Relaciones Borrosas Conceptuales para la Construcción de Ontologías con GUMSe

Jesús Serrano-Guerrero
Departamento de Informática, Universidad de Castilla la Mancha
Ciudad Real, Castilla la Mancha, 13071, España

José A. Olivas
Departamento de Informática, Universidad de Castilla la Mancha
Ciudad Real, Castilla la Mancha, 13071, España

y

Javier de la Mata
Departamento de Informática, Universidad de Castilla la Mancha
Ciudad Real, Castilla la Mancha, 13071, España

RESUMEN

El estudio aquí realizado trata sobre el descubrimiento de relaciones borrosas entre términos para la construcción automática de ontologías. Se proponen dos tipos de relaciones para la construcción de las ontologías: "relaciones físicas" y "relaciones semánticas", y un nuevo factor de influencia entre términos: "la distancia física". La distancia física entre dos términos determina el grado de relación entre dichos términos, así pues, dos términos que coexisten en la misma frase tienen más probabilidad de estar altamente relacionados, que dos términos situados en distintos párrafos.

La construcción de estas ontologías tiene dos fases principales: "el procesamiento de textos" y "el cálculo de los coeficientes de relación entre términos".

Esta aproximación está siendo incorporada en el metabuscador GUMSe (Gum Search) desarrollado en el marco del grupo de Investigación SMILe (Soft Management of Internet e-Laboratory, ORETO-UCLM, Spain) como medio de ayuda para el refinamiento de la fase "*query-expansion*".

Palabras Claves: ontología, lógica borrosa, agente, metabuscador, Internet.

1. INTRODUCCIÓN

En la actualidad Internet alberga una gran cantidad de información desordenada que dificulta la labor de los buscadores, provocando una masiva recuperación de información irrelevante para los usuarios. El principal problema reside en la búsqueda de información basada únicamente en aspectos léxicos, es decir, la aparición o no de los términos de una pregunta del usuario en el conjunto de documentos indexados por el motor de búsqueda utilizado. Los aspectos semánticos son difícilmente abordables, y para intentar de cubrir este vacío surgen distintas técnicas entre las que destaca el uso las ontologías.

A día de hoy existen muchas ontologías referentes a distintos dominios que mejoran algunos aspectos de ciertas aplicaciones, pero todas ellas han sido creadas de forma manual siguiendo distintas metodologías como las de Gruninger [2] o Fernández [1]. Por el contrario la construcción automática de ontologías constituye un foco de investigación actual donde los logros obtenidos no han sido demasiado satisfactorios.

El sistema desarrollado por el agente constructor de ontologías de GUMSe constituye un cambio sobre la filosofía de construcción de ontologías basada únicamente en la co-ocurrencia de términos y la falta de una definición de "relación conceptual" consistente. Se propone una separación entre dos tipos de relaciones entre términos, las puramente semánticas y las físicas, las cuales influyen de distinta manera en el modelado de la ontología.

En las siguientes secciones se describen algunos de los principales aspectos más relevantes de la generación de ontologías según GUMSe. La sección 2 presenta la fase de tratamiento de textos para poder extraer los términos que modelarán la ontología buscada. La sección 3 trata el cálculo de los coeficientes físico y semántico necesarios para establecer las relaciones entre términos.

El correcto funcionamiento del sistema matemático se demuestra en la sección 4, la sección 5 habla sobre la plataforma GUMSe donde se está implementando este modelo, mientras que en la sección 6 se presentan las conclusiones finales.

2. PROCESAMIENTO DE TEXTOS

Uno de los mayores problemas que presenta la construcción automática de ontologías es la extracción del conocimiento verdaderamente relevante de los textos que se procesan, especialmente si los textos candidatos para su construcción no han sido seleccionados previamente por personas expertas en el dominio sobre el que versa la ontología, sino que la

información ha sido seleccionada directamente por otro proceso automático como son los buscadores.

La solución utilizada para la reconstrucción de textos, que facilita tanto el almacenamiento de información relevante como su posterior procesamiento consta de las siguientes fases:

Simplificación de Textos
La mayoría de los textos utilizan un gran número de recursos lingüísticos como frases de relativo, formas pasivas, metáforas, etc., que aumentan el tamaño en líneas de texto pero que semánticamente no suelen aportar nada especialmente relevante. Sin embargo, todas las frases por muchos recursos lingüísticos que usen no pierden su estructura básica salvo las formas impersonales:

SUJETO INFINITIVO PREDICADO (1)

Esta idea es la base del funcionamiento del procesador de texto del *agente constructor de ontologías*.

Los verbos de las frases pueden aparecer en distintas formas verbales pero el aspecto semántico que aquí interesa es independiente de dicha forma verbal.

Jesús vivió en Puertollano (2)

La idea de tiempo no es relevante en esta frase, sólo nos interesa que Jesús independientemente de quien sea Jesús, está relacionado con Puertollano, independientemente de qué ciudad sea Puertollano.

Los conceptos son propios de la mente humana y la única forma de expresarlos es mediante palabras por lo que los conceptos que representan el predicado y el sujeto deben ser palabras, que tratadas adecuadamente mediante algoritmos de Stemming (ej: Porter [3]), una StopList, y WordNet [5] quedan reducidos a una única palabra, salvo excepciones que se comentarán a continuación. Así pues, la estructura de cada frase simplificada es:

TERMINO INFINITVO TERMINO *(3)*

Lo cual se inserta fácilmente en una base de hechos (por ejemplo en formato tipo PROLO G):

Vivir (Jesús, Puertollano) (4)

que permite extraer conocimientos posteriores y añadir más información al texto.

Adquisición de Conocimiento Implícito
No todo el conocimiento procesable es extraído únicamente de los textos, también se pueden usar *theasurus*, en este caso, el utilizado es WordNet [5] mediante un agente especializado (*WordNet Agent*) de GUMSe.

Expansión del Contexto: En el apartado anterior se hace alusión a que un concepto es más fácilmente procesable si es únicamente una palabra, pero no siempre es óptimo y en el caso en el que se propone a continuación más aún.

La mayoría de las frases en el lenguaje habitual abusan de términos implícitos y sobreentendidos por el contexto, lo cual es el principal problema para la creación automática de ontologías.

La solución aquí propuesta es la adquisición de dicho conocimiento a partir de la base de hechos anteriormente creada. Así pues, el texto simplificado en la fase anterior pasará a ser completado con el conocimiento implícito, por lo que ya no quedarán conceptos de una única palabra sino de más de una palabra. Este hecho empeora el almacenamiento y tiempo de procesamiento pero es sin duda la principal mejora que aporta este sistema ya que es esencial dicho nuevo conocimiento y su situación en el texto para el cálculo del coeficiente semántico que se verá a continuación.

Resumiendo, una frase de la forma:

"Sus obras eran escuchadas por la aristocracia con gran devoción"

en la fase de simplificación de textos quedaría de la forma:

Obras Conocer Aristocracia *(5)*

lo cual es muy simple pero no aporta un conocimiento real. Sin embargo, si le preguntamos a la base de hechos de quién o de qué se estaba hablando en el último párrafo, y está nos responde que se hablaba de Beethoven, tendríamos:

Beethoven Obras Conocer Aristocracia (6)

lo cual es mucho más intuitivo y acerca mucho más la relación entre Beethoven y la aristocracia tanto semánticamente como en distancia física (ambas palabras están en el misma frase).

3. COEFICIENTES DE RELACIÓN

La principal mejora que propone el *agente de ontologías* es la separación de los tipos posibles de relaciones entre conceptos en dos:

- Relaciones puramente conceptuales y
- Relaciones puramente físicas

En la frase:

Las personas tienen piernas *(7)*

sin duda alguna, existe una relación muy fuerte entre los conceptos personas y piernas, ya que la relación es física además de ser un axioma. Por el contrario en la frase:

Las mujeres llevan joyas *(8)*

relaciona a las mujeres con las joyas, lo cual ni es un axioma ni es una verdad que se cumple en todos los casos y es muy dudoso que aporte un conocimiento de una gran relevancia para ningún contexto, por lo tanto este tipo de relación es calificada de puramente semántico.

Partiendo de la base de los dos tipos de relaciones anteriores y sus diferencias aparentes, se calculan los coeficientes relativos a cada uno de ellas de la siguiente manera.

Coeficiente Semántico
Widyantoro [4] propone que las relaciones entre palabras están basadas en la cantidad de veces que co-ocurren las palabras que forman la relación en un documento, pero esta teoría presenta un problema que es la gran cantidad de documentos necesarios para poder estimar con cierta credibilidad la relación entre 2 conceptos, ya que sólo se procesan palabras sin contexto alguno.

El *Ontology Agent* de GUMSe trata de solventar esa gran demanda de documentos relevantes ya que mediante el descubrimiento de conocimiento implícito, se añaden al texto aquellas ideas y conceptos que verdaderamente predominan y que son los conceptos que determinan las relaciones más fuertes. Así en unos pocos documentos se tiene esa gran cantidad de palabras que demanda el cálculo propuesto por Widyantoro.

Pero además, tal y como se ha podido ver en el ejemplo 1 las relaciones semánticas entre conceptos pueden ser potenciadas por la distancia física que hay entre las palabras. En un documento no podrá haber la misma relación entre dos palabras que aparecen en la misma frase, como entre dos palabras que estén en párrafos distintos.

Así, dada una colección de páginas P = {a_1, a_2, a_3, …., a_n,..} recuperadas por el agente *Page Retriever Broker* ante la consulta de un usuario, donde cada página a = {t_1, t_2, t_3, ….. t_n } está formada por un conjunto de términos t_js, el *Ontology Agent* calcula el coeficiente semántico como:

$$S = \sum_{a \in A} \left(\frac{\sum_{occur\,(ti,tj)} \frac{1}{\exp(\ frase\ *\ parrafo\ )}}{\max(\ ocurr\ (t_j), ocurr\ (t_i))} \right) \qquad (9)$$

Frase y párrafo son medidas de distancia física, donde si los términos t_j, t_i están el párrafo actual su valor es 1 y si está en un párrafo posterior vale 2, y así sucesivamente. Por otro lado, si t_j, t_i están en la misma frase están a distancia 0, y si aparece t_i en una frase y t_j en la siguiente están a distancia 1 y así sucesivamente. Así la distancia física afecta de forma exponencial al coeficiente semántico, es decir, a mayor distancia menor valor.

La función *occur* (t_j, t_i) denota las veces que aparecen conjuntamente t_j, $_y$ t_i, mientras que *occur(t_j)* denota el número de veces que aparece el término t_j.

La división por *max (ocurr(t_j),ocurr(t_i))* es simplemente para normalizar, ya que el término que más veces se repite es el que probablemente más influya en el contexto ya que será sobre lo que versa el texto.

Coeficiente Físico
La idea de tratar las frases como estructuras de la forma:

TERMINO INFINITIVO TERMINO *(10)*

permite identificar aquellos verbos que pueden expresar relaciones físicas, como agregaciones ("*have*", *"is part of*") o herencia ("*is a*", *"belong to*"), y como consecuencia aquellos conceptos que están relacionados físicamente.

El valor numérico de este coeficiente se calcula

$$F = \sum_{a \in P} \frac{ocurr\ \ (t_i, t_j)}{N\ (t_j)} \qquad (11)$$

donde N(t_j) es el número de documentos evaluados que contiene a t_j, donde t_j es el sujeto de las estructuras:

TERMINO INFINITIVO TERMINO *(12)*

Este término N(t_j) simplemente vale para normalizar el valor del coeficiente ya que t_j es el término que más puede influir en el contexto al ser el que tiene más posibilidades de definir sobre lo que se está hablando, al igual que en el caso del coeficiente semántico donde el término que más influye se calcula como

max(ocurr(t_j), ocurr (t_i)).

La elección del sujeto como término normalizador es evidente ya que si:

El conejo tiene patas *(13)*

El término más genérico es *"conejo"* y es sobre el que más posibilidades hay que el resto de páginas hablen.

Esta fórmula (11) es una simplificación de la anterior (9) ya que aquí la distancia física no existe porque los términos t_j, t_i se encuentran en la misma frase y en el mismo párrafo, así la influencia del la distancia viene medida por:

*Exp(frase * parrafo)* *(14)*

y estando en la misma frase y en el mismo párrafo:

*exp(1 * 0) = 1* *(15)*

por lo que la influencia por la distancia no hace falta ponerla.

Ambas fórmulas están normalizadas por el término más influyente en el contexto, y ambas representan básicamente el sumatorio de las co-ocurrencias de los términos t_j, t_i, pero matizados unas veces por la distancia física y otras veces por el tipo de predicados en los que se encuentran dichos términos.

4. SOLAPAMIENTO ENTRE AMBOS COEFICIENTES

Supongamos una frase en la que existiera aislada una relación de agregación y nada más relacionado en el documento:

Jesús Tener Maleta *(16)*

Donde t_j = *Jesús* y t_i = *Maleta*.

Según el coeficiente físico

$$F = \sum_{a \in P} \frac{ocurr\ (t_i, t_j)}{N\ (t_j)} \qquad (17)$$

Ambas co-ocurren una vez en un documento entonces $F = 1$.

Y según el coeficiente semántico

$$S = \sum_{a \in A} \left(\frac{\sum_{occur\ (t_i, t_j)} \frac{1}{\exp(\ frase\ *\ parrafo\ )}}{\max(\ ocurr\ (t_j), ocurr\ (t_i))} \right) \qquad (18)$$

Entonces tj = *Jesús* y t_i = *Maleta*, están en el mismo párrafo y en la misma frase, sólo aparecen una vez, luego $S = 1$.

Este ejemplo pone de manifiesto que ambas medidas están claramente relacionadas y no pueden trabajar aisladamente si se quiere tener todo el conocimiento de un conjunto de páginas. Pero además denota que la relación medida por el coeficiente físico absorbe a la medida por el coeficiente semántico ya que como es lógico, si un término está relacionado físicamente con otro implica que semánticamente también lo está.

Para procesar todo el conocimiento extraído debemos relacionar ambos coeficientes y la forma elegida por el *Ontology Agent* es mediante la unión borrosa (T-conorma de Luckasiewicz en este caso), ya que es la medida idónea para eliminar la zona de solapamiento entre ambos coeficientes.

Relación = S ö F
Fig. 1

Donde ö representa la unión borrosa.

$$Relación = S + F - S*F, \qquad (19)$$

donde $S*F$ representa el grado de solapamiento que hay que eliminar de ambas relaciones.

5. PRUEBAS

En el ejemplo (16), donde las relaciones eran puramente físicas S valía 1 y F también valía 1, por lo que

$$Relación = S + F - S*F = 1+1 -1*1 = 1, \qquad (20)$$

es decir se sigue demostrando la fuerte relación entre ambos términos mediante los dos coeficientes.

Si por ejemplo no hay apenas relación física *(F = 0,05)* y toda la carga sea puramente semántica *(S = 0.8)*, no habría apenas zona de intersección por lo que todo el valor sería el de la carga semántica:

$$Relación = S + F - S*F = 0.8+ 0.05 -0.8*0.05 = 0.81, \qquad (21)$$

que es el valor de carga semántica encontrado.

Si ambos coeficientes son similares *(S = F = 0,4)*, entonces deberían influirse positivamente hasta cierto grado como se demuestra:

$$Relación = S + F - S*F = 0.4+ 0.4 -0.4*0.4 = 0.64 \qquad (22)$$

6. GUMSe

GUMSe es un meta-buscador basado en agentes que, partiendo de los buscadores *google* y *altavista*, pretende abordar uno de los principales problemas de los buscadores actuales como es el factor *"recall"*. La precisión de los buscadores actuales es buena, pero sin embargo hay muchas páginas que no son recuperadas por la falta de capacidades semánticas de los buscadores y que sin embargo son relevantes ante las preguntas del usuario.

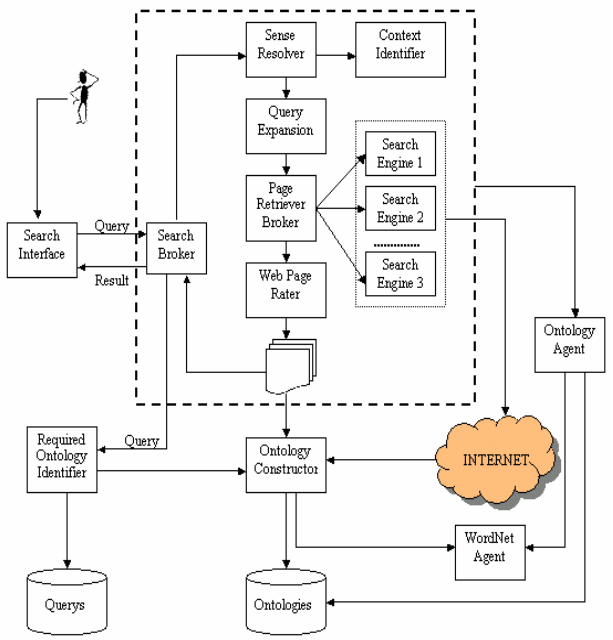

Fig. Arquitectura del buscador GUMSe

GUMSE pretende dotar de estas capacidades semánticas mediante el modelo FIS-CRM, usado ya por el meta-buscador FISS [6], y el *Ontology Agent* aquí propuesto.

La finalidad del *Ontology Agent* es la de mantener un repositorio de ontologías generadas a partir de las páginas recuperadas de los distintos buscadores utilizados (*google* y *altavista*) en función de las consultas de los usuarios para poder

descubrir relaciones entre los términos de la consulta y los documentos recuperados, y redefinir así la pregunta del usuario en la fase "*query expansion*".

7. FUTUROS TRABAJOS

Las medidas actuales para el cálculo de los coeficientes tanto semántico como físico están basadas en los factores "frase" y "párrafo", pero también se proponen diversas alternativas que deben ser estudiadas como la distancia a nivel de frases simples o frases compuestas.

El procesamiento de las frases puede ser guardado como grafos conceptuales tal y como propone Cao [7].

El cálculo de los coeficientes relacionales puede ser refinado mediante la identificación de otras relaciones entre términos como la hiponimia, antonimia, sinonimia [6, 8], etc., o el descubrimiento y tratamiento de axiomas.

La aplicación de distintos algoritmos de poda en función del valor de cada coeficiente trabajando independientemente o conjuntamente puede permitir la construcción de distintos tipos de ontologías.

La aplicación actualmente está desarrollada para lengua inglesa pero también debe ser extensible a otros idiomas distintos donde la forma de escribir puede ser distinta, como el caso del español, donde las frases son mucho más largas y la cantidad de recursos lingüísticos es mayor.

Así mismo son mejorables muchos de los algoritmos utilizados, y se pueden plantear nuevas formas de afrontar el problema.

REFERENCIAS

[1] M. Fernández López, A. Gómez-Pérez, J. Pazos Sierra, A. Pazos Sierra (1999): "Building a Chemical Ontology Using Methontology and the Ontology Design Environment", IEEE Intelligent Systems, v.14 n.1, 37-46.

[2] M. Gruninger,. M.S Fox, (1995), "Methodology for the Design and Evaluation of Ontologies", Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal.

[3] M.F Porter (1980), An Algorithm for Suffix Stripping. *Program*, 14(3): 130-137.

[4] D. Widyantoro, J. Yen, (2001): "Incorporating fuzzy ontology of term relations in a search engine", Proceedings of the BISC Int. Workshop on Fuzzy Logic and the Internet, 155–160.

[5] WORDNET. A lexical database for the English language. http://www.cogsci.princeton.edu/~wn/

[6] J. A. Olivas; P. Garcés; F. P. Romero (2003): "An Application of the FIS-CRM Model to the FISS Metasearcher: Using Fuzzy Synonymy and Fuzzy Generality for Representing Concepts in Documents". Int. Jour. of Approx. Reasoning (Soft Computing in Recognition and Search) 34, 201-219.

[7] T. H. Cao (2001): "Fuzzy Conceptual Graphs for the Semantic Web", Proceedings of the BISC, International Workshop on Fuzzy Logic and the Internet, 74–79.

[8] S. Fernández (2001): Una contribución al procesamiento automático de la sinonimia usando Prolog, Tesis Doctoral, Universidad de Santiago de Compostela, España.