Validación de conceptos ontológicos usando métodos de agrupamiento

Mireya Tovar[1,2], David Pinto[2], Azucena Montes[1,3], Gabriel González[1], Darnes Vilariño[2], Beatriz Beltrán[2]

[1]Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, México

[2] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, México

[3] Instituto de Ingeniería, Universidad Nacional Autónoma de México, México

{mtovar, amontes, gabriel}@cenidet.edu.mx, {dpinto, darnes,bbeltran}@cs.buap.mx

Resumen. En este artículo proponemos un enfoque para validar la información existente en una ontología de dominio mediante la identificación de conceptos sobre un corpus asociado. El mecanismo propuesto está basado en la determinación del grado de cercanía de los conceptos existentes en la ontología. Para llevar a cabo este proceso, inicialmente se representa cada concepto usando la información contextual, y posteriormente se usa dicha representación para agrupar la información obtenida. Dado que la matriz de representación suele ser de alta dimensionalidad, se usan técnicas de análisis semántico latente para reducir la dimensionalidad y permitir un proceso más eficiente de la información. El proceso de agrupamiento se lleva a cabo usando la técnica conocida como "agrupamiento por comités". Los resultados experimentales muestran un comportamiento satisfactorio para las dos ontologías revisadas en este trabajo.

Palabras clave: Ontologías de dominio, conceptos, análisis semántico latente, agrupamiento.

1. Introduccin

En los últimos años la representación del conocimiento y las ontologías han ganado importancia. Las ontologías se han utilizado en una gran diversidad de aplicaciones, entre las cuales podemos mencionar las siguientes: la comunicación de agentes [13], en el descubrimiento de servicios web [7], en sistemas de recuperación de información [8], en sistemas de pregunta-respuesta [1], y en el procesamiento del lenguaje natural [15].

Una ontología se define como "una especificación explícita y formal de una conceptualización compartida" [6]. En general, este tipo de recurso semántico está formado por conceptos o clases, relaciones, instancias, atributos, axiomas, restricciones, reglas y eventos. Las ontologías de dominio son un sistema de

representación del conocimiento que se pueden organizar en estructuras taxonómicas y ontológicas de conceptos de algún área o dominio de conocimiento específico. Análogamente, podemos decir que un corpus de dominio es aquel que está formado por textos de carácter específico.

El aprendizaje de ontologías o generación automática de ontologías, es un proceso que puede facilitar la construcción automática o semiautomática de las mismas. El término aprendizaje de ontologías se atribuye originalmente a Alexander Mädche y Steffen Staab [12] y se describe como la adquisición de un modelo de dominio desde los datos. El aprendizaje de ontologías necesita datos de entrada, como textos estructurados o no estructurados, desde los cuales se aprenden conceptos relevantes para un dominio específico, sus definiciones y relaciones establecidas entre estos. El aprendizaje de ontologías a partir de textos no estructurados se le conoce simplemente como "aprendizaje de ontologías desde textos" [4].

La construcción automática o semi-automática de ontologías es una área de trabajo ampliamente estudiada, sin embargo, la validación de la información contenida en los recursos obtenidos no lo es tanto. En la mayoría de los casos se asume que, por ejemplo, los conceptos encontrados son correctos en su mayoría. De ahí la intención de construir métodos que permitan validar la calidad de este tipo de recursos construidos.

En particular, en este trabajo estamos interesados en la identificación de conceptos de dominio en textos no estructurados. Partimos de la suposición de que los conceptos que están semánticamente relacionados, tienden a estar "cercanos" en un texto. Por lo tanto, un concepto se define como una idea que forma el entendimiento[1]. Desde el punto de vista de la filosofía, un concepto es una unidad de ideas que consiste de dos partes, la extensión y la intensión [9]. Cimiano [4] concibe la intensión como una definición no extensional de un cierto concepto o relación. Es decir, describir intuitivamente el significado de un concepto en lenguaje natural, como las glosas de recursos léxicos como WordNet [14]. La parte extensional es proporcionada por una base de conocimiento que contiene afirmaciones acerca de las instancias de los conceptos y las relaciones como se definen en la ontología. Por ejemplo, un "animal es un ser orgánico que vive, siente y se mueve por propio impulso"[2], una instancia para este concepto es "araña", una relación léxica entre un par de conceptos como mamífero y animal es, por ejemplo, hiperonímia ("un mamífero es un animal").

Para la extracción o descubrimiento de conceptos, algunos autores han considerado algoritmos como el análisis de conceptos formales (FCA, *Formal Concept Analysis*) y construyen jerarquías de conceptos al mismo tiempo [5]. Algunos otros autores han considerado enfoques de agrupamiento y consideran a los grupos de términos relacionados como conceptos [11], [18]. Otros aplican técnicas de reducción de dimensiones tales como el análisis semántico latente (LSA) [10], que revelan conexiones inherentes entre palabras, lo que conduce a la formación de grupos. En particular, el enfoque propuesto verifica los conceptos existentes

[1] Definición de concepto en la real academia española; (`http://www.rae.es`).
[2] Definición de animal en la real academia española; (`http://www.rae.es`).

en dos ontologías de dominio y en sus correspondientes corpus de dominio que están formados por documentos no estructurados. Para la identificación de conceptos se utiliza primeramente LSA para reducir la dimensionalidad de una matriz de representación, cuya versión reducida es introducida en un método de agrupamiento basado en comités (CBC, *Clustering By Committee*). Se parte de la hipótesis de que los conceptos que estan semánticamente relacionados, tienden a estar "cercanos" en un contexto y/o diferentes contextos.

El resto de este artículo se organiza como sigue: en la sección 2 se presenta el método LSA y algunos trabajos relacionados con la identificación de conceptos. En la sección 3 se muestra el algoritmo de agrupamiento basado en comités. En la sección 4 se presenta nuestra propuesta para la identificación de conceptos. En la sección 5 se muestran los resultados experimentales de la propuesta. Finalmente, las conclusiones se presentan en la sección 6.

2. Análisis semántico latente

El análisis semántico latente o *Latent Semantic Analysis* (LSA) es un modelo computacional utilizado en procesamiento de lenguaje natural, considerado en sus inicios como un método de representación del conocimiento [22].

LSA se considera una herramienta no supervisada de reducción de la dimensionalidad, como el análisis de los componentes principales (PCA, *principle component analysis*) [20]. Parte de la idea de que palabras en el mismo campo semántico tienden a aparecer juntas o en contextos similares [10], [21].

LSA tiene su origen en una técnica de recuperación de información llamada LSI (*Latent Semantic Indexing*) cuyo propósito es reducir la dimensión de una matriz de términos-documentos utilizando una técnica de algebra lineal llamada descomposición de valores singulares (SVD, *Singular Value Decomposition*). La diferencia con LSA, es que ésta utiliza una matriz de palabra-contexto. El contexto puede ser una palabra, una oración, un párrafo, un documento, un ensayo, etc.

Venegas [22] considera que LSA se caracteriza por ser una técnica matemático-estadística que permite la creación de vectores multidimensionales para el análisis semántico de las relaciones existentes entre los diferentes contextos.

El propósito de la reducción de la dimensionalidad es eliminar el ruido presente en las relaciones existentes entre los términos y los contextos, dado que generalmente es posible expresar el mismo concepto con distintos términos.

LSA no considera la estructura lingüística de los contextos, sólo las frecuencias de aparición y co-ocurrencia de los términos. Sin embargo, usando LSA se ha logrado en algunos casos identificar relaciones semánticas como sinonímia [10].

3. Agrupamiento por comités

El algoritmo de agrupamiento por comités (CBC, *Clustering By Committee*) permite descubrir automáticamente conceptos a partir de textos [11,18]. Inicialmente descubre un conjunto de grupos estrictos llamados comités que estan

dispersos en el espacio de similitud. El vector de características del grupo es el centroide de los miembros del comité y se procede a asignar elementos a sus grupos más similares.

El algoritmo CBC consiste de tres fases:

1. Encontrar los elementos más similares. Para calcular las palabras más similares de una palabra w, primero se ordenan las características de la palabra w de acuerdo a su información mutua con w.
2. Encontrar los comités. Cada comité que se descubre en esta fase define uno de los grupos finales de la salida del algoritmo.
3. Asignar elementos a los grupos. Cada elemento se asigna al grupo que contiene al comité más similar.

CBC también se ha utilizado para encontrar los sentidos de una palabra w [16] (algoritmo en su versión flexible), y para agrupamiento de textos (algoritmo en su versión fuerte) [17]. Otros autores, como Chatterjee y Mohan [3], han utilizado con éxito este algoritmo en su versión flexible para el descubrimiento de sentidos de las palabras, incluyendo además *Random Indexing* para disminuir la dimensionalidad de la matriz de contextos.

4. Enfoque propuesto para la identificación de conceptos

El enfoque que se propone en este artículo, para la identificación de conceptos de ontologías de dominio en sus correspondientes corpus de dominio, realiza los siguientes pasos:

1. Preprocesamiento del corpus de dominio y de las ontologías de dominio. El corpus de dominio se divide en oraciones y se eliminan palabras cerradas o vacías (como preposiciones, artículos, etc). El algoritmo de truncamiento de Porter [19] también se aplica sobre las palabras contenidas en estas oraciones. Los conceptos y relaciones de las ontologías son extraídos utilizando Jena[3]. El mismo proceso es aplicado a cada uno de los conceptos de la ontología con la finalidad de mantener consistencia en la representación terminológica (eliminación de palabras vacías y el algoritmo de truncamiento de Porter).
2. Aplicación del algoritmo LSA para disminuir la dimensionalidad de la matriz de contextos. En este caso, se utiliza del paquete S-Space[4], el algoritmo LSA[5]. El algoritmo recibe como párametros las oraciones del corpus de dominio, la lista de conceptos de la ontología y la cantidad K de dimensiones (que en nuestro caso fue definida como 300). La salida del algoritmo LSA son vectores semánticos de dimensión K para cada palabra o concepto identificado por LSA en el corpus.
3. Aplicación del algoritmo CBC en su versión flexible. La salida de LSA (entrada del algoritmo CBC) son las palabras y conceptos agrupados.

[3] `http://jena.apache.org/`
[4] `https://github.com/fozziethebeat/S-Space`
[5] http://code.google.com/p/airhead-research/wiki/LatentSemanticAnalysis

4. Identificación de los conceptos de la ontología en los grupos generados por el algoritmo de agrupamiento por comités.

En la siguiente sección se presentan los resultados obtenidos del enfoque propuesto.

5. Experimentos

En esta sección se presentan los datos utilizados y los resultados obtenidos en los experimentos. Primeramente mostramos la información asociada con los conjuntos de datos usados en los experimentos (ontologías y corpora asociado) para posteriormente mostrar los resultados obtenidos. El criterio de evaluación que se considera en la validación de conceptos es el de exactitud [2]. Es decir, se determina la cantidad de conceptos generados por el enfoque que existen en las ontologías de dominio.

5.1. Conjunto de datos

En los experimentos, por el momento, sólo se consideran dos ontologías debido a que están libremente disponibles y sus corpora contienen información asociada a los conceptos y relaciones que permiten afirmar o negar su validez. Como trabajo a futuro se considera incluir más ontologías y formar sus corpora correspondientes.

Los dominios de las ontologías son: inteligencia artificial (AI) y estándar e-Learning SCORM (SCORM)[6] [23].

Cada ontología contiene un número determinado de conceptos (C), relaciones tipo class-inclusion (S) y relaciones ontológicas (R). Los documentos (D) de los corpora de dominio asociados a cada ontología fueron utilizados para determinar la cantidad de palabras (P), el vocabulario (V) que excluye de P las palabras vacías como artículos, preposiciones, etc., y el número de oraciones (O) (ver Tabla 1).

Tabla 1. Conjunto de datos

Dominio	Ontología			Corpora			
	C	S	R	D	P	V	O
AI	276	205	61	8	10,797	2,180	519
SCORM	1,461	1,038	759	36	32,572	2,154	1,779

[6] Las ontologías y sus corpora correspondientes están disponibles en la página http://azouaq.athabascau.ca/goldstandards.htm

5.2. Resultados

El algoritmo de agrupamiento CBC es capaz de obtener grupos de conceptos relacionados como los que se muestran en la Tabla 2 para el dominio de Inteligencia Artificial. Algunos conceptos relacionados con el concepto "Agent" son por ejemplo: *neural network, action*, que en la ontología son identificados como relaciones ontológicas, y *entire agent, reflex agent, abstract intelligent agent, individual agent* los cuales son identificados como relaciones tipo "class-inclusion".

Tabla 2. Ejemplo de conceptos agrupados por CBC.

Concepto	Conceptos relacionados
Agent	Neural network, action, entire agent, reflex agent, abstract intelligent agent, rational agent, individual agent, planning problem, system, intelligent action, etc.
Artificial Intelligence	Strong ai, intelligence, field of science, human reasoning, etc.

El total de vocabulario (palabras y conceptos) encontrado por el algoritmo LSA es de 1,537 para la ontología de Inteligencia Artificial (AI) y 2,049 para la ontología SCORM (LSA_V). Los conceptos encontrados ($C\_Exactos$) en los grupos generados por el algoritmo CBC es de 222 de los 276 conceptos existentes en la ontología de AI, logrando así un 80 % de exactitud (ver Tabla 3). En el caso de la ontología SCORM, al utilizar el algoritmo LSA, sólo se logró identificar 539 conceptos de los 1,461 conceptos definidos en la ontología de dominio, obteniendo sólo el 37 % de exactitud.

Tabla 3. Resultados experimentales

Dominio	C	LSA_V	$C\_Exactos$	%
AI	276	1,537	222	0.80
SCORM	1,461	2,049	539	0.37

En base a los resultados observamos que el método LSA logra identificar conceptos relacionados en cada ontología de dominio. El algoritmo de agrupamiento permite realizar una búsqueda más profunda de aquellos conceptos que están asociados con otros conceptos o relaciones de la ontología, por lo cual puede resultar útil en la tarea de validación.

6. Conclusiones

En este artículo se presenta un enfoque que permite validar conceptos de dos ontologías de dominio (inteligencia artificial y estándar e-Learning SCORM) a partir de la identificación de los mismos en un corpus de referencia asociado a

cada ontología analizada. El enfoque propuesto utiliza un algoritmo de reducción de la dimensionalidad de las características de cada concepto de las ontologías (LSA). Además se incluye un algoritmo de agrupamiento (CBC) que permite realizar una asociación más directa entre los conceptos identificados por el algoritmo LSA. En base a los resultados experimentales, se observa que la ontología de inteligencia artificial es más estable al encontrar por lo menos el 80 % del total de los conceptos. La ontología SCORM presenta una disminución en el cantidad de conceptos obtenidos por el enfoque, debido a que el algoritmo LSA sólo produce 2,049 palabras y/o conceptos de esta ontología que es una cantidad mucho menor comparada con los 1,537 que se obtienen para la ontología IA. También se observa que al analizar los conceptos asociados a cada concepto de la ontología se definen relaciones de tipo "class-inclusion" y ontológicas. Lo que permitiría extender las ontologías de dominio al realizar un análisis más profundo de cada concepto agrupado.

Como trabajo a futuro, consideramos revisar a profundidad las asociaciones existentes entre cada concepto agrupado y hacer una revisión de los conceptos que no fueron detectados para la ontología SCORM.

Agradecimientos. Este trabajo de investigación ha sido parcialmente financiado por el Consejo Nacional de Ciencia y Tecnología (CONACYT) con el número de becario 54371, por el Programa para el Mejoramiento del Profesorado (PROMEP) con número de convenio PROMEP/103.5/12/4962 BUAP-792 y a través del proyecto CONACYT 106625.

Referencias

1. Beale, S., Lavoie, B., McShane, M., Nirenburg, S., Korelsky, T.: Question answering using ontological semantics. In: Proceedings of the 2Nd Workshop on Text Meaning and Interpretation. pp. 41–48. TextMean '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
2. Cantador, I., Ferández, M., Castells, P.: A collaborative recommendation framework for ontology evaluation and reuse. In: Actas de International Workshop on Recommender Systems, en la 17th European Conference on Artificial Intelligence (ECAI 2006), Riva del Garda, Italia. pp. 67–71 (2006)
3. Chatterjee, N., Mohan, S.: Discovering word senses from text using random indexing. In: Gelbukh, A.F. (ed.) CICLing. Lecture Notes in Computer Science, vol. 4919, pp. 299–310. Springer (2008)
4. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Studies in philosophy and religion, Springer (2006)
5. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. J. Artif. Int. Res. 24(1), 305–339 (Aug 2005)
6. Gruber, T.R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In: Guarino, N., Poli, R. (eds.) Formal Ontology in Conceptual Analysis and Knowledge Representation. Kluwer Academic Publishers, Deventer, The Netherlands (1993)

7. Ji, X.: Research on web service discovery based on domain ontology. In: Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on. pp. 65–68 (Aug 2009)
8. Jimenez Muñoz, R.J.: Un sistema de búsqueda semántica de información para su uso en el dominio de recuperación mejorada en yacimientos petroleros. Master's thesis, Fac. Ciencias de la Computación, BUAP, Puebla, Mex. (2013)
9. Krings, H. (ed.): Handbuch philosophischer Grundbegriffe. Kösel, München, studienausg. edn. (1973)
10. Landauer, T.K., Dutnais, S.T.: A solution to platoś problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review pp. 211–240 (1997)
11. Lin, D., Pantel, P.: Concept discovery from text. In: Proceedings of the 19th International Conference on Computational Linguistics - Volume 1. pp. 1–7. COLING '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
12. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems 16(2), 72–79 (Mar 2001)
13. Malucelli, A., Costa Oliveira, E.: Ontology-services to facilitate agents interoperability. In: Lee, J., Barley, M. (eds.) Intelligent Agents and Multi-Agent Systems. Lecture Notes in Computer Science, vol. 2891, pp. 170–181. Springer Berlin Heidelberg (2003)
14. Miller, G.A.: Wordnet: A lexical database for english. COMMUNICATIONS OF THE ACM 38, 39–41 (1995)
15. Nirenburg, S., Raskin, V.: Ontological Semantics. Language, speech, and communication, MIT Press (2004)
16. Pantel, P., Lin, D.: Discovering word senses from text. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 613–619. KDD '02, ACM, New York, NY, USA (2002)
17. Pantel, P., Lin, D.: Document clustering with committees. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 199–206. SIGIR '02, ACM, New York, NY, USA (2002)
18. Pantel, P.A.: Clustering by committee. Ph.D. thesis, University of Alberta (2003)
19. Porter, M.F.: An algorithm for suffix stripping. In: Sparck Jones, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)
20. Sidorov, G.: Non-linear construction of n-grams in computational linguistics: syntactic, filtered, and generalized n-grams. Sociedad Mexicana de Inteligencia Artificial, Mexico (2013)
21. Vázquez Pérez, S.: Resolucin de la ambigedad semntica mediante mtodos basados en conocimiento y su aportacin a tareas de PLN. Ph.D. thesis, Universidad de Alicante (abril 2009)
22. Venegas V., R.: Análisis Semántico Latente: una panorámica de su desarrollo. Revista signos 36, 121 – 138 (00 2003)
23. Zouaq, A., Gasevic, D., Hatala, M.: Linguistic patterns for information extraction in ontocmaps. In: Blomqvist, E., Gangemi, A., Hammar, K., del Carmen Suárez-Figueroa, M. (eds.) WOP. CEUR Workshop Proceedings, vol. 929. CEUR-WS.org (2012)