ISSN 1870-4069

Identificación de relaciones taxonómicas de dominio usando métricas textuales

Yuridiana Alemán, María Somodevilla, Darnes Vilariño

Benemérita Universidad Autónoma de Puebla (BUAP),
Facultad de Ciencias de la Computación,
México

{yuridiana.aleman,mariajsomodevilla,dvilarinoayala}@gmail.com

Resumen. El proceso de aprendizaje de ontologías comprende tres pasos fundamentales: creación de clases y relaciones, población y evaluación. Este documento se enfoca en la creación de clases y relaciones, realizando un estudio sobre la detección de subclases para la ontología. Como caso de estudio se seleccionó un dominio pedagógico, donde se construyó un corpus semiautomático, a partir de artículos escritos en español publicados en el área de Ciencias Sociales. Para la detección de subclases fueron implementadas cuatro métricas de similitud textual basadas en términos, con las cuáles se construyó una heurística para determinar cuáles de los conceptos tienen posibilidades de convertirse en una subclase de la ontología y tienen una relación taxonómica con la clase principal. La evaluación se realizó mediante un *gold* verificado por un experto en el dominio y el contexto teórico de las clases y se obtuvo el recuerdo de cada clase. Los resultados muestran un recuerdo de 100 % 72 % y 67 % respectivamente para cada una de las clases, además de que se recuperaron conceptos relacionados a la clase principal mediante relaciones no taxonómicas.

Palabras clave: métricas semánticas, ontología, conceptos principales, relación *Is_ a*, pedagogía, dice, Jaccard, traslape, coseno.

Domain Taxonomical Relationships Identification Using Textual Metrics

Abstract. The ontology learning process comprises three fundamental steps: creation of classes and their relationships, population and evaluation. This document focuses on the first step, by performing a study on the detection of ontology subclasses. As a case study was selected a pedagogical domain. A semiautomatic corpus, based on articles written in Spanish published in Social Sciences journals was built. Four textual similarity metrics based on terms for the detection of subclasses were implemented, with which a heuristic was conducted to determine which of these concepts have the potential to become in a subclass of the

ontology and at the same time keep a taxonomic relationship with the superclass. The evaluation was carried out through a gold, which was validated by an expert in the theoretical context domain. The results show a recall of 100 % ,72 % and 67 % for each class respectively. In addition, concepts related to the super classes were recovered through non-taxonomic relationships.

Keywords: semantic metrics, ontology, class, is_a relation, pedagogy, dice, Jaccard, overlap, cosine.

1. Introducción

La información disponible en diversos repositorios va en aumento, especialmente en la Web. Por lo tanto, es necesario implementar técnicas para procesar esa información y relacionarla con otros repositorios a fin de incrementar el conocimiento extraído. Las ontologías se presentan como una opción para procesar esta información, que se pueden utilizar para gestión del vocabulario, aplicaciones de procesamiento del lenguaje natural, búsquedas, sistemas de recomendación, e-learning, entre otros [5]. El proceso de aprendizaje de la ontología integra la detección de clases, la creación, la población y la evaluación [6].

Este documento se centra en el primer paso del proceso de aprendizaje ontológico: la detección de clases. En investigaciones previas se trabajó con la detección y validación de clases principales, por lo que este artículo se centra en la detección de las subclases y relaciones entre conceptos. Para los experimentos se utiliza un corpus formado por documentos pedagógicos en español. El dominio pedagógico es extenso, por lo que la investigación consiste en la creación de herramientas que respalden las clases de los profesores en el aula. El corpus contiene tres clases principales: estilos de aprendizaje, tipos de inteligencias y estrategias de enseñanza-aprendizaje. Cada una de estas clases principales se subdivide de acuerdo a enfoques teóricos propuestos en la literatura pedagógica. El método propuesto utiliza métricas de similitud textual para extraer los términos más relacionados con cada una de las clases principales, tomando estos como subclases. En el proceso también se recuperan términos importantes que figuran como relaciones no taxonómicas con la clase principal.

El artículo está organizado en siete secciones que se describen a continuación. La sección 2 presenta los trabajos relacionados con la detección de clases y relaciones en el proceso de construcción de ontologías. La sección 3 describe teóricamente las ontologías, así como las tres clases principales del corpus. La sección 4 presenta las métricas de similitud textual utilizadas en los experimentos. La sección 6 presenta la metodología propuesta y la sección 7 muestra el análisis de los resultados. Finalmente, la sección 8 presenta las conclusiones y el trabajo futuro de la investigación.

2. Trabajos relacionados

Es importante mencionar que antes de iniciar el proceso de extracción de elementos principales, se debe tener un corpus para el dominio a trabajar, por lo que se analizaron algunas investigaciones sobre la construcción de corpus en diferentes dominios. EL trabajo que se discute en [12] se centra en la creación de un corpus lingüístico relevante escrito en lengua serbia, en dicha investigación, el enfoque es el análisis de sentimientos de los contenidos generados por los estudiantes en la educación superior. En el trabajo realizado en [22] se analizó el problema de crear un corpus de referencia para la clasificación de artículos de noticias en escenarios de etiquetas múltiples. Los autores proponen un enfoque semiautomático para crear un corpus de referencia que utiliza tres métodos de clasificación auxiliares: máquinas de vectores de soporte, clasificadores de vecinos más cercanos y otro basado en un diccionario.

En investigaciones como [20] se presentan métodos para la extracción de clases de manera semiautomática, utilizan una base de datos de verbos, alternancias de diátesis y esquemas sintáctico-semánticos del Español (ADESSE) [7] la cual contiene aproximadamente 160,000 cláusulas recuperadas de un corpus; con la ayuda de ADESSE se extraen patrones semánticos que llevan a la determinación de las clases para una ontología. Esta metodología fue aplicada en un subdominio educativo y replicada en ámbito financiero [19]. La extracción de clases se complementó con la opinión de expertos en el dominio. En [16] se presenta un método para la extracción de conceptos utilizando extracción de patrones lingüísticos y cálculo de pesos con métricas de procesamiento de lenguaje natural como el etiquetado morfológico.

Dentro del ámbito pedagógico, en la investigación de [11] se propone el proyecto OURAL (*Ontologies for the Use of digital learning Resources and semantic Annotations on Line*) el cual integra las disciplinas de ciencias de la educación, informática y psicología cognitiva con el fin de crear servicios para *E-learning*. Como resultados, se muestran las clases obtenidas mediante la aplicación de técnicas de PLN a situaciones de aprendizaje descritas en lenguaje natural. Este dominio se analiza también en [6], sin embargo, al ser aplicado al idioma Chino, utilizan un preprocesamiento para analizar las características de dicho idioma: acoplamiento, relevancia y consenso.

Otras investigaciones se centran en la educación en línea como [3], [4], [15] y recientemente [14], donde las ontologías se definen manualmente a partir de recursos XML disponibles en Internet, y la evaluación también es un proceso manual. Investigaciones como [23] se enfocan en el aprendizaje automático; en este trabajo, se crea una ontología basada en Internet de las cosas utilizado en un aula, teniendo en cuenta las inteligencias estudiantiles. [18] propone utilizar un modelo ontológico para la personalización del aprendizaje que involucre el perfil de los estudiantes de acuerdo con la teoría de inteligencia múltiple de Howard Gardner, así como usar una ontología de dominio que ayude a representar el conocimiento en plataformas de aprendizaje virtuales.

3. Ontologías de dominio

En las ciencias computacionales, una ontología se define como una especificación formal de una conceptualización [13]. Es una base de datos que describe los conceptos en el mundo o algún dominio, algunas de sus propiedades y cómo los conceptos se relacionan entre sí [24]. Esta base de datos se define a partir de un corpus base, en donde se extraen los elementos principales o palabras clave. Posteriormente, del mismo texto se infieren las relaciones entre palabras clave, de esta manera, se crea una estructura de grafo donde los nodos son las palabras clave y las aristas representan la relación existente entre ellas.

Entre las aplicaciones más representativas de las ontologías se encuentran la representación formal del conocimiento, lo que facilita el manejo e integración de datos con estructuras diferentes. Formalmente, una ontología se define como la sextupla $O = (C, H, I, R, P, A)$ [5] donde:

- C es el conjunto de entidades de la ontología,
- H son las relaciones taxonómicas entre los conceptos,
- I indica las relaciones entre instancias,
- R en el conjunto de relaciones no taxonómicas,
- P es el conjunto de propiedades de la ontología,
- A representa el conjunto de axiomas y reglas que prueban la consistencia de la ontología que realizan el proceso de inferencia.

Una relación de ontología es una formalización de las manera en que las entidades están asociadas. La relación que se analiza en esta investigación es la de tipo *Is-a*, la cual es un vínculo entre clases en forma de jerarquía, una columna vertebral de una ontología. La organización jerárquica de las entidades, siempre por relaciones *Is-a*, permite la herencia de propiedades y la estructuración de la taxonomía.

3.1. Dominio pedagógico

Dado que el dominio pedagógico es muy extenso, se establecieron tres clases principales a fin de obtener una herramienta de apoyo para el docente en clases presenciales.

Estilos de aprendizaje. Los estilos de aprendizaje reflejan la forma en que el individuo aprende. Existen variaciones en cuanto a la manera en que los seres humanos captan y procesan información. Se han propuesto varias teorías para describir los distintos tipos de aprendizaje, para esta investigación se tomó como referencia el modelo de David Kolb [17], en el cual se determina un estilo de aprendizaje usando una escala denominada *Learning Style Inventory* (LSI). La teoría propone un método para describir cómo los estudiantes resuelven sus problemas y aplican conocimientos nuevos a partir de la experiencia personal dentro de su entorno de aprendizaje. Considera los procesos psicológicos de percepción y procesamiento [21]. El método propone 4 estilos de aprendizaje: activo, reflexivo, pragmático y teórico.

Inteligencias múltiples. La inteligencia se define como "la capacidad de resolver problemas o de crear productos que sean valiosos en uno o más ambientes culturales" [8]. Los seres humanos poseen una gama de capacidades y potenciales que se pueden emplear de muchas maneras productivas, tanto juntas como por separado, esta idea da origen a las inteligencias múltiples, las cuales han sido identificadas por Gardner: lógico-matemática, lingüística, espacial, musical, corporal, intrapersonal, interpersonal y naturalista.

Estrategias de aprendizaje. Una estrategia de aprendizaje es un conjunto de procedimientos que un alumno usa de manera consciente, controlada e intencional como herramientas flexibles para aprender y resolver problemas [1], también pueden ser definidas como conductas y pensamientos que un aprendiz utiliza durante el aprendizaje con la intención de influir en su proceso de codificación [25]. Aunque existen muchos enfoques para la clasificación de las estrategias de aprendizaje, [10] menciona tres principales tipos de estrategias: cognitivas, metacognitivas y las estrategias de manejo de recursos.

4. Métricas de similitud textual

La tarea de similitud textual se encarga de comparar textos para conocer el parecido entre ellos. Para lograr este objetivo, se han propuesto en la literatura métricas que comparan la proximidad entre las palabras o caracteres de dos textos.

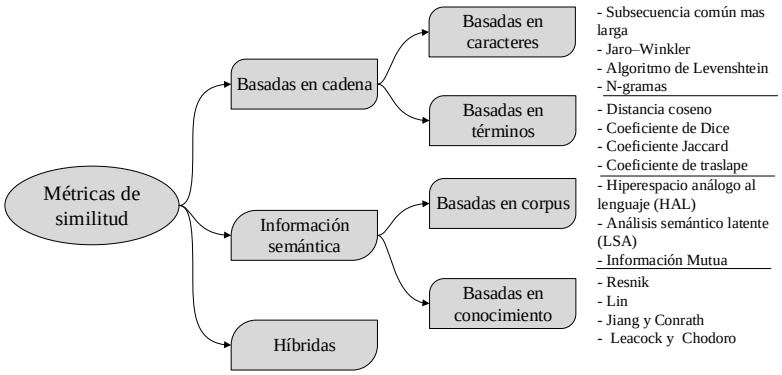

Fig. 1. Clasificación de las métricas de similitud textual según [3 y 10].

La Figura 1 muestra la clasificación propuesta por dos autores. Se presentan 3 clases principales: métricas basadas en cadenas, basadas en información semántica e híbridas.

Las métricas basadas en cadena contienen los enfoques basados en caracteres y en términos, mientras que las basadas en información semántica integran las métricas basadas en corpus y basadas en conocimiento.

Para la presente investigación se trabajan con las métricas basadas en términos. Las métricas basadas en caracteres pierden información al manejar un corpus lematizado; las métricas basadas en corpus suelen obtener resultados altos, pero son costosas en su implementación y necesitan un corpus extenso para calcular el valor de la co-ocurrencia de cada par de palabras [2]. Las métricas basadas en conocimientos están basadas en *WordNet*, y son utilizadas para el idioma inglés, por lo que no son pertinentes para esta investigación. Las métricas basadas en términos solo necesitan el corpus de entrada, aunque a mayor tamaño del texto se espera más exactitud en los resultados, aun así requieren menos recursos que las basadas en corpus. Las métricas más citadas en la literatura se describen en los siguientes párrafos.

Coeficiente de Jaccard. Se obtiene al dividir la intersección de términos entre la unión de los mismos. Su fórmula se presenta en la ecuación 1 y da por resultado el grado de superposición de dos conjuntos, en este caso, de dos N-gramas:

$$sim_J(t_1, t_2) = \frac{\mid t_1 \cap t_2 \mid}{\mid t_1 \cup t_2 \mid}. \tag{1}$$

Coeficiente de Dice. Se basa en la teoría de conjuntos. Toma el número de las palabras que comparten ambas cadenas y los divide entre el número total de la suma de las palabras del texto uno y dos. Su cálculo está determinado por la ecuación 2. El coeficiente Dice da el doble de peso a las coincidencias positivas entre los términos. El resultado está normalizado entre cero y uno donde cero es nula similitud, mientras que uno se refiere a la máxima similitud [1]:

$$sim_D(t_1, t_2) = 2\frac{\mid t_1 \cap t_2 \mid}{\mid t_1 \mid + \mid t_2 \mid}. \tag{2}$$

Coeficiente de Traslape. Considera la cardinalidad de caracteres del texto más pequeño en lugar de la unión de los caracteres [9]. Para esta métrica, una coincidencia completa de dos cadenas es cuando una es un subconjunto de la otra, ecuación 3:

$$sim_T(t_1, t_2) = \frac{\mid t_1 \cap t_2 \mid}{min(\mid t_1 \mid, \mid t_2 \mid)}. \tag{3}$$

Coeficiente de Coseno. Se obtiene dividiendo la cardinalidad de la unión de los dos conjuntos entre la raíz cuadrada del producto de las cardinalidades de los conjuntos considerados, ecuación 4:

$$sim_C(t_1, t_2) = \frac{\mid t_1 \cap t_2 \mid}{\sqrt{\mid t_1 \mid \mid t_2 \mid}}. \tag{4}$$

El coeficiente de Dice y de traslape son similares al coeficiente Jaccard, solo que el coeficiente de Dice da el doble de peso a las coincidencias positivas entre los

términos, y el coeficiente de traslape considera solo la cardinalidad de caracteres del texto más pequeño en lugar de la unión de los caracteres, como lo hace el coeficiente de Jaccard.

5. Metodología

La Figura 2 muestra la propuesta para la obtención de las subclases y las relaciones en el corpus. En la parte superior se muestra la metodología general, mientras que la flecha segmentada dirige al método seguido para esta investigación.

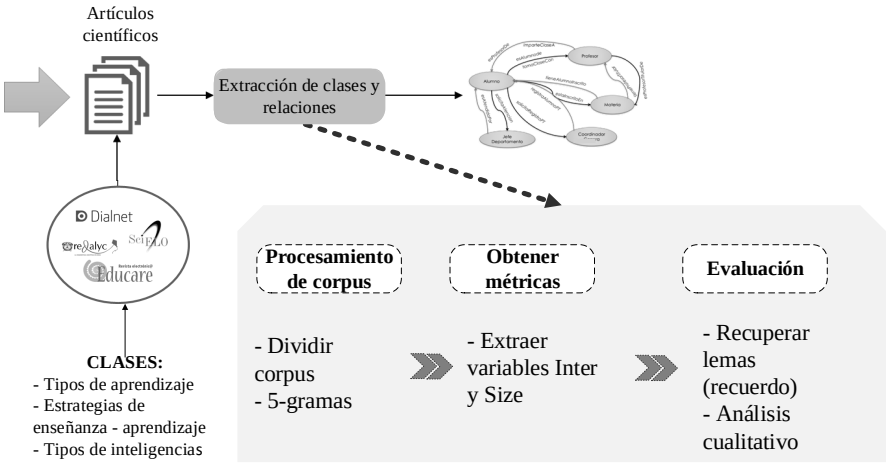

Fig. 2. Metodología general y proceso realizado para la obtención de subclases.

Previo a esta investigación, se realizó la validación de clases principales (C) mediante técnicas de agrupamiento, donde $C = \{ TiposInteligencias, EstilosAprendizaje, EstrategiasEnseñanza \}$. Para esto, se recolectó de manera manual un corpus compuesto por 51 artículos obtenidos de fuentes como Dialnet, Redalyc, SciElo y Educare.

El método propuesto tiene tres fases principales: En la primera, se tiene como entrada el corpus procesado (sin palabras cerradas ni signos de puntuación y lematizado). Dicho corpus se separa de acuerdo a la clase principal obteniendo 3 corpus diferentes de 17 instancias cada uno. Las métricas utilizadas en este experimento nos dicen en un rango de 0 a 1 que tan similar es una oración a otra, sin embargo, en este caso se utilizarán para determinar qué tan relacionado está un concepto respecto a otro, por lo que en lugar de utilizar oraciones completas se extraen 5-gramas y estos se toman como unidad de análisis.

En la segunda fase, se obtienen las variables necesarias para el cálculo de las métricas. Todas las métricas implementadas están basadas en el traslape de dos palabras a analizar: t_1 y t_2, en este caso, t_1 representa la clase principal y t_2 cada uno de los lemas del vocabulario. Se extrajeron dos variables: $Inter = x_1, x_2, x_3, \ldots, x_n$ y $Size = x_1, x_2, x_3, \ldots, x_n$ donde $Inter$ es el número de n-gramas en los que el lema y la clase principal coinciden, $Size$ es el número de n-gramas en los que el lema aparece, independientemente si coincide o no con la clase, x representa cada uno de los lemas y n es el total del vocabulario para el corpus analizado. La Figura 3 muestra un ejemplo de cómo se obtienen estos vectores en dos n-gramas, uno en el que aparece la clase y otro en el que no. Como se puede observar, en el n-grama del ejemplo 1 aparece la clase (*Estilos de aprendizaje*) por lo que los dos vectores incrementan su valor para las palabras que aparecen en el n-grama, en el ejemplo 2 no aparece la clase, por lo que solo se incrementan los valores correspondientes al vector $Size$.

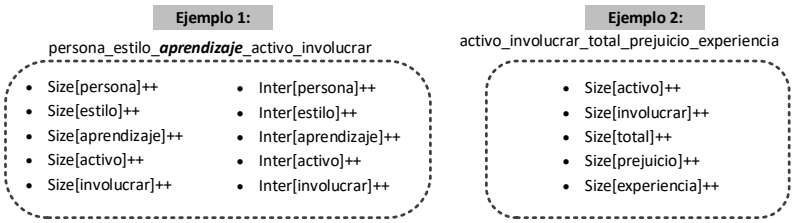

Fig. 3. Dos ejemplos en los que se obtienen los valores de los vectores Inter y Size para la clase Estilos de aprendizaje.

La tercera fase consiste en determinar cuáles lemas serán tomados como resultados, ya que los cálculos se hacen para todo el vocabulario. La hipótesis que se plantea es que si un lema tiene una similitud alta con la clase principal, significa que tiene una relación con esta, por lo que es una subclase para la ontología. Como se determinarán 4 métricas, se estableció un tope de 0.03 para determinar que un lema está relacionado con la clase, por lo que si un lema obtiene más de 0.03 en dos o más métricas, esta se toma como resultado, si no, se desecha. El tope es pequeño por el tamaño del corpus, aunque estás métricas suelen obtener resultados altos, el tamaño del corpus no es suficiente como para esperar resultados mayores al 50 %. El total de subclases recuperadas se evalúa utilizando el recuerdo, el cual es una métrica de recuperación de información que representa la fracción de los datos relevantes que son recuperados (Ecuación 5). Finalmente, se presenta un análisis cualitativo sobre los lemas relevantes recuperados por clase de acuerdo al tipo de relación que guardan con la clase principal:

$$R = \frac{items\ relevantes\ recuperados}{items\ relevantes}. \tag{5}$$

6. Resultados

Los resultados finales se muestran en la Tabla 1, en donde se observan el total de subclases a recuperar, las subclases recuperadas, el recuerdo por clase y otros conceptos que se recuperaron. Estos conceptos no son subclases, pero tienen relaciones no taxonómicas con la clase principal, ya sea que forman parte de la definición, aplicación o autor de dicho referente teórico.

Tabla 1. Análisis de los lemas obtenidos por cada clase.

Clase	Subclases			Otras
	Reales	Recuperadas	Recuerdo	relaciones
Tipos de inteligencias	8	5	0.625	9
Estrategias de enseñanza	3	2	0.667	10
Estilos de aprendizaje	4	4	1.00	8

Sólo la clase de *Estilos de aprendizaje* tuvo un recuerdo del 100 % recuperando a las dos clases restantes, *Tipos de inteligencias* recuperó cinco de ocho, bajando su recuerdo a 63 %. En cuanto a la clase de *Estrategias de enseñanza*, solo se recuperaron dos de las tres subclases, por lo que obtuvo un recuerdo de 67 %. La clase *Estilos de aprendizaje* es la más estructurada ya que sólo cuenta con cuatro subclases, por lo que están muy bien definidas, en cuanto a *Tipos de inteligencias*, aunque la clasificación propuesta por Gardner es definida y cada una de las inteligencias propuestas tiene fundamentos y características propias, en cuanto a tareas de procesamiento de lenguaje natural, es complicado manejar siete clases distintas, esto hace que en los experimentos no se separen completamente las instancias y no se recuperen la totalidad de las clases. Además, puede que alguno de los artículos se enfoque en solo una o dos inteligencias, por lo que el vocabulario asociada a estas es mayor al vocabulario de otras menos mencionadas en los artículos. La clase *Estrategias de enseñanza* presenta una clasificación muy general y en muchas ocasiones los autores usan nombres específicos para nombrarlas, no solo como metacognitivas, cognitivas o de apoyo.

En las siguientes tablas se presentan los resultados por clase. Se anexa el valor del vector *Inter*, ya que este sirve para determinar en cuantos n-gramas coincide cada lema y la clase, y los valores de cada una de las métricas, donde los resultados vas de 0 a 1. Solo se muestran los lemas que cumplen la heurística planteada: Obtener más de 0.03 en al menos dos métricas de las calculadas. Los lemas que son considerados como subclases se muestran en negritas, mientras que los que tienen otro tipo de relación con la clase principal se muestran en itálicas.

En la Tabla 2 se muestran los resultados para la clase *Tipos de inteligencias*. Como se menciona en la sección 3.1, este enfoque teórico distingue ocho tipos de inteligencias, de las cuales cinco aparecen en negritas en la tabla: musical, lingüística, interpersonal, emocional y lógico-matemática. Las inteligencias de tipo espacial, intrapersonal y naturalista no aparecen en la lista. Como se mencionó antes, si existen muchas subclases es complicado determinar las características

específicas de cada una de ellas. Además, estos resultados dependen totalmente del corpus y es posible que en algunos artículos se mencionen los 8 tipos de inteligencias, pero los experimentos o las discusiones se centren en las más predominantes dentro de la muestra estudiada. En la mayoría de los casos, los coeficientes de traslape y coseno tienen valores mayores a los coeficientes de Dice.

Tabla 2. Resultados para N-gramas en la clase *Tipos de inteligencias*.

Lema	–Intersección–	—– Dice -	—– Jaccard -	—– Traslape -	—– Coseno -
inteligencia	5702	1.0000	1.0000	1.0000	1.0000
múltiple	1332	0.3589	0.2187	0.774	0.4252
poder	163	0.0456	0.0234	0.1131	0.0569
numero	211	0.0611	0.0315	0.1758	0.0807
teoría	418	0.1231	0.0656	0.3835	0.1677
desarrollar	214	0.0632	0.0326	0.1998	0.0866
capacidad	170	0.0507	0.0260	0.1685	0.0709
ser	105	0.0314	0.0159	0.1062	0.0442
gardner	233	0.0698	0.0361	0.2385	0.0987
alumno	139	0.0421	0.0215	0.1544	0.0614
cada	209	0.0643	0.0332	0.2606	0.0977
diferente	153	0.0474	0.0243	0.2032	0.0738
persona	105	0.0328	0.0167	0.1491	0.0524
relación	235	0.0735	0.0381	0.3381	0.1180
considerar	104	0.0331	0.0168	0.1781	0.0570
rendimiento	159	0.0510	0.0262	0.2983	0.0912
tipo	294	0.0944	0.0495	0.5558	0.1693
musical	103	0.0332	0.0169	0.2060	0.0610
lingüístico	231	0.0747	0.0388	0.4793	0.1393
interpersonal	105	0.0340	0.0173	0.2253	0.0644
emocional	168	0.0552	0.0284	0.4386	0.1137
basar	120	0.0394	0.0201	0.3141	0.0813
todo	97	0.0321	0.0163	0.2820	0.0693
lógicomatemática	156	0.0521	0.0267	0.5379	0.1213

En cuanto a los otros lemas relacionados con la clase principal, se recuperaron 9 lemas que describen o forman parte del concepto de *Tipos de inteligencias*: La *teoría* de la inteligencias *múltiples* fue propuesta por *Gardner*, cada *persona* o *estudiante* tiene un *tipo* de inteligencia dominante, con la cual *desarrolla capacidades* diferentes, el correcto manejo de estas inteligencias permite aumentar el *rendimiento* académico. Si bien estos términos no son relaciones taxonómicas, son importantes para describir a la clase principal. Es importante mencionar que en esta clase es que las palabras que describen a la clase obtuvieron tienen más intersecciones que las subclases, pero en el resultado de las métricas, a excepción del coeficiente Jaccard, los resultados de las subclases son más altos que los de estos lemas.

La Tabla 3 muestra los resultados para la clase *Estilos de aprendizaje*. De acuerdo a la sección 3.1 existen cuatro tipos de aprendizaje, y todos estos se recuperan en la lista de la tabla: reflexivo, activo, teórico y pragmático. Al igual que en la clase anterior, con el coeficiente de Traslape se obtuvieron los resultados más altos. En cuanto a los otros lemas recuperados, algunos son parte de las características de esta clase por ejemplo: Para determinar el *estilo* de aprendizaje en un *estudiante*, se utiliza un *cuestionario* propuesto por *Honey-Alonso* llamado *CHAEA* (Cuestionario de Honey y Alonso de Estilos de

Aprendizaje). El conocer el estilo de aprendizaje también puede ir encaminado a la mejora del *rendimiento académico* del *alumno*, para implementar *estrategias* que ayuden al aprendizaje significativo. En esta clase, no se muestra una división entre las relaciones taxonómicas y no taxonómicas, las subclases se encuentran distribuidas a lo largo de la tabla.

Tabla 3. Resultados para N-gramas en la clase *Estilos de aprendizaje*.

Lema	–Intersección–	—– Dice -	—– Jaccard -	—– Traslape -	—– Coseno -
aprendizaje	6160	1.0000	1.0000	1.0000	1.0000
estilo	3903	0.5661	0.3948	0.6336	0.5694
estudiante	717	0.1616	0.0879	0.264	0.1753
estrategia	509	0.1422	0.0765	0.5095	0.2052
chaea	394	0.1076	0.0568	0.3382	0.1471
académico	312	0.0827	0.0431	0.2254	0.1069
reflexivo	266	0.0725	0.0376	0.2268	0.0990
activo	254	0.0688	0.0356	0.2077	0.0925
cuestionario	227	0.0641	0.0331	0.2459	0.0952
rendimiento	214	0.0607	0.0313	0.2415	0.0916
alumno	212	0.0580	0.0299	0.1839	0.0795
Vol	196	0.0524	0.0269	0.1492	0.0689
número	191	0.0501	0.0257	0.1306	0.0636
teórico	190	0.0527	0.0270	0.1798	0.0745
preferencia	184	0.0522	0.0268	0.2081	0.0788
proceso	175	0.0509	0.0261	0.2441	0.0833
alonso	158	0.0467	0.0239	0.2607	0.0818
universitario	154	0.0446	0.0228	0.2064	0.0718
análisis	140	0.0403	0.0206	0.1768	0.0634
promedio	134	0.0382	0.0195	0.1582	0.0587
pragmático	126	0.0361	0.0184	0.1518	0.0557
octubre	116	0.0361	0.0184	0.4265	0.0896
estudio	113	0.0317	0.0161	0.1177	0.0465
honey	111	0.0333	0.0169	0.2216	0.0632
relación	105	0.0319	0.0162	0.2530	0.0657
variable	104	0.0308	0.0157	0.1772	0.0547
predominante	102	0.0322	0.0163	0.5543	0.0958

Finalmente, la Tabla 6 muestra los resultados para la clase *Estrategias de enseñanza*, donde se pueden observar que se recuperaron dos de las tres subclases: cognitiva y metacognitiva. Aunque esta clase cuenta con una clasificación, esta no es adoptada por todos los autores, e incluso, aparte de esta clasificación, cada estrategia tiene un nombre, independientemente de la subclase a la que pertenezca. Por ejemplo, una estrategia para comprensión de lectura puede ser una estrategia de manejo de recursos, pero los autores se refieren a ella como estrategia de lectura, de hecho, la palabra lectura está en la lista de los lemas recuperados, pero al no ser parte de la clasificación, no se toma como relación taxonómica. Esta riqueza de nombres hace que se dificulte localizar estas subclases en el corpus.

Para esta clase también se recuperaron algunos conceptos que tienen relaciones no taxonómicas con el concepto principal como *aprendizaje*, *enseñanza*, *conocimiento*, *planificación*, *proceso*, *motivación*. Estos conceptos ayudan a la descripción de una estrategia de enseñanza aprendizaje como una herramienta para que el estudiante adquiera conocimiento, estas estrategias son planificadas

Tabla 4. Resultados para N-gramas en la clase *Estrategias de enseñanza aprendizaje.*

Lema	–Intersección–	—– Dice -	—– Jaccard -	—– Traslape -	—– Coseno -
estrategia	3803	1.0000	1.0000	1.0000	1.0000
aprendizaje	867	0.2289	0.1292	0.2297	0.2289
cognitivo	524	0.1957	0.1085	0.3378	0.2158
metacognitivas	410	0.1827	0.1005	0.5985	0.2540
estudiante	214	0.0656	0.0339	0.0785	0.0665
uso	203	0.0886	0.0463	0.2599	0.1178
conocimiento	176	0.0666	0.0345	0.1188	0.0742
emplear	175	0.0803	0.0418	0.3142	0.1202
utilizar	170	0.0760	0.0395	0.2534	0.1064
categoría	146	0.0668	0.0346	0.2575	0.0994
poder	128	0.0495	0.0254	0.0936	0.0561
estilo	116	0.0571	0.0294	0.4514	0.1173
autorregulación	106	0.0501	0.0257	0.2482	0.0832
motivación	106	0.0491	0.0251	0.2046	0.0755
enseñanza	104	0.0457	0.0234	0.1387	0.0616
permitir	89	0.0384	0.0196	0.1075	0.0502
proceso	88	0.0317	0.0161	0.0502	0.0341
numero	83	0.0315	0.0160	0.0568	0.0352
patrón	80	0.0397	0.0202	0.3493	0.0857
frecuencia	80	0.0389	0.0198	0.2564	0.0734
enfocar	76	0.0386	0.0197	0.5547	0.1053
elaboración	74	0.0369	0.0188	0.3610	0.0838
nivel	74	0.0323	0.0164	0.0956	0.0431
utilización	71	0.0354	0.0180	0.3333	0.0789
significativo	70	0.0319	0.0162	0.1207	0.0471
deber	69	0.0306	0.0155	0.0980	0.0422
tipo	68	0.0312	0.0158	0.1216	0.0466
usar	65	0.0322	0.0164	0.2778	0.0689
lectura	64	0.0303	0.0154	0.1509	0.0504
planificación	63	0.0307	0.0156	0.2059	0.0584

por los docentes procurando incentivar la motivación de los estudiantes, por ejemplo, las estrategias para la comprensión de lectura.

7. Conclusiones y trabajo futuro

En este artículo se presentó el análisis de métricas de similitud basadas en término, a fin de determinar las subclases para la construcción de una ontología. Para los procedimientos se utilizó un corpus pedagógico con tres clases principales. En los resultados se muestran los lemas recuperados por clase, entre los que se encuentran subclases, conceptos descriptivos y algunas palabras no relacionadas con la clase o que aportan poca información para la ontología.

La clase de Estilos de aprendizaje recupera todas las subclases, y los otros conceptos recuperados describen bien a la clase. En la clase Estrategias de aprendizaje solo se recuperan dos de tres mientras que en la clase Tipos de inteligencia se recuperan cinco de ocho; estas clases no presentan subclases bien definidas, por lo que el tamaño del corpus no permite establecer todas las subclases.

En cuanto a las métricas utilizadas, el coeficiente de traslape es el que obtuvo resultados más altos en las tres categorías, pero esta métrica puede ser engañosa, ya que si un lema tiene pocas apariciones, pero estas están en el mismo n-grama que la clase, el coeficiente de traslape es igual a 1. Es por eso que se utilizan las

otras métricas para recuperar los lemas. Los coeficientes de Dice y de Jaccard obtienen resultados muy bajos, incluso más que el tope establecido de 0.03. Esto tiene que ver con la longitud del corpus, que para efectos de estas métricas, puede considerarse pequeño. Sin embargo, a pesar de los resultados bajos en las métricas, los valores más altos se encuentran en los lemas importantes para la clase, ya sea como una subclase o como una relación no taxonómica.

Como trabajo futuro, se pretende incrementar el corpus a fin de obtener una mayor riqueza en el vocabulario. Al incrementar las instancias del corpus de entrada, los resultados en las métricas permitirán recuperar solo términos importantes. Además, se implementarán las métricas basadas en corpus.

Referencias

1. Barriga, F., Hernández, G.: Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista. McGraw Hill, México (2004)
2. Álvarez Carmona, M.n.: Detección de similitud semántica en textos cortos. Ph.D. thesis, Instituto Nacional de Astrofísica Óptica y Electrónica (2014)
3. Dai, X., Li, X.: Study of learning source ontology modeling in remote education. In: 2010 International Conference on Multimedia Technology. pp. 1–4 (Oct 2010)
4. Du, L., Zheng, G., You, B., Bai, L., Zhang, X.: Research of online education ontology model. In: 2012 Fourth International Conference on Computational and Information Sciences. pp. 780–783 (Aug 2012)
5. Faria, C., , Girardi, R.: A domain-independent process for automatic ontology population from text. Science of Computer Programming 95, Part 1, 26 – 43 (2014), `http://www.sciencedirect.com/science/article/pii/S0167642313003419`, special Issue on Systems Development by Means of Semantic Technologies
6. Fu, J., Jia, K., Xu, J.: Domain ontology learning for question answering system in network education. In: 2008 The 9th International Conference for Young Computer Scientists. pp. 2647–2652 (Nov 2008)
7. García-Miguel, J.M., Vaamonde, G., Domínguez, F.G.: Adesse, a Database with Syntactic and Semantic Annotation of a Corpus of Spanish. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
8. Gardner, H.: Estructuras de la Mente (Sep 2001)
9. Gomaa, W., Fahmy, A.: A survey of text similarity approaches 68 (04 2013)
10. González, M., Tourón, J.: Autoconcepto y rendimiento escolar: sus implicaciones en la motivación y en la autorregulación del aprendizaje. Eunsa (1992)
11. Grandbastien, M., Azouaou, F., Desmoulins, C., Faerber, R., Leclet, D., Quenu-Joiron, C.: Sharing an ontology in education: Lessons learn from the OURAL project. In: Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007). pp. 694–698 (July 2007)
12. Grljevic, O., Bosnjak, Z.: Development of serbian higher education corpus. In: 2015 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI). pp. 177–181 (Nov 2015)
13. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum.-Comput. Stud. 43, 907–928 (December 1995), `http://portal.acm.org/citation.cfm?id=219666.219701`

14. Hssina, B., Bouikhalene, B., Merbouha, A.: An ontology to assess the performances of learners in an e-learning platform based on semantic web technology: Moodle case study. In: Europe and MENA Cooperation Advances in Information and Communication Technologies, pp. 103–112. Springer (2017)
15. Hu, J., Li, Z., Xu, B.: An approach of ontology based knowledge base construction for chinese k12 education. In: 2016 First International Conference on Multimedia and Image Processing (ICMIP). pp. 83–88 (June 2016)
16. Kang, Y.B., Haghighi, P.D., Burstein, F.: Cfinder: An intelligent key concept finder from text for ontology development. Expert Systems with Applications 41(9), 4494 – 4504 (2014), http://www.sciencedirect.com/science/article/pii/S0957417414000189
17. Kolb, D.: Learning style inventory. MA: Hay Group, Hay Resources Direct, Boston USA (1976)
18. Méndez, N.D.D., Carranza, D.A.O., Ocampo, M.G.: Representación ontológica de perfiles de estudiantes para la personalización del aprendizaje. Revista Educación en Ingeniería 10(19), 105–115 (2015)
19. Ochoa, J.L., Hernández-Alcaraz, M.L., Almela, A., Valencia-García, R.: Learning semantic relations from Spanish natural language documents in the financial domain. In: Proceedings of the 3rd International Conference on Computer Modeling and Simulation, held at Mumbai, India. Chengdu: Institute of Electrical and Electronics Engineers, Inc. pp. 104–108 (2011)
20. Ochoa, J.L., Hernández-Alcaraz, M.L., Valencia-Garcia, R., Martinez-Bejar, R.: A semantic role-based methodology for knowledge acquisition from Spanish documents. International Journal of Physical Sciences 6(7), 1755–1765 (2011)
21. Olivos, P., Santos, A., Martín, S., Cañas, M., Gómez, E., Maya, Y.: The relationship between learning styles and motivation to transfer of learning in a vocational training programme. Suma Psicológica 23(1), 25–32 (2016)
22. Teixeira, J., Sarmento, L., Oliveira, E.: Semi-automatic creation of a reference news corpus for fine-grained multi-label scenarios. In: 6th Iberian Conference on Information Systems and Technologies (CISTI 2011). pp. 1–7 (June 2011)
23. Uskov, V., Pandey, A., Bakken, J.P., Margapuri, V.S.: Smart engineering education: The ontology of internet-of-things applications. In: 2016 IEEE Global Engineering Education Conference (EDUCON). pp. 476–481 (April 2016)
24. Weigand, H.: A multilingual ontology-based lexicon for news filtering-the TREVI project. In: Proceedings of the IJCAI Workshop on Multilingual Ontologies-Nagoya (1997)
25. Weinstein, C.E., Mayer, R.E.: The teaching of learning strategies. In: Innovation abstracts. vol. 5. ERIC (1986)