Elicitación y Comparación de Conocimiento Semántico para Ontologías Naturales

L. Cota-Gomez y J. Figueroa-Nazuno

Centro de Investigación en Computación
Instituto Politécnico Nacional
Unidad Profesional "Adolfo López Mateos"
{lcota,jfn}@cic.ipn.mx
Paper received on 28/07/10, Accepted on 27/09/10.

Resumen. Las ontologías son un componente técnico fundamental para la interoperabilidad semántica, objetivo fundamental en la Semantic Web. Se han desarrollado distintos métodos y herramientas para su construcción, que varían desde la anotación directa efectuada por ingenieros de conocimiento hasta aproximaciones de extracción semiautomática en documentos electrónicos. En una u otra forma las ontologías son la descripción por medio de grafos, de diferentes relaciones entre conceptos que a su vez sirven para definir otro concepto, esas relaciones pueden ser por ejemplo: asociaciones, sinonimias, conceptos frecuentemente relacionados, etc. El presente trabajo describe una metodología automatizada para elicitación y comparación cuantitativa de conocimiento semántico, capaz de generar estructuras computacionales para ontologías que reflejan directamente el significado utilizado en una comunidad suscrita a un dominio.

Palabras Clave: Ingeniería del Conocimiento, Elicitación de Conocimiento, Construcción de Ontologías, Comparación de Ontologías, Modelo Semántico, Redes Semánticas Naturales.

1 Introducción

Uno de los principales tópicos constantes en la Inteligencia Artificial es la necesidad de manejar contenido semántico. La búsqueda de interoperabilidad semántica se enfatizó con el surgimiento de la idea de Semantic Web [1] debido a la vasta cantidad de información contenida en medios electrónicos originalmente creada para ser únicamente leída por humanos (human readable information) y su poco potencial de explotación automática (machine-processable information).

Los componentes técnicos considerados con gran importancia para permitir el manejo de contenido semántico son las ontologías. Algunas definiciones sobre lo que es una ontología han sido dadas por Gruber: «una especificación explicita de una conceptualización»; Borst: «una especificación formal de una conceptualización compartida»; Studer: «Una ontología es una especificación formal, explícita, de una conceptualización compartida». Todas estas definiciones asumen una noción informal de «conceptualización»[19]. El propósito primordial de éstas, en su forma más general, consiste en almacenar una colección de conocimiento humano estructurado

Advances in Intelligent and Information Technologies.
Research in Computing Science 50, 2010, pp. 225-233.

para otorgar significado, por medio de relaciones entre conceptos, habilitando a las computadoras y a las personas a trabajar en cooperación.

Existen muchos tipos de ontologías, y diferentes formas de clasificarlas, que pueden utilizarse para múltiples propósitos, también se han desarrollado distintos métodos y herramientas para su construcción y administración [19].

El problema general de ontologías en el sentido computacional tiene muchas aristas que van desde metodologías totalmente formales para su construcción hasta trabajos que son recolección de datos directos en campo [6,9,13]. Uno de los aspectos más importantes, es la extracción de esas estructuras generalmente en forma de grafos que definen el significado, en dónde los nodos son conceptos con un valor numérico, sin embargo el resultado final y más representativo son las relaciones o red de conceptos.

Algunos métodos para la anotación de ontologías obtienen su conocimiento directamente de humanos a través de técnicas ”activas” como mesas redondas, entrevistas, diálogos y otras semejantes; o técnicas ”pasivas” que incluyen observación, lecturas y protocolo de ”pensar en voz alta”. En todas, resulta sumamente complicado obtener la representación de un consenso, además de que el proceso de anotación se ve afectado por la transcripción.[12,14]

Por otra parte, muchas metodologías para la construcción de ontologías implican el uso de otra auxiliar que proporcione significados representados a través de relaciones entre conceptos, ya sea de cierta área particular o de sentido común no especializado. Un ejemplo de ontología muy usada para éste propósito es WordNet, considerada como el estándar de oro (”gold standard”)[7], y catalogada como base de datos léxica o como ”upper ontology”, que irónicamente su objetivo en un principio no era representar significado.[10]

Este trabajo presenta una metodología para construcción de Ontologías Naturales implementada en software, procedente de la técnica clásica de Redes Semánticas Naturales (RSN) [11] cuya fundamentación teórica y empírica como estructuras que representan el significado en humanos, está sólidamente demostrada por las ciencias cognitivas. [3,2,5,15,21]

Otros trabajos recientes en el área, también se han basado en RSN con diferente procedimiento.[20,4] Las particularidades y sutilezas de la técnica RSN para elicitación del significado, permite generar de forma directa estructuras que contienen palabras y relaciones, proporcionando métricas de distancia y similitud intrínsecas que representan el significado de conceptos importantes con consenso de una determinada comunidad suscrita en cierto dominio de conocimiento. Lo más relevante es que estas estructuras son económicamente manejables computacionalmente, esto es demostrado mediante el ejemplo desarrollado para la construcción de una ontología formada con conceptos del dominio ”Computación”.

2. Construcción de una Ontología Natural del dominio Computación

Se presenta una aproximación para la representación del significado en humanos por medio de la técnica de RSN [11,23], que directamente puede obtener Ontologías

Naturales compartidas por una comunidad, proporcionando estructura y métricas de comparación cuantitativas de forma directa y simple, automatizable y económicamente computacional.

A continuación se describe la metodología para la elicitación y comparación de Ontologías Naturales, mediante la construcción de una ontología en el dominio ”Computación”.

2.1 Metodología para Elicitación y Obtención de la Ontología Natural

Elección de conceptos base dentro del dominio de conocimiento. Para elegir los conceptos que fungirán como base para la construcción de la ontología, se pidió a varios expertos en el área, que determinarán cuáles son los conceptos más importantes en el dominio. Después, se eligieron los 5 conceptos con mayor frecuencia de mención. Este paso puede tener variaciones, inclusive resulta útil hacer uso de la misma técnica que se realiza para obtener las definidoras.

En este caso los conceptos elegidos fueron: Algoritmo, Información, Lenguaje, Red y Sistema Operativo.

Elicitación del significado de los conceptos base. Se eligen 30 sujetos que representen una comunidad inmersa en un dominio de conocimiento, que mediante un sistema de captura, se les solicita lo siguiente:

1. Proporcionar el significado de cada uno de los conceptos base, por medio de un rango de 5 a 10 palabras sueltas, sin usar partículas gramaticales.
2. Jerarquizar las palabras proporcionadas.

Es muy importante tomar en cuenta la sutileza de solicitar el significado de conceptos; ya que esta demostrado que se pueden obtener resultados muy diferentes si se pide dar ”palabras relacionadas”[11].

Para este caso, se eligieron 3 comunidades inmersas en el dominio de Computación con diferentes enfoques para formar las RSN de cada grupo:

RSN1 Grupo de expertos con enfoque en Ciencias.
RSN2 Grupo con enfoque en Sistemas Computacionales.
RSN3 Grupo con enfoque en Telemática.

Cálculo de Pesos Semánticos. Para cada concepto definido, y a su vez, para cada una de las palabras definidoras obtenidas (cabe mencionar que para determinar las definidoras en ocasiones es necesario efectuar un procesamiento previo, en este caso manual, en el que se pueden fusionar palabras sinónimos y eliminar errores ortográficos o léxicos), se calcula la frecuencia de aparición organizada de acuerdo a la jerarquía proporcionada. Las jerarquías del 1 al 10 se convierten mediante una relación de equivalencia en valores semánticos. Es decir, la jerarquía 1, que representa la más importante en cercanía al concepto, toma el valor 10, que es el máximo valor semántico con base en nuestro análisis. La jerarquía 2 toma el valor 9, de esta forma continua hasta la jerarquía 10 que toma el valor 1. Una vez dadas las equivalencias,

se calcula el valor M [23] de cada palabra definidora, mediante, la sumatoria de la frecuencia de la jerarquización por el valor semántico correspondiente. O bien:

$$M = \sum_{i=1}^{10} (Frecuencia_i \times ValorSemantico_i) \quad (1)$$

El valor M representa el peso semántico de la definidora en relación al concepto que define. Ver Tabla 1. Este valor es uno de los más importantes ya que en base a éste, la técnica de RSN tiene muchas formas de obtener distancia semántica.

Tabla 1. Valores M del grupo RSN3 para el concepto ”Algoritmo”.

Definidora	Valor
PROGRAMACION	68
SOLUCIÓN	58
INSTRUCCIONES	53
SECUENCIA	52
COMPLEJIDAD	47
LOGICA	33

Generación del Canónico General para la Ontología Natural. Este es un paso optativo, útil cuando se tienen ontologías de diferentes grupos. Una vez obtenidas las ontologías para los grupos definidos previamente, se construye el canónico general de éstas sumando el valor M de cada relación concepto-definidora encontrado.

Ontologías Obtenidas En la Figura 1, se muestra la Ontología Natural obtenida mediante el software de elicitación que implementa RSN y otro de graficación. En el grafo que se muestra cada nodo representa una palabra “definidora” y cada arista la relación entre éstas, obtenidas a partir de la información dada por los grupos de sujetos.La Tabla 2 muestra algunas métricas de las ontologías obtenidas. La columna con el valor J indica la cantidad de palabras [23], la columna ”Links” indica la cantidad de relaciones encontradas entre las palabras.

Todas las palabras y relaciones obtenidas con la metodología descrita, se pueden aplicar directamente para la anotación de ontologías mediante algún lenguaje como OWL o RDF.

2.2 Metodología para Comparación de Ontologías Naturales

Se puede efectuar la comparación de Ontologías Naturales de diferentes formas, la metodología usada para este trabajo se describe a continuación:

Reconstrucción Automática de Matriz Los datos obtenidos después del cálculo de peso semántico, nos permiten representar las relaciones concepto-definidora de la Ontología Natural en una matriz cuadrada.

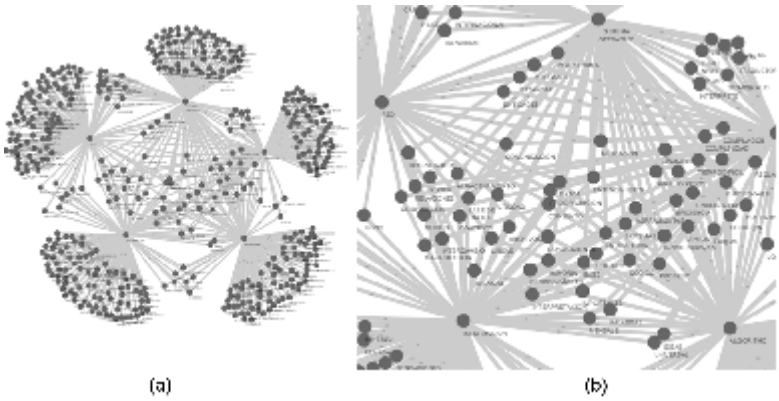

Fig. 1. (a) Visualización gráfica para la Ontología Natural obtenida del canónico formado por los grupos RSN1, RSN2, RSN3. (b) Acercamiento

Tabla 2. Número de palabras y relaciones obtenidas para cada ontología natural

Ontología	ValorJ	Links
RSN1	68	73
RSN2	374	471
RSN3	172	212
Canónico	471	617

Cálculo de Eigenvalores Los Eigenvalores son conocidos como una forma poderosa de representar matrices y con ellos se puede hacer comparación de tipo elástica, ya que el vector de *Eigenvalores* obtenido es una representación en R1 mucho más fácil de manejar que una matriz completa. Se calcularon los 6 Eigenvalores con mayor magnitud, como se muestra en la Tabla 3. [16]

Table 3. *Eigenvalores* para las matrices de cada ontología.

Canónico	RSN1	RSN2	RSN3
38.53132229	-0.000957142	21.6079831	5.916079783
-24.28726989	-0.000957142	-13.93788685	-5.916079783
-24.28726989	0.000957142	-7.67009625	-2.19342E-14
10.04321749	0.000957142	-9.04058E-05	-2.19342E-14
3.74373E-14	1.55378E-16	4.52029E-05	-5.15254E-16
3.74373E-14	1.55378E-16	4.52029E-05	-5.15254E-16

Es necesario utilizar todo el lexicón usado en las Ontologías para construir las matrices que nos permitirán compararlas dónde cada [i][j]ésimo elemento representa el valor M de la relación concepto-definidora. De tal forma que se obtuvieron 4 matrices esparcidas de 471x471.

Cálculo de Distancias Se obtuvo el cálculo de distancia entre los Eigenvalores correspondientes a cada Ontología Natural. Se pueden aplicar diferentes medidas, para este trabajo las distancias calculadas automáticamente fueron Distancia Chebyshev (2), Distancia Euclideana al Cuadrado (3), Distancia Minkowski (4):

$$dis(U,V) = \max\left(|u_i - v_i|_{i=1}^{n}\right) \quad (2)$$

$$dis(U,V) = \sum_{i=1}^{n}(u_i - v_i)^2 \quad (3)$$

$$dis(U,V) = \sqrt[\lambda]{\sum_{i=1}^{n}|u_i - v_i|^{\lambda}} \quad (4)$$

Donde:

$U =$ Muestra 1
$V =$ Muestra 2
$u =$ Valor de la variable i en la muestra U
$v =$ Valor de la variable i en la muestra V
$n =$ Número total de variables.
$\lambda =$ Órden.

Las Tablas 4, 5 y 6 muestran los resultados del cálculo de distancias para comparar las matrices de las Ontologías Naturales. Cada métrica de distancia nos proporciona diferente información, no obstante todas coinciden en que los valores de distancia para las matrices más parecidas son menores, y viceversa.

Tabla 4. Distancia Chebyshev

Chebyshev	**Canónico**	**RSN1**	**RSN2**	**RSN3**
Canónico	0	38.532279	16.923339	32.615243
RSN1	38.532279	0	21.60894	5.917037
RSN2	16.923339	21.60894	0	15.691903
RSN3	32.615243	5.917037	15.691903	0

Tabla 5. Distancia Euclideana al Cuadrado

Euclideana^2	**Canónico**	**RSN1**	**RSN2**	**RSN3**
Canónico	0	2765.3265	770.50763	2091.9924
RSN1	2765.3265	0	720.02937	70.000004
RSN2	770.50763	720.02937	0	369.4156
RSN3	2091.9924	70.000004	369.4156	0

Tabla 6. Distancia Minkowski

Minkowski	**Canónico**	**RSN1**	**RSN2**	**RSN3**
Canónico	0	44.289372	22.60896	38.3119
RSN1	44.289372	0	23.662343	7.453794
RSN2	22.60896	23.662343	0	16.905289
RSN3	38.3119	7.453794	16.905289	0

Como se puede observar en el paso 1, se obtiene una matriz global de todos los conceptos generados. Estas matrices esparcidas son la base para encontrar, dados los grupos de referencia que se utilizaron, por un lado, cuales son los conceptos con más altas relaciones y, por otro lado, permiten comparar las matrices de resultados de los grupos de expertos con los no expertos.

En la matriz de cada grupo están todas las relaciones entre conceptos dadas por los sujetos. Y dado que se aceptan los diferentes grados de conocimiento que tienen estos sujetos por su origen, entonces se puede ver que la técnica permite encontrar que tan diferentes son entre sí. Lo cual es muy importante para encontrar acuerdo entre expertos y conocer cómo la estructura de red es diferente en la de no expertos.

Una forma más específica de ver la diferencia entre los grupos es el uso de tres medidas diferentes de distancia, lo que nos ayuda a demostrar la diferencia entre grupos tanto en el espacio euclidiano como en el de Minkowski y también nos indica la posible operación de la red en un espacio n-dimensional.

3 Discusión

La metodología presentada hace un énfasis en obtener la información directamente de especialistas en el tema y estos datos se obtienen en forma sistemática y con diferentes restricciones lo cual elimina la posibilidad de textos ambiguos o de tipo narrativo y hace énfasis en la restricción de conceptos que son las "definidoras" que dan los sujetos. Lo más importante de esta técnica y que se ha demostrado en muchos trabajos es la repetibilidad sistemática de definidoras.

El caso más extremo de demostración del uso de conceptos para definir fenómenos de alta complejidad es el trabajo de Dravnieks sobre olores presentado en la revista Science[8]. Así mismo existen muchos trabajos en dónde las respuestas generadas por los sujetos son altamente consistentes y repetibles en diferentes situaciones [24,22,17].

El aspecto central de la organización de los conceptos dados por los sujetos es el hecho de la alta consistencia que tienen los pesos semánticos generados por éstos, donde es muy fácil ver que hay un grupo de definidoras que es constante y repetitivo. Estos conceptos altamente usados como definidoras son la base para encontrar las redes de relación.

Usar la técnica de RSN como método de elicitación para la anotación de conocimiento en la construcción de ontologías, asegura que se está representando el significado mediante los conceptos y relaciones entre éstos, basándose en las teorías modernas del significado además de estar sólidamente demostrada por las ciencias cognitivas.

Es un método que obtiene el significado directamente de humanos, logrando encontrar consenso y proporcionando estructura y valores semánticos. Estos valores semánticos pueden procesarse para proporcionar mayor información, como distancia entre conceptos, y a su vez distancia entre el conocimiento de grupos.

La representación en grafo, aunque da una idea visual de las relaciones, es incompleta debido a que las longitudes de los links en el esquema sólo son una aproximación respecto a su peso semántico. No obstante, sus datos son fácilmente representados en una matriz esparcida cuadrada, en la que se pueden identificar las relaciones y sus pesos semánticos. A su vez esta matriz puede ser reducida a sus eigenvalores, y de esta forma aplicar diferentes técnicas para calcular distancia, y de esta forma comparar conocimiento humano.

La ontología obtenida fue desarrollada mediante un software que implementa esta metodología. Es un modelo de elicitación y representación plausible computacionalmente, que puede ser de gran utilidad en el objetivo de interoperabilidad semántica planteado por la Semantic Web.

4 Conclusiones

La metodología de elicitación y comparación de Ontologías Naturales propuesta en este trabajo, es una aproximación que puede ser de utilidad para las tecnologías destinadas a la interoperabilidad semántica; ya que además de generar una representación del significado manejado por humanos basada en una sólida teoría cognitiva (RSN), es fácilmente automatizable, tal como se demostró mediante la utilización del software para obtener la Ontología ejemplificada.

La técnica RSN permite identificar los conceptos, sus relaciones y pesos semánticos que usa una comunidad. Su estructura se forma directamente de los datos proporcionados por grupos de sujetos, permitiendo normalizar y comparar el conocimiento. Pueden ser la base para la construcción y/o anotación de ontologías.

Referencias

1. Tim Berners-Lee, Jim Hendler, and Ora Lassila. The semantic web. Scientific American, 284(5):34–43, 2001.
2. R. J. Brachman. What is-a is and isn't: An analysis of taxonomic links in semantic networks. IEEE Computer, 16(10):30–36, 1983.
3. R.J. Brachman. On the epistemological status of semantic networks. In Associative Networks: Representations and Use of Knowledge by Computers. Findler, N. V., 1979.
4. C. Bustillo-Hernandez y J. G. Figueroa. Procedimiento para la extracción y normalización de conocimiento en humanos: Una aproximación para la anotación de ontologías en la semantic web. Memorias ROC&C 2009, IEEE Sección México., 2009.
5. Allan M. Collins and M. Ross Quillian. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8:240–247, 1969.
6. M. Cristani and R. Cuel. A survey on ontology creation methodologies. International Journal on Semantic Web and Information Systems, 2005.
7. Yang Dongqiang. Lexical Semantic Similarity: Word Similarity in Semantic Networks and Distributional Structures. Verlag Dr. Müller, 2009.

8. Andrew Dravnieks. Odor quality: semantically generated multidimensional profiles are stable. Science, 218:799–801, 1982.
9. Marc Ehrig. Ontology Alignment: Bridging the Semantic Gap. Springer, 2007.
10. Christiane Fellbaum, editor. WordNet: An Electronic Lexical Database. MIT Press, 1998.
11. Jesús G. Figueroa, Esther G. González, and Victor M. Solís. An approach to the problem of meaning: Semantic networks. Journal of Psycolinguistic Research, 5(2):107–115, 1976.
12. A. Hammed, D. Sleeman, and A. Preece. Detecting mismatches among experts ontologies acquired through knowledge elicitation. Elsevier Knowledge Based Systems, 15:265–273, 2002.
13. K. Jung-Min, Byoung-II C., S. Hyo-Phil, and K. Hyoung-Joo. A methodology for constructing of philosophy ontology based on philosophical texts. Elsevier Computer Standards and Interfaces, 29:302–315, 2007.
14. Y.V. Kapitonova. General principles of construction of knowledge computer systems. Cybernetics and Systems Analysis, 42:531–546, 2006.
15. David E. Meyer and Roger W. Schvaneveldt. Meaning, memory structure and mental processes. Science, 192:27–33, Abril 1976.
16. E.V. Ortega-Gonzalez. Una técnica para el análisis de similitud entre imágenes. Tesis de Maestría. Centro de Investigación en Computación. Instituto Politécnico Nacional., 2007.
17. V. A. Ramírez, A. Zermeño, y A. C. Arellano. Redes semánticas naturales: Técnica para representar los significados que los jóvenes tienen sobre televisión, internet y expectativas de vida. Estudios sobre las culturas contemporáneas, 22:305–334, 2005.
18. John Sowa, editor. Principles of Semantic Networks. The Morgan Kaufman Series in Representation and Reasoning. Morgan Kaufmann Publishers, Inc., 1991.
19. S. Staab and R. Studer, editors. Handbook on Ontologies. International Handbooks on Information Systems. Springer, 2 edition, 2009.
20. J. Gamboa Suasnavart, J. Osorio Reséndiz, R. Lom Romero, O. Tapia Leyva, E. Vargas Medina, y J. Figueroa Nazuno. Las redes semánticas naturales como procedimiento para extracción de conocimiento, para su uso en interoperabilidad semántica. Memorias ROC&C 2008, IEEE Sección México., pag. 42, 2008.
21. Endel Tulving and Daniel L. Schacter. Priming and human memory systems. Science, 247:301–306, 1990.
22. J. L. Valdez y Reyes L. Las categorías semánticas y el autoconcepto. Revista Psicología Social en México, IV:193–199, 1992.
23. J. L. Valdez-Medina. Las redes semánticas naturales, uso y aplicaciones en psicología social. Universidad Autónoma del Estado de México, 2004.
24. E. Vargas-Medina y C. Calzada-Ugalde. Evolución de la representación conceptual de la física en estudiantes universitarios y preuniversitarios. Revista del Centro de Investigación, Universidad La Salle, 1(2):41–61, 1994.