisión "óptima". En el cerebro se evalúan entradas como "Qué están haciendo los coches que tengo delante?", "A qué velocidad van?", "Hay camiones que impidan proceder?", etc
Esta misma capacidad de análisis ha sido clave desde el inicio de los tiempos para organizar la caza de mamuts, etc… Y seguirá jugando un papel importantísimo en todos los sistemas que intenten emular la inteligencia humana.

La web semántica representa el advenimiento de una nueva era para Internet. Hace ya más de una década Tim Berners-Lee [12] formuló la visión del movimiento de web semántica. En estos 10 años se ha avanzado mucho en la conciencia semántica, pero la aplicabilidad está empezando a traspasar las barreras del ámbito académico con paso firme hacia el terreno industrial.

Según la W3C [13], la web semántica trata de dar un significado bien definido a los datos para permitir que humanos y máquinas trabajen en cooperación. En este sentido, la web semántica extiende a la web actual (o sintáctica) en dos aspectos:

- La información viene expresada en un lenguaje preparado para ser "entendido" por una máquina – agente, etc-, frente al rango de lenguajes naturales para el consumo humano de la web actual
- Los datos están formalmente interconectados por medio de links, mientras que en la web actual los vínculos entre los datos carecen de este formalismo.

Este trabajo se centrará en una herramienta básica para la consecución de la web semántica, la ontología, y cómo la lógica difusa extiendo las ontologías "crisp" para mejorar la recuperación de la información basada en ontologías.

El primer capítulo introducirá la lógica difusa para pasar a definir formalmente la ontología y aclarar el papel que juega en la web semántica. Luego se revisará el problema concreto en la recuperación de la información por medio de ontologías difusas y su aplicabilidad. La sección final precisa los campos en los que las ontologías difusas se están siendo objeto de investigación y se presentará una valoración

II. Lógica Difusa

Las raíces de la lógica difusa hay que buscarlas en la antigua Grecia. El filósofo Platón ya indicó que entre los conceptos "verdadero" y "falso" existía otro tercer concepto (contradiciendo a su contemporáneo Aristóteles, que fundamentó la precisión de las matemáticas en el hecho de que una afirmación sólo podía ser verdadera o falsa)

La teoría de conjuntos difusos fue introducida por L.A. Zadeh, pero no fue hasta los 80 cuando encontró su apogeo, en concreto en Japón, en lo que se llamó la "ola difusa" –fuzzy wave-, que a Europa llegó en el terreno práctico a mitad de los 90.

La teoría de conjuntos difusos [14] hay que diferenciarla de la lógica polivalente, descrita en lo años 20 por Jan Lukasiewicz. En realidad presenta ciertas similitudes con la lógica polivalente en tanto que la veracidad de una afirmación se encuentra en la serie de números reales en el intervalo de 0 a 1, pero Zadeh enfatiza el matiz difuso de pertenencia de los objetos como elementos de un conjunto. En concreto mueve el foco del problema al ámbito lingüístico y a la problemática de los conceptos definidos de manera difusa.

Los fundamentos de la lógica difusa son los llamados conjuntos difusos. En contraste con los conjuntos tradicionales o crisp, en los que un elemento pertenece o no a ellos, puede un elemento en un conjunto difuso pertenecer en un cierto

grado a un conjunto difuso. El grado de pertenencia viene dado por la función de pertenencia μ, que asigna un número real –o valor de pertenencia- a los elementos del conjunto.

Estos conjuntos permiten también operaciones algebraicas normales, pero en el sentido difuso, como intersección, unión y complemento.

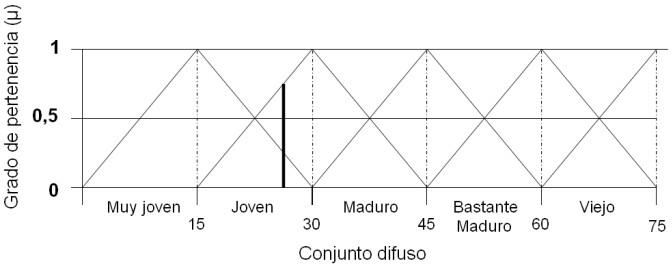


Figure 1

La Figure 1 muestra un ejemplo de función difusa de la edad de los hombres. Un hombre de 25 anos pertenecería al conjunto de gente "Joven" con un valor de 0,75 (relativamente alto), mientras que tendría un valor de pertenencia de 0,25 al conjunto de la gente "madura"(relativamente bajo). Hay diferentes tipos de función de pertenencia, no siempre tiene que ser triangular, aunque simplifica la computación de los valores de pertenencia.

En el ámbito de la recuperación de la información, la lógica difusa se ha venido aplicando en data bases extendiendo comando SQL [9][10]

Entre las ventajas que presenta, podemos resaltar un mapeado más eficiente de términos combinados en una consulta, como "de alguna manera relacionado con", o "fuertemente relacionado con" y nuevas posibilidades de tratar la relevancia de los documentos recuperados "cierta relevancia", "muy relevante",etc.

III. ONTOLOGIAS

Las ontologías se conciben como descripciones jerárquicas de un conjunto de conceptos, un conjunto de propiedades y sus relaciones y un conjunto de reglas de inferencia. El concepto de ontología es crucial en la web semántica. En este contexto, la web semántica consiste en una serie de bases de conocimiento distribuidas y una serie de agentes inteligentes que pueden leer y razonar sobre conocimiento público basándose en ontologías [2]

El objetivo de las ontologías es expresar formalmente una interpretación de la información [3]. Pese a su papel decisivo en la web semántica como instrumento para relacionar conocimiento almacenado en diferentes fuentes, la misma idea de ontología tiene un aspecto paradójico: a medida que aumenta la especialización de una ontología para una parte del dominio, decrece su capacidad comunicativa, porque la audiencia se vuelve más especializada. Este razonamiento llevado al límite, implica que una ontología que expresa perfectamente la interpretación de una persona particular sobre el mundo, es inservible para todas las demás personas con una visión levemente diferente.

En última instancia, las ontologías son una guía para expresar la relación entre los conceptos de un dominio, y el hecho de conocer esta relación simplifica la interacción entre los objetos, las operaciones que pueden aplicarse a esos objetos y el proceso de posicionar los nuevos objetos en la ontología.

Las ontologías representan un poderoso instrumento en la recuperación de información. Se pueden por ejemplo usar para asignar valores de relevancia a los documentos según las relaciones existentes entre los conceptos. Por ejemplo, una consulta preguntando por "manzanas" podría recuperar documentos en los que los términos "reineta" y "Granny Smith" aparecen, incluso si la palabra "manzana" no está presente (véase el ejemplo simple de ontología de manzanas dado en Figure 2)

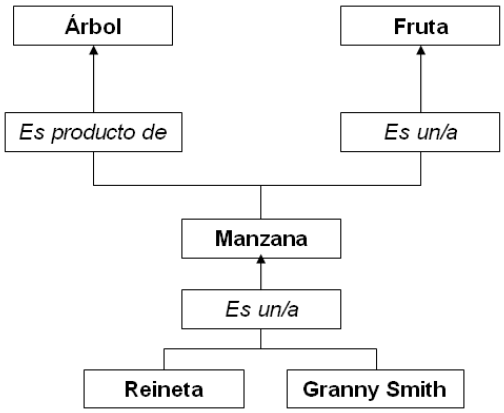


Figure 2

Las ontologías codifican una estructura jerárquica, donde las propiedades se heredan desde la raíz a las hojas, pudiendo darse atributos adicionales en cada nivel. La ventaja que presenta frente a una jerarquía fija de objetos es que las ontologías se pueden modificar incluyendo nuevas ramas, y precisamente la localización de un nuevo item puede dar información sobre él mismo a usuarios que no están al tanto de los nuevos términos. Además la variedad de relaciones permite refinar la representación del conocimiento.

En cierto modo las ontologías recuerdan a un diagrama de clases UML, pero mientras que la programación orientada a objetos se centra más bien en las operaciones (lo que hace la clase), las ontologías se centran más en la estructura que es reusable, interpretable unívocamente, de alcance limitado y traducible.

a) Representación de las ontologías

La organización WWW ha desarrollado un estándar para la representación de ontologías: OWL [6] (Web Ontology Language). En realidad, se trata de una familia de lenguajes que se basan principalmente en 2 semánticas (compatibles pero no al 100%): OWL DL y OWL Lite basadas en lógica descriptiva –con propiedades computacionales bien definidas- y OWL Full, compatible con el Schema RDF, más expresiva pero más complicada para la computación.
La semántica en OWL viene dada por la traducción a una lógica descriptiva particular. Así OWL es al mismo tiempo

una herramienta para describir ontologías y aportar compatibilidad, y un conjunto de semánticas definidas formalmente que otorgan significado. OWL DL se corresponde con $\mathcal{SROIQ}(D)$ logic.

Umberto Straccia propone en [5] una extensión de esta lógica para suportar el procesamiento difuso.

IV. Ontologias Difusas

La ontología difusa (FuzzyOnt) se basa en una modificación de la ontología normal o "crisp". Partiendo de las relaciones entre los términos de la ontología, se procede de manera incremental añadiendo valores de pertenencia a las relaciones existentes y en algunos casos añadiendo también nuevas relaciones. Otra característica importante es la normalización de los valores de pertenencia, de tal forma que la suma de los valores de pertenencia de cualquier término de la ontología es igual a 1. Esta normalización juega un papel importante en el mapeo de consultas a la ontología: para cada término de la consulta sólo una interpretación (un mapeo) es elegido y el resto de posibles mapeos son excluidos.

El proceso de fuzzificación se observa perfectamente en la Figure 3 [4]

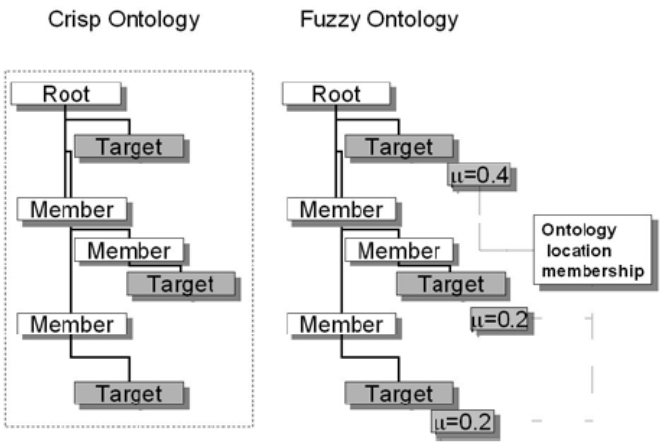


Figure 3

La ontología difusa se define formalmente [7] como una cuádrupla:

$$O_F = (C, R, F, A)$$

Donde:

C es un conjunto de conceptos difusos o Entidades E

R es un conjunto de relaciones difusas en E^n

$R = T \cup T_{not}$ donde T representa las relaciones taxonómicas y T_{not} las que no se derivan de la taxonomía de la ontología.

F es un conjunto de relaciones difusas, de tal manera que una f de F se define como

$$E^{n-1} \times P \longrightarrow [0,1]$$

Siendo P un conjunto de enteros, Strings, etc

A es un conjunto de Axiomas expresados en un lenguaje lógico apropiado (e.j.: predicados para restringir el significado de conceptos, relaciones, funciones)

David Parry propone 2 maneras de asignar los valores de pertenencia. A continuación se explica la manual:

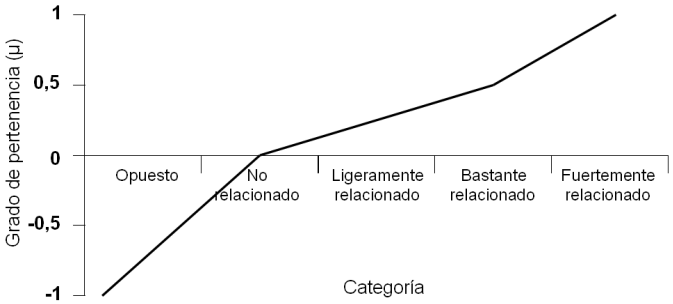


Figure 4

Una vez recuperado el conjunto de documentos en respuesta a una consulta lanzada por el usuario, los términos que representan a esos documentos han de asignarse a una de las categorías presentadas en la Figure 4, dependiendo de la relación que presentan con los términos empleados en la consulta. Así por ejemplo, si se ha empleado el término "Sierra" en la búsqueda con el significado de accidente orográfico, "Serrucho" sería asignado por el usuario en la categoría "Opuesto".

La ontología se es actualizada como se indica a continuación:

1) La localización en la ontología del término buscado se calcula. Sólo los documentos que reciben un valor de relevancia mayor que un umbral determinado se incluyen en el proceso. Los términos empleados en la consulta se comparan entonces con los términos que han sido asignados a las categorías de relación según su grado de pertenencia (como hemos indicado previamente)
2) Se calcula una puntuación para cualquier significado potencial de cada término de la consulta sumando los valores de pertenencia en las categorías que están relacionadas con cualquier localización potencial del término de la consulta
3) La localización con la puntuación más elevada es la se supone la pretendida por el usuario

El valor de pertenencia de una localización para un determinado término de la consulta, se calcula según la siguiente igualdad:

$$\mu_{total} = \frac{\sum_{0}^{i=n} \mu_i}{n}$$

Tomando μ_i como el valor de pertenencia de cada término que el usuario ha asignado a una de las categorías de pertenencia. Sólo los términos que tienen una relación de parentesco (padres o hijos) con el significado inducido para los términos de la consulta del usuario son incluidos en el cálculo. Si la ocurrencia de un término en un documento recuperado es mayor que 1, entonces cada instancia del término se incluye en el cálculo.

Si el valor de pertenencia obtenido es menor que 0, entonces éste se hace 0.

V. Recuperación de la información

Los sistemas de recuperación de información se pueden dividir a groso modo en los basados en el filtrado de contenido y aquéllos basados en el filtrado colaborativo.

Los primeros filtran y recomiendan –recuperan- información basándose en el matching de los términos usados para representar los documentos, con los términos que se formularon en la consulta

Los segundos usan las preferencias del usuario que lanza la consulta (bien explícita o implícitamente)

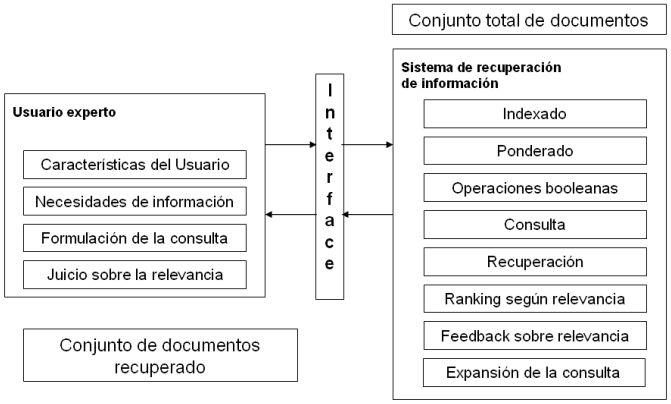


Figure 5

A. *Recuperación basada en redes de conceptos difusas*

La noción de ontología difusa se emplea para formar lo que en [8]se denomina base de conocimiento difusa (Fuzzy Knowledge Base), definida como

$KB := (O_F, I)$ donde O es la ontología difusa y I un conjunto de instancias asociadas.

Cualquier concepto c ϵ C es una conjunto difuso de I, por ejemplo

$$c : I \longrightarrow [0,1]$$

Las instancias pueden ser documentos, imágines digitales, notas, etc. Estas instancias se mantienen a parte de la ontología difusa, para habilitar que la misma ontología sirva para múltiples juegos de instancias.

La autora introduce las correlaciones semánticas como nuevas relaciones difusas entre los términos usados en una consulta o cuando un documento se inserta en la base de datos.

La idea es definir un conocimiento dinámico sobre un dominio que se adapta al contexto… Lo que representa un compromiso entre la definición de un objeto dada por la estructura de la ontología y el significado que el usuario le asigna –conocimiento basado en la experiencia que el usuario haya adquirido en ese dominio-.

El resultado de establecer las correlaciones semánticas entre las entidades de una ontología y dar pesos a esas correlaciones es lo que la Sánchez denomina una red de conceptos difusa (Fuzzy Concept Network):

$$N_f = (E, F, m)$$

Las aristas entre las entidades E son las correlaciones, definidas por

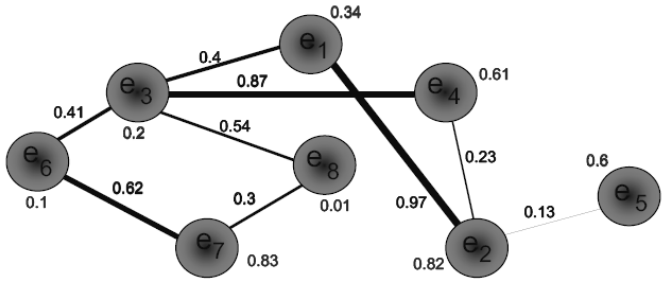
$$F : E \times E \longrightarrow [0,1]$$

Las entidades que no están co-relacionadas presentan un peso igual a 0.

Además, cada nodo o entidad se caracteriza por un valor de pertenencia que determina su importancia dentro de la ontología, definido por la función:

$$m : E \longrightarrow [0,1]$$

Por definición $F(e_i, e_i) = m(e_i)$



$F(e_3, e_8) = 0.54$ (*correlation*)

$m(e_1) = 0.34$

Figure 6

La red de conceptos difusa se puede aplicar a diferentes instancias de objetos descritos por esos conceptos para recuperar información. La estructura

$$N_{f_0} : (O_{DB}, N_f)$$

Donde O_{DB} es un conjunto de objetos almacenados en la base de datos, N_f la red de conceptos difusa, de tal modo que cada objeto o_i viene descrito por las entidades de la red:

$$o_i = \{e_1, ..., e_n\} \in E$$

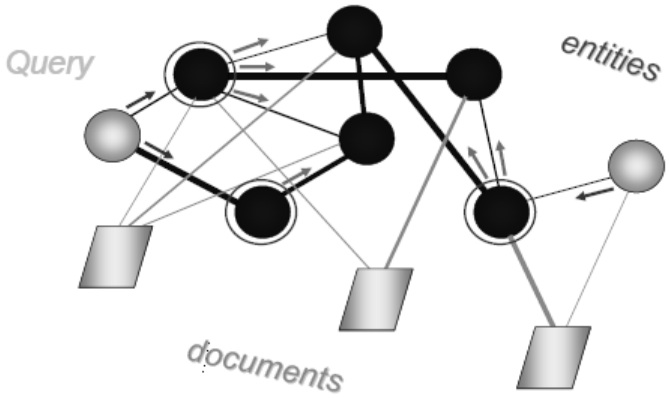


Figure 7

Un camino único se define para cada término empleado en la consulta en cada paso, correspondiendo al máximo valor de correlación.

B. *Midiendo la eficacia de recuperación en términos difusos*

Las mediciones clásicas de los sistemas de recuperación de información se pueden definir en términos difusos.

1) Precisión

Mide la proporción de los documentos recuperados que son relevantes

Ordinariamente se define como:

$$P = \frac{\left|R_{\text{Re}cuperados} \cap R_{\text{Re}levantes}\right|}{\left|R_{\text{Re}cuperados}\right|}$$

La misma fórmula se puede expresar para el dominio difuso de la siguiente manera:

$$P_F = \frac{\sum_{d \in Q_\theta} \mu_Q(d)}{\left|Q_\theta\right|}$$

Donde Q es el conjunto difuso de documentos obtenidos como resultado gradual de la consulta. Es necesario establecer un umbral θ para considerar un documento como recuperado:

$$Q_\theta = \{d \in D / \mu_Q(d) \geq \theta\} = R_{recuperados}(\theta)$$

2) Exhaustividad

Define la proporción de documentos relevantes que han sido recuperados

$$R = \frac{\left|R_{\text{Re}cuperados} \cap R_{\text{Re}levantes}\right|}{\left|R_{\text{Re}levantes}\right|}$$

Y desde el punto de vista difuso:

$$R_F = \frac{\sum_{d \in Q_\theta} \mu_Q(d)}{\sum_{d \in D} \mu_Q(d)}$$

Donde todos los documentos son relevantes, pero con un cierto grado de relevancia oscilando entre 0 y 1

C. Derivación de relaciones difusas

En [11] Widyantoro presenta un método para el refinado de consultas basado en una ontología difusa de término asociados. Dado un término empleado en la consulta, el sistema es capaz de expandir la consulta sugiriendo otros términos con significado más amplio o más

Si definimos C={a_1,…,a_n} como una colección de documentos donde cada documento viene representado por ciertos términos a = (t_1,…,t_n).

Si representamos la ocurrencia de un término ti en un artículo como *occur(t_j,a)* y su función de pertenencia se define como $\mu_{occur} = f(|tj|)$, donde la función f recibe como argumento la frecuencia de ocurrencia de un término t_j en un documento a.

Si usamos la notación RT (t_i,t_j) como que ti *es más restringido que* t_j, la función de pertenencia de RT(t_i,t_j) se definiría como:

$$\mu_{RT(t_i,t_j)} = \frac{\sum_{a \in C} \mu_{occur(t_i,a)} \otimes \mu_{occur(t_j,a)}}{\sum_{a \in C} \mu_{occur(t_i,a)}}$$

Donde el operador del denominador representa la intersección difusa. Lo que en realidad indica esta función, es la proporción entre el número de co-ocurrencias de los dos términos juntos, frente al número de ocurrencias de un término t_i. Mientras más co-ocurran los términos juntos y menos ocurra el término ti solo, *RT(t_i,t_j)* tendrá mayor grado de confianza.

Los casos extremos son 1, cuando siempre se da co-ocurrencia, o 0, cuando los términos ocurren siempre por separado.

Tomando esta relación como referencia, se puede computar la contraria, la que indica que un término tiene un significado más amplio que otro AP (t_i, t_j)

$$AP(t_i, t_j) \Leftrightarrow RT(t_j, t_i)$$

El autor propone la construcción de una ontología difusa basada en estas relaciones *más amplio que* y *más restringido que* en dos etapas: creación de la ontología basada en la relación más restringido que y eliminación de relaciones innecesarias en un segundo estado.

El uso práctico de esta ontología reside en la posibilidad de sugerir términos con un significado más amplio, y con un significado más restringido.

CONCLUSIONES Y POSIBLES LÍNEAS DE INVESTIGACIÓN

En este trabajo se ha mostrado que la lógica difusa va a jugar un papel protagonista en la recuperación de la información con las tecnologías venideras, como la inminente web semántica.

En concreto, se ha hablado de cómo uno de los pilares de la web semántica, las ontologías, se puede expandir en el terreno de lo difuso para ganar en flexibilidad en la formulación de consultas y en la manera de interpretar los resultados obtenidos (e.j.: mediante índices de relevancia), etc.

En todo caso, la recuperación de la información por medio de ontologías y la introducción de la lógica difusa en este proceso abren muchas líneas de investigación, entre las que cabe reseñar:

- La incorporación de información en documentos y consultas en las ontologías
- La mejora de la recuperación de la información mediante el uso de relaciones entre las entidades de la ontología
- La incorporación de relaciones difusas a la ontología
- Recuperación de documentos basada en el uso de entidades asociadas semánticamente en relaciones, no simplemente la ocurrencia de términos en la consulta

REFERENCES

[1] Fuzzy-Logic available at http://www.fuzzy-logic.com/Ch1.htm [accessed Nov 2008]

[2] M. Lu, F. Dong, F. Fotouhi, The Semantic Web. Opportunities and Challenges for next Generation Web Applications. Information Research (7), 2002

[3] Noy, N. F. and D. McGuinness (2001). Ontology Development 101: A guide to Creating your First Ontology. Stanford Medical Informatics Technical Report SMI-2001-0880 . Stanford University

[4] Parry, D. ACM International Conference Proceeding Series; Vol. 54 Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32, New Zealand 2004
[5] U. Straccia. A fuzzy description logic for the semantic web. In Capturing Intelligence: Fuzzy Logic and the Semantic Web, E. Sánchex, ed, Elsevier, 2006.
[6] OWL features. http://www.w3.org/TR/owl-features/ (accedido en enero del 2009)
[7] Calegari S. and Sanchez E.. A Fuzzy Ontology-Approach to improve Semantic Information Retrieval. Proceedings of the Third ISWC Workshop on Uncertainty Reasoning for the Semantic Web - URSW'07.
[8] Calegari S., Sanchez E. "Object-fuzzy concept network: An enrichment of ontologies in semantic information retrieval" Journal of the American Society for Information Science and Technology 59
[9] J. Kacprzyk, P. Bosc, Fuzziness in Database Managenzent Systems, Studies in Fuzziness, Physica-Verlag, 2 Heidelberg, 1995.
[10] P. Bosc, M. Galiboug, G. Hamon, Frizzy querying with SQL: e.xtensions and inrplemenmtiotr aspects, Fuzzy Sets and Systems 28 (3) (1988), 333-349.
[11] D.H. Widyantoro, J. Yen, Using Fuzzy ontology for query refinement in a personalized abstract search engine, in: IFSA World Congress and 20th NAFIPS International Conference, 200 1. Joint gth, 2001.
[12] Herman, Ivan (2008-03-07). "Semantic Web Activity Statement". W3C. Retrieved on 2008-03-13.
[13] http://www.w3.org/2001/sw/SW-FAQ
[14] George J. Klir, Bo Yuan: Fuzzy Sets and Fuzzy Logic: Theory and Applications, 1995, ISBN 0131011715