Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Metodología para el aprendizaje ontológico semiautomático de dominio pedagógico

Tesis presentada para obtener el grado de

Doctor en Ingeniería del Lenguaje y del Conocimiento

Autor:

Candy Yuridiana Alemán Muñoz

Director:
Dra. María Josefa
Somodevilla García

Co-director:
Dra. Darnes
Vilariño Ayala

Diciembre 2020

Resumen

En este trabajo de tesis de experimenta con distintas técnicas de PLN para la construcción semiautomática de ontologías en el dominio pedagógico. Se eligen tres subdominios relacionados con el aprendizaje significativo, y con estos datos, se utiliza como entrada un conjunto de artículos escritos en español, alusivos a dichos temas. En la realización de los experimentos propuestos se diseñan procedimientos que pueden ser útiles para otras tareas propias del área.

El documento se divide en 6 capítulos que abarcan tres secciones principales: En los capítulos I, II y III se desarrolla la fundamentación de la investigación, abarcando el planteamiento del problema y los conceptos teóricos relacionados con este. Finalmente, se describen algunas investigaciones relacionadas con dicha investigación, tanto en las áreas de pedagogía como inteligencia artificial.

La segunda sección integra los capítulos IV y V, donde se explica la metodología propuesta para resolver el problema planteado en la primera sección. Se describen las técnicas implementadas y el objetivo de éstas, se analizan los resultados obtenidos de acuerdo a los recursos creados para la evaluación, además de generar ontologías manuales que ayuden a este proceso.

Finalmente, en el capítulo VI se presentan las conclusiones sobre la investigación, integrando comentarios sobre los experimentos y contribuciones para las ciencias computacionales y pedagogía. Al final se presenta el trabajo en progreso y posibles modificaciones al tema principal.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Desde mediados del siglo XIX, el área de estudio de la Inteligencia Artificial (IA) se ha enfocado en representar con modelos formales el conocimiento que el humano genera en las diferentes ramas de la ciencia. Con el Procesamiento del Lenguaje Natural (PLN) se generan técnicas que permiten integrar modelos computacionales con la lingüística aplicada. Uno de los enfoques que se explora con interés especial, es el estudio y análisis de la información disponible en Internet, la cual, al no ser estructurada, requiere técnicas de representación que permitan integrarla a fin de lograr la formalización y dar respuesta a preguntas que involucren lenguaje natural por parte del usuario.

La Web ha ido evolucionando, desde ser un repositorio meramente estático compuesto de texto (Web tradicional) hasta contribuir al aprendizaje colaborativo mediante redes sociales, blogs y otras herramientas que permiten publicar contenido sin necesidad de tener conocimientos sobre diseño o programación Web. Aunque este cambio ha sido sustancial, se tiene el problema del exceso de información no utilizable. Una de las primeras propuestas para manejar este problema es la Web Semántica, donde se adopta una visión del contenido de la *World Wide Web* no solo como un repositorio, sino que trata de representar el contenido para que sea accesible a las máquinas y éstas sean capaces de interpretar la información ofrecida (Rettinger et al., 2012).

La Web Semántica y el Procesamiento del Lenguaje Natural son áreas de investigación en las que convergen, entre otras ramas del conocimiento, la Ingeniería Informática, la Lingüística y la Documentación. Las ontologías surgen como una alternativa para el etiquetado manual del

contenido de la Web siguiendo un estándar rígido y ayudan a que éste se haga de manera automática por medio de modelos computacionales.

1.1. Planteamiento del problema

Las ontologías representan un modelo ideal para describir de manera formal el contenido de los recursos en la Web, las cuales se construyen mediante el proceso de aprendizaje ontológico. La anotación semántica permite a las aplicaciones obtener una visión precisa del significado del contenido. En este sentido, en el presente trabajo se analiza la aplicación del aprendizaje ontológico en el tratamiento semiautomático de textos no estructurados. La propuesta se centra en el uso de técnicas de PLN y Recuperación de Información (RI) aplicadas en un subdominio de la Pedagogía en el idioma español. El problema de investigación se basa en el impacto que han tenido las ontologías en el área computacional. En la Figura 1.1 se pueden observar algunos aspectos clave que llevaron a la elección y fundamentación del tema.

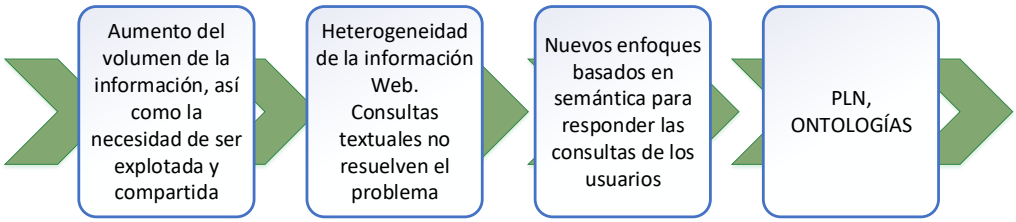


Figura 1.1: Problema de investigación

En los últimos años, el volumen de información disponible en diversos repositorios se ha incrementado de manera exponencial. Esto ha llevado a investigadores de las Ciencias de la Información a buscar estrategias para recuperar información, además de procesarla, analizarla y generar respuestas acordes con los requerimientos de los diferentes usuarios (López & Castillo, 2013). La característica de heterogeneidad en la información Web hace difícil su procesamiento tanto manual como automático. Las técnicas de RI tradicionales recuperan información de acuerdo a palabras clave y el tipo de información a procesar presenta un alto contenido de relaciones semánticas, en algunos casos estas búsquedas textuales generan desambiguación. Cuando un usuario realiza una búsqueda es difícil determinar automáticamente en qué sentido o dominio quiere obtener resultados, el uso de las ontologías para integrar la información puede ayudar a tratar este problema.

Las ontologías surgieron como recurso de formalización de información para la Web Semántica. Considerando que una ontología aborda una temática específica, una metodología de diseño

actual propone la integración de las mismas en sistemas para representar la información en un dominio específico (Rodríguez et al., 2012). De esta manera, se facilita la gestión de información de diferentes repositorios y formatos de representación.

1.1.1. Definición del problema

Las ontologías ofrecen oportunidades para modelar a los usuarios y su interacción con los sistemas y su entorno, teniendo en cuenta su capacidad de captar conocimientos complejos en representaciones formales reutilizables, (Bouamrane et al., 2008). La construcción manual de ontologías, presenta un problema de tiempo para la Web Semántica, ya que es una tarea que consume muchos recursos, lo cual se traduce en un trabajo tedioso y difícil.

Considerando la problemática planteada anteriormente, se propone desarrollar el proceso de construcción semiautomático de ontologías en el dominio pedagógico. Para cumplir con este reto se tomarán textos pedagógico en idioma español. Dichos textos son artículos publicados en revistas del área, los cuales permitirán analizar la situación actual de México en el ámbito educativo. La Figura 1.2 muestra los subdominios de la pedagogía utilizados para la investigación.

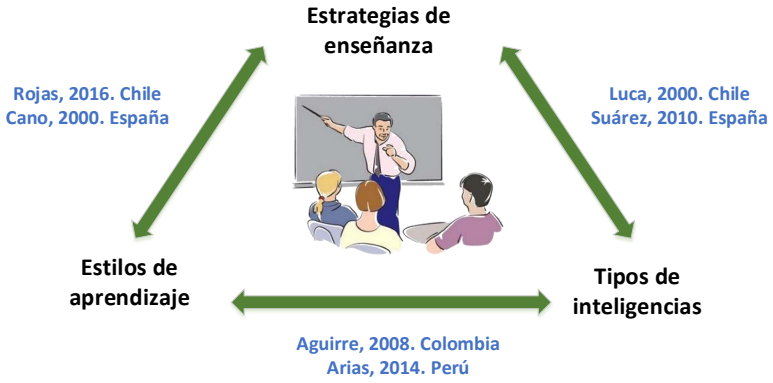


Figura 1.2: Subdominios de la pedagogía utilizados en la investigación

Para el dominio bajo estudio se consideran tópicos relacionados con la implementación de técnicas de enseñanza dentro del salón de clases; por lo tanto, además de las estrategias de enseñanza, se toman en cuenta características de los estudiantes como los estilos de aprendizaje y tipos de inteligencia. En la figura 1.2 se muestran también algunas investigaciones que han utilizado dos de los tres subdominios analizadas para poder implementar técnicas que permitan incrementar el desempeño de un determinado grupo, especialmente en el nivel superior: Suárez

et al. (2010) y De Luca (2000) analizan la relación entre estrategias de enseñanza y tipos de inteligencias, Aguirre (2008) y Arias (2014) estudian la interacción entre estilos de aprendizaje y tipos de inteligencias; finalmente, Rojas et al. (2016) y Cano García (2000) investigan los estilos de aprendizaje y estrategias de enseñanza.

1.1.2. Objetivos y preguntas de investigación

A partir de esta investigación se planea generar las bases para una herramienta que pueda ser utilizada en clases presenciales. Se inicia con la hipótesis de que mediante el uso de técnicas de PLN es posible implementar el proceso de aprendizaje ontológico en un subdominio específico de la pedagogía. Las ontologías creadas serán útiles para determinar una estrategia de enseñanza según el tipo de inteligencia y estilo de aprendizaje de un alumno o un grupo de alumnos. Esto colaborará en la personalización de la educación, lo que, aunado con los nuevos paradigmas educativos, puede ser un factor en el incremento del nivel educativo del país. Cabe resaltar que el producto resultante no suplirá al docente en el aula, ni a los nuevos modelos educativos vigentes, si no que será una herramienta que coadyuve a alcanzar los objetivos del proceso de enseñanza aprendizaje.

El objetivo general consiste en desarrollar una metodología de aprendizaje ontológico que comprenda desde la creación de corpus hasta la construcción semiautomática de ontologías en el dominio pedagógico, basada en la aplicación y análisis de técnicas de PLN así como de aprendizaje automático, utilizando textos no estructurados en la Web. Para alcanzar este objetivo se desglosan cada una de las fases del aprendizaje ontológico y se aplican en un rubro específico del dominio pedagógico. Este objetivo general se desglosa en los siguientes objetivos particulares:

1. Determinar los elementos teóricos pedagógicos necesarios para la creación de una herramienta de apoyo al docente, considerando la opinión de un experto en Pedagogía.
2. Construir un corpus con documentos extraídos de la Web, de acuerdo a tres subdominios: estilos de aprendizaje, tipos de inteligencias y estrategias de aprendizaje.
3. Determinar las herramientas y algoritmos que permitan identificar los conceptos, así como sus relaciones en los corpus.
4. Construir semiautomáticamente una ontología por cada uno de los subdominios.
5. Crear ontologías manuales de cada uno de los dominios, que junto con un conjunto de evaluación (*Gold standard*), permitan evaluar los resultados de los experimentos.

Analizando los objetivos de la investigación, se propone la siguiente pregunta general: ¿Es posible desarrollar una metodología para el aprendizaje ontológico utilizando técnicas de PLN

a partir de textos no estructurados en la Web?

En relación con los objetivos particulares, se trabajan las siguientes preguntas de investigación:

1. ¿Qué elementos teóricos son necesarios para la creación de una herramienta de apoyo al docente?
2. ¿Cuál procedimiento se debe utilizar para crear un corpus con textos de la Web de acuerdo a los tres subdominios pedagógicos planteados?
3. ¿Qué herramientas y algoritmos permiten identificar los conceptos y relaciones en un corpus de texto no estructurado?
4. ¿Qué pasos debe considerar una metodología de diseño de una ontología por subdominio?
5. ¿Cuáles técnicas se deben utilizar para evaluar la metodología?

1.1.3. Justificación

Con el desarrollo del PLN, se han realizado investigaciones que proponen una creación ontológica automática o semiautomática; estas propuestas suponen que las funciones de los expertos en dominio y en Lingüística se realicen por un sistema capaz de procesar automáticamente el corpus inicial, obtener las palabras clave, las relaciones entre ellas y finalmente generar la estructura formal. Esto supone que el experto en el dominio se dedicará únicamente a validar la salida del sistema, logrando con ello agilizar los procesos de construcción de las ontologías. A partir del 2001 surgió el término de aprendizaje ontológico propuesto por Maedche & Staab (2001) para un campo emergente de investigación que buscaba la generación automática de ontologías. Desde la perspectiva de hoy, el aprendizaje ontológico es un caso de uso para el aprendizaje conceptual, pero la colaboración y el intercambio entre estos todavía es limitado.

Tradicionalmente, el proceso de construcción de una ontología es manual, ya que para su elaboración es necesario contar con expertos en el dominio de interés que identifiquen y definan las relaciones y palabras clave existentes en el texto analizado. También es importante contar con un experto en lingüística que formalice las estructuras encontradas por el experto en dominio y un experto en computación que traduzca las representaciones propuestas por los dos expertos en estructuras computacionales. Este procedimiento es costoso, ya que requiere de una interacción muy alta de los expertos, lo que conlleva en muchos casos a requerir mucho tiempo para la creación de la ontología cuando el dominio es extenso (Faria et al., 2014; Ferreira et al., 2016).

Dentro de la Pedagogía, la investigación se centra en la creación de una herramienta de apoyo para el docente para las clases presenciales. De acuerdo con investigaciones en el área, tres conceptos principales se pueden estudiar para poder construir dicha herramienta: tipos de

inteligencias, estrategias de aprendizaje y estilos de aprendizaje. El desarrollo de la metodología se basa en artículos publicados referentes a cada uno de los conceptos, a partir de los cuales se integra un corpus para poder extraer las clases, así como las relaciones entre ellas.

1.1.4. Alcances y límites

Considerando los objetivos que se proponen, se genera una metodología que pueda ser aplicada al dominio pedagógico, específicamente al subdominio de herramientas educativas presenciales considerando su posible extensión a otros subdominios de la Pedagogía. La investigación involucra las fases de creación del corpus, preprocesamiento del mismo y extracción de conceptos principales y sus relaciones.

El corpus inicial estará conformado por artículos científicos en formato PDF, por lo que tanto la creación y poblado estarán sujetos al contenido de los mismos. Una característica importante de este corpus, es que los conceptos que se consideran principales tienen diferentes enfoques teóricos en el dominio, donde cada enfoque tiene una subclasificación distinta. Por lo tanto, cada artículo debe tener el mismo enfoque que el resto de los que integran el corpus. Esta característica hace que al menos en una fase inicial, la creación del corpus tenga algunos procesos manuales, para posteriormente utilizar técnicas automáticas que incrementarán el número de instancias.

Uno de los problemas al desarrollar una ontología de dominio es que pueden existir muchas posibilidades para el diseño de la misma, ya que es casi imposible especificar en la ontología todos los aspectos de un dominio (Wu, 2008), por lo cual, el diseñador debe seleccionar los conceptos y relaciones que de acuerdo con su comprensión subjetiva del problema puedan modelarse por medio de la ontología. Dada esta situación, en los objetivos se muestra una evaluación cualitativa analizando los resultados desde el punto de vista pedagógico, aparte de las métricas de recuperación de información.

1.2. Aportaciones

Las aportaciones de la presente investigación, se darán en dos ámbitos: el computacional y el enfocado al dominio elegido para las pruebas. En el campo computacional, se espera obtener una metodología para aprendizaje ontológico, enfocada principalmente en la detección de conceptos mediante técnicas de PLN. Como se menciona en la sección del estado del arte, las investigaciones previas se basan principalmente en metodologías manuales, con evaluaciones enfocadas en la decisión de expertos en el dominio analizado. En esta investigación, la validación del experto en el dominio servirá para formalizar los procedimientos automáticos implementados.

De los artículos analizados, más del 80 % se refieren a investigaciones centradas en el idioma inglés, al ser esta una investigación con enfoque en el idioma español, se profundiza en la búsqueda, análisis y creación de herramientas y recursos léxicos que permitan procesar automáticamente los textos. Otro aspecto importante es que los dominios analizados automáticamente son generales, por lo que no hay hasta el momento investigaciones que se centren en varios temas específicos que compartan un gran porcentaje de vocabulario.

La metodología propuesta presenta una interacción más profunda de técnicas de PLN y RI, así como funciones semánticas para la extracción de clases principales. Esto da por resultado un proceso semiautomático desde la creación del corpus hasta la evaluación, mientras que las investigaciones analizadas se centran únicamente en una fase del aprendizaje ontológico. Además, aunque se trabajará con un subdominio específico, se considera que la metodología resultante pueda ser aplicada a otros subdominios pedagógicos.

En el plano pedagógico, la educación en México está experimentando cambios en todos los niveles educativos (Díaz Barriga, 2008; Diaz-Barriga & Hernández-Rojas, 2010); si bien el docente aún forma parte del proceso enseñanza-aprendizaje, lo que se requiere es que sea capaz de innovar en la manera en que organiza y dirige sus cursos para crear verdaderas experiencias de aprendizaje significativo. La superación del paradigma conductista a un paradigma centrado en procesos cognitivos significó una reestructuración de los procesos de enseñanza-aprendizaje. Esta reestructuración implica una sinergia entre las inteligencias múltiples, los estilos de aprendizaje y las metodologías de enseñanza-aprendizaje las cuales son precisamente los subdominios que serán utilizados en el conjunto de prueba. Por lo tanto, en el ámbito pedagógico, se espera que el sistema de ontologías resultante de la investigación sea propicio en el análisis de las relaciones entre conceptos en las distintas clases (estilos de aprendizaje, tipos de inteligencias, estrategias de enseñanza) para su aplicación en escenarios específicos.

1.3. Contenido de la tesis

La tesis está estructurada de la siguiente manera:

- El capítulo II muestra el análisis del estado del arte, haciendo énfasis en las investigaciones más recientes respecto a la creación de ontologías y la aplicación de técnicas de PLN en otras áreas del conocimiento.
- El capítulo III integra el marco teórico dividido en dos secciones: la primera se relaciona con los conceptos principales relacionados con el aprendizaje ontológico y el dominio elegido. La segunda parte integra los conceptos de PLN y recuperación de información, así como técnicas y herramientas utilizados para la realización de los experimentos.

- En el capítulo IV se analiza el dominio pedagógico, los conceptos teóricos más importantes, así como las investigaciones recientes sobre el tema.
- En capítulo V se describe la metodología propuesta para la solución del problema. Se analizan de manera general los procesos y herramientas utilizadas en cada fase de dicha metodología, además del método para la obtención de las distintas versiones del corpus.
- El capítulo VI está conformado por los resultados obtenidos en cada una de los experimentos realizados. Los resultados se muestran en forma de tablas y gráficas, al final se analizan los más representativos.
- En el capítulo VII se muestran las conclusiones obtenidas del proceso de investigación, tanto en el ámbito computacional como en el pedagógico.
- Finalmente, en la sección de apéndices se muestran algunos instrumentos relacionados con el dominio pedagógico, y los resultados de algunos experimentos relacionados con la metodología principal.

Capítulo 2

Marco teórico

En este capítulo se muestran los conceptos teóricos necesarios para el tema de investigación. Se inicia con una definición de aprendizaje ontológico y los elementos que están involucrados en este proceso. Posteriormente, se da una referencia teórica de las ontologías, así como su proceso de construcción, poblado y evaluación. Dentro de estas etapas, se utilizan diversas técnicas de PLN y de aprendizaje automático las cuales se detallan en la misma sección. Finalmente, se analizan los temas seleccionados dentro del dominio pedagógico.

El aprendizaje ontológico se ha beneficiado de los problemas a los que se enfrenta el PLN, uno de los cuales es el tratamiento de texto el cual impide la construcción a gran escala de las ontologías. Existe un cuello de botella en la elaboración manual de fuentes de conocimiento estructuradas, por ejemplo, diccionarios, taxonomías, bases de conocimiento (Cullen & Bryman, 1988) así como en datos de capacitación entre los que se encuentran los corpus de textos anotados.

Poco a poco se vuelve aparente que para minimizar los esfuerzos humanos en el proceso de aprendizaje y mejorar la escalabilidad y robustez del sistema, es posible que los recursos estáticos y diseñados por expertos ya no sean adecuados. Reconociendo esta problemática, una cantidad cada vez mayor de esfuerzo de investigación se está dirigiendo gradualmente hacia el aprovechamiento de la inteligencia colectiva en la Web, con la esperanza de abordar este importante cuello de botella (Wong et al., 2011).

2.1. Aprendizaje ontológico

El término “aprendizaje ontológico” fue acuñado por Maedche & Staab (2001), el cual puede describirse como la adquisición de un modelo de dominio a partir de datos. Históricamente está relacionado con la Web Semántica, que se basa en modelos de ontología y lógica descriptiva. Por lo tanto, los modelos de dominio que deben aprenderse también están restringidos en su complejidad y expresividad.

El aprendizaje ontológico necesita datos de entrada para detectar los conceptos relevantes de un dominio dado, sus definiciones y las relaciones que se mantienen entre ellos. Los datos de entrada deben ser representativos del dominio y estar en forma de esquemas tales como XML, diagramas de Lenguaje Unificado de Modelado (UML) o esquemas de bases de datos. El aprendizaje ontológico basado en texto se realiza cuando la entrada está representada por recursos textuales no estructurados. Por lo tanto, el aprendizaje ontológico se define como el proceso de identificar términos, conceptos, relaciones y, opcionalmente, axiomas a partir de información textual y usarlos para construir y mantener una ontología (Wong et al., 2011).

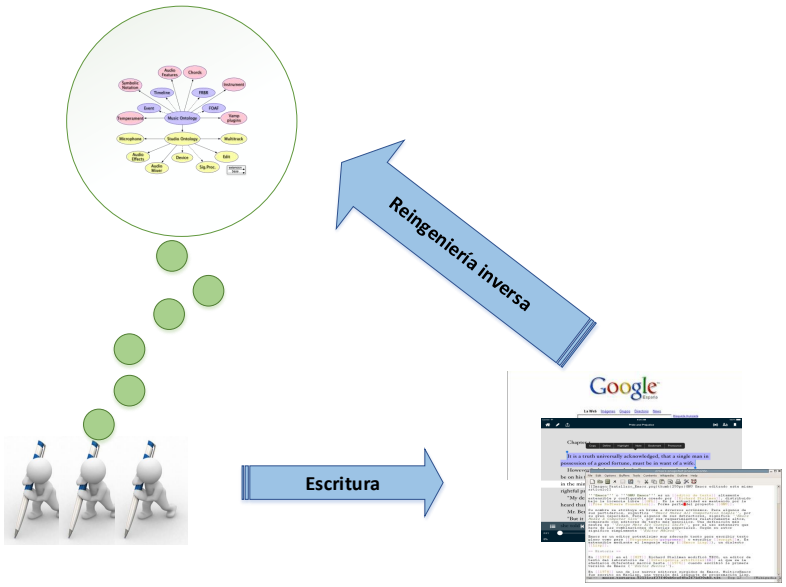


Figura 2.1: Aprendizaje ontológico visto como un proceso de reingeniería inversa (Cimiano, 2006).

La figura 2.1 muestra de manera general el proceso de aprendizaje ontológico. De acuerdo a este esquema, el proceso se puede considerar hasta cierto punto como un proceso de ingeniería inversa (Cimiano, 2006). El autor de un determinado texto o documento tiene en mente un modelo de mundo que comparte hasta cierto punto con otros autores que escriben textos sobre el mismo

dominio. Este modelo da forma al contenido del texto resultante. La tarea de reconstruir el modelo del autor o incluso del modelo compartido por diferentes autores puede verse como una ingeniería inversa.

Históricamente, el aprendizaje ontológico está conectado a la Web Semántica, que se basa en modelos de ontología o formalismo lógico restringido a lógica de primer orden, en particular, las lógicas de descripción (Staab & Studer, 2009). Por lo tanto, los modelos de dominio están restringidos en su complejidad y expresividad. La Figura 2.2 muestra las tareas del aprendizaje ontológico propuestas por Cimiano (2006). Cada capa de la figura representa una tarea y se describe a continuación:

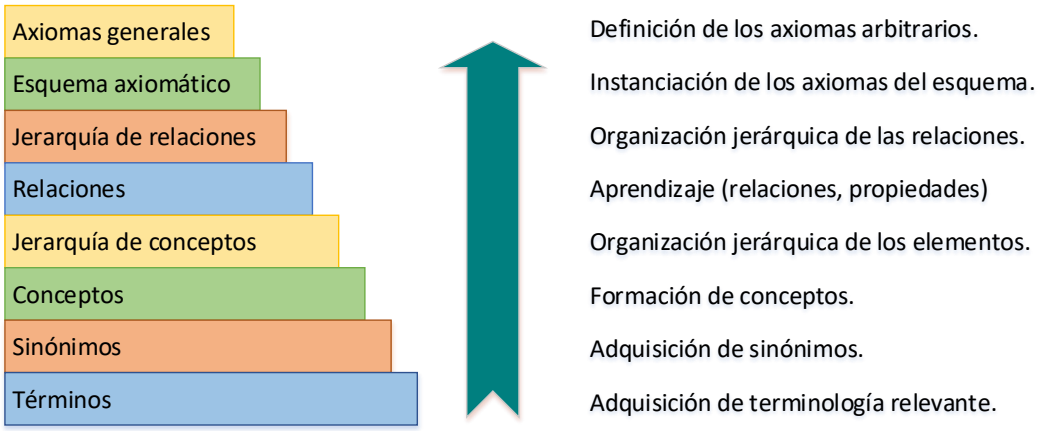


Figura 2.2: Capas del aprendizaje ontológico.

- **Términos:** Son realizaciones lingüísticas de conceptos específicos del dominio y, por lo tanto, son fundamentales para otras tareas más complejas. En esta subtarea, se determinan un conjunto de términos o signos relevantes para conceptos y relaciones, que son característicos para el dominio representado en la colección de texto y que proporcionan la base para definir un léxico para una ontología.
- **Sinónimos:** El descubrimiento de sinónimos consiste en encontrar palabras que denotan el mismo concepto y que aparecen así en el mismo conjunto para un concepto dado. Se consideran dos palabras como sinónimos si comparten un significado común que puede usarse como base para formar un concepto relevante para el dominio en cuestión.
- **Conceptos:** En la formación de una ontología, los conceptos contienen tres elementos: una definición intencional de conceptos, extensión y los signos léxicos que se utilizan para referirse a ellos. El léxico también puede contener estructuras más complejas enriquecidas con información estadística como lo describe o incluso analizar árboles, marcos de sub categorización, entre otros.

- **Jerarquía de conceptos:** En esta subtarea, se presentan tareas relacionadas con inducir, extender y refinar la ontología. Consiste en, dado un conjunto de conceptos, y su realización léxica se forman pares para inducir una jerarquía conceptual desde cero. Por ejemplo, partiendo de un conjunto de conceptos $C := ciudad, pais, capital, ...$, se deriva la relación $ciudad < pais$.
- **Relaciones:** Tarea de aprender los identificadores de relaciones o las etiquetas, así como su dominio apropiado y rango.
- **Jerarquía de relaciones:** Dada una cierta relación, determinar el nivel correcto de abstracción con respecto a la jerarquía de conceptos para el dominio.
- **Esquema axiomático:** Se determinan los diferentes tipos de axiomas dentro del aprendizaje ontológico. Para los conceptos se definen los axiomas de disjunción o equivalencia, mientras que para las relaciones se definen axiomas que describen las propiedades de la relación, es decir, transitividad, simetría, etc. En esta tarea, se analizan qué conceptos, relaciones o pares de conceptos se aplican.
- **Axiomas generales:** En esta subtarea, los axiomas tienen que aprenderse y no simplemente crearse una instancia. Aquí el tipo de axiomas depende en gran medida del formalismo lógico utilizado en el fondo. Los axiomas generales pueden considerarse como implicaciones lógicas que limitan la interpretación de conceptos y relaciones. Estos difieren de los esquemas de axiomas en que no ocurren con tanta frecuencia.

2.2. Ontologías

La palabra *Ontología* se deriva del griego *ontos* (estudio del ser) y *logos* (palabra). Filosóficamente, es la ciencia de *qué es*, es una explicación sistemática de la existencia, de los tipos de estructuras, categorías de objetos, propiedades, eventos, procesos y relaciones en cada área de la realidad (Smith, 2004). En las aplicaciones de la vida real, una ontología es una entidad computacional, y no ha de ser considerada como una entidad natural que se descubre, sino como un recurso artificial que se crea (Mahesh, 1996).

En las ciencias computacionales, una ontología se define como “una especificación formal de una conceptualización” (Gruber, 1995, p. 1). Otra definición es la de Weigand (1997), el cual la define como “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.” (p. 1). Esta base de datos se define a partir de un corpus base, del cual 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 $O = (C, H, I, R, P, A)$ (Faria et al., 2014) 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 es 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 y que realizan el proceso de inferencia.

En una ontología, los conceptos representan la base para la descripción de la información. Esta descripción se realiza mediante tres componentes: Términos, atributos y relaciones. Los términos son nombres utilizados para referirse a un concepto específico que puede incluir un conjunto de sinónimos que especifican los mismos conceptos. Los atributos describen el concepto a detalle utilizando características, y las relaciones se utilizan para representar correspondencias entre diferentes conceptos y proveer una estructura general de la ontología (Sánchez López, 2007).

2.2.1. Creación

Gruber (1995) propone el siguiente conjunto preliminar de criterios de diseño para ontologías:

1. **Claridad:** Una ontología debe comunicar efectivamente el significado pretendido de los términos definidos y las definiciones deben ser objetivas. Para especificar una conceptualización se necesita establecer axiomas que restrinjan las posibles interpretaciones para los términos definidos.
2. **Coherencia:** Las inferencias deben ser consistentes con las definiciones. La coherencia también se aplica a los conceptos que se definen informalmente, como los que se describen en la documentación y ejemplos de lenguaje natural. Si una oración que puede inferirse de los axiomas contradice una definición o ejemplo dado de manera informal, entonces la ontología es incoherente.
3. **Extensibilidad:** Una ontología debe diseñarse para anticipar los usos del vocabulario compartido. Debe ofrecer una base conceptual para una gama de tareas anticipadas, y la representación se elabora de manera que se pueda extender y especializar la ontología monótonamente.
4. **Mínimo sesgo de codificación:** La conceptualización debe especificarse en el nivel de conocimiento sin depender de una codificación de nivel de símbolo particular. Se produce un sesgo de codificación cuando las elecciones de representación se realizan puramente para la comodidad de la notación o la implementación.

5. **Compromiso ontológico mínimo:** Una ontología debe hacer el menor número posible de afirmaciones sobre el mundo que se está modelando, permitiendo a las partes comprometidas con la libertad ontológica especializarse e instanciar la ontología según sea necesario.

Para el proceso de creación de una ontología, Noy & Mcguinness (2001) propone los siguientes pasos:

1. **Determinar el dominio y alcance de la Ontología:** Se determina para qué se desarrolla la ontología, quién la usará y qué tipo de información proporcionará.
2. **Considerar reutilizar ontologías existentes:** Investigar si es posible extender fuentes de conocimientos ya existentes, y que puedan ser de utilidad para el dominio del problema.
3. **Enumerar términos importantes en la ontología:** Elaborar una lista de los términos proporcionados por el usuario, indicando las propiedades de cada uno, de la manera más precisa y carente de ambigüedades.
4. **Definir clases y jerarquía de clases:** De la lista creada en el paso anterior, seleccionar aquellos términos independientes para constituir las clases.
5. **Definir las propiedades de las clases:** Describir la estructura de los conceptos, los términos que no fueron seleccionados como clases, pasan a considerarse como propiedades de la clase (denominados *slots*).
6. **Definir las características de los *slots*:** Definir los diferentes tipos de valores que describan a los *slots*, por ejemplo el tipo de valor asociado, cardinalidad, valores permitidos, entre otros.

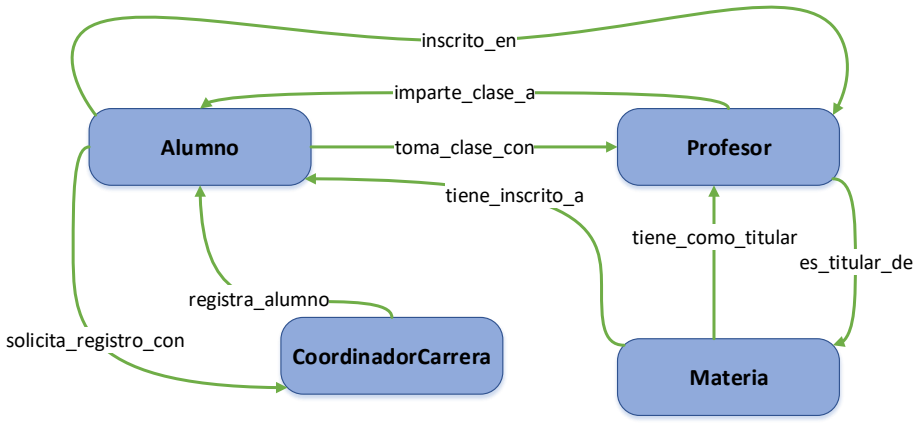


Figura 2.3: Ejemplo de ontología.

La figura 2.3 representa un ejemplo de una ontología en el dominio de la educación, donde se representan dos de los tres conceptos representativos de las mismas. Los conceptos

clave son representados por el conjunto {*Alumno, Profesor, CoordinadorCarrera, Materia*} y las relaciones están representadas en las flechas que unen a las entidades, conformando el conjunto {*inscrito_en, imparte_clase_a, toma_clase_con, tiene_inscrito_a, tiene_como_titular, es_titular_de, registra_alumno*}. Los atributos de las clases son los que describen los conceptos, aunque no se presentan en la figura, se puede conformar el conjunto de atributos para la clase *Alumno* como: {*nombre, matricula, año_ingreso, promedio*}, por mencionar algunos.

2.2.2. Evaluación

Una vez creada y poblada una ontología, es necesario evaluar la estructura obtenida de acuerdo al dominio. Algunos métodos en la literatura son los clásicos en la recuperación de información, los cuales involucran dos métricas estándar para valorar la ontología: precisión (proporción de material relevante recuperado) y recuerdo (capacidad de la ontología para recuperar objetos) .

En los trabajos que aplican estos criterios se contempla tanto la dimensión semántica, inherente a cualquier ontología, como la vertiente de RI. Otros métodos están basados en diferentes teorías matemáticas (Senso et al., 2011), analizan fundamentalmente, la profundidad de la descripción en función del número de clases principales, subclases, notaciones formales y profundidad media del árbol taxonómico. La evaluación puede considerar los siguientes criterios:

- **Compilación:** Consistencia de la ontología (sin errores de estructura lógica).
- **Preguntas de competencia:** Proporción de preguntas respondidas correctamente de acuerdo al criterio de un experto en el dominio.
- **Tiempo de respuesta:** Tiempo que utiliza el razonador de las ontologías en realizar la inferencia.
- **Axiomas:** Son teoremas que se declaran sobre relaciones que deben cumplir los elementos de la ontología.

Por su parte, Raad & Cruz (2015) dividieron las técnicas de evaluación en cuatro grupos, de acuerdo a los métodos utilizados en el proceso y el enfoque de las ontologías. Los grupos se describen a continuación:

1. **Enfoques basados en *gold standard*:** También se conoce como alineación ontológica o mapeo de ontologías, es el enfoque más directo ya que permite obtener métricas de RI para evaluar la efectividad de los experimentos realizados, se obtienen algunas métricas utilizadas comúnmente en la evaluación de sistemas de recuperación de información.
2. **Enfoques basados en corpus:** También conocidos como enfoques impulsados por datos, se usan para evaluar el grado en que una ontología cubre un dominio dado. Sus métodos consisten en comparar la ontología aprendida con el contenido de un corpus de texto que cubre significativamente un dominio dado. La ventaja es comparar una o más ontologías

con un corpus, en lugar de comparar una ontología con otra existente. Entre las técnicas implementadas en este enfoque se encuentra la de realizar una extracción automática de términos en el corpus y contar el número de conceptos que se superponen entre la ontología y el corpus.

3. **Enfoques basados en tareas:** Los enfoques basados en tareas intentan medir hasta qué punto una ontología ayuda a mejorar los resultados de una determinada tarea. Este tipo de evaluación considera que una determinada ontología está destinada a una tarea en particular, y solo se evalúa de acuerdo con su desempeño en esta tarea, independientemente de todas las características estructurales (Raad & Cruz, 2015).
4. **Enfoques basados en criterios:** Estos enfoques miden hasta qué punto una ontología o taxonomía se adhiere a ciertos criterios deseables. Se puede distinguir entre medidas relacionadas con la estructura de una ontología y medidas más sofisticadas.

2.3. PLN y aprendizaje automático

El Procesamiento de Lenguaje Natural "es un área de investigación y aplicación que explora cómo las computadoras pueden ser usadas para entender y manipular el lenguaje natural tanto hablado como escrito" (Chowdhury, 2003, p. 51). Con el PLN se generan técnicas que permiten integrar modelos computacionales con la lingüística aplicada. Uno de los enfoques que se explora con interés especial, es el estudio y análisis de la información disponible en Internet, la cual, al no ser estructurada, requiere técnicas de representación que permitan integrarla a fin de lograr la formalización y dar respuesta a preguntas que involucren lenguaje natural por parte del usuario. Tanto en el procesamiento de lenguaje natural como en otras áreas de la inteligencia artificial, se utilizan varias técnicas para el procesamiento y análisis de textos. Las más representativas se describen en las siguientes subsecciones.

2.3.1. Preprocesamiento de texto

Tokenización Dada una secuencia de caracteres y una unidad definida para el documento (generalmente la unidad seleccionada es la palabra), tokenización es la tarea de "dividir" en partes, denominadas *tokens* en el documento, tomando en cuenta la unidad definida y al mismo tiempo, eliminando caracteres determinados, como los signos de puntuación (Manning & Schütze, 1999). La tokenización presenta distintas reglas según el idioma que se esté trabajando, por ejemplo, algunas palabras en inglés son compuestas, o unidas con apóstrofes. Un token puede ser una palabra, una oración, un párrafo, un documento o un n-grama.

Teniendo los documentos de la colección divididos en *tokens*, se pueden presentar casos en los que un mismo término sea representado con diferentes cadenas de símbolos, por ejemplo, el

término *U.S.A* y *USA*. Esto trae la necesidad de realizar un proceso de normalización, donde se toman en cuenta las posibles formas de escritura de una palabra. Entre las modificaciones que se realizan en el proceso de normalización se encuentran las siguientes:

- Determinar las posibles formas de escritura de una palabra (México y República Mexicana, abreviaciones, entre otras).
- Capitalización de las palabras (generalmente todas se pasan a minúsculas, lo mismo sucede con las consultas realizadas).
- En algunos casos también se toman en cuenta los sinónimos de los términos, por ejemplo *carro* y *automóvil*.
- Sustituir las palabras por sus lemas. Este proceso se refiere a realizar un análisis morfológico de las palabras, con el objetivo de eliminar desinencias[1] y devolver la base de una palabra (conocida como "lema").

Palabras cerradas Este tipo de palabras (también denominadas como *stopwords* por su término en inglés) son aquellas que son tan comunes en todos los documentos que dificultan la tarea de excluir documentos en una búsqueda Manning & Schütze (1999). La estrategia general para determinar una lista de palabras cerradas es determinar el número de apariciones del término en la colección de documentos, posteriormente se omiten en el índice las palabras que tengan una frecuencia mayor. A continuación, se muestran algunos ejemplos de estas palabras en el idioma español.

de	*y*	*las*	*no*	*como*	*ya*	*esta*
la	*a*	*por*	*una*	*más*	*o*	*entre*
que	*los*	*un*	*su*	*pero*	*este*	*cuando*
el	*del*	*para*	*al*	*sus*	*sí*	*muy*
en	*se*	*con*	*lo*	*le*	*porque*	*sin*

Utilizando una lista de palabras cerradas, se reduce significativamente el número de términos que el sistema tiene que almacenar, por lo tanto, también se reduce el tiempo de indexación. Otra ventaja del uso de las palabras cerradas es que en cada documento quedan las palabras que no son comunes en toda la colección.

Frecuencia de términos y ponderación

La frecuencia de términos se utiliza como característica al momento de aplicar alguna técnica de clasificación o de agrupamiento, esta también es utilizada al momento de recuperar documentos en un sistema de recuperación de información. Aunque es una manera de tomar en cuenta la aparición de las palabras, no es suficiente, una solución es calcular un *score* para cada documento, el cual calcula la coincidencia entre cada término de la consulta y el documento.

[1] Morfema flexivo añadido a la raíz de adjetivos, sustantivos, pronombres y verbos.

La manera de calcular el *score* es la siguiente: A cada término en los documentos de la colección se le asigna un peso, el cual depende del número de ocurrencias del término en el documento. Posteriormente se calcula el *score* ente un término t y un documento d (basado en el peso de t en d). Este *score* se conoce como "frecuencia del término" y es denotado como $Tf_{t,d}$ donde los subíndices indican la clave del término y el documento (fórmula 2.1).

$$Tf_{t,d} = Apariciones\ de\ t\ en\ d \tag{2.1}$$

Frecuencias Maximales Una secuencia frecuente maximal es una secuencia de palabras que debe aparecer en un número dado (umbral) de ejemplos (por ejemplo, documentos, oraciones, etc.) y además, no debe estar contenida en otra secuencia de palabras (Orta Palacios, 2008). Existen dos enfoques para la extracción de estas secuencias: por documento y por colección. La diferencia entre estos dos enfoques se basa en la unidad a tomar para el conteo, ya sea extraer las frecuencias tomando en cuenta la aparición por instancia (documento) o en todo el corpus (colección). En Vázquez Cuchillo (2008) se presentan varios métodos utilizando la extracción por documento, dicho proceso consiste en obtener las frecuencias maximales por instancia y usarlas para construir los vectores de cada documento. El tamaño de cada vector es igual al número de frecuencias por documento encontradas en la colección.

Colocaciones

El procesamiento de lenguaje natural (PNL) define una colocación como una expresión que consta de dos o más palabras que corresponden a alguna forma convencional de nombrar cosas. Las colocaciones tienen importancia en PNL, por ejemplo en la recuperación de información, el análisis y la detección de conceptos principales, entre otros. Además, la extracción de colocaciones es el primer paso en muchas tareas específicas. Algunas de las técnicas para la detección de colocaciones incluyen métricas semánticas Dinu et al. (2014), ontologías Li et al. (2015), modelos estadísticos Yu et al. (2003) y el uso de recursos externos dependiendo del idioma Pazos & Pamies (2006).

2.3.2. 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, ya sea mediante la aparición de caracteres o utilizando corpus, o recursos semánticos externos.

La Figura 2.4 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.

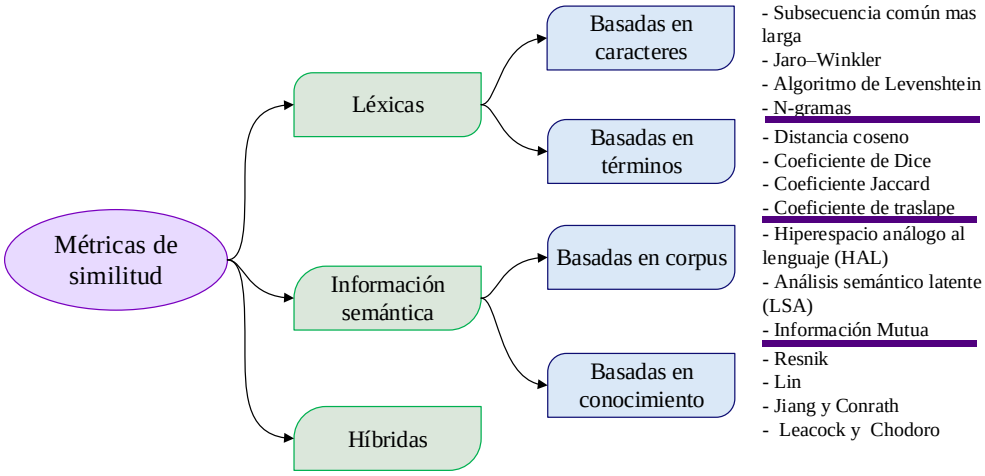


Figura 2.4: Clasificación de las métricas de similitud textual según Álvarez Carmona (2014) y Gomaa & Fahmy (2013)

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 (Álvarez Carmona, 2014). 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 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 2.2. El resultado está normalizado entre cero y uno donde cero es nula similitud mientras que uno se refiere a la máxima similitud (Al-Shamri, 2014).

$$sim_D\left(t_1, t_2\right) = 2\frac{\mid t_1 \cap t_2 \mid}{\mid t_1 \mid + \mid t_2 \mid} \tag{2.2}$$

Coeficiente de Jaccard. Parecido al coeficiente de Dice, este se obtiene al dividir la intersección de términos entre la unión de los mismos. Su fórmula se presenta en la ecuación 2.3 (Huang et al., 2011).

$$sim_J\left(t_1, t_2\right) = \frac{\mid t_1 \cap t_2 \mid}{\mid t_1 \cup t_2 \mid} \tag{2.3}$$

Coeficiente de Traslape. Similar al coeficiente de Jaccard pero considera solo la cardinalidad de caracteres del texto más pequeño en lugar de la unión de los caracteres (Gomaa & Fahmy, 2013). Este cambio se especifica en la ecuación 2.4.

$$sim_T\left(t_1, t_2\right) = \frac{\mid t_1 \cap t_2 \mid}{min\left(\mid t_1 \mid, \mid t_2 \mid\right)} \tag{2.4}$$

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 2.5).

$$sim_C\left(t_1, t_2\right) = \frac{\mid t_1 \cap t_2 \mid}{\sqrt{\mid t_1 \mid \mid t_2 \mid}} \tag{2.5}$$

Medidas de distancia

Una medida de distancia satisface las siguientes propiedades matemáticas (Han et al., 2000):

- Valores no negativos: La distancia es un número no negativo $d(i, j) \geq 0$
- Indiscernibles: La distancia de un objeto a sí mismo es 0. $d(i, i) = 0$
- Simetría: La distancia es una función simétrica. $d(i, j) = d(j, i)$.
- Desigualdad triangular: El ir directamente de objeto i con el objeto j en el espacio no es más que hacer un desvío por encima de cualquier otro objeto k. $d(i, j) \leq d(i, k) + d(k, j)$.

Una medida que cumple las últimas tres condiciones se conoce como métrica (la propiedad de no-negatividad está implícita en las tres propiedades), la cual se utiliza en tareas de procesamiento de texto para calcular la similitud entre documentos. Una de las métricas más utilizada en la literatura es la distancia coseno, en la cual los objetos se consideran vectores y su similaridad se mide por el ángulo que los separa usando el coseno. Su fórmula se determina por la ecuación 2.6, donde:.

$$S_{cos}(d_1, d_2) = \frac{(d_1 * d_2)}{\|d_1\|\|d_2\|} \tag{2.6}$$

- d_1 y d_2 son vectores que representan documentos.
- $\|d\|$ es la longitud del vector d.

2.3.3. Clasificación

Dentro de la minería de datos se pueden distinguir diversos tipos de tareas, una de las más utilizadas es la clasificación, la cual consiste en un proceso de dos pasos, como lo indica la figura 2.5. Se construye un modelo describiendo un conjunto de datos predeterminado (paso de aprendizaje o entrenamiento), posteriormente, se utiliza el conjunto de evaluación para estimar la precisión de las reglas de clasificación creadas anteriormente. Si la precisión se considera aceptable, las reglas se pueden aplicar a la clasificación de nuevos datos. En los experimentos realizados se utilizan principalmente dos tipos de clasificación, los cuales se describen a continuación.

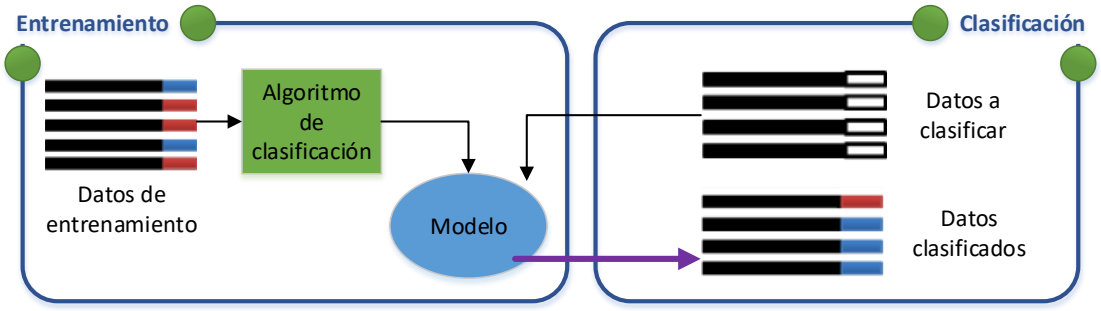


Figura 2.5: Etapas del proceso de clasificación.

Clasificación unaria También llamada OCC por sus siglas en inglés (*One Class Classification*). En un problema de clasificación, las clases están disponibles y la decisión es soportada por la presencia de ejemplos para cada clase; mientras que en un problema OCC, los datos negativos son nulos o escasos, por lo tanto solo una clase puede ser determinada usando esos datos. La tarea en OCC es definir un límite de clasificación alrededor de la clase positiva, de manera que acepte tantas instancias como sea posible de la clase positiva, mientras que minimiza la posibilidad de aceptar instancias atípicas. En OCC, es difícil decidir, sobre la base de una sola clase, qué tanto sesgo se le debe dar a cada uno de los valores de los atributos (Khan & Madden, 2010).

Clasificación no supervisada También denominada agrupamiento o *clustering*, por su término en inglés, es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio, el cual por lo general es la distancia o similitud. La cercanía se define en términos de una determinada función de distancia, como la euclídea, aunque existen otras más robustas o que permiten extenderla a variables discretas (Han et al., 2000).

De estos procedimientos se obtienen n grupos, cada uno es una región continua del espacio que contiene una densidad relativamente alta de puntos, y que se encuentra a su vez separada de otras regiones de alta densidad por regiones cuya densidad de puntos es relativamente baja. La medida más utilizada para medir la similitud entre los casos es la matriz de correlación entre los nxn casos. Sin embargo, también existen muchos algoritmos que se basan en la maximización de una propiedad estadística llamada verosimilitud.

Generalmente, los vectores de un mismo grupo comparten propiedades comunes. El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. La Tabla 2.1 muestra las principales características de algunos algoritmos de agrupamiento reportados en la literatura.

Tabla 2.1: Principales características de algunos algoritmos de agrupamiento. Fuente: Pedregosa et al. (2011)

Método	Parámetros	Escalabilidad	Usos	Agrupamiento
K-means	Número de grupos	Muestras muy grandes, valor de n mediano	Propósito general, no aplicable para n grande. Geometría plana	Distancia entre puntos
Espectral	Número de grupos	Muestras medianas, valos de n pequeño	Pocos grupos, geometría no plana	Gráfico de distancia
Aglomerativo	Número de grupos, tipo de vinculación, distancia	Muestras grandes, valor de n grande	Para muchos grupos, restricciones de conectividad, distancias no euclidíneas	Cualquier distancia
Birch	Factor de ramificación, umbral (número de grupos)	Muestras grandes, valor de n grande	Conjunto de datos grande, eliminación de *outlayers*	Distancia euclidínea

Uno de los algoritmos de agrupamiento más utilizados es *K-means* propuesto por Arthur & Vassilvitskii (2007), el cual asigna cada punto con el grupo cuyo centro es el más cercano. El centro es el promedio de todos los puntos del grupo, es decir, sus coordenadas son la media aritmética de cada dimensión de todos los puntos del grupo. En el proceso, crea subconjuntos de los datos de entrada originales al tratar de separar muestras en n grupos de igual varianza. Este algoritmo requiere el número de grupos como parámetro de entrada y se adapta bien a una gran cantidad de muestras.

El algoritmo de agrupamiento espectral define una matriz de afinidad entre muestras, seguido de un *K-means* en un espacio dimensional bajo. La agrupación espectral requiere que se especifique la cantidad de grupos. Funciona bien para un pequeño número de grupos, pero no se aconseja cuando n es mayor (Luxburg, 2007).

El agrupamiento por el método aglomerativo realiza una agrupación jerárquica utilizando un enfoque ascendente: cada instancia comienza en su propio grupo y los grupos se fusionan

sucesivamente (Pedregosa et al., 2011). Mientras que el método *Birch* (*Balanced Iterative Reducing and Clustering using Hierarchies*) almacena para cada grupo una tripleta de datos que contiene el número de objetos que pertenecen a ese grupo, el valor de la suma de todos los valores de los atributos de todos los objetos pertenecientes al grupo, y la suma de los cuadrados de los atributos de los objetos que pertenecen al grupo. Con esta información se construye un árbol de grupos (Zhang et al., 1996).

La evaluación de agrupamiento evalúa la viabilidad del análisis de agrupación en un conjunto de datos y la calidad de los resultados generados por dicho método. Las tareas incluyen evaluar la tendencia de agrupamiento, determinar el número de grupos y medir la calidad de los grupos (Han et al., 2000). Algunas métricas utilizadas para la evaluación de grupos se describen en los siguientes párrafos (Pedregosa et al., 2011).

Información mutua. Es una función que mide la concordancia entre dos asignaciones, ignorando las permutaciones (etiquetas verdaderas y las etiquetas predicen). El rango de resultados es de 0 a 1, donde 1 es la combinación perfecta. El valor de la información mutua no se ajusta por casualidad y tenderá a aumentar a medida que aumente el número de grupos diferentes. La fórmula general se define en la ecuación 2.7. Donde U y V son las etiquetas reales y predichas por el algoritmo de agrupamiento, respectivamente. Este concepto usa la entropía para realizar el cálculo de la métrica.

$$MI\left(U,V\right)=\sum_{i=1}^{|U|}\sum_{j=1}^{|V|}P\left(i,j\right)log\left(\frac{P\left(i,j\right)}{P\left(i\right)P'\left(j\right)}\right) \tag{2.7}$$

Índice Rand. Es una función que mide la similitud entre las etiquetas verdaderas y las obtenidas por el algoritmo de agrupamiento; es simétrico, por lo que no afecta el orden en que se procesan las etiquetas. El rango de resultados va de -1 a 1, donde los valores negativos se consideran malos y 1 es la similitud perfecta. Si C es una tarea de clasificación y K es el resultado de aplicar un algoritmo de agrupamiento, a y b se definen como:

- a: El número de pares de elementos que están en el mismo conjunto en C y en el mismo conjunto en K.
- b: El número de pares de elementos encontrados en diferentes conjuntos en C y en diferentes conjuntos en K.

El índice Rand se determina mediante la ecuación 2.8, donde el denominador indica el total de pares posibles en el conjunto de datos que se ordenarán.

$$RI = \frac{a+b}{C_2^{n_{samples}}} \tag{2.8}$$

Homogeneidad. Es la métrica en la que cada grupo contiene solo los miembros de una sola clase, este concepto se complementa por exhaustividad, en el que todos los miembros de una clase dada se asignan al mismo grupo. Los valores de la homogeneidad están entre 0 a 1 y su fórmula matemática está dada por la ecuación 2.9, donde H es la entropía condicional de las clases.

$$h = 1 - \frac{H\left(C \mid K\right)}{H\left(C\right)} \tag{2.9}$$

Fowlkes. Esta puntuación se define como la media geométrica de precisión y el recuerdo. Su expresión matemática se representa en la ecuación 2.10, donde TP, FP y FN son verdaderos positivos, falsos positivos y falsos negativos, respectivamente.

$$FMI = \frac{TP}{\sqrt{\left(TP+FP\right)\left(TP+FN\right)}} \tag{2.10}$$

2.3.4. Filtrado de documentos

Information filtering por su nombre en inglés, es un sistema que proporciona a los usuarios solo la información relevante para ellos mediante la eliminación de datos redundantes o no deseados, ya sea mediante métodos semiautomáticos o automáticos, previo a la presentación de la información (Hanani et al., 2001). Existen tres tipos de sistemas de filtrado de información, los cuales dependen de la forma en que se seleccionan los documentos: los cognitivos en los cuales la selección se basa en las características de su contenido, los sociales en donde la selección se realiza a partir de recomendaciones de otros usuarios, y los económicos, en donde la selección se basa en algún cálculo de costo/beneficio para el usuario (Ferreira & Atkinson Abutridy, 1998).

2.4. Dominio pedagógico

En este apartado se muestra el dominio elegido para la implementación de los experimentos. Como se mencionó en la sección 1.1.1, se utiliza el dominio pedagógico, dado que este dominio es muy extenso, se delimita el análisis a los elementos relacionados al aprendizaje significativo dentro de clases presenciales, integrando aspectos del docente y del estudiante.

Ausubel & Novak (1983) define al aprendizaje significativo como el proceso que se genera en la mente humana cuando subsume nuevas informaciones de manera no arbitraria y sustantiva y que requiere como condiciones: predisposición para aprender y material potencialmente significativo

que, a su vez, implica significatividad lógica de dicho material y la presencia de subsumidores o ideas de anclaje en la estructura cognitiva del que aprende.

La teoría del aprendizaje significativo es una teoría que, probablemente por ocuparse de lo que ocurre en el aula y de cómo facilitar los aprendizajes que en ella se generan, ha impactado profundamente en los docentes Rodriguez (2011). Para abordar esta temática, se analiza la personalización del aprendizaje, donde un estudiante aprende mejor con ciertos esquemas y técnicas de acuerdo a sus características intrínsecas. Cakula & Sedleniece (2013) definen la personalización como la adaptación de la experiencia de aprendizaje a cada estudiante, todo ello mediante el análisis del conocimiento, las habilidades y las preferencias de aprendizaje de cada individuo.

La personalización en el proceso de enseñanza-aprendizaje puede ser conducido siguiendo diferentes teorías que relacionan varios aspectos, dadas estas reflexiones, para investigar esta área, se propone analizar tres tópicos específicos a fin de detectar conceptos importantes que permitan establecer la relación entre estos. Dos de estos tópicos están relacionados con el estudiante (inteligencias múltiples y estilos de aprendizaje) y uno está relacionado con los docentes (estrategias de enseñanza aprendizaje). En las siguientes subsecciones se analizarán cada uno de estos, mientras que la Figura 2.6 muestra la relación que existe entre los tópicos analizados.

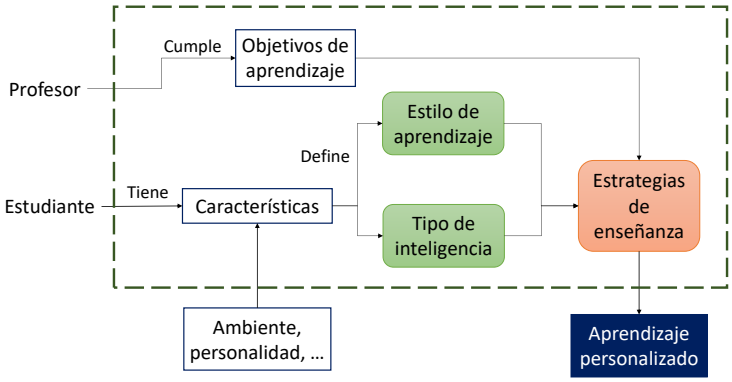


Figura 2.6: Papel de los elementos seleccionados en el aprendizaje significativo

2.4.1. Estrategias de enseñanza 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 (Barriga & Hernández, 2004), 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" (Weinstein & Mayer, 1986, p. 315). Según Genovard & Gotzens (1990), las estrategias de aprendizaje son comportamientos que el estudiante desarrolla durante su proceso de aprendizaje, influyendo en su proceso de codificación de la información.

Los rasgos característicos más destacados de las estrategias de aprendizaje son los siguientes (Valle et al., 1998):

- Su aplicación no es automática sino controlada.
- Para que un estudiante pueda poner en marcha una estrategia debe disponer de recursos alternativos, entre los que decide utilizar, en función de las demandas de la tarea, aquellos que él cree más adecuados.
- Las estrategias están constituidas de otros elementos más simples, que son las técnicas o tácticas de aprendizaje y las destrezas o habilidades.

Aunque existen muchos enfoques para la clasificación de las estrategias de aprendizaje, González & Tourón (1992) define tres tipos principales, los cuales se describen a continuación:

- **Estrategias cognitivas:** Hacen referencia a la integración del material nuevo con el conocimiento previo. En este sentido, serían un conjunto de estrategias que se utilizan para aprender, codificar, comprender y recordar la información al servicio de metas de aprendizaje determinadas. Dentro de este grupo, se distinguen tres clases de estrategias: estrategias de repetición, de elaboración, y de organización. La estrategia de repetición se trata de un mecanismo de la memoria que activa los materiales de información para mantenerlos en la memoria a corto plazo y, a la vez, transferirlos a la memoria a largo plazo (Beltrán, 1993). La estrategia de elaboración trata de integrar los materiales informativos relacionando la nueva información con la información ya almacenada en la memoria mientras que la estrategia de organización intenta combinar los elementos informativos seleccionados en un todo coherente y significativo.
- **Estrategias metacognitivas:** Hacen referencia a la planificación, control y evaluación por parte de los estudiantes de su propia cognición. Son un conjunto de estrategias que permiten el conocimiento de los procesos mentales, así como el control y regulación de los mismos con el objetivo de lograr metas de aprendizaje determinadas. El conocimiento metacognitivo requiere consciencia y conocimiento de variables de la persona, de la tarea y de la estrategia. En relación con las variables personales está la consciencia y conocimiento que tiene el sujeto de sí mismo y de sus capacidades y limitaciones cognitivas. Las variables de la tarea se refieren a la reflexión sobre el tipo de problema que se va a tratar de resolver; averiguar el objetivo de la tarea, si es familiar o novedosa, cuál es su nivel de dificultad, entre otras actividades. En cuanto a las variables de estrategia, incluyen el conocimiento acerca de las estrategias que pueden ayudar a resolver la tarea (Valle et al., 1998).

- **Estrategias de manejo de recursos:** También denominadas estrategias de apoyo, son una serie de estrategias de apoyo que incluyen diferentes tipos de recursos que contribuyen a que la resolución de la tarea se lleve a buen término. Tienen como finalidad sensibilizar al estudiante con lo que va a aprender; y esta sensibilización hacia el aprendizaje integra tres ámbitos: la motivación, las actitudes y el afecto. Este tipo de estrategias, en lugar de enfocarse directamente sobre el aprendizaje tienen como finalidad mejorar las condiciones materiales y psicológicas en que se produce el aprendizaje. Incluyen aspectos claves que condicionan el aprendizaje como son, el control del tiempo, la organización del ambiente de estudio y el manejo y control del esfuerzo (Valle et al., 1998).

Otros autores como Diaz-Barriga & Hernández-Rojas (2010) y Ramón (2006) analizan las estrategias de diferentes formas, en las que destacan cuatro grupos:

- **Cognoscitivas:** Son capacidades internamente organizadas de las cuáles hace uso el estudiante para guiar su propia atención, aprendizaje, recuerdo y pensamiento. Emplea estrategias cognoscitivas para pensar acerca de lo que ha aprendido y para la solución de problemas.
- **Enseñanza:** Se concretan en una serie de actividades de aprendizaje dirigidas a los estudiantes y adaptadas a sus características, a los recursos disponibles y a los contenidos objeto de estudio. Las actividades deben favorecer la comprensión de los conceptos, su clasificación y relación, la reflexión, el ejercicio de formas de razonamiento.
- **Didácticas:** Son el sistema de acciones y operaciones, tanto física como mentales, que facilitan la confrontación del sujeto que aprende con el objeto de conocimiento.
- **Aprendizaje:** Son un conjunto de pasos o habilidades que un estudiante adquiere y emplea de forma intencional como instrumento flexible para aprender significativamente y solucionar problemas y demandas académicas.

2.4.2. Estilos de aprendizaje

El aprendizaje se define como un cambio en la conducta debido a la experiencia (Chance, 2001), otras definiciones integran otros elementos relacionados con la didáctica, como la de Alonso et al. (2007) en la que se describe como el proceso de adquisición de una disposición, relativamente duradera, para cambiar la percepción o la conducta como resultado de una experiencia.

Los autores previamente mencionados asumen el aprendizaje como un proceso en el que el sujeto que aprende lo hace de manera dinámica, de acuerdo con unas disposiciones y características particulares y en el que están involucrados un orden y un procedimiento lógico. De este análisis surge el estudio de los estilos de aprendizaje, los cuales reflejan la forma en que el individuo aprende, describen las condiciones bajo las que un discente se encuentra en la mejor situación para aprender, o qué estructura necesita para mejorar el proceso de

aprendizaje. 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 (1976), 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 (Olivos et al., 2016). El método propone 4 estilos de aprendizaje, los cuales se resumen en la Tabla 2.2 y se explican posteriormente.

Tabla 2.2: Características principales y preguntas clave de los estilos de aprendizaje. Fuente: Alducin-Ochoa & Vázquez (2016)

Estilo	Características	Preguntas clave
Activo	Animador Improvisador Descubridor Arriesgado Espontáneo	¿Aprenderé algo nuevo, algo que no sabía o no podía hacer antes? ¿Habrá una amplia variedad de actividades diversas? ¿Se aceptará que intente algo nuevo, cometa errores, me divierta? ¿Encontraré algunos problemas y dificultades que signifiquen un reto para mí? ¿Habrá otras personas de mentalidad semejante a la mía con las que pueda dialogar?
Reflexivo	Ponderado Concienzudo Receptivo Analítico Exhaustivo	¿Tendré tiempo suficiente para analizar, asimilar y preparar? ¿Habrá oportunidades y facilidad para reunir información pertinente? ¿Habrá posibilidades de oír los puntos de vista de otras personas de enfoques diferentes? ¿Me veré sometido a presión para actuar precipitadamente o improvisar?
Teórico	Metódico Lógico Objetivo Crítico Estructurado	¿Habrá muchas oportunidades de preguntar? ¿Los objetivos y las actividades del programa revelan una estructura y finalidad clara? ¿Encontraré ideas y conceptos complejos capaces de enriquecerme? ¿Son sólidos y valiosos los conocimientos y métodos que van a utilizarse? ¿El nivel del grupo será similar al mío?
Pragmático	Experimentador Práctico Directo Eficaz Realista	¿Habrá posibilidades de practicar y experimentar? ¿Habrá suficientes indicaciones prácticas y concretas? ¿Se abordarán problemas reales y me ayudarán a resolver algunos de mis problemas?

- *Activo:* Personas que se involucran con experiencias nuevas, tienden a actuar primero y luego piensan en las consecuencias. Muestra como principales conductas al momento de aprender la animosidad, la improvisación, la búsqueda y el descubrimiento de novedad, el riesgo y la espontaneidad (Esguerra & Guerrero, 2010).
- *Reflexivo:* Personas que son observadores y analizan sus experiencias desde diferentes perspectivas. Recopilan y analizan datos en detalle antes de tomar una conclusión. Este perfil tiene conductas de receptividad, ponderación, análisis, exhaustividad y toma de

conciencia, y entre las otras menos centrales pero presentes en él, la observación, la identificación de pequeños detalles, la elaboración de argumentos, la previsión, la habilidad para redactar informes y la prudencia (Esguerra & Guerrero, 2010).

- *Teórico:* Muestra dentro de las principales características la lógica, la metódica, la objetividad, la criticidad y la estructuración en las acciones (Esguerra & Guerrero, 2010). Las personas con este estilo de aprendizaje adaptan e integran sus observaciones en teorías complejas y lógicamente fundadas. Su sistema de valores prioriza la lógica y la racionalidad antes del análisis y la síntesis.
- *Pragmático:* Incluye personas que prueban sus ideas, teorías y técnicas nuevas y tratan de ver si funcionan en la práctica. No les gustan las discusiones largas sobre el mismo tema. Son prácticos y se adhieren a la realidad. Las cinco principales características de este estilo de aprendizaje se hallan la experimentación, la practicidad, el dirigirse a situaciones y a personas de manera directa, la eficacia y el realismo (Esguerra & Guerrero, 2010).

Para determinar el estilo de aprendizaje predominante, se utiliza el Cuestionario Honey Alonso de Estilos de Aprendizaje (CHAEA), el cual contiene 80 reactivos, 20 relacionadas con cada estilo. Dicho cuestionario se muestra en el apéndice B.

Varios autores como Tapias (2018) relacionan estos estilos de aprendizaje con estrategias de enseñanza, las cuales son idóneas para las características de cada estudiante. La tabla 2.3 muestra un extracto de dicho análisis.

2.4.3. Inteligencias múltiples

Una inteligencia implica "implica la habilidad necesaria para resolver un problema o para elaborar productos que son importantes en un contexto cultural" (Gardner, 2001, p. 10). Esta definición, según el autor, contiene dos elementos principales:

1. **La resolución de problemas:** El tener un problema para resolver, significa que la actividad mental siempre tiene una meta. Es importante considerar que los problemas van desde los simples a los complejos.
2. **La creación de un producto cultural:** Creaciones cuya importancia están demarcadas por las culturas, igualmente se puede decir que van desde productos rudimentarios pero útiles, pasando por tecnologías sociales, hasta el desarrollo de la llamada tecnología dura, todas ellas en función del mejoramiento de la calidad de vida de las sociedades humanas.

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 de Gardner, (2001). Se distinguen ocho tipos de inteligencias, las cuales se muestran en la figura 2.7 y se describen a continuación:

Tabla 2.3: Estrategias de aprendizaje y estilos que favorecen.

Estrategia	Estilo que favorece
Lluvia o tormenta de ideas. Forma de trabajo que permite la libre presentación de ideas, sin restricciones ni limitaciones, con el objetivo de producir ideas originales o soluciones nuevas.	Activo
Estudio de un caso. Descripción escrita de un hecho acontecido en la vida de una persona, grupo u organización. La situación descrita puede ser real o hipotética, pero construídas con características análogas a las presentadas en la realidad.	Teórico
Situación problema. El profesor selecciona una situación problema tomada de la realidad y relacionado con los contenidos del curso que se espera sean abordadas por el alumno de manera grupal. Lo fundamental en la forma de trabajo que se genera está en que los alumnos puedan identificar lo que requieren para enfrentar la situación problemática y las habilidades que se desarrollan para llegar a resolverla.	Pragmático
Concordar - Discordar. Se fundamenta en presentar a los alumnos un mínimo de 10 y máximo de 20 enunciados breves y redactados de forma tal que provoque en los discentes la reflexión (de manera individual y después en equipos de cuatro integrantes). El alumno debe contestar si está de acuerdo o en desacuerdo con lo que se escribió.	Reflexivo
Lámina/foto mural. Se basa en la presentación de una fotografía, lámina o caricatura (sin texto) proyectada como entrada a un tema de la lección que se quiere ver.	Pragmático, Activo
Elaboración de blogs y wikis. Se utilizan para plasmar ideas propias sobre temas entendidos a través de medios electrónicos interactivos.	Activo, Reflexivo
Elaboración de mapas conceptuales. Como un medio de representación que permite visualizar los conceptos y proposiciones de un texto, así como la relación que existe entre ellos.	Teórico,Pragmático

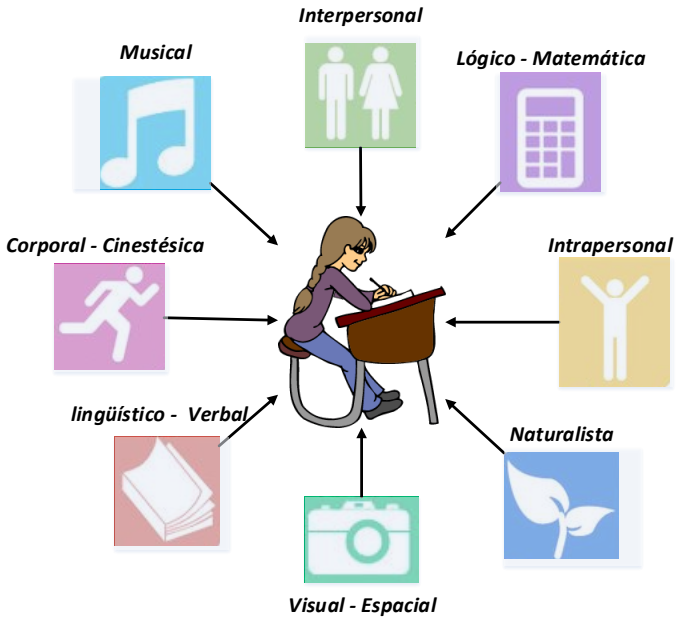


Figura 2.7: Tipos de inteligencias según Gardner.

➥ *Lógico-matemática:* Utilizada para resolver problemas de lógica y matemáticas. Es la inteligencia que tienen los científicos. Se corresponde con el modo de pensamiento del

hemisferio lógico y con lo que la cultura ha considerado siempre como la única inteligencia.

- *Lingüística:* La que tienen los escritores, los poetas, los buenos redactores. Utiliza ambos hemisferios.
- *Espacial:* Consiste en formar un modelo mental del mundo en tres dimensiones. Es la inteligencia que tienen los marineros, los ingenieros, los cirujanos, los escultores, los arquitectos o los decoradores.
- *Musical:* Se relaciona con la capacidad de percibir, discriminar, transformar y expresarse mediante las formas musicales. Asimismo, incluye las habilidades en el canto dentro de cualquier tecnicismo y género musical, tocar un instrumento a la perfección y lograr con él una adecuada presentación, dirigir un conjunto y tener apreciación musical.
- *Corporal-kinestésica:* Capacidad de utilizar el propio cuerpo para realizar actividades o resolver problemas. Es la inteligencia de los deportistas, los artesanos, los cirujanos y los bailarines.
- *Intrapersonal*: Es la que permite a una persona entenderse a sí misma. No está asociada a ninguna actividad concreta.
- *Interpersonal:* Como entender a los demás, se suele encontrar en los buenos vendedores, políticos, profesores o terapeutas.
- *Naturalista:* La que se utiliza cuando se observa y estudia. Es la que demuestran los biólogos o los herbolarios.

Para determinar el tipo de inteligencia predominante, se utiliza el *test* de inteligencias múltiples de Howard Gardner el cual contiene 36 reactivos, dicho *test* se muestra en el apéndice C

Capítulo 3

Estado del arte

En el presente capítulo se muestra un análisis sobre las investigaciones en torno a cada una de las fases del aprendizaje ontológico. Para esto, se analizan los artículos relacionados con la creación, uso de herramientas y dominios utilizados en esta investigación. Al final, se agregan algunas investigaciones sobre técnicas PLN que son utilizadas en la investigación y un análisis sobre los puntos más relevantes y limitantes encontrados en las investigaciones revisadas.

Dentro del proceso de revisión de la literatura, se encontraron autores como Maedche & Staab (2001), quienes presentan una metodología para el aprendizaje ontológico utilizando herramientas semiautomáticas. El método involucra cinco etapas que conforman un ciclo: unión de estructuras existentes o definición de reglas, extracción de los modelos para la ontología, bosquejo inicial de la ontología objetivo, refinamiento y la aplicación de la ontología para su evaluación. Dichos procedimientos son aplicados a diversos entornos como el texto libre, diccionario y ontologías heredadas.

Investigaciones como esta, Tovar et al. (2014) y recientemente Gong & Gao (2016) trabajan en el proceso de aprendizaje ontológico como un todo, mientras que otras investigaciones se centran en una o dos fases del proceso. Además cada investigación maneja diferentes dominios y procedimientos para la obtención de sus resultados. Algunos autores se enfocan en el análisis del dominio utilizando diferentes técnicas, otros, analizan el comportamiento de técnicas y métricas independientemente del dominio. Rodríguez & Simón (2013) presenta un método para la extracción de información estructurada desde textos escritos en idioma español. Dicho método

combina el análisis sintáctico superficial y profundo o de dependencias, el reconocimiento de entidades, patrones lingüísticos y conocimientos de referencia almacenado en un corpus de mapas conceptuales. Analizando el dominio comercial, Vasilateanu et al. (2015) presentan un motor de búsqueda semántico para documentos relevantes de una empresa, en donde realizan una ontología basados en las técnicas de Cimiano (2006) y la librería *Text2Onto* (Cimiano & Völker, 2005)

La Figura 3.1 muestra las etapas generales del aprendizaje ontológico, además del dominio elegido. En cada una de las etapas se anexan las características y técnicas más representativas de los artículos encontrados en cada una de ellas.

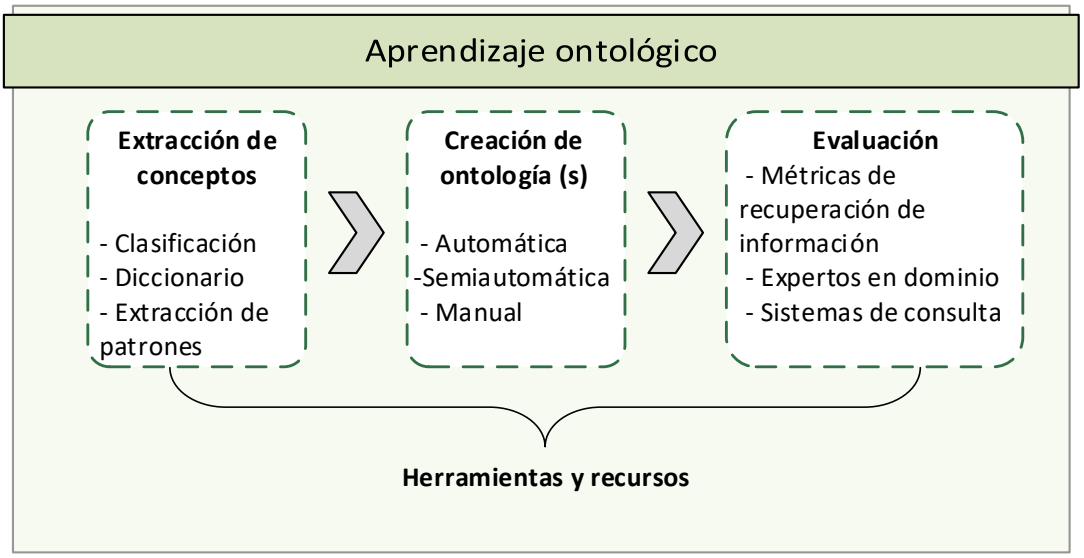


Figura 3.1: Rubros de investigación para el estado del arte

3.1. Detección de conceptos principales y relaciones

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. Grljevic & Bosnjak (2015) se centran en la creación del 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. Teixeira et al. (2011) se analizaron 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 la de Ochoa et al. (2011) 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) (García-Miguel et al., 2010) 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 (Ochoa et al., 2011). La extracción de clases se complementó con la opinión de expertos en el dominio. Kang et al. (2014) presentan un método para la extracción de conceptos utilizando extracción de patrones lingüísticos y cálculo de pesos con métricas de PLN como el etiquetado morfológico.

Aguilar et al. (2016) formulan una metodología para extraer contextos definitorios desde un corpus de biomedicina en español mediante un análisis semiautomático. El método consta de tres pasos principales: listado de términos, listado de definiciones, y una taxonomía basada en relaciones de hiponimia e hiperonimia, los autores utilizan WordNet y medidas para relevancia de palabras. Kaushik & Chatterjee (2018) proponen un esquema de dos pasos para diseñar una ontología del dominio de la agricultura. El esquema propuesto funciona en dos pasos. Utilizan expresiones regulares dependientes del dominio y técnicas de procesamiento del lenguaje natural para la extracción automática de vocabulario, posteriormente se identifican las relaciones entre los términos extraídos y las frases. Otras investigaciones utilizan la extracción de relaciones como parte de una tarea mayor, por ejemplo, los autores Jing et al. (2013) usan la clasificación abstracta para extraer conceptos y relaciones en la tarea de creación de ideas.

Una vez obtenidos los conceptos importantes, se sigue con el proceso de extracción de relaciones entre estos. Algunos autores separan estos procedimientos en dos fases, mientras que otras investigaciones recuperar conceptos y relaciones con un solo método. En Ortega-Mendoza et al. (2011) se presenta un enfoque para la extracción automática de información léxica en forma de pares hipónimo-hiperónimo, se utilizan un conjunto de patrones léxicos del español y un esquema de calificación de patrones. Las búsquedas se realizan directamente en la Web, donde el método extrae los pares de palabras. Ontras investigaciones como la de Barciela & Padilla (2012) utilizan un corpus compuesto de noticias, conversaciones telefónicas y texto web para la representación de textos utilizando grafos, donde los nodos son conceptos y las aristas relaciones. Los autores manejan 3 tipos de conceptos y 4 de relaciones.

Dorantes et al. (2017) presentan un proyecto centrado en la extracción de definiciones analíticas

e hiperónimos. La metodología consiste en la búsqueda automática de información con patrones construidos manualmente basados en la estructura léxica de definiciones analíticas en lenguaje natural. Los resultado de la extracción de definiciones analíticas superan a los del estado del arte, sin embargo, la extracción de hiperónimos se encuentra en las etapas iniciales.

Investigaciones más recientes como la de Das et al. (2019) realizan una búsqueda de relaciones en un corpus sobre crímenes, mediante la creación de un grafo. La propuesta considera un corpus textual que contiene información sobre el crimen contra las mujeres en la India y extrae relaciones sustanciales entre las entidades nombradas mediante una técnica de agrupación basada en gráficos jerárquicos. Para extraer las relaciones, se utilizan métricas se similitud y se establece un umbral para dividir el gráfico en función de los pesos de los bordes. Otras investigaciones utilizan técnicas de lógica para extraer dichas relaciones, Lima et al. (2019) presenta el enfoque de aprendizaje relacional llamado *OntoILPER*, el cual, utiliza la programación lógica inductiva para generar modelos de extracción en forma de reglas de extracción simbólica.

3.2. Creación de ontologías

De acuerdo con el proceso de aprendizaje ontológico, el siguiente paso corresponde a la creación de la ontología. La Tabla 3.1 muestra las investigaciones analizadas tanto para la creación automática como para la creación manual de ontologías. Además, se anexa una columna para especificar el dominio en cada investigación.

Tabla 3.1: Investigaciones y dominios analizadas para la creación de ontologías

Método de construcción	Autores	Dominio
Automática	Alani et al. (2003)	Biografías de pintores
	Valencia (2005)	Textos técnicos y médicos
	Lee et al. (2007)	Documento no estructurado (FIFA[1])
	Ochoa et al. (2011)	Oncología
	De la Villa Moreno (2016)	Biografías
Manual	Vargas-Vera & Celjuska (2004)	Noticias sobre nivel básico
	Corda et al. (2008)	Historia de la ciencia
	Somodevilla et al. (2015)	Análisis de personas con NCD[2]

Alani et al. (2003) presentan el proyecto *Artequackt*, un sistema generador de biografías de acuerdo a parámetros establecidos por el usuario. En este proyecto la ontología se realiza de manera automática utilizando herramientas como WordNet[3]. Otras investigaciones como Lee et al. (2007) muestran un mecanismo para la construcción de ontologías basado en la extracción de episodios en un domino de documentos no estructurados. Dado que los proyectos se realizan para el idioma chino, el enfoque principal es estudiar las características de dicho idioma antes

[3] `https://wordnet.princeton.edu/`

de la construcción de la ontología. Se hacen pruebas con noticias de la FIFA evaluándolas con métricas de recuperación de información como precisión y recuerdo.

Además de las investigaciones previamente citadas, se analizaron tesis enfocadas en la creación de ontologías. En Valencia (2005), se desarrolla un entorno para la extracción incremental de conocimiento desde texto natural. Se propone una metodología híbrida que utiliza *POSTagging* y WordNet para la extracción de elementos clave. El proceso final es semiautomático y requiere entrenamiento previo de un corpus del dominio analizado; además, se extraen relaciones taxonómicas y conceptos semánticos.

Ochoa et al. (2011) propone una metodología de obtención de información para la construcción automática de ontologías en español a partir de texto libre, esto principalmente para la extracción de conocimiento de la Web. La metodología está basada en tres etapas secuenciales: búsqueda de conceptos, extracción de relaciones y construcción de ontologías. De la Villa Moreno (2016) realiza una investigación similar pero con un método que no analiza la estructura sintáctica superficial del lenguaje, sino que estudia su nivel semántico profundo (lo que permite escenarios multilingües); se utilizan técnicas de resolución de anáforas, agrupamiento y extracción de patrones léxico-sintácticos.

Vargas-Vera & Celjuska (2004) presenta una ontología para el reconocimiento de eventos utilizando 200 artículos de noticias de un conjunto de reportajes que describen la vida académica del nivel básico. Se realiza extracción de patrones y evaluación con métricas de RI, reportando resultados superiores al 90 % de precisión. En la investigación propuesta por Corda et al., (2008) se implementa una aplicación Web para ayudar a los usuarios a examinar un espacio conceptual histórico y explorar relaciones temporales entre eventos científicos. La ontología se formuló utilizando predicados generales (semánticos y fácticos) y se presenta un análisis detallado de las reglas y relaciones que la componen. En Somodevilla et al. (2015) se presenta una ontología construida manualmente sobre el estilo de vida en personas con enfermedades no transmisibles utilizando herramientas de Web Semántica. El enfoque principal de la investigación se centra en el uso de técnicas para la integración de ontologías.

En Hajiabadi (2014) se propone un enfoque de minería de datos basado en ontologías para clasificar documentos Web con el fin de facilitar las aplicaciones basadas en documentos de texto clasificados, como los motores de búsqueda. La ontología es generada por la minería Wikipedia. Debido a los esfuerzos colaborativos de muchos usuarios para agregar nuevos artículos, Wikipedia se expande enormemente y, en consecuencia, contiene casi todos los campos y subcampos. La ontología se denominó WikiOnt y contiene todas las categorías y subcategorías existentes en Wikipedia. Otras investigaciones se enfocan en el uso de ontologías

para la obtención de conceptos clave o toma de decisiones respecto a un corpus. Mala & Lobiyal (2015) utiliza una ontología lingüística para extraer conceptos de documentos de texto del ámbito de la medicina. Por medio de WordNet se crean grupos con un peso semántico asignado y técnicas de minería de datos para la extracción de conceptos relacionados.

3.3. Técnicas PLN

En este apartado se discuten las investigaciones relacionadas con técnicas implementadas a lo largo de la investigación. Entre dichas técnicas utilizadas se encuentra el filtrado de documentos y el análisis de varianza por grupos.

En un sistema de Recuperación de Información, el filtrado de documentos se puede aplicar para calcular la relevancia entre un documento y una clase determinada, ya sea mediante un conjunto de palabras clave (Li et al., 2018) o mediante algún atributo extra que determina si un documento es importante o no. Por ejemplo, puede ser usado en aplicaciones donde se desea seleccionar sitios web de acuerdo al estilo de aprendizaje de un estudiante (Bergasa-Suso et al., 2005). Recientemente se han propuesto modelos basados en redes neuronales, donde el filtrado de documentos se implementa con autoaprendizaje y capacidad de adaptación (Li, 2018).

Aunque el filtrado de documentos suele usarse para tareas como la selección de características (Zorić et al., 2018), en otras investigaciones es un paso previo cuyo resultado es la entrada de un sistema de búsqueda de respuestas (Noguera Robles et al., 2006). Ésta técnica también se utiliza dentro del proceso de clasificación, especialmente cuando se tienen pocas o nulas instancias de la clase que interesa a la investigación; los ejemplos más comunes es la clasificación binaria (Yu et al., 2003) y problemas de clasificación sin datos (Guan et al., 2009). Es importante hacer notar que la mayoría de los trabajos realizados en el área de filtrado de documentos suelen aplicarse a dominios específicos, por lo que si se desea hacer un cambio de dominio es necesario hacer ajustes a los métodos.

En el análisis de la varianza Dinu et al. (2014) propuso un método para la tarea de detección de colocaciones, el cual consiste en aplicar algunos métodos como la métrica Dice, prueba de ji-cuadrado, y la razón de probabilidad. En el área de detección de elementos compuestos, Li et al. (2015) proponen un enfoque de cálculo de similitud semántica basada en ontología, donde el concepto se descompone en dos conjuntos para extrae la similitud.

En el área de procesamiento de imágenes, el análisis de varianza se utiliza para la segmentación, especialmente para la estimación de un umbral, es decir, para separar la imagen en dos clases de pixéles: blanco y negro. En Zhang & Hu (2008) se utilizó el método bidimensional (2D) de

Otsu para corregir el umbral al segmentar imágenes de baja relación señal / ruido, este método eliminó completamente el ruido en las imágenes de muestras de biopsia renal.

3.4. Investigaciones en el dominio pedagógico

En este apartado se analizan algunas investigaciones en el dominio pedagógico, enfocadas en una o más fases del proceso ontológico. La tabla 3.2 muestra algunos de los dominios específicos analizados.

Grandbastien et al. (2007) proponen 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. Fu et al. (2008) también analizan el dominio educativo, sin embargo, al ser aplicado al idioma Chino, utilizan un preprocesamiento para analizar las características de dicho idioma: acoplamiento, relevancia y consenso.

Tabla 3.2: Investigaciones en el dominio pedagógico

Fase	Método	Autores	Dominio
Detección de clases	Automático	Grandbastien et al. (2007)	*E-learning*
		Fu et al. (2008)	Sistemas de preguntas
Creación	Manual o semiautomática	Wu (2008)	Manuales de cursos en línea
		Fan et al. (2008)	Material para enseñanza del inglés
		Zhu & Yao (2009)	Secuencias de aprendizaje
		Bucos et al. (2010)	Contenidos en *E-learning*
		Dai & Li (2010)	Informática
		Du et al. (2012)	Educación en línea
		Bagiampou & Kameas (2012)	Cursos de ingeniería de software
		Ameen et al. (2012)	Cursos de ingeniería de software
		Silva Sprock & Ponce (2013)	Objetos de aprendizaje
		Méndez et al. (2015)	Inteligencias múltiples de Gardner
		Uskov et al. (2016)	Niveles de inteligencia
		Hu et al. (2016)	Libros para nivel K12
		Hssina et al. (2017)	E-learning
		Aminah et al. (2017)	Evaluación académica

En Wu (2008) se desarrolla un sistema de educación en Internet basado en ontologías, que implementa el intercambio y la reutilización del material de aprendizaje en diferentes sistemas. Es una investigación cualitativa, donde se aborda un ejemplo con un curso básico de computación en línea que describe los módulos del sistema: aprendizaje, interfaz y recursos. Fan et al. (2008) presentan ENGOnto; la ontología integra múltiples ontologías relevantes para que los agentes personalizados aborden los cambios dinámicos del proceso de aprendizaje de los alumnos, la interacción entre el instructor y los recursos de aprendizaje en el entorno de la enseñanza del idioma inglés. La ontología se construyó manualmente, pero los autores describen el proceso

de generación de dependencia de puntos de conocimiento (integración de clases). En Zhu & Yao (2009) se presenta el diseño de secuencias de aprendizaje utilizando la formalización por medio de ontologías (lenguaje OWL); dicho diseño se enfoca en el aprendizaje individual con el propósito de establecer una secuencia personalizada de acuerdo con el nivel y características de cada alumno en la educación superior en ambientes virtuales.

En Bucos et al. (2010) se aborda una estructura de ontología de dominio que desempeña un papel importante en la representación de conceptos de educación superior y en la asistencia a sistemas de *e-learning* especializados. A medida que aumenta el número de clases que forman parte de la estructura de ontología y las propiedades asociadas a ellas, la ontología se divide en un conjunto de ontologías más pequeñas.

Una ontología creada a partir de diagramas CASE para la educación en línea se presenta en Bagiampou & Kameas (2012); su evaluación es abordada por expertos en un proceso manual. En este trabajo, el enfoque está en la fase de construcción, donde las clases se extraen manualmente. El proceso de creación de ontologías a partir de la información de los cursos ofrecidos en niveles avanzados se explica en Ameen et al. (2012), donde los estudiantes pueden elegir cursos de acuerdo con sus antecedentes académicos. Ambos trabajos presentan la estructura, información y jerarquía de las clases de forma manual.

Autores como (Uskov et al., 2016) se centran en el aprendizaje autónomo en línea, proponen una ontología en base al Internet de las cosas (IoT); más que aprendizaje en línea, se centra en el aprendizaje dentro del aula con la ayuda de la tecnología tomando como referencia los tipos de inteligencia de los estudiantes. En Ameen et al. (2012) se explica el proceso de creación de una ontología a partir de la información de cursos que se ofertan en nivel superior, donde el estudiante puede elegir los cursos a tomar de acuerdo a su historial académico. Se presenta la estructuración y jerarquización de las clases de manera manual.

Otras investigaciones se centran en la educación en línea como Dai & Li (2010), Du et al. (2012), Hu et al. (2016) y recientemente Hssina et al. (2017), donde las ontologías se definen manualmente a partir de recursos XML disponibles en Internet, y la evaluación también es un proceso manual. Investigaciones como Uskov et al. (2016) se enfocan en el aprendizaje automático; en este trabajo, se crea una ontología basada en IoT utilizado en un aula, teniendo en cuenta las inteligencias estudiantiles. Fan et al. (2008) y Hu et al. (2016) presentan ontologías para el aprendizaje en el aula, la primera investigación propone una ontología para la interacción entre el estudiante y el maestro en la enseñanza del idioma inglés; mientras que en la segunda investigación la ontología involucra el uso del Internet para la mejora del aprendizaje. Se utiliza una ontología por cada entidad que participa en el proceso enseñanza aprendizaje; la evaluación

se realiza manualmente por expertos en el dominio. Bagiampou & Kameas (2012) presenta una ontología de dominio sobre diagramas de casos de uso creada para ambientes *online*, específicamente para la materia de ingeniería de software, este trabajo es evaluado por expertos en el dominio.

Aminah et al. (2017) se desarrolla una ontología para la evaluación académica en una Universidad de Indonesia. Este trabajo toma en cuenta que los datos académicos evolucionan con el tiempo según el desarrollo de los criterios de evaluación académica. La universidad gestiona el proceso de evaluación interna y determina los parámetros y criterios correspondientes. Se definieron 26 clases con 21 propiedades de objeto. La ontología desarrollada satisface la Guía de Evaluación Académica y podría ser extendido con más elementos de vocabulario.

La Figura 3.2 muestra las investigaciones realizadas en el dominio pedagógico de acuerdo al enfoque utilizado. Algunos autores crean una ontología para alguna materia en particular, como por ejemplo ingeniería de software o matemáticas, otros autores se enfocan en la creación de herramientas para el docente o los alumnos, ya sea el uso de una plataforma o secuencias didácticas para el aprendizaje. Además, la aplicación de las ontologías creadas es para clases *online* en algunos casos y en otros para clases presenciales.

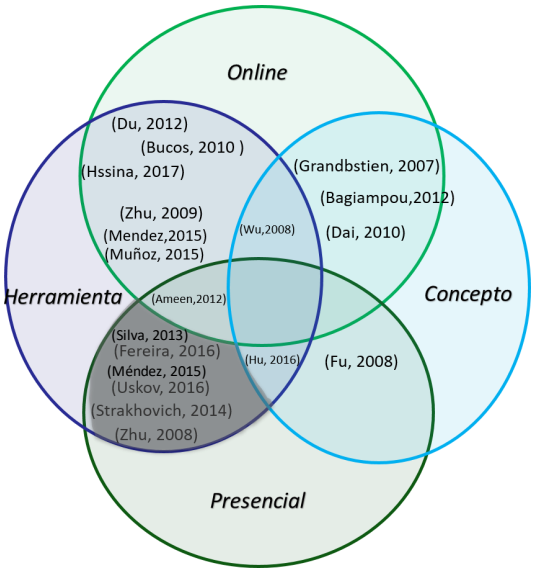


Figura 3.2: Investigaciones sobre ontologías en el dominio pedagógico

En investigaciones más relacionadas con los subdominios estudiados, Méndez et al. (2015) 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. Este modelo es utilizado en la educación online, donde se infiere el tipo de inteligencia del estudiantes y de acuerdo a esto se recomiendan contenidos pertinentes. Silva Sprock & Ponce (2013) presenta una reingeniería de una ontología de estilos de aprendizaje en varios enfoques, la cual tiene el objetivo de apoyar la creación, adaptabilidad y uso de objetos de aprendizaje.

El trabajo de investigación propuesto se centra en el área sombreada, es decir, la construcción de la ontología será como una herramienta para clases presenciales. Esta área contiene varias investigaciones analizadas en el estado del arte, sin embargo, todas adoptan un enfoque de construcción manual.

3.5. Análisis y limitantes

A manera de resumen, la Tabla 3.3 muestra las técnicas, herramientas y dominios analizados en cada una de las fases del aprendizaje ontológico (filas). Como se puede observar, algunas herramientas se repiten en dos fases, sobre todo en creación y poblado, además, se muestra una variedad de dominios y herramientas. En cuanto a la evaluación de las metodologías analizadas, en todas las fases se encuentra la intervención de un experto en el dominio analizado.

Entre las limitantes detectadas en los trabajos relacionados se encuentran las siguientes:

- No se encontraron estudios que incluyan el proceso ontológico y de poblado en un dominio específico, sino una parte de éste (detección de clases, creación o poblado).
- Dentro de la fase de la creación de ontologías, son pocas las investigaciones que abordan un enfoque automático o semiautomático. La mayoría trabajan una creación manual, especialmente las del dominio pedagógico.
- No se detectaron investigaciones dentro del campo de la creación de un corpus enfocado exclusivamente al proceso de aprendizaje ontológico. Por lo tanto, el corpus creado representa una aportación inicial para la investigación.
- Dentro del dominio pedagógico, las investigaciones no involucran más de una clase principal en el proceso.

El paso de evaluación contiene varias técnicas, donde el mejor método puede ser determinado por el objetivo de la investigación. En el trabajo presentado por Wong et al. (2011), los autores dividieron las técnicas en tres categorías principales:

Tabla 3.3: Técnicas y métodos más utilizados en cada una de las fases del aprendizaje ontológico de acuerdo al análisis del estado del arte

Fase	Dominios	Técnicas	Herramientas	Evaluación
Detección de clases	Textos financieros Inteligencia artificial Wikipedia Gestión de emergencias	Manual Patrones lingüísticos Expertos o *Gold* Herramientas de etiquetado (POS)	Protégé Etiquetado Patrones léxicos Clasificador supervisado	Manual *Gold* Experto Recuperación de información
Creación	Manuales de cursos Informática básica Aprendizaje social Personas con ETN Biografías personalizadas Oncología	Manual Relaciones semánticas Anáforas Agrupamiento Similitud Herramientas POS	Swoogle Protégé XML SPARQL	Recuperación de información Experto Relaciones entre conceptos Relaciones taxonómicas Contraste de ontologías Cualitativa
Evaluación	Sin dominio	Comparación Métricas: cohesión entre elementos, distancias Análisis de jerarquías	Ontología base Análisis cualitativo Software especializado en el dominio	Cuadro comparativo entre técnicas analizadas
Dominio	Aprendizaje social	Combinación de ontologías	Protégé	Expertos en dominio

1. El primer enfoque evalúa la idoneidad de las ontologías en el contexto de otras aplicaciones.
2. El segundo enfoque utiliza fuentes de datos específicas del dominio para determinar en qué medida las ontologías pueden cubrir el dominio correspondiente.
3. El tercer enfoque evalúa las ontologías al determinar qué tan bien se adhieren a un conjunto de criterios. Por ejemplo, se puede establecer el número promedio de términos que se agregaron para formar un concepto en la ontología.

Capítulo 4

Propuesta metodológica

En este capítulo se describe la metodología propuesta para esta investigación, la cual, inicialmente se divide en tres fases. En la fase uno, se observa el proceso de creación del corpus mediante una recopilación manual de un conjunto de artículos pedagógicos; aunado a un proceso de extracción automática y un método para integrar nuevas instancias. La segunda fase se centra en en el proceso de formalización de la ontología analizando las características estructurales del corpus. En esta fase tienen una presencia predominante las técnicas de PLN para la detección de conceptos principales, que es donde se presentan más experimentos, y la extracción de relaciones, principalmente mediante el análisis de métricas de similitud textual.

En cada experimento realizado, se necesita implementar algunas métricas de evaluación a fin de validar los resultados obtenidos. Por lo tanto, se integra un módulo de evaluación presente en todas las fases, en el cual se analizarán los resultados mediante técnicas de recuperación de información y una evaluación manual, la cual está basada en el análisis cualitativo y la comparación con ontologías creadas manualmente. La Figura 4.1 muestra el procedimiento general.

4.1. Creación del corpus

En el proceso de creación de ontologías de dominio, es necesario tener un repositorio de información del cual se puedan extraer los componentes de dichas ontologías. Como texto de

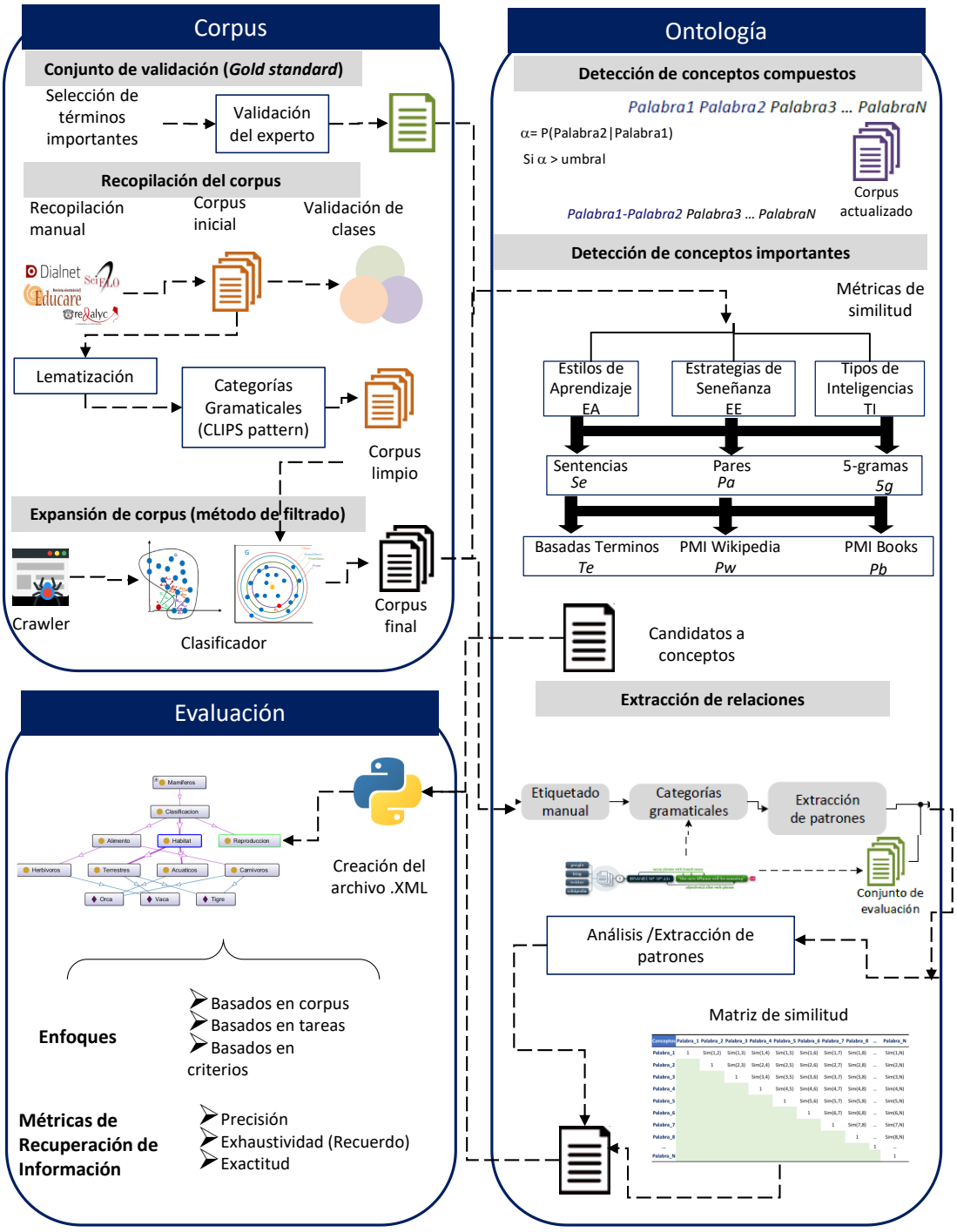


Figura 4.1: Metodología general propuesta

entrada se utiliza un corpus creado con instancias de cada uno de los temas analizados, unas obtenidas de manera manual y otras mediante un proceso de filtrado de documentos. La Figura

4.2 muestra los procesos implementados para esta fase.

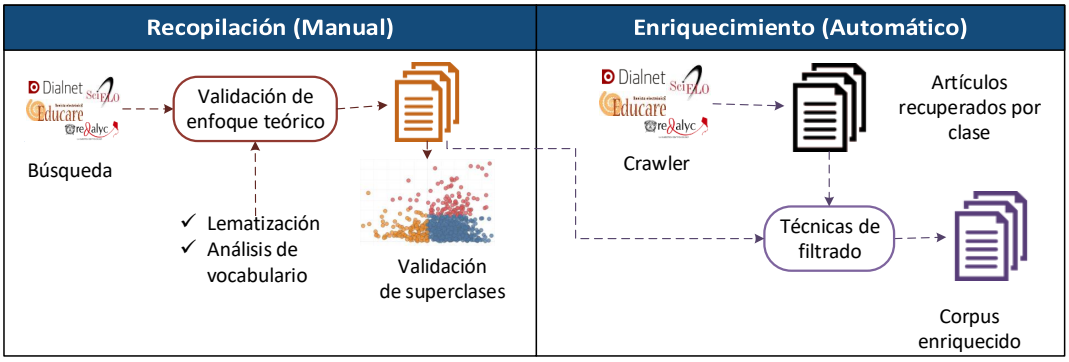


Figura 4.2: Procesos implementados en la fase 1 de la metodología

Considerando el dominio elegido, se tomaron artículos publicados que traten de uno de los temas analizados. Es importante aclarar que los artículos recabados son del área de Pedagogía o Psicología y que están publicados en el idioma español. Se realizó un etiquetado manual, en donde la clase es el subdominio del que trata dicho artículo. Con esto, se asegura la riqueza del vocabulario y de los conceptos principales, los cuales son indispensables para la creación de la ontología.

En lo que respecta a la recolección de artículos, se realizó una búsqueda manual utilizando los siguientes recursos académicos:

- **Educare Electronic Journal**[1]**:** Forma parte del Centro de Investigación y Enseñanza de la Educación (CIDE) de la Universidad Nacional de Costa Rica. Es una revista electrónica internacional de publicación trimestral cuyo objetivo es difundir la producción científica y promover el análisis académico en todos los campos de la educación.
- **Scientific Electronic Library Online (SciElo)**[2]**:** Es un modelo para la publicación electrónica cooperativa de revistas científicas en Internet. Su objetivo es responder a las necesidades de la comunicación científica en los países en desarrollo, particularmente de América Latina y el Caribe.
- **Dialnet**[3]**:** Es un proyecto de cooperación bibliotecaria que comenzó en la Universidad de La Rioja. Se constituye como un portal que recopila y proporciona acceso fundamentalmente a documentos publicados en España en cualquier lengua o que traten sobre temas hispánicos.

[1] http://www.revistas.una.ac.cr/index.php/EDUCARE/index
[2] http://www.scielo.org/php/index.php?lang=es
[3] https://dialnet.unirioja.es/

- **Latindex**[4]**:** Es el Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal. Surgió en 1995 en la Universidad Nacional Autónoma de México (UNAM), la misión del sistema es difundir, hacer accesible y elevar la calidad las revistas académicas editadas en la región, a través del trabajo compartido.

El corpus se representa mediante el nombre $Inicial$ y se representa como: $Inicial = \{K, T, C\}$ donde:

- K Es la clave del artículo ($\{1..,51\}$).
- T Es todo el texto del artículo, eliminando las palabras cerradas[5].
- C Es la clase del artículo. Ésta se representa como un atributo nominal de acuerdo al título del artículo. $C = \{$ *EA, TI, EE*$\}$, representando a las clases estilos de aprendizaje, tipos de inteligencia y estrategia de enseñanza respectivamente.

La Tabla 4.1 muestra el vocabulario y número de palabras por clase en el corpus inicial. La clase estrategias de enseñanza contiene más vocabulario, sin embargo, la clasificación de las estrategias no está del todo definida, por lo que se justifica el número de palabras. La clase estilos de aprendizaje contiene menos vocabulario, al tener clases bien definidas de acuerdo al enfoque teórico elegido. En el apéndice A se muestra la lista de artículos que integra dicho corpus.

Tabla 4.1: Vocabulario del corpus $Inicial$

Clase	Palabras	Vocabulario
Estilos de aprendizaje	60,551	8,587
Estrategias de enseñanza	68,397	10,145
Tipos de inteligencias	56,090	9,863
Total	**185,038**	**18,563**

4.1.1. Expansión del corpus

Considerando que el corpus cuenta solamente con 51 instancias se hace necesario incrementar éstas de forma semiautomática. Las clases principales tienen un enfoque teórico determinado, por lo que además de extraer un artículo que se refiera a cualquiera de ellas, se debe determinar si dicho artículo corresponde o no al enfoque teórico seleccionado.

Para la extracción automática de artículos se implementó un *crawler* en cada uno de los recursos académicos utilizados en la primera búsqueda manual. Se descargaron todos los artículos recuperados de las consultas utilizando las clases principales. Hay algunas revistas que están

[4] `http://www.latindex.org/latindex/inicio`

[5] Palabras que se filtran antes o después del procesamiento de texto. Generalmente el término se refiere a las palabras más comunes en un idioma aunque no existe una sola lista universal

indexadas en más de un buscador, por lo que el *crawler* también incluye una comparación entre los artículos iniciales y los descargados automáticamente, a fin de evitar elementos repetidos.

Dado el funcionamiento del *crawler*, cada artículo ya tiene asignada una de las tres clases, la Tabla 4.2 muestra el número de documentos recuperados por clase y buscador, sin considerar los repetidos. Haciendo una eliminación de los textos que no son legibles y aquellos cuyo texto no puede ser extraído se obtienen 800 artículos a procesar. Éstos artículos no pueden entrar directamente al corpus inicial, ya que es necesario determinar si cada instancia maneja el mismo enfoque teórico que los artículos del corpus inicial.

Tabla 4.2: Número de artículos recuperados por buscador académico

Clase	Dialnet	Educare	Redie	SciElo	Total
EA	28	20	30	130	246
EE	6	20	45	354	425
TI	118	20	6	23	167

Se propone una arquitectura de filtrado semisupervisada aplicable a cualquier dominio (Figura 4.3). Dicha arquitectura tiene por objetivo servir como una herramienta para que un especialista pueda recolectar artículos de un área y tema concreto, así como proporcionar apoyo en el análisis de trabajo relacionado. Por ejemplo, al investigar el tema de desarrollo dentro del dominio de psicología educacional, existen dos enfoques principales: el desarrollo visto como un proceso cognoscitivo psicogenético de Piaget y el enfoque de Vygotsky, con una perspectiva sociocultural. Ambas teorías están vigentes, y los expertos tienden a elegir una de ellas al realizar sus investigaciones.

La arquitectura consta de dos métodos de filtrado soportados en ideas diferentes, el primero asume que un documentos es relevante a la temática de interés si este es parecido a algún documento del conjunto de entrenamiento. El segundo método toma como relevantes aquellos documentos que se encuentren cerca de un vector representativo del conjunto. Ambos métodos fueron estructurados para filtrar instancias dentro de un ámbito reducido, donde los elementos importantes no tienen grandes diferencias en contenido con los elementos no importantes.

Para la aplicación de estos métodos, el conjunto de entrenamiento es el corpus inicial, mientras que el conjunto de evaluación es un subconjunto de artículos descargados con la ayuda del *crawler*, el cual fue etiquetado de manera manual. La Tabla 4.3 muestra el número de artículos por clase y cuántos de estos son etiquetados como importantes.

Solo un subconjunto pequeño fue etiquetado como elementos positivos; esto se debe a que en algunas clases existen más de dos enfoques que las describen. Para la clases *Estilos de*

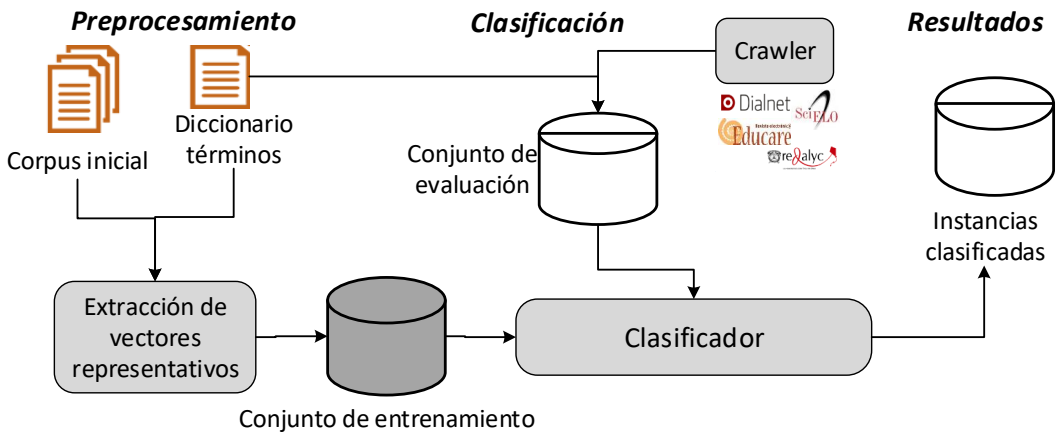


Figura 4.3: Propuesta de arquitectura de filtrado

Clase	Total	Importantes
Estilos de aprendizaje	54	17
Estrategias de enseñanza	73	15
Tipos de inteligencias	39	13

Tabla 4.3: Tamaño del conjunto de evaluación y número de artículos importantes por clase.

aprendizaje, los artículos no importantes se enfocan en estudios sobre el aprendizaje en general, ya sea el diseño de materias para ciertos grupos o enfoques motivacionales. Para la clase *Tipos de inteligencias* los documentos no importantes tratan principalmente sobre la inteligencia emocional en diferentes niveles educativos, mientras que en la clase *Estrategias de enseñanza* se centran en la descripción de estrategias de estudio muy particulares o una comparación epistémica entre distintos enfoques.

Representación de documentos

Para la representación de los documentos se utilizan vectores integrados por palabras representativas de cada dominio. Se generó un diccionario, compuesto por la definición de las clases principales (*inteligencia, estilos, aprendizaje, estrategias, enseñanza*). Estas definiciones fueron recuperadas de manera manual utilizando 5 recursos: 2 diccionarios pedagógicos, el diccionario de la Real Academia Española (RAE), un diccionario enciclopédico y una lista en línea de sinónimos. Por lo tanto, se obtuvo el diccionario inicial $Dic_i = \{W, C_1, C_2, C_3, C_4, S\}$, donde:

- W es el concepto principal (lista inicial).
- C_1, C_2, C_3, C_4 son las definiciones en cada uno de los diccionarios utilizados. Algunas palabras, no tiene definición en algún recurso por lo que se omite ese elemento del arreglo.

Los textos fueron preprocesados eliminando las palabras cerradas, signos de puntuación y palabras con una longitud menor a tres caracteres.

➥ S una lista de los sinónimos del concepto.

D_i se expandió usando las palabras de las definiciones encontradas y la lista de sinónimos para obtener la lista de conceptos Dic. La Figura 4.4 muestra un ejemplo con una palabra de las incluidas en la lista inicial.

Figura 4.4: Ejemplo de la expansión del diccionario para el término “Corporal”

Aplicando el proceso de expansión, la lista se incrementó a 990 palabras. Al final se lematizaron los conceptos y se eliminaron aquellos que solo aparecen una vez en el corpus inicial, quedando $\|Dic\| = 303$. Este diccionario es una lista simple que solo contiene los conceptos recuperados. A partir de la lista de conceptos Dic y con la ayuda del experto en el dominio, se eliminan aquellas palabras que no son consideradas como importantes o que no son viables para conformar la ontología. De este proceso se obtiene una lista depurada, con 202 palabras.

4.1.2. Detección de conceptos compuestos

Uno de los aspectos más importantes en este dominio es que la mayoría de los conceptos principales están compuestos por más de una palabra, lo cual no permite la implementación de análisis automáticos en donde el primer paso es la tokenización (división por palabras). Para la detección de estos conceptos se implementó un algoritmo que calcula la probabilidad condicional de cada par de palabras. La Figura 4.5 muestra el método general, la cual tiene el objetivo de encontrar un umbral que permita separar los conceptos compuestos de los simples.

El corpus se divide por clases, y cada uno de éstas se analiza por separado, de cada instancia se extraen las palabras adjuntas y se calcula su frecuencia y su probabilidad condicional. De

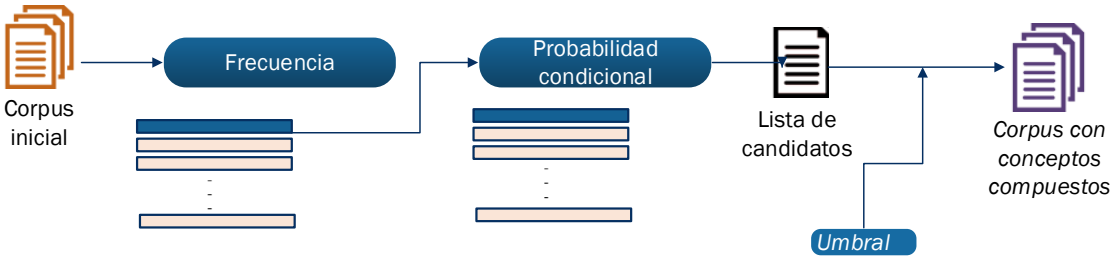


Figura 4.5: Método en dos fases para la detección automática de conceptos compuestos.

acuerdo a la figura, primero se analizan las frecuencias, dicho análisis consiste en la división de cuartiles, a fin de analizar cada segmento por separado y determinar si todos son significativos. La Figura 4.6 muestra un ejemplo del proceso.

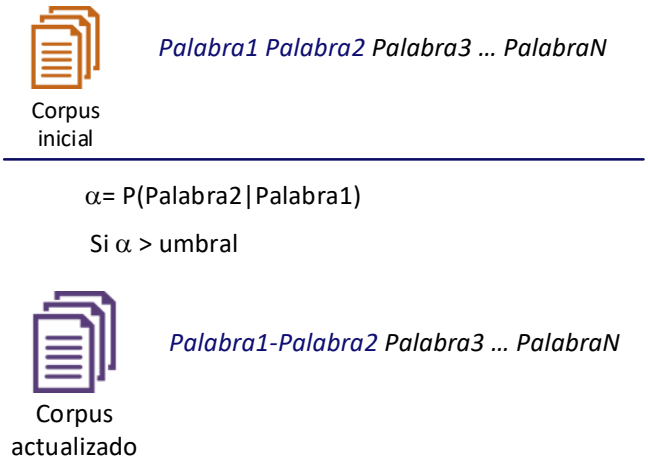


Figura 4.6: Probabilidad condicional para la extracción de conceptos compuestos

Al calcular la probabilidad condicional, se compara con un umbral previamente establecido, si la probabilidad es mayor o igual a dicho umbral, este par de palabras se considera como un concepto compuesto. En este proceso se analizan todos los posibles escenarios, por ejemplo si los pares $P(Palabra2|Palabra1)$ y $P(Palabra3|Palabra2)$ superan el umbral, se toma el par de la probabilidad mayor, quedando el corpus como $Palabra\_Palabra2\ Palabra3$ o $Palabra1\ Palabra2\_Palabra3$.

4.2. Conceptos importantes

Esta fase integra dos procesos secuenciales: detección de conceptos principales y extracción de relaciones. Para la detección de conceptos se realizó un análisis con las métricas de similitud textual. La hipótesis es que si dos palabras tienen un alto grado de similitud, estas guardan una relación entre sí, ya sea taxonómica o no taxonómica. Primero se realizan experimentos utilizando conceptos unipalabra, a fin de tener resultados preliminares, posteriormente se utiliza un *gold standard*. Los resultados de estos experimentos generan una lista de conceptos por clase, las cuales se utilizarán en la fase de detección de relaciones entre dichos conceptos.

Entre la fase de detección de conceptos y detección de relaciones se realizan algunos experimentos para la extracción de segmentos de texto, los cuales integren dos o más conceptos y palabras clave que permitan establecer una relación entre estos. En esta fase se utilizan corpus previamente etiquetados de otros dominios, y eso origina la necesidad del etiquetado manual de instancias. Otros métodos implementados en esta fase involucran el análisis manual y automático de patrones de categorías gramaticales, extracción de métricas de similitud en un proceso iterativo y la implementación de metodologías externas para complementar resultados.

Para esta etapa, el primer experimento se utilizaron tres representaciones del corpus $Final$ y métricas de similitud textual. Además, el corpus fue dividido de acuerdo a la clase, obteniendo tres subcorpus: *Estrategias de enseñanza*, *Estilos de aprendizaje* y *Tipos de inteligencias*. La Figura 4.7, muestra todas las representaciones junto con las métricas utilizadas para su análisis. El proceso a realizar fue el siguiente:

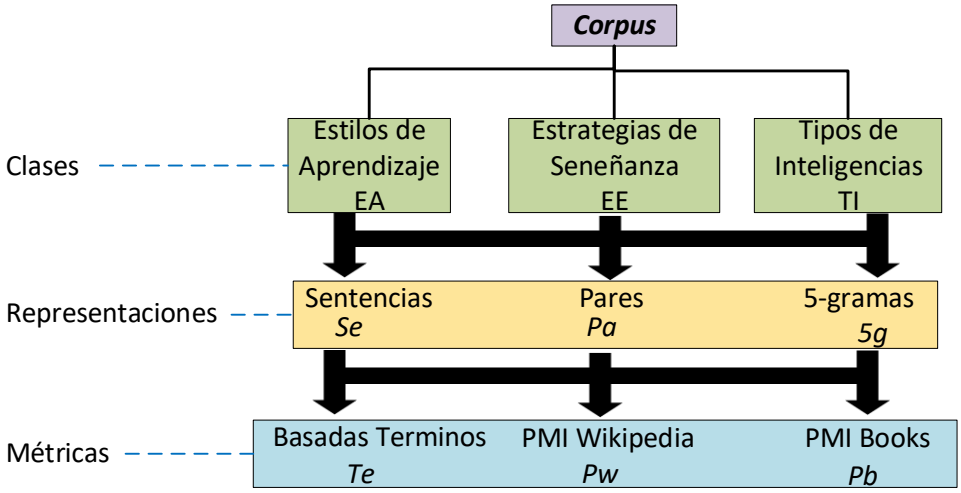


Figura 4.7: Clases, representaciones y métricas utilizadas en el experimento inicial.

1. Se extrajeron los lemas de cada subcorpus (con la ayuda de la herramienta *TreeTagger*). Las palabras cerradas fueron eliminadas y se extrajeron las representaciones de la Figura 4.7: Se, Pa y $5g$.
2. Con las representaciones del corpus se extrajeron los coeficientes de Dice, Jaccard, Traslape y Coseno para obtener el conjunto Te. Además, se calculó la métrica de PMI utilizando un corpus extraído de Wikipedia (Pw) y un conjunto de libros de Pedagogía disponibles en la red (Pb).
3. Se utilizó el *gold standard* con un valor tope de las métricas para determinar si una palabra es considerada como recuperada. Los topes en los valores son representados por γ.
4. Se calcularon la precisión y el recuerdo para cada experimento. Se da una mayor importancia a la precisión por que no es necesario recuperar todas las palabras del *gold*, y la precisión muestra el porcentaje de palabras correctamente recuperadas en cada experimento.

Las métricas utilizadas en este experimento fueron modificadas en la interpretación de las variables. En la literatura consultada se utilizan para obtener la similitud entre dos oraciones, sin embargo, en este experimento es necesario obtener la similitud entre dos palabras. Usualmente, las fórmulas de este tipo de métricas involucran dos elementos: t_1 y t_2 que representan la oración 1 y 2 respectivamente. Por ejemplo, en la fórmula del coeficiente de Jaccard, $\mid t_1 \cap t_2 \mid$ es el número de palabras que aparecen en la sentencia 1 y 2, y $\mid t_1 \cup t_2 \mid$ es el número de palabras entre ambas sentencias. Para este experimento, el significado de las variables cambiaron de acuerdo a la representación del corpus analizada: t_1 es la clase, y t_2 es cada una de las palabras del vocabulario, por lo tanto, el valor de $\mid t_1 \cap t_2 \mid$ and $\mid t_1 \cup t_2 \mid$ es diferente para cada representación:

➥ En la representación Se:

$\mid t_1 \cap t_2 \mid$**:** Número de sentencias donde la clase (t_1) y la palabra (t_2) aparecen juntas. No es importante si las palabras están separadas, solo tienen que aparecer en la misma oración.
$\mid t_1 \cup t_2 \mid$**:** Total de oraciones donde aparece la clase o la palabra.

➥ En la representación Pa:

$\mid t_1 \cap t_2 \mid$**:** Total de apariciones de las dos palabras juntas, sin importar el orden ($t_1\, t_2$ o$t_2\, t_1$
$\mid t_1 \cup t_2 \mid$**:** Frecuencia de t_1 más la frecuencia de t_2 en el subcorpus.

➥ En la representación $5g$:

$\mid t_1 \cap t_2 \mid$**:** Número de 5-gramas donde aparecen juntas t_1 y t_2
$\mid t_1 \cup t_2 \mid$**:** Número de 5-gramas donde aparece t_1 o t_2

Las dos versiones de la métrica PMI también se calcularon utilizando estas modificaciones a las variables. Para esta métrica, fueron necesarios dos corpus externos: El primero (Pw) está compuesto por artículos aleatorios de Wikipedia y al segundo (Pb) lo componen libros de Pedagogía, Filosofía y Psicología. Pw fue extraído automáticamente mediante un *crawler* web, y Pb fue extraído manualmente. La Tabla 4.4 muestra los componentes de cada corpus. Pw es más rico en vocabulario, pero el dominio no está relacionado como con el corpus Pb. La hipótesis inicial es que un corpus relacionado con el dominio puede obtener mejores resultados. Una descripción detallada de dicho corpus se muestra en el apéndice D.

Tabla 4.4: Corpus obtenidos para el cálculo del PMI

	Wikipedia	Libros
Instancias	174,605	113
Palabras	21,529,363	4,289,894
Vocabulario	498,866	116,986

4.3. Relaciones entre conceptos

Para la detección de relaciones se presentan métodos basados en frecuencias maximales y en el etiquetado manual del corpus. Para las frecuencias maximales se toma en cuenta el análisis de partes de texto en las cuales se pueden detectar relaciones entre los conceptos que aparecen en dichos fragmentos. Algunos de los experimentos iniciales se basan en frecuencias maximales, obteniendo $n - gramas$ con $n = \{3, 4, 5\}$. El proceso integra una parte automática y una evaluación manual por parte del experto en el dominio, como texto de entrada se utiliza el *corpus* $Final$.

Para el cálculo de dichas frecuencias maximales se implementó el siguiente algoritmo :

1. Del corpus inicial, se extraen $n - gramas$ de lemas (2 a 5).
2. De cada $n - grama$ se calcula una métrica α que relacione la frecuencia y el número de artículos en los que aparece.
3. Se eligen los $n - gramas$ con el valor de α más alto (mayor a 100) y se contrastan con los conceptos del conjunto de evaluación. Si al menos dos palabras del $n - grama$ se encuentran dentro de la lista, el $n - grama$ pasa a formar parte de los resultados.
4. Con la ayuda del experto en el dominio, se obtienen los $n - gramas$ que contienen relación teórica entre dos o más conceptos.

Este proceso ahorra tiempo al experto del dominio, ya que en lugar de revisar de manera manual todos los artículos del corpus inicial, se revisan las listas de $n - gramas$ que son más representativos dentro de los artículos. Cabe mencionar que, al tener un fuerte enfoque manual, estos experimentos se consideran como un *baseline* para a partir de estos incrementar

los resultados y el nivel de automatización.

Además de este algoritmo inicial, se presentan dos métodos para la detección de frases importantes del texto, en las cuales se encuentran los conceptos que conformarán la ontología y las relaciones entre ellos. Para encontrar los textos se definen patrones de categorías gramaticales mediante una lista de palabras y un etiquetado manual de un documento. Aunque existen diversas técnicas para este procedimiento, la mayoría de estas se enfocan en el idioma inglés, y en diferentes dominios como medicina y noticias.

Se propone el uso de patrones de categorías gramaticales que puedan detectar los términos importantes del texto, de los cuales se extraerán los conceptos y relaciones entre ellos. Para la extracción de las categorías gramaticales se utiliza el módulo *pattern* de la herramienta CLIPS. De la información proporcionada por el módulo se utilizan los primeros dos elementos por palabra para la construcción de los patrones, separados por el símbolo _.

4.3.1. Extracción de patrones

Este método se basa en la búsqueda de una lista de términos considerados importantes para el experto en el dominio. La Figura 4.8 muestra el procedimiento general, donde una instancia del corpus funge como entrenamiento y el resto como conjunto de evaluación.

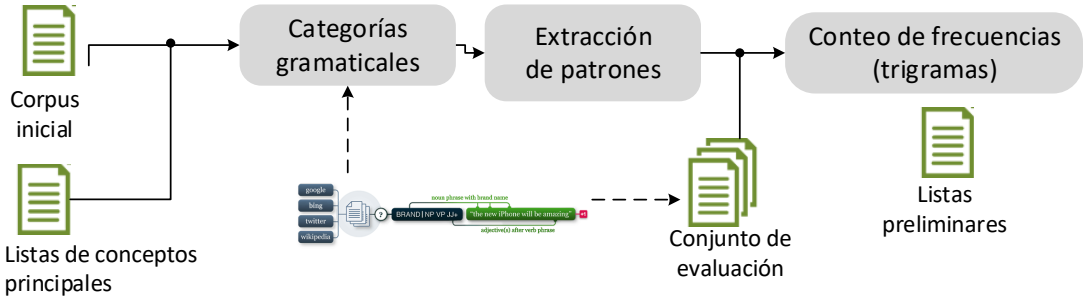


Figura 4.8: Método automático.

Los pasos a seguir se mencionan a continuación:

1. Se recuperan todas las incidencias de los conceptos de la lista dentro del conjunto de entrenamiento, junto con sus palabras cercanas a fin de obtener trigramas de palabras.
2. Se etiqueta el conjunto de entrenamiento, extrayendo los trigramas de categorías gramaticales junto con la frecuencia de cada uno de ellos. Para el cálculo de la frecuencia solo se toman en cuenta las categorías, no las palabras, por ejemplo, las frases *la necesidad de* y *la existencia de* son representadas por el mismo trigrama de categorías gramaticales: $DT\_B - NP, NN\_I - NP, IN\_B - PP$.

3. Bajo las mismas condiciones se etiqueta el conjunto de evaluación, se recuperan los patrones del punto anterior y se muestran los trigramas pertenecientes a dichos patrones.
4. Para mejorar el análisis de los resultados, se recuperan también las palabras antecesoras y sucesoras de los patrones, a fin de obtener 5-gramas.

4.3.2. Matriz de similitud

Otro método utilizado para la detección de conceptos relacionados es una variante del procedimiento explicado en la sección 4.2, con la diferencia de utilizar solo dos métricas de similitud, y hacer el proceso iterativo. Se calcula una matriz de similitud como la de la Figura 4.9. Esto no solo permite tener una aproximación de los conceptos principales, además, se extraen posibles relaciones entre estos.

Conceptos	Palabra_1	Palabra_2	Palabra_3	Palabra_4	Palabra_5	Palabra_6	Palabra_7	Palabra_8	...	Palabra_N
Palabra_1	1	Sim(1,2)	Sim(1,3)	Sim(1,4)	Sim(1,5)	Sim(1,6)	Sim(1,7)	Sim(1,8)	...	Sim(1,N)
Palabra_2		1	Sim(2,3)	Sim(2,4)	Sim(2,5)	Sim(2,6)	Sim(2,7)	Sim(2,8)	...	Sim(2,N)
Palabra_3			1	Sim(3,4)	Sim(3,5)	Sim(3,6)	Sim(3,7)	Sim(3,8)	...	Sim(3,N)
Palabra_4				1	Sim(4,5)	Sim(4,6)	Sim(4,7)	Sim(4,8)	...	Sim(4,N)
Palabra_5					1	Sim(5,6)	Sim(5,7)	Sim(5,8)	...	Sim(5,N)
Palabra_6						1	Sim(6,7)	Sim(6,8)	...	Sim(6,N)
Palabra_7							1	Sim(7,8)	...	Sim(7,N)
Palabra_8								1	...	Sim(8,N)
...									1	...
Palabra_N										1

Figura 4.9: Matriz de similitud de palabras (Ejemplo)

Una vez extraída la matriz de similitud, se utiliza para realizar una búsqueda exhaustiva, de elementos cuya similitud es similar. Se inicia con los conceptos representativos de cada clase (estrategia de enseñanza, tipo de inteligencia, estilo de aprendizaje), a partir de este punto, se hacen búsquedas iterativas en donde la salida de una iteración es la entrada de la siguiente. Se tiene la hipótesis de que los elementos recuperados están relacionados con el concepto que se utilizó para la búsqueda, ya sea como subclase o atributos de las clases. La utilidad de este método es que no solo se pueden detectar conceptos principales, sino que se pueden analizar las relaciones existentes entre ellos.

4.4. Evaluación

Para la evaluación de la ontología y en general de los experimentos realizados a lo largo de la investigación se utilizarán enfoques basados en *gold standard*. En estos sistemas se tienen dos

tipos de documentos (ó instancias):

- **Documentos relevantes (o documentos positivos)**: Total de documentos que el sistema debe recuperar en una consulta.
- **Documentos recuperados**: Total de documentos que son recuperados en la consulta.

Utilizando estos datos, existen dos medidas de uso frecuente en la evaluación de los sistemas: Precisión (P) y recuerdo (R), llamadas respectivamente *precision* y *recall* por sus términos en inglés. Estas son calculadas de la siguiente manera:

1. ***Precisión:*** Fracción de los datos recuperados que son relevantes.

$$P = \frac{items\ relevantes\ recuperados}{items\ recuperados} \tag{4.1}$$

2. ***Recuerdo:*** Fracción de los datos relevantes que son recuperados.

$$R = \frac{items\ relevantes\ recuperados}{items\ relevantes} \tag{4.2}$$

Estas métricas pueden ser definidas en función de una matriz de confusión (figura 4.10), la cual es característica de los modelos de clasificación supervisada, en esta, las instancias evaluadas por el clasificador pueden caer en las siguientes opciones:

- **Verdaderos positivos (TP)**: Instancias que pertenecen a la clase y se clasifican en ella.
- **Falsos positivos (FP)**: Instancias que no pertenecen a la clase, pero son clasificadas en ella.
- **Falsos negativos (FN)**: Instancias que pertenecen a la clase, pero no clasifican en ella.
- **Verdaderos negativos (TN)**: Instancias que no pertenecen a la clase y no se clasifican en ella.

	Relevante	No relevante
Recuperada	**TP**	**FP**
No recuperada	**FN**	**TN**

Figura 4.10: Matriz de confusión para evaluar los resultados de una clasificación

Por lo tanto, precisión y recuerdo también pueden ser expresadas como se muestra en las fórmulas 4.3 y 4.4 respectivamente.

$$P = \frac{TP}{TP + FP} \tag{4.3}$$

$$R = \frac{TP}{TP + FN} \tag{4.4}$$

Dado que las dos medidas explicadas se complementan, es necesario tener en cuenta las dos para hacer un análisis completo, ya que una precisión alta no nos asegura que se recuperen todos los datos relevantes, mientras que un recuerdo alto puede indicar que entre los elementos recuperados hay basura. Por lo tanto, se considera una tercera medida para comparar los resultados de los experimentos: el *F-score*, el cual es la media armónica entre la precisión y el recuerdo (fórmula 4.5).

$$F - score = 2 * \frac{P * R}{P + R} \tag{4.5}$$

Gold standard

El *gold standard* se generó manualmente con la ayuda del experto en el dominio y contiene una serie de conceptos por clase considerados como importantes. A diferencia del diccionario, este recurso cuenta con conceptos formados por una o más palabras, además de que considera los sinónimos, mismos que son reemplazados en el corpus original. Las listas contienen 104 elementos para los tipos de inteligencias, 79 para las estrategias de enseñanza y 144 para los estilos de aprendizaje.

Parra facilitar el análisis y evaluación de los experimentos realizados, se creó una segunda versión de esta lista, integrando solamente conceptos simples (conformados por una sola palabra). Para esta versión, se incrementó el número de elementos, solo para las estrategia de aprendizaje el número de elementos queda igual. Las listas extraídas tienen elementos en común ya que pertenecen a un mismo tema pedagógico. La Figura 4.11 muestra el diagrama de Venn con el número de elementos que se comparten por clase.

Las tres clases comparten 12 conceptos, mientras que la clase de estilos de aprendizaje es la que comparte un mayor número de elementos con otras clases (22 con estrategias de enseñanza y 26 con tipos de inteligencias). La clase estrategias de enseñanza, a pesar de tener más texto que las demás, tiene menos conceptos importantes y comparte pocos con otras clases.

Generador de ontologías

Finalmente, se genera un creador de ontologías en Python, utilizando las salidas de los experimentos realizados. La Figura 4.12 presenta el proceso detallado.

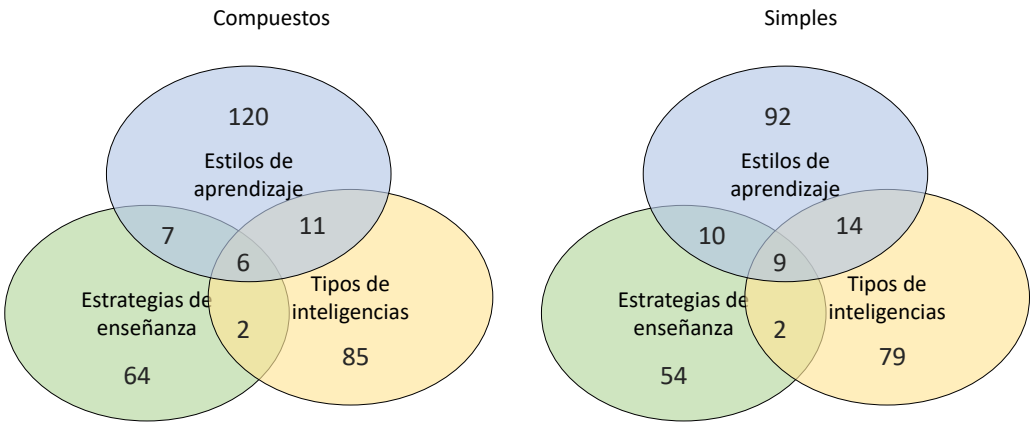


Figura 4.11: Número de elementos del conjunto de evaluación por clase

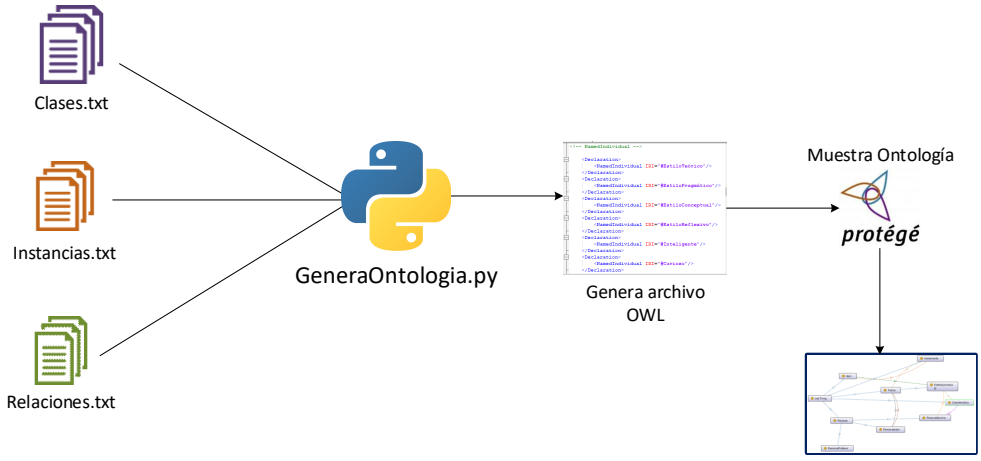


Figura 4.12: Proceso para la generación de ontologías mediante la salida de los experimentos realizados

En la figura se observa que el programa necesita 3 entradas, las cuales se describen a continuación:

- **Clases:** Contiene los candidatos a clases, así como sus atributos (si es que los tiene). Cada línea es una clase de la forma: $NombreClase; Atributo1, Atributo2, AtributoN;$
- **Instancias:** Lista las instancias de cada una de las clases.
- **Relaciones:** Contiene las relaciones entre conceptos, agrregando relaciones inversas (si las hay)

El código lee las entradas y estructura la información en forma de un archivo $.OWL$. Este archivo es compatible con la herramienta *Protégé*, la cual genera el grafo representativo de la ontología.

Capítulo 5

Resultados de experimentos

En este capítulo se detallan los experimentos realizados a lo largo de la investigación para todas las fases del aprendizaje ontológico. Se analizan algunos aspectos de los métodos utilizados y los resultados obtenidos.

5.1. Validación de clases

Los métodos de agrupamiento mencionados en la sección 2.3.3 fueron utilizados para estos experimentos. En todos los casos, solo se modificó el número de grupos ($n = 3$), y el resto de los parámetros se establecieron con su valor predeterminado. Para los experimentos, las clases de corpus reales se representaron usando los números 0, 1 y 2 para *TI*, *EA* y *EE* respectivamente. Se extrajo una frecuencia de palabras, así como la métrica Tf-idf. Se utilizaron dos conjuntos de características como atributos: el vocabulario del corpus y las palabras del diccionario.

La Tabla 5.1 muestra los grupos creados en cada algoritmo y el corpus utilizado. Se observa un mayor equilibrio al usar las palabras del diccionario como características. Dado que el corpus inicial está equilibrado en sus tres clases, se esperaría que los resultados del agrupamiento fueran similares. En este punto, una menor cantidad de atributos genera grupos más equilibrados dentro del corpus.

Para validar las clases principales, se contrastó el etiquetado manual con los resultados obtenidos por técnicas no supervisadas. Para esto se utilizó la herramienta *Scikit-learn* (Pedregosa et al.,

Tabla 5.1: Grupos creados en cada algoritmo y conjunto de características

Características	Algoritmo	Grupos creados
Real		0 1 0 0 1 0 1 1 2 2 2 2 1 1 0 1 **0** 0 2 2 2 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2
Diccionario	Aglomerativo	0 1 0 0 1 0 1 1 1 1 1 1 1 2 0 1 1 0 1 1 1 1 2 2 1 1 2 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
	Birch	0 1 0 0 1 0 1 1 2 2 2 2 1 1 0 1 **2** 0 2 2 2 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2
	K-Means	0 1 0 0 2 0 1 1 2 2 2 2 2 1 0 2 2 0 2 2 2 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2
	Espectral	0 1 1 1 0 1 1 1 1 1 1 1 1 2 1 1 1 0 1 1 1 0 1 1 0 1 2 0 1 1 1 0 1 1 0 0 1 1 1 0 0 0 1 0 1 1 1 1 1 1 0 1
Vocabulario	Aglomerativo	0 0 0 0 0 0 0 0 0 2 2 1 1 0 0 0 0 0 0 0 2 1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 0 2 1 1 2 0 2 1
	Birch	0 0 0 0 0 0 1 0 2 2 1 1 0 0 0 0 0 0 0 2 1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 0 2 1 1 2 0 2 1
	K-Means	0 0 0 0 0 0 0 0 1 1 1 2 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 1 1 0 1 0
	Espectral	0 0 0 1 0 0 0 0 0 0 0 0 0 2 0 1 0 0 0 0 0 0 1 0 0 1 2 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0

2011) para implementar algunos algoritmos de agrupamiento y métricas de evaluación. La Tabla 5.2 muestra los resultados obtenidos.

Tabla 5.2: Resultados de las métricas de agrupamiento

Algoritmo	Métrica	Exactitud
Agglomerative	Fowlkes	0.6924
	Homogeneity	0.5548
	Mutual information	0.5355
	Rand Index	0.4861
Birch	Fowlkes	0.9596
	Homogeneity	0.9311
	Mutual information	0.9284
	Rand Index	0.9406
K Means	Fowlkes	0.8445
	Homogeneity	0.7774
	Mutual information	0.7686
	Rand Index	0.7695
Spectral	Fowlkes	0.4193
	Homogeneity	0.0617
	Mutual information	0.0205
	Rand Index	0.0066

Los mejores resultados en todas las métricas se obtuvieron utilizando la frecuencia invertida de los lemas del diccionario como características y el algoritmo de Birch, principalmente considerando la métrica de Fowlkes (96 %). Los peores resultados se obtuvieron utilizando el vocabulario como características y el algoritmo espectral, logrando en algunas métricas menos del 5 %. La figura 5.1 presenta los mejores resultados obtenidos, contrastados con la asignación real. Los números representan las claves de las instancias (artículos).

Primero, las etiquetas reales se representan con tres conjuntos equilibrados. A la derecha, se

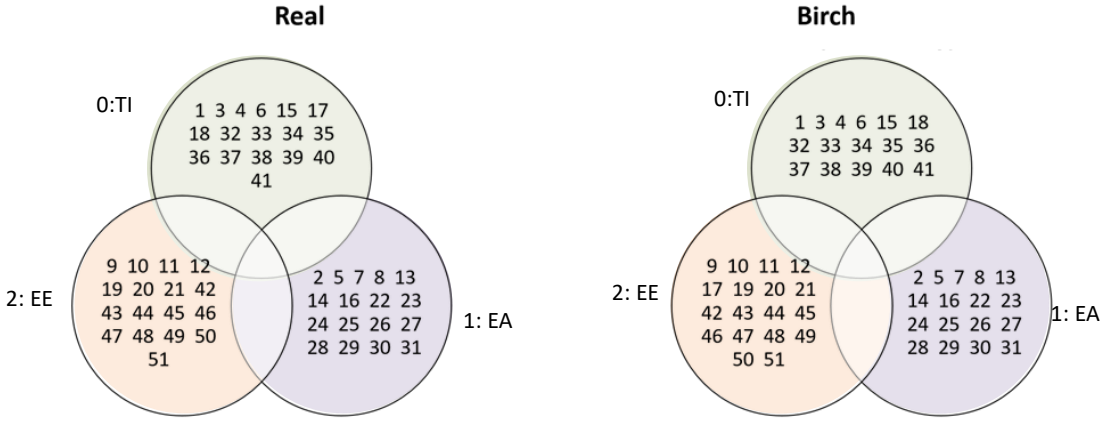


Figura 5.1: Comparación entre los grupos reales y los creados por el algoritmo Birch

muestran los resultados del algoritmo de Birch (diccionario), que son muy similares a las etiquetas reales. Por ejemplo, el primer grupo solo tiene una instancia menos que el conjunto real, y no contiene ningún elemento adicional. Este grupo corresponde a la categoría TI. La instancia 17 se agrupa en la categoría de EE junto con todas las instancias de esta clase. Esto se puede ver en la Tabla 5.2, donde la fila de etiquetas reales y la que corresponde al algoritmo de Birch con el diccionario son muy similares, excepto el elemento 17 (resaltado en negrita). Finalmente, la categoría de EA (grupo inferior derecho en el diagrama de Venn) es la que está mejor definida, ya que esta clase no presenta intersección con las otras dos.

5.2. Expansión del corpus

Para la expansión del corpus se proponen dos métodos de filtrado y para obtener un punto de referencia se genera un *baseline* con un algoritmo para *One Class Classification*. Este algoritmo maneja un método similar al objetivo planteado en esta investigación, ya que reduce el análisis a una sola clase para aprender solo de esas instancias; si el entrenamiento contiene datos de otra clase, los toma como valores atípicos pero no influyen en la construcción del modelo. Como algoritmo base para la clasificación se utilizaron las máquinas de vectores de soporte (SMO).

Clase	P	R	F_1	***Rec***
Estilos de aprendizaje	0.857	0.375	0.522	5
Estrategias de enseñanza	0.400	0.267	0.320	4
Tipos de inteligencias	0.357	0.385	0.370	5

Tabla 5.3: Resultados obtenidos con algoritmo SMO para *one class classification*

La Tabla 5.3 muestra los resultados obtenidos por clase, especificando precisión (P), recuerdo (R) y medida F (F_1), además se anexan el número de instancias recuperadas (*Rec*). En la clase *Estilos de aprendizaje* los resultados son mejores, especialmente la precisión, sin embargo, el recuerdo es bajo para todas las clases (menos de 0.4). Respecto a la métrica F_1, solo la clase *Estilos de aprendizaje* superó 0.5.

5.2.1. Métodos propuestos

El método local se basa en el análisis de distancias entre vecinos. Se comparan las distancias entre la nueva instancia y sus vecinos más cercanos, con las distancias entre estos elementos y sus propios vecinos. Con este antecedente, se formulan dos variantes cuya diferencia se centra en la métrica que se utilizará para agrupar la distancia de los vecinos más cercanos. Sea $T = \{t_1, t_2, ..., t_n\}$ un conjunto de n textos del tema de interés, y:

- t_x el nuevo texto, que debe determinarse si pertenece a T.
- $d_{x,i}$ la distancia de t_x con la instancia conocida t_i.
- $d_{i,j}$ la distancia entre los textos t_i y t_j del conjunto T.
- $knn(d_j)$ el conjunto de k vecinos más cercanos de d_j en T.

La variante **L-Prom** obtiene la distancia promedio de los k vecinos mas cercanos de t_x, y lo compara con los k vecinos de estos.(Ecuación 5.1).

$$\frac{1}{k} \sum_{\forall d_i \in knn(d_x)} d_{x,i} < \frac{1}{k} \sum_{\forall d_i \in knn(d_x)} \frac{1}{k} \sum_{\forall d_j \in knn(d_i)} d_{i,j} \tag{5.1}$$

L-Max sigue una estructura similar, solo que compara las distancias máximas en lugar de los promedios (Ecuación 5.2).

$$max_{\forall d_i \in knn(d_x)}(d_{x,i}) < max_{\forall d_i \in knn(d_x)} \left(max_{\forall d_j \in knn(d_i)}(d_{i,j}) \right) \tag{5.2}$$

El método global toma en cuenta todas las instancias del conjunto de entrenamiento para evaluar a t_x. Se obtiene el vector promedio de los vectores de T, llamado centroide (denotado por C) y un área dentro de T denotada por el centroide y una distancia. Teniendo en cuenta el promedio de distancias y la desviación estándar (Ecuaciones 5.3 y 5.4 respectivamente) se crean áreas en las cuales se encontrarán las nuevas instancias importantes.

$$\mu(T) = \frac{1}{\mid T \mid} \sum_{\forall t_i \in T} d_{i,c} \tag{5.3}$$

$$\sigma(T) = \sqrt{\frac{1}{\mid T \mid} \sum_{\forall d_{i,c} \in T} (d_{i,c} - \mu(T))^2} \tag{5.4}$$

La primera distancia utilizada para la comparación es μ, si las nuevas instancias entran dentro del área conformada por C como centro y un radio de μ, son consideradas como importantes, a esta variante se le denomina **G-Prom** (Ecuación 5.5).

$$d_{x,c} < \mu(T) \tag{5.5}$$

Con μ y σ se genera otras que detecta los elementos importantes siguiendo la Ecuación 5.6, con $k = 1$ (variante llamada **G-PromDesv**) y $k = 2$ (variante llamada **G-Prom2Desv**). Finalmente, la variante **G-Max** crea un área entre C y la máxima distancia entre este punto y T (Ecuación 5.7). La Figura 5.2 muestra la comparación entre la nueva instancia t_x con sus vecinos, con $k = 3$ con el método local y las áreas creadas en el método global.

$$d_{x,c} < \mu(T) + k\sigma(T) \tag{5.6}$$

$$d_{x,c} < max_{\forall d_i \in T}(d_{i,c}) \tag{5.7}$$

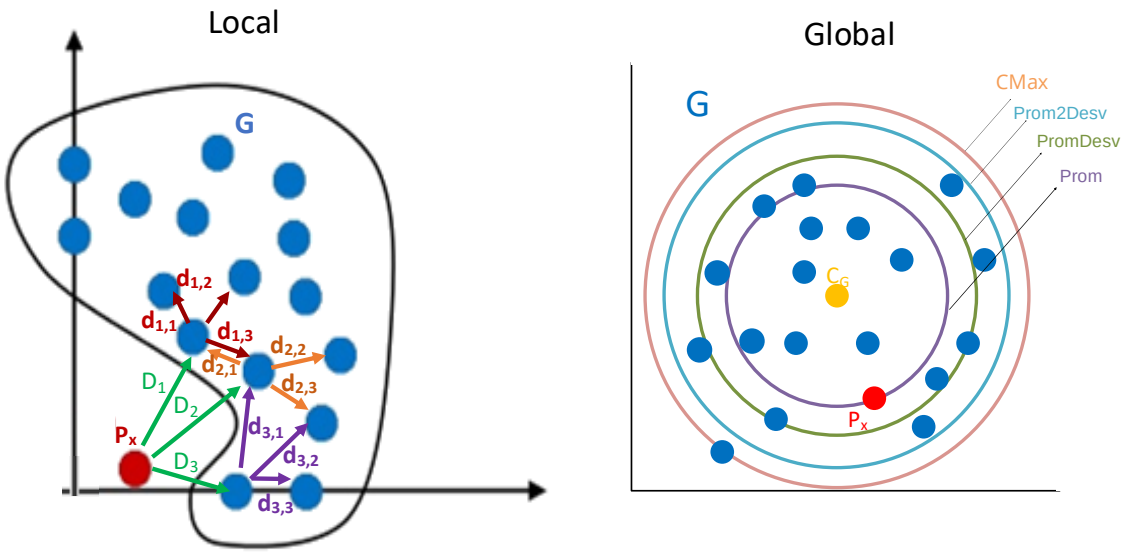


Figura 5.2: Método local que compara una nueva instancia con sus k vecinos más cercanos (izquierda) y áreas creadas con el método global (derecha).

Para medir la distancia $d_{i,j}$ entre un par de texto t_i y t_j se utiliza $d_{i,j} = 1 - sim(t_i, t_j)$, donde $sim()$ es la similitud coseno entre los vectores de t_i y t_j. ngitud del vector d.

La Tabla 5.4 muestra los resultados del método local, separados de acuerdo a la variante del experimento y clase analizada. Los experimentos que superan al *baseline* en F_1 se muestran en negritas.

Método	**Clase**	P	R	F_1	**Rec.**
L-Max	*Estilos de aprendizaje*	1.000	0.071	0.133	1
	Estrategias de enseñanza	0.625	0.333	**0.435**	5
	Tipos de inteligencias	1.000	0.077	0.143	1
L-Prom	*Estilos de aprendizaje*	1.000	0.214	0.353	3
	Estrategias de enseñanza	0.500	0.267	**0.348**	4
	Tipos de inteligencias	1.000	0.077	0.143	1

Tabla 5.4: Resultados obtenidos con el método local para filtrado

Dos de las cuatro clases obtuvieron una precisión de 1, la diferencia se centra en el recuerdo, la cual es mayor para las clases NA y *Estrategias de enseñanza*. La clase *Tipos de inteligencias* presenta los resultados más bajos, sobre todo al analizar el recuerdo. En cuanto al número de documentos recuperados, en todos los casos se recuperan menos de 10 documentos. Al momento de comparar las distancias, los elementos de T están muy unidos entre sí, por lo que al comparar una nueva instancia, esta no se encuentra tan cerca de sus vecinos, esto hace que el número de elementos recuperado sea mínimo.

Los resultados del método global se muestran en la Tabla 5.5. Al analizar la distancia de las nuevas instancias con el centroide y compararlas con la distancia promedio (variante $Prom$), los resultados son 0 en todos los experimentos, por lo que esta variante no se anexa en la tabla. Esto refuerza la idea de que los documentos de entrenamiento son muy cercanos, por lo que al crear un área de radio igual al promedio de distancias, esta es muy restrictiva para que las nuevas instancias se puedan recuperar. En cambio, cuando el área a considerar es desde el centroide hasta la distancia máxima (variante *G-Max*), la mayoría de los nuevos artículos son etiquetados como importantes. Por lo tanto, la exhaustividad es alta, pero la precisión es baja, especialmente en las clases homogéneas (*Estilos de aprendizaje*); aún así, los resultados son mejores que las propuestas locales.

La versión *G-Max* recuperó más documentos, pero la versión *G-PromDesv* tiene mejores resultados en cuando a F_1. Aunque los documentos no se encuentran cerca del centroide, si se amplía mucho el área se recuperan muchos falsos positivos cuando las clases no son homogéneas. En el caso de las clases homogéneas, la mejor variante fue *G-Prom2Desv*.

Contrastando los resultados con el *baseline*, el método local obtiene mejores resultados para la clase *Estrategias de enseñanza*, mientras que el método global lo supera en todas las

Método	Clase	P	R	F_1	Rec.
G-Max	*Estilos de aprendizaje*	0.929	0.813	**0.867**	13
	Estrategias de enseñanza	0.273	0.8	**0.407**	12
	Tipos de inteligencias	0.563	0.692	**0.621**	9
G-PromDesv	*Estilos de aprendizaje*	1	0.375	**0.545**	6
	Estrategias de enseñanza	0.571	0.267	**0.364**	4
	Tipos de inteligencias	1	0.308	**0.471**	4
G-Prom2Desv	*Estilos de aprendizaje*	0.929	0.813	**0.867**	13
	Estrategias de enseñanza	0.293	0.8	**0.429**	12
	Tipos de inteligencias	0.615	0.615	**0.615**	8

Tabla 5.5: Resultados del método global para filtrado

clases. Todos los documentos relacionados con las estrategias de enseñanza aprendizaje, aunque comparten el mismo enfoque epistémico, no todos abarcan todas las estrategias, hay documentos que solo analizan las estrategias metacognitivas o incluso estrategias de lectura, que se consideran como de apoyo. Esta característica hace que al menos para esta clase, sea más factible comparar un documento con sus similares en lugar de compararlo con todo el conjunto.

Los documentos catalogados como falsos positivos son artículos que hablan del tema y que mencionan el enfoque analizado (generalmente en el trabajo relacionado o marco teórico del artículo) pero no se centran en dicho enfoque. Los falsos negativos son aquellos que manejan un fundamento teórico muy pequeño y se centran más en el análisis de resultados.

En la clase *Tipos de inteligencias* los falsos positivos son artículos que hablan sobre inteligencia emocional (la cual no está dentro de las 8 inteligencias de Gardner), pero al momento de definir el término inteligencia utilizan conceptos que están dentro de la lista de palabras utilizada como atributos. Incluso algunos describen la teoría de Gardner, como parte de un sustento teórico y para comparar las bases de los dos planteamientos. Los falsos negativos abordan la temática desde un punto de vista epistémico, por lo que utilizan palabras no tan comunes y las comparan con otros enfoques, esto hace que el vocabulario se expanda y no se obtenga una frecuencia adecuada para las palabras de conforman los atributos. Otros ejemplos de falsos negativos son artículos que analizan una sola inteligencia (espacial o corporal).

Las clases *Estrategias de enseñanza* y *Estilos de aprendizaje* tienen comportamientos similares, por ejemplo si un artículo habla de los estilos de aprendizaje bajo el enfoque de la programación neurolingüística pero en su marco teórico desarrolla ampliamente la teoría de Kolb, este quedará como falso positivo, y si el documento solo analiza el estilo de aprendizaje pragmático pero no menciona los otros tres estilos, quedará como falso negativo.

Finalmente, se aplicaron los métodos creados a cada clase para clasificar la totalidad de las instancias del *corpus* B. La Figura 5.3 muestra el número de documentos catalogados como importantes por clase y por variante. La relación entre el total recuperado por clase se mantiene en las variantes (la clase *Estrategias de enseñanza* recupera más documentos en todas las variantes y la clase *Tipos de inteligencias* es la que menos documentos recupera), esto también se relaciona con el número de documentos que se analizan.

La efectividad de los métodos depende de la estructura de las clases, el método local es mejor cuando los documentos son heterogéneos, es decir, cada uno de los documentos analizados se enfocan en una pequeña parte del dominio analizada. Él método global funciona mejor cuando los documentos son uniformes y todos abarcan la totalidad de las subclases analizadas. En cuestión de los algoritmos, el método global conlleva menos tiempo de ejecución, ya que el centroide se obtiene una sola vez para posteriormente sacar la distancia entre éste y todos los documentos a analizar. El método local implica calcular los vecinos de cada instancia y a su vez los vecinos de estos vecinos en cada iteración.

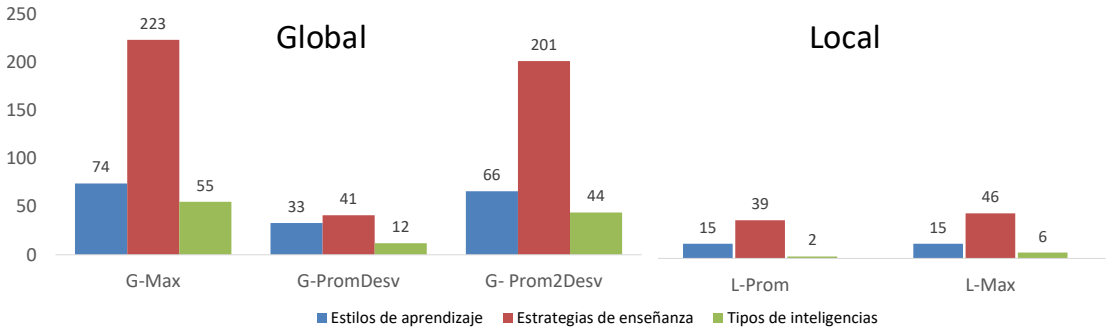


Figura 5.3: Número de documentos recuperados por clase y método utilizando todo el corpus auxiliar.

Tomando en cuenta los resultados más altos en los experimentos en cuanto a F_1, las instancias recuperadas se unen al *corpus* A para conformar el *corpus* $Final$. La Tabla 5.6 muestra el número de instancias por clase, así como el método que se tomó en cuenta para su creación. El corpus $Final$ queda con 61 instancias de la clase *Tipos de inteligencias*, 63 instancias en la clase *Estrategias de enseñanza* y 83 instancias en la clase *Estilos de aprendizaje*.

5.3. Conceptos compuestos

De acuerdo a la Figura 4.5 del capítulo de metodología, el método para la detección automática de conceptos compuestos consta de dos fases: Análisis de frecuencias y cálculo de probabilidad

Tabla 5.6: Total de instancias por clase del *corpus* $Final$.

Clase	Método	*Corpus* A	Recuperados (B)	*Corpus* $Final$
Tipos de inteligencias	$G - Prom2Desv$	17	44	61
Estrategias de enseñanza	$L - Max$	17	46	63
Estilos de aprendizaje	$G - Prom2Desv$	17	66	83
TOTAL		51	156	**207**

condicional. En las siguientes tablas de analiza el método para la clase estilos de aprendizaje.

La Tabla 5.7 muestra los datos por cuartil de la clase estilos de aprendizaje. El total de pares de palabras adyacentes asciende a 96,000 pares aproximadamente, los cuales se dividen en 4 cuartiles, quedando cada uno en 24,000 pares. En la tabla se muestran además los límites mínimo y máximo de cada cuartil, así como el promedio y la desviación estandard de cada uno de ellos. La mayoría de los pares de palabras tienen una frecuencia de uno, quedando 3 cuartiles sin cambios en sus frecuencias, y quedando los pares más importantes en el cuartil 1. El promedio y desviación de esos cuartiles también permanecen sin cambios.

Tabla 5.7: Análisis de frecuencias en pares de palabras.

Quartil	**FrecMin**	**FrecMax**	**Núm instancias**	**Promedio**	**Desviación estándar**
1	2	3,054	24,022	4.729	22.8531
2	1	2	24,022	1.038	0.1912
3	1	1	24,022	1.000	0
4	1	1	24,020	1.000	0
General	1	3,054	96,086	1.9418	11.5397

Tomando en cuenta este análisis, solo se calcula la probabilidad condicional de los pares pertenecientes al cuartil 1 (24,022 palabras). Éstos pasan a una segunda etapa en la cual se vuelven a calcular cuartiles pero analizando la probabilidad condicional; los resultados se muestran en la Tabla 5.8. Esta etapa presenta el mismo comportamiento que el análisis de frecuencias, es decir, el mayor rango de probabilidades se concentra en el cuartil 1. Además, en este cuartil se presenta el promedio y desviación más altos, mientras que en los otros cuartiles estas dos métricas no varían respecto a sus máximos y mínimos.

Tabla 5.8: Análisis de probabilidad condicional en pares de palabras.

Quartil	**ProbMin**	**ProbMax**	**Núm instancias**	**Promedio**	**Desviación estándar**
1	0.0952	1.000	6,006	0.4211	0.3172
2	0.0228	0.0952	6,006	0.0474	0.0196
3	0.0071	0.0228	6,006	0.0134	0.0044
4	0.0003	0.0071	6,004	0.0034	0.002
General	0.0003	1	24,022	0.1213	0.2355

Al igual que en el análisis de frecuencias, solo se toman los pares de palabras del cuartil uno

para compararlos con un umbral. Dicho umbral queda determinado por el rango $\chi - \sigma$ a $\chi + \sigma$. Por lo tanto, en una primera versión del corpus, todos los pares de palabras que tengan una probabilidad condicional entre *0.1038* y *0.7382* se consideran como conceptos compuestos. Con esta regla se recuperan algunos conceptos relevantes para el dominio de longitud 2 como *estilo pragmático* y *estilo aprendizaje*.

Los experimentos anteriores detectan conceptos de dos palabras. Para obtener conceptos de longitud 3 (por ejemplo *estrategias enseñanza aprendizaje* se vuelve a ejecutar el mismo procedimiento con el corpus obtenido de la primera iteración. La Tabla 5.9 muestra los análisis de frecuencias y probabilidades de este experimento.

Tabla 5.9: Análisis de frecuencias y probabilidad para extracción de conceptos de longitud 3

Análisis	Quartil	Min	Max	Núm instancias	Promedio	Desviación estándar
Frecuencia	1	1	259	25,263	3.3199	5.4717
	2	1	1	25,263	1.0000	0
	3	1	1	25,263	1.0000	0
	4	1	1	25,263	1.0000	0
	General	1	259	101052	1.5800	2.9144
Probabilidad Condicional	1	0.1020	1	6,316	0.5910	0.3637
	2	0.0247	0.1017	6,316	0.0521	0.0219
	3	0.0079	0.0247	6,316	0.0146	0.0047
	4	0.0003	0.0079	6,315	0.0040	0.0021
	General	0.0003	1	25,263	0.1654	0.3064

Al igual que en la primera iteración, cuartil 1 es el que concentra los conceptos con mayor frecuencia, por lo que solo se trabaja con ese. En cuando las probabilidades condicionales, los promedio son más altos que en la primera iteración, por lo que el rango a tomar como umbral de mueve un poco a la derecha.

Contrastando la lista de conceptos encontrados con el *gold standard* se detectan 175 de los 275 conceptos, a parte de otros que si bien no son significativos para el dominio, si cumplen la función de un concepto compuesto. Dentro de los conceptos que no fueron detectados se encuentran aquellos que son importantes para el dominio pero no tan relacionados con la clase por ejemplo:

- método docente
- estudio universitario
- característica ambiental
- equipo trabajo
- trayectoria académica
- binomio enseñanza aprendizaje
- estilo aprender

- ➥ institución educativa
- ➥ área conocimiento

5.4. Detección de conceptos

En esta sección se mencionan los resultados de los experimentos descritos en el capítulo de metodología para la detección de conceptos compuestos. El método principal utilizado en esta fase del aprendizaje ontológico consta del análisis de métricas de similitud textual, las cuales se aplican al vocabulario de cada corpus. Los resultados que se muestran pertenecen al procesamiento de los documentos con conceptos simples, aunque se ejecutan experimentos para las dos versiones de dichos documentos (simples y compuestos, obtenidos con la metodología previamente descrita)

Se realizaron los experimentos con las tres representaciones y métricas del corpus inicial. Los resultados de cada experimento fueron señalados con la notación $Clase_{Representacion,Metricas,\gamma}$ donde γ es el tope del experimento para recuperar palabras. Por ejemplo, el experimento de PMI utilizando Wikipedia para la clase de estilos de aprendizaje con 5-gramas y un tope de $0{,}1$ es representado por $EA_{5g,Pw,0,1}$ y la precisión de las métricas basadas en términos para los tipos de inteligencias con oraciones y un tope de $0{,}15$ es representado por $TI_{Se,Te,0,15}$. Primeramente se tomarán como ejemplo algunos experimentos para analizar la lista de palabras recuperadas y describir como se obtuvieron los valores de la precisión.

La Tabla 5.10 muestra la lista de palabras recuperadas en el experimento $EE_{Se,Te,0,27}$. Te está conformado por 4 métricas de traslape de términos, por lo que se usó el sistema de voto para determinar las palabras que se recuperarían. Para este experimento en particular, el valor de γ determina el valor de cada métrica que se tomará en cuenta para recuperar una palabra.

Tabla 5.10: Resultados de métricas basadas en términos para el experimento $EE_{Se,Te,0,27}$.

Palabra	Dicce	Jaccard	Overlap	Coseno
estudio	0.1655	0.0902	0.3802	0.2005
conocimiento	0.2333	0.132	0.4973	0.2753
proceso	0.2418	0.1375	0.4531	0.2733
cognitivo	0.3253	0.1943	0.6701	0.3794
metacognitivas	0.2556	0.1465	0.9271	0.3707
poder	0.1752	0.096	0.3971	0.2113
aprendizaje	0.5156	0.3474	0.5656	0.5177
estudiante	0.3134	0.1858	0.438	0.3269

En la Tabla se muestran 8 palabras, cuyas métricas asociadas presentan valores mayores de $0{,}16$ en 3 o 4 de ellas. Por ejemplo, la palabra *conocimiento* al compararse con la clase *Estilos de aprendizaje* obtuvo un Jaccard de $0{,}132$, el cual es menor a γ, sin embargo, las otras tres

métricas obtienen valores mayores, por lo que se considera como recuperada. De estas ocho palabras recuperadas, solo 3 no se encuentran dentro del *gold* (*estudio*, *proceso*, *poder*) por lo que la precisión para este experimento es de $0{,}625$. El resto de las palabras representan un tipo de estrategias (*metacognitivas*) y las características de estas estrategias en la teoría del aprendizaje. Además, se anexa la restricción de que un experimento debe recuperar al menos 5 palabras, de lo contrario, se considera una cantidad muy pequeña para la construcción de la ontología y la precisión pasa a ser 0.

La Tabla 5.11 muestra la precisión utilizando la representación del corpus Te con un valor de γ de $0{,}02$ a $0{,}30$, con intervalos de $0{,}02$. Se aplica la condición explicada anteriormente sobre el número de palabras, por ejemplo, el experimento $EE_{5g,Te,0{,}16}$ recuperó dos palabras, de las cuales solo una es relevante, esto representa una precisión de $0{,}5$, pero una palabra no es suficiente para considerarla como concepto en una ontología, por lo que la precisión se consideró 0.

Tabla 5.11: Precisión obtenida utilizando las métricas basadas en términos.

γ	$EA_{5g,Te}$	$EA_{Se,Te}$	$EA_{Pa,Te}$	$EE_{5g,Te}$	$EE_{Se,Te}$	$EE_{Pa,Te}$	$TI_{5g,Te}$	$TI_{Se,Te}$	$TI_{Pa,Te}$
0.02	0.6452	0.2523	0.8889	0.4286	0.0951	0.625	0.5758	0.1379	0.7500
0.04	0.9000	0.3918	0	0.7000	0.1486	0	0.7273	0.2105	0
0.06	0.8333	0.4464	0	0.6250	0.1882	0	0	0.2766	0
0.08	0	0.6857	0	0	0.2432	0	0	0.3636	0
0.10	0	0.7200	0	0	0.3467	0	0	0.4426	0
0.12	0	0.7500	0	0	0.3774	0	0	0.4444	0
0.14	0	0.8889	0	0	0.4000	0	0	0.5143	0
0.16	0	0.8333	0	0	0.4483	0	0	0.5652	0
0.18	0	0	0	0	0.4400	0	0	0.5882	0
0.20	0	0	0	0	0.5000	0	0	0.6923	0
0.22	0	0	0	0	0.4615	0	0	0.8000	0
0.24	0	0	0	0	0.6000	0	0	0.7500	0
0.26	0	0	0	0	0.5556	0	0	0.7500	0
0.28	0	0	0	0	0	0	0	0.7143	0
0.30	0	0	0	0	0	0	0	0.8333	0

Para los estilos de aprendizaje, la precisión más alta fue $EA_{Pa,Te,0{,}02}$, obteniendo una precisión de $0{,}89$ con nueve palabras recuperadas correctas, entre las que se encuentran los 4 tipos de estilos de aprendizaje, y los autores que crearon el cuestionario para detectarlos. En la clase de estrategias de enseñanza, el mejor experimento fue $EE_{5g,Te,0{,}04}$, con una precisión de $0{,}7$. Finalmente, para la clase tipo de inteligencias, el mejor resultado fue $TI_{Se,Te,0{,}30}$ con precisión $0{,}83$. En este experimento se recuperaron solo seis palabras, todas asociadas al concepto de inteligencia y a los tipos de inteligencias. En las tres clases se observa que la representación en oraciones (Se) tiene menos experimentos con precisión 0, y precisiones más altas con esta representación. En general, los resultados se van incrementando, hasya llegar al punto en que no se recuperan palabras o se hacen con una frecuencia muy baja.

La Tabla 5.12 muestra los resultados obtenidos comparando la métrica PMI con libros de pedagogía (Pb). Los valores de γ van de $-0{,}5$ a $0{,}5$ en intervalos de $0{,}05$. La tabla solo muestra los resultados hasta $\gamma = 0{,}3$, ya que, para valores mayores, en todos los experimentos el resultado es 0. En estos experimentos los resultados no varían mucho a medida que γ incrementa su valor.

Tabla 5.12: Precisión obtenida en los experimentos utilizando la representación Pb.

γ	$EA_{5g,Pb}$	$EA_{Se,Pb}$	$EA_{Pa,Pb}$	$EE_{5g,Pb}$	$EE_{Se,Pb}$	$EE_{Pa,Pb}$	$TI_{5g,Pb}$	$TI_{Se,Pb}$	$TI_{Pa,Pb}$
-0.5	0.0205	0.0191	0.0262	0.0259	0.0239	0.0357	0.0371	0.0349	0.0554
-0.45	0.0206	0.0191	0.0262	0.0259	0.0239	0.0357	0.0371	0.0349	0.0554
-0.4	0.0206	0.0191	0.0262	0.0259	0.0239	0.0357	0.0371	0.0349	0.0554
-0.35	0.0197	0.0182	0.0262	0.0259	0.024	0.0357	0.0371	0.0349	0.0554
-0.3	0.0197	0.0185	0.0262	0.0254	0.0235	0.0357	0.0366	0.035	0.0554
-0.25	0.0201	0.0186	0.0263	0.0237	0.0219	0.0357	0.0321	0.0307	0.0554
-0.2	0.0206	0.0188	0.0258	0.0205	0.0196	0.0357	0.0308	0.0281	0.0554
-0.15	0.0233	0.0208	0.0248	0.0186	0.0182	0.0358	0.0301	0.028	0.0555
-0.1	0.0265	0.0228	0.0257	0.0198	0.0183	0.034	0.0336	0.0297	0.0528
-0.05	0.0302	0.0274	0.0286	0.0243	0.02	0.0308	0.03	0.0266	0.0469
0	0.033	0.0347	0.0275	0.0256	0.0215	0.0242	0.0385	0.029	0.0443
0.05	0.0481	0.0425	0.0371	0.027	0.0223	0.0196	0.0431	0.0374	0.0496
0.1	0.0909	0.0796	0.0571	0.0417	0.0302	0.0191	0.0667	0.0504	0.0517
0.15	0.1429	0.1169	0.0701	0	0	0	0.1149	0.0902	0.0751
0.2	0.2917	0.1944	0.1013	0	0	0	0.1458	0.1333	0.1034
0.25	0	0	0	0	0	0	0.25	0.186	0.1739
0.3	0	0	0	0	0	0	0	0	0.2414

La clase de tipos de inteligencias obtuvo más palabras a medida que γ se incrementa, mientras que la clase de estrategias de enseñanza obtiene ceros a partir de $\gamma = 0{,}05$. Para la clase de estilos de aprendizaje la precisión más alta fue de $0{,}29$ con el experimento $EA_{5g,Pb,0{,}20}$, para la clase de tipos de inteligencia la precisión más alta fue de $0{,}25$, con el experimento $TI_{5g,Pb,0{,}25}$. La clase de estrategias de enseñanza obtiene un valor mayor en $EE_{5g,Pb,0{,}10}$, pero la precisión apenas alcanza $0{,}041$. Aunque utilizando la métricas Pb los resultados decrecieron en todas las representaciones con respecto a las métricas Te. Un cambio respecto a los experimentos anteriores es que los mejores resultados se dieron en la representación $5g$.

La Tabla 5.13 muestra los resultados obtenidos utilizando la métrica PMI con el corpus de Wikipedia (Pw), el rango de γ es el mismo que el utilizado en la representación Pb. En general los resultados son más bajos que en los experimentos con la representación Pb. Para la clase de estilos de aprendizaje la precisión más alta se obtuvo con el experimento $EA_{5g,Pw,0{,}20}$, obteniendo una precisión de $0{,}1951$. Para la clase de estrategia de enseñanza la mayor precisión fue $0{,}0286$ con $EE_{5g,Pw,0{,}05}$, mientras que para la clase de tipos de inteligencia la mayor precisión fue $0{,}1778$ con el experimento $TI_{Pa,Pw,0{,}25}$. La clase de tipos de inteligencias fue la única que obtuvo mejores resultados utilizando la representación en pares, mientras que las otras dos clases los obtuvieron con la representación de 5-gramas.

Tabla 5.13: Precisión obtenida en los experimentos utilizando la representación Pw.

γ	$EA_{5g,Pw}$	$EA_{Se,Pw}$	$EA_{Pa,Pw}$	$EE_{5g,Pw}$	$EE_{Se,Pw}$	$EE_{Pa,Pw}$	$TI_{5g,Pw}$	$TI_{Se,Pw}$	$TI_{Pa,Pw}$
-0.5	0.0339	0.0343	0.0403	0.0224	0.0253	0.024	0.0352	0.0357	0.0475
-0.45	0.0339	0.0343	0.0403	0.0224	0.0253	0.024	0.0352	0.0357	0.0475
-0.4	0.0339	0.0343	0.0403	0.0224	0.0253	0.024	0.0352	0.0357	0.0475
-0.35	0.0339	0.0343	0.0403	0.0224	0.0253	0.024	0.0352	0.0357	0.0475
-0.3	0.0339	0.0343	0.0403	0.0224	0.0253	0.024	0.0352	0.0357	0.0475
-0.25	0.0339	0.0343	0.0403	0.0224	0.0253	0.024	0.0352	0.0357	0.0475
-0.2	0.0339	0.0343	0.0403	0.0224	0.0253	0.024	0.0352	0.0357	0.0475
-0.15	0.0339	0.0344	0.0403	0.0224	0.0254	0.024	0.0352	0.0358	0.0475
-0.1	0.0339	0.0346	0.0403	0.0224	0.0255	0.0241	0.0352	0.036	0.0477
-0.05	0.0339	0.0359	0.0406	0.0224	0.0254	0.0243	0.0352	0.038	0.0483
0	0.0388	0.0407	0.0425	0.0236	0.0244	0.0245	0.0398	0.0346	0.0493
0.05	0.0567	0.0439	0.0491	0.0286	0.0233	0.0212	0.0405	0.0411	0.0468
0.1	0.0804	0.0503	0.0466	0	0.0177	0.0174	0.0634	0.0375	0.0579
0.15	0.1429	0.0599	0.0556	0	0	0.0211	0	0.0465	0.0701
0.2	0.1951	0.0885	0.0621	0	0	0	0	0	0.0982
0.25	0	0.1522	0.1139	0	0	0	0	0	0.1778

La precisión obtenida tanto para la representación con Pw y Pb es mucho más baja en las tres clases. Sin embargo, en un análisis del vocabulario de los dos corpus utilizados en PMI y el corpus original, hay varias palabras que no comparten, incluso en el *gold standard* hay palabras que no aparecen relacionadas con ninguna de las clases. La Figura 5.4 muestra el número de palabras del *gold* que no aparecen dentro de los pares de palabras recuperados en las métricas Pw y Pb.

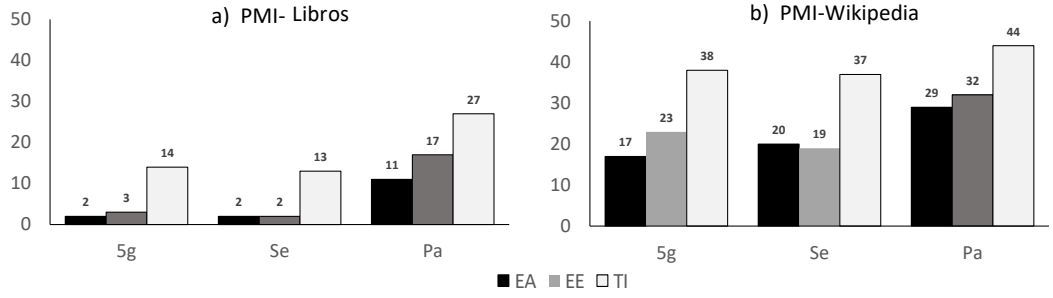


Figura 5.4: Número de palabras por clase que no aparecen en el corpus de $Wikipedia$ y $Libros$

Se puede apreciar que el número de palabras que no aparecen dentro del corpus de wikipedia es mucho mayor al número de palabras que no aparecen en el corpus de libros, especialmente en la representación de pares de palabras. La clase tipos de inteligencias es la que tiene más palabras perdidas dentro de los corpus,sin embargo, esta clase es de las que tiene la precisión más alta en las representaciones, especialmente en la de $5g$. Analizando las representaciones, Pa tiene el mayor número de palabras que no aparecen en el *gold*, esto se justifica por el método usado para obtener las representaciones. La representación Pa solo relaciona dos palabras

cuando aparecen juntas en el corpus, mientras que las representaciones Se y $5g$ relacionan dos palabras si aparecen en la misma oración y si están a una distancia de cuatro o menos palabras respectivamente.

En cuanto al análisis del recuerdo, la Figura 5.5 muestra los resultados de todos los experimentos para las representaciones Pw y Pb. En estos experimentos, todos los resultados se encuentran en el segundo y tercer cuadrante, los que se encuentran cerca del eje vertical corresponden a la representación pa, por lo tanto la precisión es alta, pero el recuerdo bajo. Los experimentos usando la representación $5g$ muestran resultados más alejados a este eje.

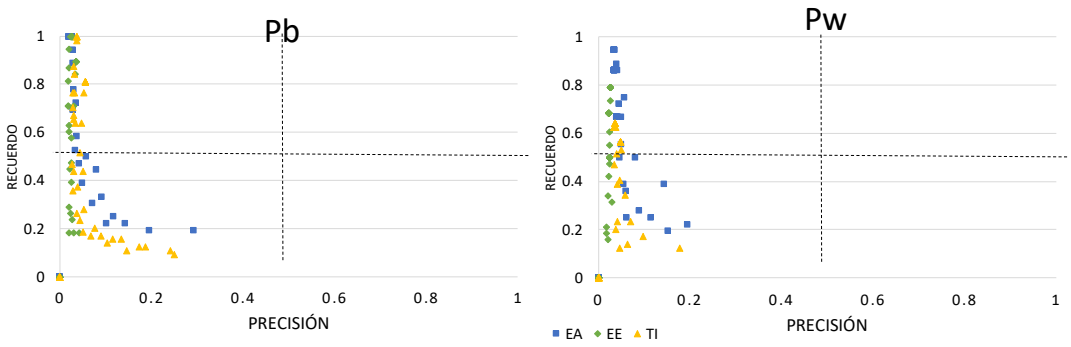


Figura 5.5: Precisión y recuerdo para las representaciones Pb y Pw.

La Figura 5.6 muestra los resultados para todos los experimentos del conjunto Te, separados por representación y clases. En la gráfica a y c, la mayoría de los experimentos se encuentran en el tercer cuadrante (recuerdo y precisión bajos). En la representación Se se muestra un comportamiento similar en las tres clases, donde los experimentos tienen recuerdo bajo y precisión alta o recuerdo alto y precisión baja.

En estos experimentos, se recuperan varios conceptos principales, pero también algunos que no tienen relación con el concepto principal, es por esto que el recuerdo es muy bajo yy a su vez, disminuye el F_1.

5.4.1. Frecuencias maximales

De acuerdo con al algoritmo descrito en el marco metodológico (sección 4.3), en la clase *Tipos de inteligencias* se recuperan 9 $n-gramas$, mientras que en la clase *Estrategias de enseñanza* se recuperan 10 $n-gramas$; finalmente, en la clase *Estilos de aprendizaje* se recuperan 43 $n-gramas$. La Tabla 5.14 muestra algunos de los ejemplos recuperados.

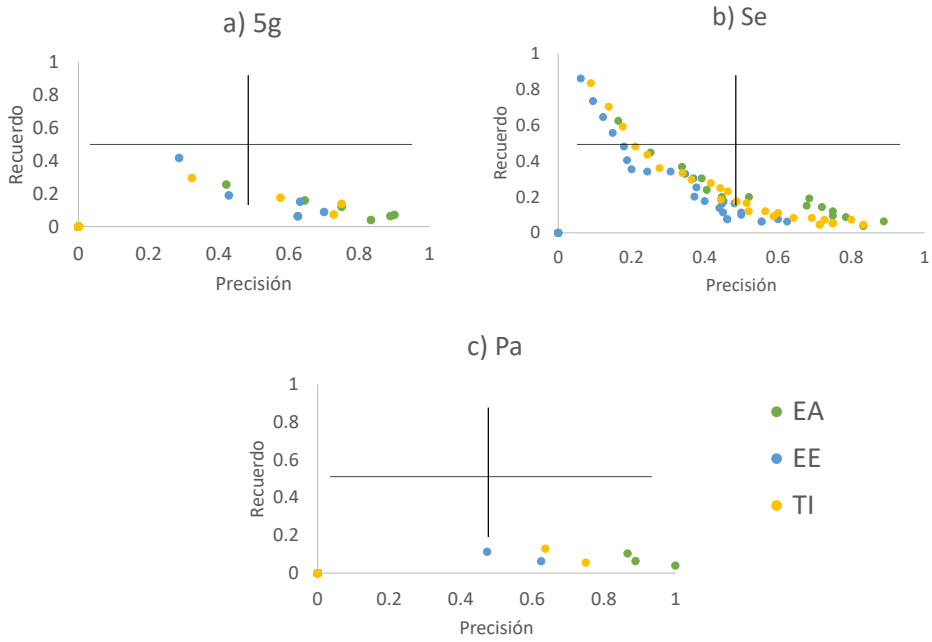


Figura 5.6: Precisión y recuerdo por clases para el conjunto de métricas Te.

Tabla 5.14: $N-gramas$ recuperados utilizando frecuencias maximales y validados por el experto en el dominio.

Clase	Ejemplo	
Tipos de inteligencias	rendimiento académico estudiante teoría inteligencia múltiple	inteligencia lógico matemático habilidad social
Estrategias de enseñanza	proceso enseñanza aprendizaje proceso de aprendizaje uso tic conocimiento metacognitiva	aprendizaje estudiante cognitivo y metacognitiva proceso cognitivo proceso aprendizaje
Estilos de aprendizaje	estilo aprendizaje activo activo pragmático cuestionario honey alonso estilo aprendizaje teórico reflexivo estilo estrategia aprendizaje estilo aprendizaje reflexivo estilo aprendizaje enseñanza estilo reflexivo teórico estilo aprendizaje predominante pragmático activo estrategia enseñanza activo reflexivo teórico pragmático reflexivo pragmático de aprendizaje en estudiante el rendimiento académico	rendimiento académico estudiante estilo activo reflexivo teórico y pragmático estilo aprendizaje alumno proceso enseñanza aprendizaje activo teórico preferencia estilo aprendizaje estilo aprendizaje rendimiento académico estilo aprendizaje estudiante universitario aprendizaje teórico aprendizaje activo reflexivo cada estilo aprendizaje estrategia de aprendizaje

Estas listas fueron analizadas con la ayuda del experto en el dominio a fin de descartar elementos no relevantes. La clase *Estilos de aprendizaje* presenta una mayor consistencia en cuando a los conceptos recuperados; además, tanto el concepto principal como su clasificación están en varios de los $n-gramas$ recuperados. La clase *Estrategias de enseñanza* aunque tiene pocos

elementos recuperados, entre estos se encuentran los conceptos más importantes para esta clase principal ya que con estos resultados ya se pueden definir las estrategias de aprendizaje como un proceso relacionado con los alumnos, incluso se mencionan las estrategias cognitivas y metacognitivas. La clase *Tipos de inteligencias* tiene pocos elementos, pero importantes, sin embargo, no se logran recuperar conceptos que permiten describirla ampliamente.

5.4.2. Método para detección de patrones

Como se mencionó anteriormente, en el método automático se buscaron trigramas de categorías gramaticales, pero se recuperaron quintigramas de palabras, tomando en cuenta el antecesor y sucesor de cada patrón. En total, para la clase de tipos de inteligencias se recuperaron 100 frases y de la clase estilos de aprendizaje 52, eliminado aquellos con una frecuencia menor, quedan los mostrados en la Tabla 5.15.

Tabla 5.15: Conceptos y patrones de categorías encontrados con el método automático.

Patrón	Ejemplo
Tipos de inteligencias	
$DT\_B - NP, NN\_I - NP, IN\_B - PP$	de la lengua como una, con la enseñanza de las
$DT\_B - NP, NN\_I - NP, JJ\_I - NP$	de una manera racional a
$DT\_B - NP, NN\_I - NP, DT\_B - NP$	como el ritmo el tono, con la música la actividad
$DT\_B - NP, NN\_I - NP, IN\_B - PP$	por una combinación de inteligencias
$DT\_B - NP, NN\_I - NP, IN\_B - PP$	es la inteligencia para entender
$DT\_B - NP, NN\_I - NP, IN\_B - PP$	por otro lado según richards
$DT\_B - NP, NN\_I - NP, JJ\_I - NP$	de una filosofía educativa que
$DT\_B - NP, NN\_I - NP, IN\_B - PP$	entre el alumnado debido a, para la música como en
$IN\_B - PP, NN\_B - NP, VBN\_B - VP$	racional a menudo encontrada en
$DT\_B - NP, NN\_I - NP, IN\_B - PP$	definen el dominio de la, posibilita el desarrollo de las
Estilos de aprendizaje	
$DT\_B - NP, NN\_I - NP, JJ\_I - NP$	en la experimentación activa y
$DT\_B - NP, NN\_I - NP, IN\_B - PP$	para el procesamiento de la,a la comprensión de los
$DT\_B - NP, NNS\_I - NP, JJ\_I - NP$	con las preferencias individuales de
$IN\_B - PP, NN\_B - NP, VBN\_B - VP$	reflexivo de aprendizaje centrado en
$IN\_B - PP, NN\_B - NP, VB\_B - VP$	estilo de aprendizaje implica diferentes
$DT\_B - NP, NN\_I - NP, IN\_B - PP$	conclusiones este estilo de aprendizaje
$IN\_B - PP, NN\_B - NP, DT\_B - NP$	ambientes de aprendizaje estos rasgos

Se aprecian algunos patrones repetidos porque se toman en cuenta los sucesores y sucesores, sin embargo solo se muestra el patrón original (trigrama). En la clase tipos de inteligencias se recuperan más patrones ya que la instancia de entrenamiento utilizada pertenece a esta clase. En la columna de ejemplos se muestran algunas de las frases que coinciden con el patrón, de estás, el concepto importante (el que pertenece a la lista inicial) es la palabra del centro, por ejemplo *lengua, enseñanza, experimentación, procesamiento*.

Analizando solo las categorías principales, la mayoría de los patrones encontrados tienen una estructura similar: $DT, NN, \_$ y otra categoría como IN, JJ, DT, por mencionar algunas. De acuerdo a la nomenclatura de CLIPS, esto es un determinante, seguido de un sustantivo y posteriormente

puede ir una conjunción, adjetivo u otro determinante. Estos serían los que presentan mayor frecuencia, sin embargo, analizando sus ejemplos, solo en la clase de tipos de inteligencias tienen algo de información importante, se recuperan características de la inteligencia musical, autores que las han estudiado, incluso el término *filosofía educativa*. En la clase de estilos de aprendizaje sólo se recuperan algunas alusiones vagas a los estilos de aprendizaje.

La Tabla 5.16 muestra algunos de los patrones detectados con el método semiautomático, en donde las dos etiquetas trabajadas se analizan indistintamente. Al igual que en la tabla anterior, solo se muestran los patrones con mayor frecuencia, y un ejemplo de una frase recuperada con estos. Se observa que predominan los patrones compuestos por sustantivos y determinantes como en el método automático, sin embargo, en estos aparecen otros elementos como pronombres personales (PRP), verbos (VB) y conjunciones (IN).

Tabla 5.16: Conceptos y patrones de categorías encontrados con el método semiautomático.

Concepto	Patrón
Tipos de inteligencias	
$DT\_B-NP, NN\_I-NP, IN\_B-PP$	ese tipo de,la existencia de
$DT\_B-NP, NN\_I-NP, VB\_B-VP$	las inteligencias son
$IN\_B-PP, NN\_B-NP, IN\_B-PP$	En función de
$PRP\_B-NP, VB\_B-VP, IN\_B-PP$	se consideraba que,se tratarán en
$IN\_B-PP, VB\_B-VP, DT\_B-NP$	para desarrollar otra
$DT\_B-NP, NN\_I-NP, JJ\_I-NP, IN\_B-PP$	un talento natural por
$IN\_B-PP, VB\_B-VP, NNS\_B-NP$	a plantear cuestiones
IN_B-PP, VB_B-VP, CC_O	de reconocer y,de actuar y
$JJ\_I-NP, JJ\_I-NP, JJ\_I-NP$	matemática corporal musical
CC_O, VB_B-VP, NNS_B-NP	e introducir conceptos
$IN\_B-PP, NN\_B-NP, VB\_B-VP$	de pensamiento identificar
$JJ\_I-NP, NN\_I-NP, NN\_I-NP$	artísticas armando rompecabezas
Estilos de aprendizaje	
$DT\_B-NP, NN\_I-NP, IN\_B-PP$	la información entre,el aprendizaje en
$IN\_B-PP, NN\_B-NP, IN\_B-PP$	como construcción de, De acuerdo a
$DT\_B-NP, NN\_I-NP, VB\_B-VP$	todo estudiante debe, la experiencia es
$IN\_B-PP, VB\_B-VP, DT\_B-NP$	pues define la, que existe una
$PRP\_B-NP, VB\_B-VP, IN\_B-PP$	Se caracteriza por, la toma de
$IN\_B-PP, NN\_B-NP, VB\_B-VP$	de aprendizaje aporta
$DT\_B-NP, NN\_I-NP, JJ\_I-NP, IN\_B-PP$	la relación existente entre
VB_B-VP, VB_B-VP, CC_O	perciben analizan y, aprende conoce y
$IN\_B-PP, VB\_B-VP, NNS\_B-NP$	a establecer relaciones
$DT\_B-NP, NN\_I-NP, NN\_I-NP, IN\_B-PP$	la formación pregrado como
$VB\_B-VP, VB\_B-VP, VB\_B-VP$	experimentar reflexionar elaborar

En estas frases se observa una mayor descripción de los dominios, sobre todo en la clase de tipos de inteligencias en donde los patrones encontrados mencionan las características de los distintos tipos de inteligencias: *talento natural, planear cuestiones, reconocer y actuar, introducir conceptos* son frases recuperadas que al momento se extraer las palabras adyacentes. En el dominio de estilos de aprendizaje, aunque no se recuperan suficientes patrones referentes a la clasificación, si se observan aquellos que describen las características de este concepto: *la experiencia es, se caracteriza por, relación existente entre, experimentar reflexionar y elaborar* son frases que describen teóricamente a los estilos de aprendizaje. Con eso se aprecia que

aunque el conjunto de entrenamiento sea de un tópico distinto (tipos de inteligencias) si se logran obtener patrones representativos en otro tema similar (estilos de aprendizaje).

Analizando la frecuencia y la pertinencia de los patrones extraídos por ambos métodos, se proponen algunos patrones generales para la detección de textos en el dominio pedagógico:

- $DT\_B-NP, NN\_I-NP, VB\_B-VP$: Inicia la descripción de un concepto principal.
- $IN\_B-PP, VB\_B-VP, NNS\_B-NP$: Describe las características de un concepto.
- $JJ\_I-NP, JJ\_I-NP, JJ\_I-NP$: Listado de subclases.
- $JJ\_I-NP, NN\_I-NP, NN\_I-NP$: Características de personas en una subclase específica.
- $IN\_B-PP, NN\_B-NP, VB\_B-VP$: Descripción de conceptos.

5.5. Análisis de matriz de similitud

Para este experimento se utilizó la matriz de similitud explicada en el capítulo de metodología, dicha matriz de similitud se generó con el promedio de la similitud coseno y el coeficiente de traslape. El objetivo de este experimento no solo es detectar conceptos importantes, sino las posibles relaciones entre estos.

Como entrada, este proceso admite el número de iteraciones ($iter$) que se realizarán, y un valor de γ, el cual determinará el tope utilizado para considerar a dos conceptos como relacionados. La primera iteración se realiza con el nombre de la clase, y a partir de ahí, los conceptos recuperados se transforman en las entradas de la siguiente iteración. Al igual que en el resto de las propuestas, se realizaron varios experimentos con diferentes valores para γ e $iter$.

A medida que γ se incrementa, el número de conceptos recuperados disminuye, por lo que es necesario determinar un valor óptimo, este valor debe tomar en cuenta tanto los elementos recuperados como el *gold* diseñado para la evaluación de experimentos. La Tabla 5.17 muestra los conceptos recuperados de acuerdo a la clase, valor de γ y representación de los textos (5-gramas y oraciones).

La tabla solo muestra los valores del 5 % al 50 %, ya que en este rango se encuentran los valores más altos en cuanto a conceptos recuperados. Analizando el número de elementos recuperados, la representación en oraciones tiene valores más altos y la clase de estilos de aprendizaje tiene un mayor número de conceptos correctos recuperados. En la tabla solo se muestran los resultados con tres iteraciones dentro de la matriz principal, aunque se hicieron experimentos con 3, 5 y 10 iteraciones. Los resultados con 5 y 10 iteraciones no difieren en gran medida con los mostrados en la tabla.

Tabla 5.17: Número de conceptos recuperados por experimento de acuerdo a la clase y representación.

Clase	γ	Oraciones		5Gramas	
		Recuperados	Correctos	Recuperados	Correctos
Estilos de aprendizaje	0.05	924	112	924	112
	0.10	924	112	629	94
	0.15	924	112	210	44
	0.20	922	112	53	18
	0.25	912	111	28	11
	0.30	797	105	8	2
	0.35	517	86	5	2
	0.40	243	51	3	1
	0.45	92	27	1	1
	0.50	41	19	1	1
Tipos de inteligencias	0.05	1,073	97	1,057	97
	0.10	1,073	97	372	51
	0.15	1,070	97	68	21
	0.20	1,042	96	28	10
	0.25	869	90	11	4
	0.30	483	61	6	3
	0.35	192	35	3	2
	0.40	39	19	3	2
	0.45	17	10	1	1
	0.50	6	5	1	1
Estrategias de aprendizaje	0.05	1,053	71	1,035	71
	0.10	1,053	71	294	37
	0.15	1,051	71	79	17
	0.20	1,019	70	30	7
	0.25	796	66	8	2
	0.30	408	50	2	1
	0.35	134	19	1	0
	0.40	43	12	1	0
	0.45	12	3	1	0
	0.50	6	1	1	0

La Figura 5.7 muestra la relación entre precisión y recuerdo de cada experimento para $iter = 3$. Se presenta un comportamiento similar al de el experimento con una sola iteración, pero al ser un proceso recursivo, el valor de γ es más alto, al igual que la precisión. Analizando las clases, los estilos de aprendizaje siguen teniendo un mejor comportamiento que las otras dos clases.

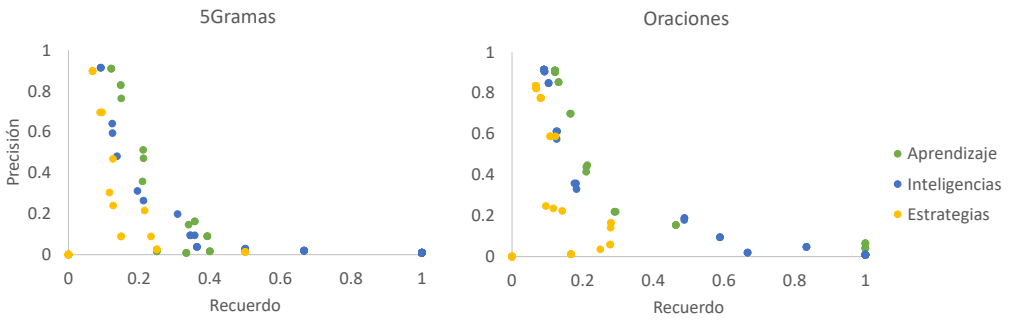


Figura 5.7: Relación entre la precisión y el recuerdo de los experimentos basados en la matriz de similitud

Para describir el proceso, se muestran los resultados obtenidos en el experimento de estilos

de aprendizaje, con un $\gamma = 0{,}45$. En la iteración uno, solo entra al sistema la palabra *estilo* y se recuperan todas aquellas que tengan una métrica mayor a 0.45. Entre algunos de los pares recuperados se encuentran los siguientes:

- estilo pragmático: 0.618537577
- estilo reflexivo: 0.615692804
- estilo activo: 0.612243827
- estilo teórico: 0.604470391
- estilo preferencia: 0.568669998
- estilo estudiante: 0.551556301
- estilo honey: 0.54218769
- estilo estrategia: 0.541823132
- estilo chaea: 0.506143719

De los resultados, los cuatro más altos corresponden a los estilos de aprendizaje, y según la hipótesis inicial, estas serían las relaciones taxonómicas de la ontología. Los siguientes resultados, corresponderían a relaciones no taxonómicas entre estos conceptos (esto se debe de tomar en cuanta al momento de realizar el código XML para la visualización en *Protégé*). En la segunda iteración, las entradas del sistema son las salidas de la primera (*pragmático,reflexivo,activo,preferencia, etc.*)

- estudiante cursar 0.502870331
- estudiante docente 0.481137678
- estudiante argentino 0.475225248
- estudiante posgrado 0.467618153
- honey lsq 0.487503323
- chaea instrumento 0.497060752

En la segunda iteración, la mayoría de las relaciones recuperadas son no taxonómicas pero aún existe esa relación entre conceptos, por ejemplo chaea e instrumento, y honey y lsq (Honey reestructuró el cuestionario LSQ). Finalmente, en la última iteración, hay algunas relaciones formales, pero sd introducen más casos negativos en la lista, y no todas las palabras tienen valores de γ tan altos como para llegar a esta. Es por eso que los resultados de 5 y 10 iteraciones no cambian demasiado, al no tener resultados más altos que γ.

- octavo semestre 0.497638114
- universidad simón 0.587322805
- universidad ríos 0.579143013
- universidad nacional 0.53977489
- universidad bolívar 0.494816676
- plazo largo 0.584559237

5.6. Ontologías manuales

A fin de tener un referente de comparación para los experimentos, se realizó una ontología de manera manual. Siguiendo la metodología de Noy & Mcguinness, para el diseño se dividen los siete pasos generales en dos secciones: el análisis de los elementos teóricos (pasos 1, 2 y 3) y la estructuración final (pasos 4 a 7). La primera parte se analiza conjuntamente, mientras que la segunda parte es presentada clase por clase.

5.6.1. Análisis de los elementos teóricos

Como primer paso, se determina el dominio y alcance de cada una de las ontologías. En la tabla 5.18 se muestran los datos generales. En la tabla se mencionan los enfoques teóricos para cada una de las clases, básicamente, los autores de las teorías que se analizan.

Tabla 5.18: Dominio y alcance de las ontologías diseñadas.

Nombre de la ontología	Estilos de aprendizaje	Tipos de inteligencias	Estrategias de enseñanza
Enfoque teórico	Peter Honey y Catalina Alonso	Howard Gardner	Frida Díaz-Barriga. Gerardo Hernández, María Gonzalez, Javier Turrón
Propósito	Representar y formalizar el conocimiento del dominio.		
Alcance	Apoyo para lograr el aprendizaje significativo por medio de la personalización del mismo.		
Usuarios finales	Docentes de clases presenciales, preferentemente nivel medio superior y superior.		
Fuentes de conocimiento	Expertos en el dominio y artículos especializados sobre el tema.		
Reutilización de ontologías	No se tomó ninguna ontología creada, pero se tomaron en cuenta las investigaciones de Silva y Ponce Silva Sprock & Ponce (2013)	No se utilizan otros archivos OWL o estudios base de otros autores.	

El propósito, alcance, usuarios finales y fuentes de conocimiento son los mismos para las tres: Se busca formalizar el conocimiento a fon de apoyar a los alumnos para que logren un aprendizaje significativo. Para esto, se utilizan artículos especializados en el tema, aunados con el análisis de expertos en el dominio. Se espera que los usuarios finales sean docentes de clases presenciales de nivel superior y medio superior.

El segundo punto consiste en determinar si se reutilizarán ontologías. Para la clase estilos de aprendizaje se mencionan los trabajos de Silva Sprock & Ponce (2013), no para reutilización del archivo fuente pero si para ayuda y consulta en cuanto a la estructuración de los conceptos. Las clases de tipos de inteligencias y estrategias de enseñanza no tienen otras ontologías o investigaciones para consulta.

Después de realizar el análisis del material recopilado y la lista de conceptos importantes previamente generada, se definen los términos candidatos a clases o subclases. A continuación se describen algunos de los conceptos de cada clase, en caso de que tengan sinónimos, estos se agregan entre paréntesis.

Estilos de aprendizaje:

- **Actividad** (Tarea,Deber). Conjunto de operaciones o tareas propias de una persona o entidad.
- **Alonso**. Autor de la teoría de estilos de aprendizaje
- **Aprendizaje** (Conocimiento). Cambio en la conducta debido a la experiencia.
- **CHAEA** (CuestionarioCHAEA,CuestionarioHoneyAlonso). Cuestionario sobre estilos de aprendizaje que consta de ochenta preguntas (veinte items referentes a cada uno de los cuatro Estilos) a las que hay que responder manifestando acuerdo o desacuerdo.
- **Discente** (Alumno,Aprendiz,Alumna,Estudiante,Sujeto). Que recibe enseñanza.
- **EstiloActivo**. Estilo de aprendizaje de personas que se involucran con experiencias nuevas, tienden a actuar primero y luego piensan en las consecuencias.
- **EstiloPragmático**. Estilo de aprendizaje que incluye personas que prueban sus ideas, teorías y técnicas nuevas, tratando de ver si funcionan en la práctica.
- **EstiloReflexivo**. Estilo de aprendizaje de personas que son observadores y analizan sus experiencias desde diferentes perspectivas.
- **EstiloTeórico**. Estilo de aprendizaje que muestra dentro de las principales caracteríticas la lógica, la metódica, la objetividad y la estructuración en las acciones.
- **Honey**. Autor de la teoría estilos de aprendizaje
- **Instrumento** (Cuestionario). Aquel que plantea una serie de preguntas para extraer determinada información de un grupo de personas.
- **Item** (Pregunta,Reactivo). Distintas preguntas que se plantean en una evaluación.
- **Item1**. Item para determinar si el estudiante tiene un estilo predominante pragmático: Tengo fama de decir lo que pienso claramente y sin rodeos.

Tipos de Inteligencias:

- **CorporalKinestésica** (InteligenciaCorporalKinestésica, Cinética, CinéticaCorporal). Tipo de inteligencia en donde la persona tiene la capacidad de utilizar el propio cuerpo para realizar actividades o resolver problemas,
- **Discente** (Alumno,Aprendiz,Alumna,Estudiante,Sujeto). Persona que recibe enseñanza.
- **Espacial** (InteligenciaEspacial,Visual,VisualEspacial,InteligenciaVisual). Tipo de inteligencia que consiste en formar un modelo mental del mundo en tres dimensiones.
- **Gardner** (HowardGardner). Autor de la teoría de inteligencias múltiples
- **InteligenciaMúltiple** (Inteligencia,TipoInteligencia). Implica la habilidad necesaria para resolver un problema o para elaborar productos que son importantes en un contexto cultural.
- **Interpersonal** (InteligenciaInterpersonal). Tipo de inteligencia que consiste en entender a los demás.
- **Intrapersonal** (InteligenciaIntrapersonal). Tipo de inteligencia que consiste en entenderse a sí misma.

Estrategias de enseñanza:

- **AprendizajeEstrategico**. Aquellos procesos internos (cognitivos, motivacionales y emocionales) y conductas que promueven un aprendizaje efectivo y eficiente.
- **Cognitiva** (EstrategiaCognitiva). Hacen referencia a la integración del material nuevo con el conocimiento previo. En este sentido, serían un conjunto de estrategias que se utilizan para aprender, codificar, comprender y recordar la información al servicio de metas de aprendizaje determinadas.
- **Constructivismo**. Corriente pedagógica creada que postula la necesidad de entregar al alumno herramientas que le permitan crear sus propios procedimientos para resolver una situación problemática.
- **Deductivo**. Método por el cual se procede lógicamente de lo universal a lo particular.
- **Hipotético**. Que está basado o fundamentado en una hipótesis o en una suposición.

- **Instrumento**. Aquel que plantea una serie de preguntas para extraer determinada información de un grupo de personas.
- **Investigación**. Proceso intelectual y experimental que comprende un conjunto de métodos aplicados de modo sistemático.
- **Lluvialdeas** (Tormentaldeas). Es una herramienta de trabajo grupal que facilita el surgimiento de nuevas ideas sobre un tema o problema determinado.
- **MapaConceptual**. Representaciones gráficas de varias ideas interconectadas, que se confeccionan utilizando dos elementos: conceptos (o frases breves, cortas) y uniones o enlaces
- **Metacognitiva** (EstrategiaMetacognitiva). Hacen referencia a la planificación, control y evaluación por parte de los estudiantes de su propia cognición. Son un conjunto de estrategias que permiten el conocimiento de los procesos mentales, así como el control y regulación de los mismos con el objetivo de lograr metas de aprendizaje determinadas.

En general, la lista de estilos de aprendizaje es más extensa, ya que integra muchos de los conceptos que comparten las tres clases, por ejemplo: *teoría*, *discente*, *autor*, *aprendizaje*, *asignatura*, etc. Las clases estilos de aprendizaje y tipos de inteligencias son las que comparten más términos, esto en gran parte se debe a que están basadas en teorías ya establecidas que proponen una clasificación y un instrumento para poder detectarlo. Las estrategias de enseñanza aprendizaje aún no están definidas del todo y varios autores proponen varias clasificaciones, si bien el enfoque analizado es de los más recurrentes en la literatura aún no se establece un método generalmente aceptado como el caso de las otras dos clases.

5.6.2. Estructuración de las ontologías

Una vez seleccionados los conceptos considerados importantes, se procede a realizar las estructuras pertinentes a cada clase analizada. En las siguientes subsecciones se mencionan los elementos más significativos de cada una de ellas.

Estilos de aprendizaje

La Figura 5.8 muestra el proceso para la estructuración de la ontología estilos de aprendizaje, de acuerdo a la metodología utilizada. Se inicia a partir del punto 4, el cual consiste en definir las clases y subclases.

En este punto, la figura muestra un esquema de tres fases, primero se da un extracto de la lista de conceptos, especificando si fueron catalogados como clases o subclases. En el ejemplo se muestran las principales clases y subclases (los tipos de estilos de aprendizaje), posteriormente,éstas se estructuran en un pequeño esquema para representar la estructura taxonómica. Precisamente este tipo de estructuras son las que nos permiten visualizar adecuadamente las relaciones *Isa*. En el ejemplo, *IsA (Teórico, EstiloAprendizaje)*, y *IsA (EstiloAprendizaje, Teoría)*.

Finalmente, en este mismo punto aparte de las relaciones *IsA*, se anexan otras posibles relaciones entre diversos conceptos por ejemplo *TieneEstilo (Alumno, Pragmático)* y *DescribeEstilo (Pragmático, Item2)*. En este último ejemplo, se muestra la relación que tiene cada uno de los

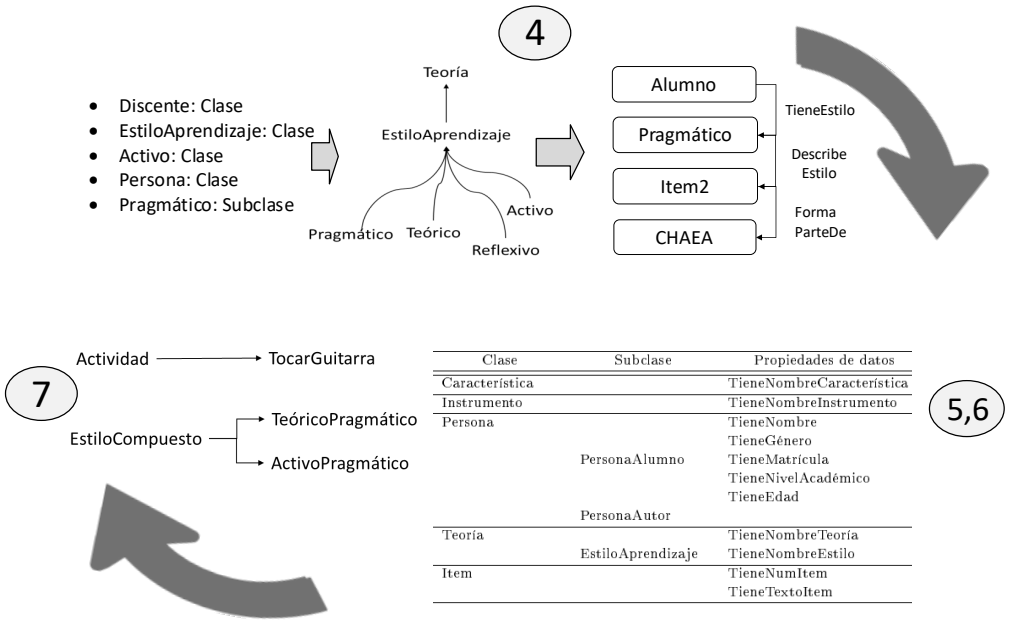


Figura 5.8: Proceso de creación de la ontología de estilos de aprendizaje

ítems del cuestionario CHEA con los estilos de aprendizaje. Al manejar una escala nominal para responder dicho cuestionario, cada pregunta va enfocada a determinar si cierto estilo de aprendizaje es el dominante en los estudiantes.

Tabla 5.19: Relaciones en ontología de estilos de aprendizaje (extracto)

Nombre	Dominio	Rango	Inversa
Nombre	Dominio	Rango	Inversa
TieneAutor	TeoríaEstiloAprendizaje	Honey	EsAutor
TieneCoautor	TeoríaEstiloAprendizaje	Alonso	EsCoautor
DescribeTeoría	CHAEA	EstilosAprendizaje	EsDescritaPorInstrumento
EsDefinidoPor	EstiloTeórico	Directo	DescribeAEstilo
TieneEstilo	Alumno	EstiloAprendizaje	DescribeEstudiante
TieneItem	CHAEA	ItemE1	FormaParteDe
DescribeEstilo	EstiloPragmático	ItemE1	SeDescribePor
DescribeEstilo	EstiloReflexivo	ItemE16	SeDescribePor
Estudia	TeoríaEstiloAprendizaje	EstiloAprendizaje	EsEstudiadoPor

La Tabla 5.19 se muestran más relaciones, especificando el dominio, rango y la relación inversa entre dichos conceptos. Las relaciones que se muestran analizan principalmente el instrumento CHAEA, las items que se relacionan con cada estilo de aprendizaje y las características presentes en los alumnos de cada estilo. También se muestra la relación *TieneAutor*, para relacionar a los creadores del enfoque estudiado. En el caso de las relaciones inversas, el rango y el dominio cambian, por ejemplo para la relación inversa *DescribeAEstilo*, un ejemplo de rango es *EstiloActivo* y el dominio *Espontáneo*.

Regresando a la Figura 5.8, los puntos 5 y 6 definen las propiedades de los datos, especificando las clases y subclases. En este fragmento se sigue trabajando en las clases relacionadas con los estilos y el análisis del cuestionario CHAEA. Finalmente, en el punto siete se dan algunos ejemplo de instancias, como el nombre de los estilos compuestos y un ejemplo de actividad (en la ontología, *Actividad* representa lo que los alumnos con cierto estilo de aprendizaje prefieren hacer.

La Figura 5.9 muestra algunas de las clases integradas, además de las estadísticas de la ontología, entre las que destacan el número de axiomas, clases, instancias y subclases.

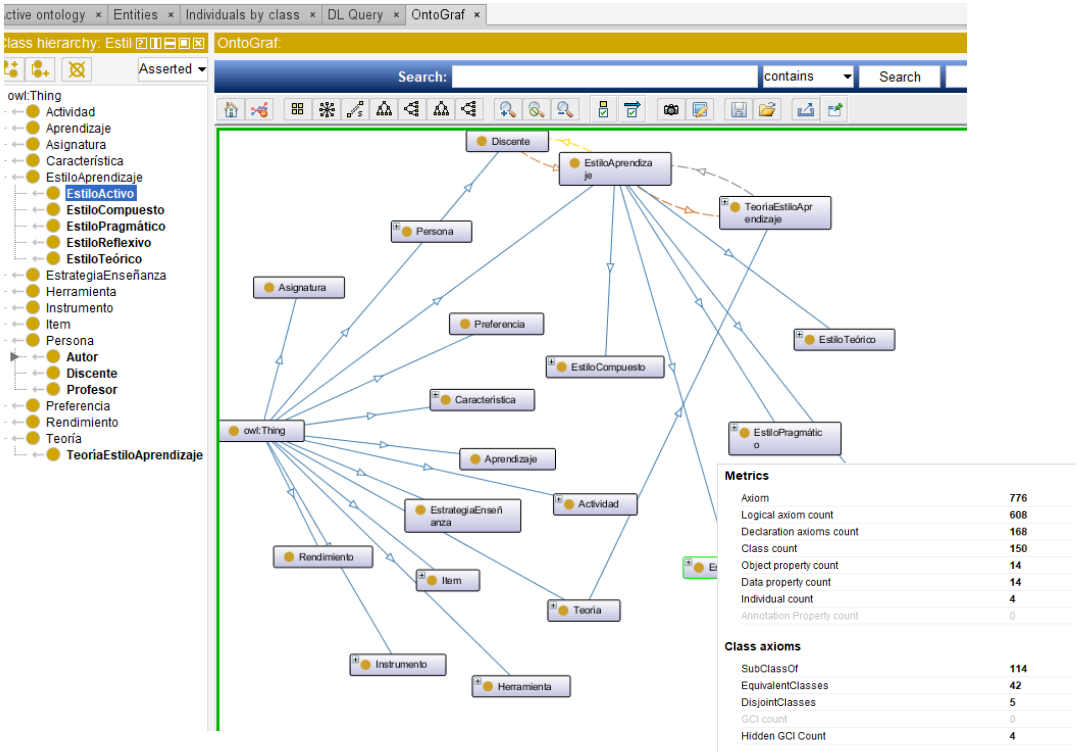
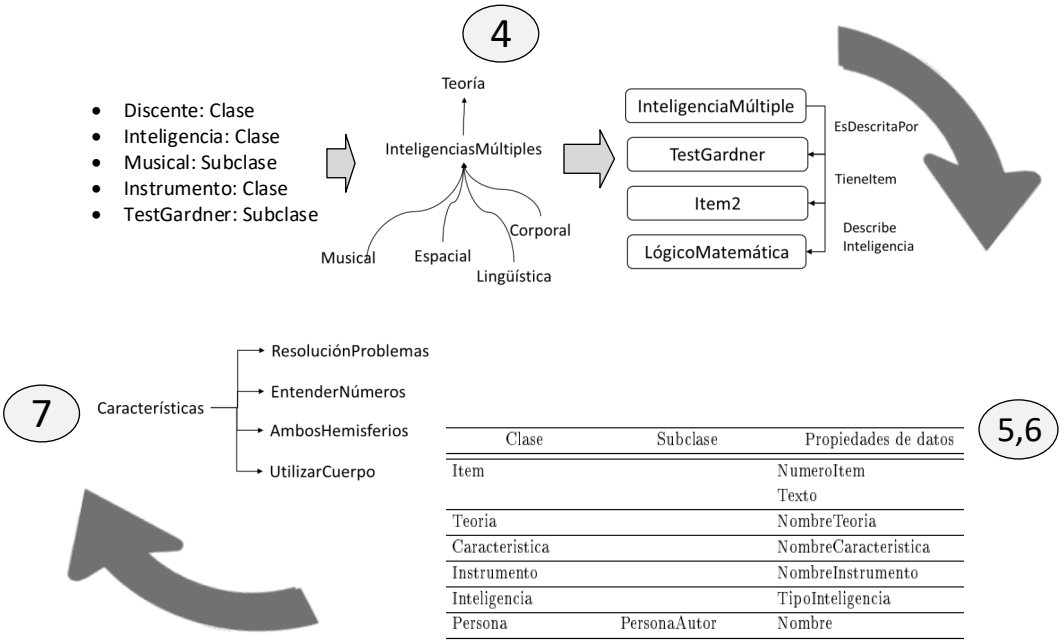


Figura 5.9: Grafo representando la ontología de estilos de aprendizaje (extracto)

Tipos de inteligencias:

Analizando los tipos de inteligencias, la Figura 5.10 muestra el resto del proceso para construcción de la ontología de este tema. En el paso correspondiente a definir las clases y jerarquías entre ellas (punto 4) se agregan tres elementos:

- ➥ Se inicia con una lista de conceptos etiquetados como clases y subclases. El dominio pedagógico contiene algunos conceptos importantes, por lo que es común que dichos conceptos se encuentren presentes en dos o las tres clases.
- ➥ Posteriormente se representa una parte de la estructura taxonómica creada. En este punto se observa la similitud de esta clase con los estilos de aprendizaje. Ambas parten de una teoría donde hay una clasificación y cada una de la subclasificaciones tiene características que la describen. Cada alumno tiene una característica predominante, pero no implica que no posea las demás en la clasificación, aunque en menor grado.
- ➥ Finalmente, se anexan otras posibles relaciones, también relacionadas con el instrumento utilizado para detectar el tipo de inteligencia en un alumno (test de Gardner). En este esquema, el item 2 está relacionado con la inteligencia lógico matemática.



Clase	Subclase	Propiedades de datos
Item		NumeroItem
		Texto
Teoria		NombreTeoria
Caracteristica		NombreCaracteristica
Instrumento		NombreInstrumento
Inteligencia		TipoInteligencia
Persona	PersonaAutor	Nombre

Figura 5.10: Proceso de creación de la ontología tipos de inteligencias

La Tabla 5.20 complementa las relaciones planteadas en la figura, se sigue la misma estructura de los estilos de aprendizaje, relacionando la teoría con los autores, los tipos de inteligencias que componen la teoría y los ítems relacionados con ellas.

En los siguientes puntos de la Figura 5.10 se definen las propiedades de las clases. En el extracto se agregan características como nombre de la teoría, del instrumento y una descripción de los tipos de inteligencia analizada. Finalmente, en el último punto se agregan instancias a las clases

Tabla 5.20: Relaciones entre conceptos en la ontología de tipos de inteligencias (extracto)

Nombre	Dominio	Rango	Inversa
TieneAutor	TeoríaInteligenciaMúltiple	Gardner	EsAutor
DescribeTeoría	TestGarder	InteligenciaMúltiple	EsDescritaPorInstrumento
TieneItem	TestGarder	ItemI1	FormaParteDe
TieneItem	TestGarder	ItemI2	FormaParteDe
DescribeInteligencia	ItemI2	LógicoMatemática	SeDescribePor
Estudia	TeoríaGardner	InteligenciaMúltiple	EsEstudiadoPor
TieneInteligencia	Alumno	InteligenciaMúltiple	EstáPresenteEn

y subclases, en el ejemplo, se muestran características propias de los estudiantes, los cuáles hacen que se desarrolle un tipo de inteligencia en particular.

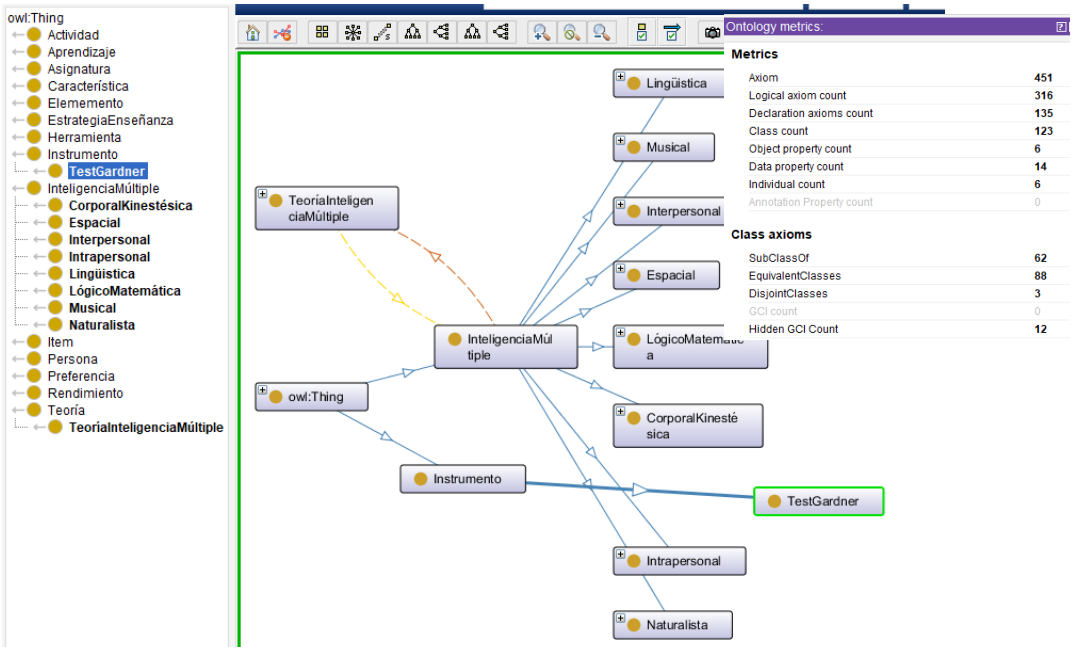


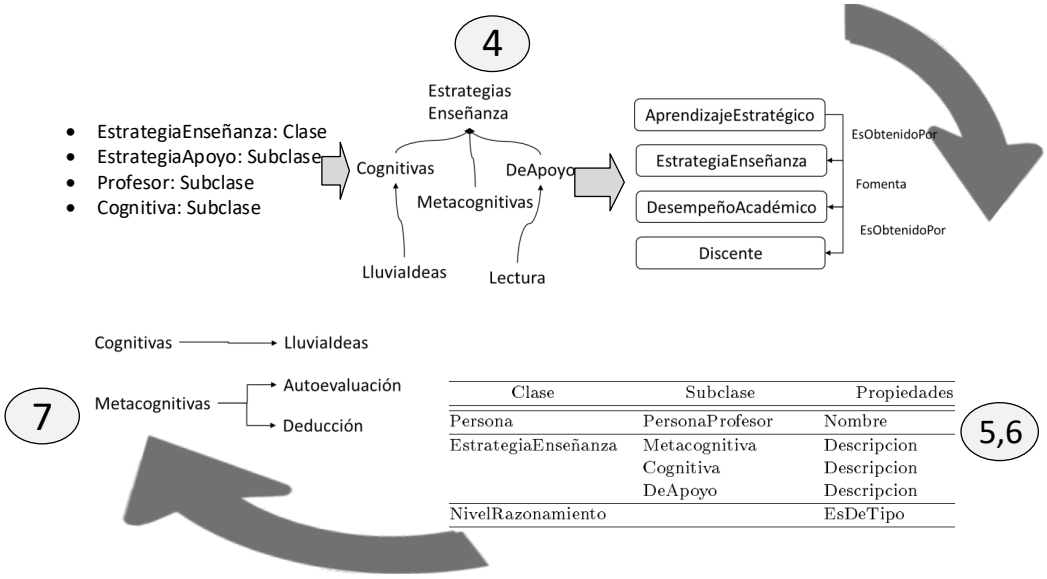
Figura 5.11: Grafo representando la ontología de estilos tipos de inteligencias (extracto)

La Figura 5.11 muestra el la ontología diseñada en *Protégé*, además de la lista de clases y las estadísticas de dicha ontología. En general, tiene menos elementos que los estilos de aprendizaje, ya que en los estudios de Gardner, no se toman en cuenta tantas características como en el cuestionario CHAEA.

Estrategias de enseñanza:

La Figura 5.12 muestra el resto del proceso para las estrategias de enseñanza. Como se comentó en otras secciones en esta clase los conceptos utilizados y en general el fundamento teórico aún no está tan definido como en las otras dos por lo que al momento de relacionarlos y estructurar

la ontología, las clases y subclases encontradas son menos.



Clase	Subclase	Propiedades
Persona	PersonaProfesor	Nombre
EstrategiaEnseñanza	Metacognitiva	Descripcion
	Cognitiva	Descripcion
	DeApoyo	Descripcion
NivelRazonamiento		EsDeTipo

Figura 5.12: Proceso de creación de la ontología estrategias de enseñanza aprendizaje

En el punto 4 se estructura la taxonomía de las clases, especificando primer un extracto de clases y subclases, las relaciones *IsA* y otras relaciones entre conceptos. En este punto se menciona la clasificación de las estrategias de enseñanza propuesta por varios autores: Cognitivas, metacognitivas y de apoyo, además de un ejemplo en el caso de dos de ellas. Al relacionar los conceptos, en este caso el análisis se enfoca en el impacto de las estrategias de enseñanza en el estudiante (discente). La Tabla 5.21 muestra un extracto un poco más extenso de estas relaciones.

Tabla 5.21: Relaciones entre conceptos en la ontología de la clase de estrategias de enseñanza (extracto)

Nombre	Dominio	Rango	Inversa
Tiene	Discente	PerfilCognitivo	DescribeA
Desarrolla	EstrategiaEnseñanza	AprendizajeEstrategico	EsObtenidoPor
Implementa	Profesor	EstrategiaEnseñanza	EsImplementadaPor
AprendeCon	Discente	EstrategiaEnseñanza	SeRealizaPor
Fomenta	EstrategiaEnseñanza	DesempeñoAcademico	EsFomentadoPor
Obtiene	Discente	DesempeñoAcademico	EsObtenidoPor
Diseña	Profesor	Planeación	EsDiseñadaPor
InvolucraActividad	Cognitiva	CentrarAtención	EsParteDeEstrategias
InvolucraActividad	Cognitiva	RecogerInformación	EsParteDeEstrategias
InvolucraActividad	Metacognitiva	Planificar	EsParteDeEstrategias
InvolucraActividad	Metacognitiva	Supervizar	EsParteDeEstrategias

La estructura presentada es distinta a la de las otras dos clases, aquí se presentan más

relaciones enfocadas a la estructura de las estrategias, no a su clasificación. Por ejemplo, se mencionan los elementos de una estrategia de enseñanza, las actividades que involucran. Además, ya se tienen relaciones entre la clase *Docente* con el alumno y las estrategias diseñadas.

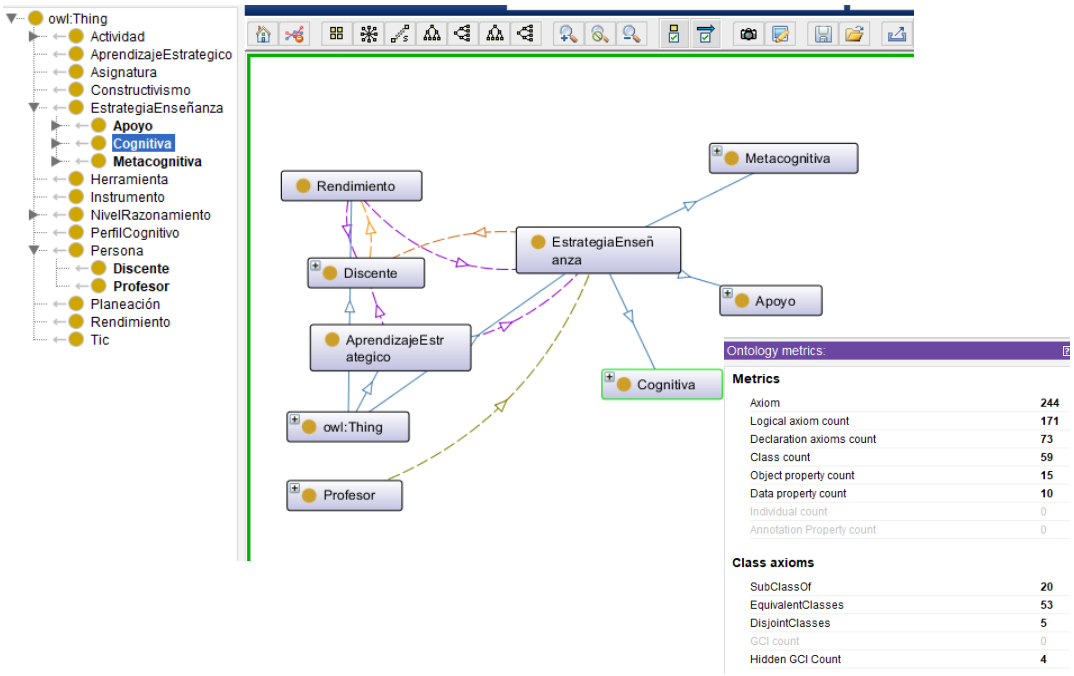


Figura 5.13: Grafo representando la ontología de estrategias de enseñanza (extracto)

La Figura 5.13 muestra las estadísticas de la ontología, así como la lista de clases y una vista del grafo generado. Respecto al número de clases utilizadas, es todavía menor a los otros dos temas, pero esto es lógico si se compara con la lista inicial de conceptos elaborada con la ayuda de expertos en el dominio.

La tabla 5.22 muestra una recopilación de datos relacionados con las ontologías y el conjunto de validación creado. Al final, se anexa el porcentaje de conceptos abarcados en el diseño. Los tipos de inteligencias son los que tienen un porcentaje mayor, con un 95 %, mientras que las estrategias de enseñanza abarcan solo un 68 %.

Otros datos que se integran en la tabla es el total de subclases (relación *IsA*), clases equivalentes (sinónimos) y clases disjuntas (clases en el mismo nivel taxonónico de la ontología). A nivel general, se observa una cobertura extensa respecto al conjunto de validación, además de que teóricamente se cubren los aspectos más importantes de acuerdo a un análisis cualitativo.

Tabla 5.22: Comparación entre métricas de las ontologías y total de elementos del conjunto de validación

	Estilos de aprendizaje	Tipos de inteligencias	Estrategias de enseñanza
Axiomas	776	451	244
Clases	150	123	59
Subclases	114	62	20
Clases equivalentes	42	88	53
Clases disjuntas	5	3	5
Conjunto de validación	184	130	87
Cobertura	82 %	95 %	68 %

Capítulo 6

Reflexiones finales

En esta capítulo se presentan las conclusiones finales acerca de los distintos experimentos para la creación de ontologías de manera semiautomática, así como las contribuciones hechas en el área y posibles trabajos futuros una vez concluida esta tesis.

6.1. Conclusiones generales

En esta investigación se cumplió el objetivo planteado en el capítulo uno, el cual consiste en desarrollar una metodología semiautomática para la creación de ontologías en el dominio pedagógico. En general, se hicieron experimentos para cada una de las fases del proceso de creación de ontologías, utilizando herramientas y procedimientos propios del área de procesamiento de lenguaje natural. Dentro de los recursos que más se utilizaron fueron el lematizador y listas de palabras alusivas al subdominio analizado, los procesos predominantes en el desarrollo de dicha investigación fueron el análisis de métricas se similitud textual.

Creación de un corpus.

En esta fase de la metodología se construye un conjunto de análisis para la extracción de los datos de la ontología. Al no haber recursos disponibles en la literatura, surgió la necesidad de recopilar fuentes de información sobre los temas analizados. Se decidió trabajar con artículos en español publicados en revistas con un enfoque pedagógico o psicológico.

Otra actividad importante en esta etapa fue la elección de los tópicos a analizar. Se decidió trabajar con el aprendizaje significativo, por lo que se seleccionaron tres conceptos o clases importantes para ese tópico. Se propusieron dos métodos en esta etapa, la creación de un corpus inicial de manera manual y el enriquecimiento de manera automática. El reto en esta etapa fue discriminar automáticamente los elementos recuperados, ya que las instancias (artículos) aparte de hablar de las clases analizadas, deberían hablar bajo el mismo enfoque teórico utilizado en el corpus generado manualmente.

También se trabajó en la creación de un conjunto de validación (*gold estándard*), el cual integra conceptos considerados importantes para un experto en el dominio, además de tomar en cuenta los sinónimos de dichos conceptos. Finalmente, en la realización de esta investigación se publicaron varios artículos, los cuales se listan en el apéndice E

Detección de conceptos principales

En esta etapa se incluye el preprocesamiento del corpus, el cual se integró de actividades básicas en tareas del idioma español:

- Eliminar las palabras cerradas
- Sustituir los conceptos por su lema
- Eliminar símbolos no correspondientes al idioma
- Eliminar el *abstract* y las referencias de los artículos que los contengan
- Reemplazar los sinónimos por un solo conceptos
- Generar varias versiones del corpus con ngramas de distinto tamaño

Para la detección de conceptos importantes se utilizaron métricas de similitud textual en varios experimentos, en los cuales se buscó la combinación de parámetros más efectiva. Los experimentos recuperaron varios conceptos a medida que se cambiaba la métrica mínima para considerarse como importante. La precisión en estos experimentos fue alta, pero con un número muy pequeño de conceptos recuperados, por otro lado, un recuerdo alto implica que se recuperan muchos elementos que no son considerados como importantes.

Extracción de relaciones

En esta etapa se trabaja con varias propuestas para determinar si dos conceptos están relacionados entre sí. En esta parte es donde se proponen más métodos, involucrando el análisis de patrones de categorías gramaticales, detección de conceptos definitorios y métricas de similitud textual. Se detectan relaciones que están integradas en las ontologías manuales de cada clase principal y otras que aunque no fueron tomadas en cuenta, son teóricamente correctas.

6.2. Contribuciones y trabajo en progreso

Dadas las características de la investigación y los objetivos propuestos, fue necesario trabajar más la parte de la recolección de información, sobre todo en los siguientes aspectos:

- ➥ Elección de clases principales: Al determinar el objetivo de diseñar una herramienta de apoyo para el aprendizaje significativo, se trabajó en la elección de tópicos que se relacionaran teóricamente con este objetivo.
- ➥ Al trabajar en el idioma español, se limita el uso de herramientas semánticas, por lo que se tienen que diseñar recursos que ayuden a facilitar el procesamiento de los conceptos.
- ➥ Se analiza el dominio pedagógico, pero no todo, solo lo relacionado al aprendizaje significativo. Esto trajo la necesidad de generar procesos que permitan analizar texto con un vocabulario muy específico, lo que limita aún mas los recursos externos que se pueden utilizar.

Por lo tanto, un aspecto importante de esta investigación es la transversalidad, y aunque las aportaciones van encaminadas principalmente al área de ciencias computacionales, también se tienen importantes avances en la pedagogía. La Tabla 6.1 muestra algunos de los recursos léxicos construidos para esta investigación.

Tabla 6.1: Recursos generados a lo largo de la investigación

Método	Parámetros	Escalabilidad
Gold de términos	Lista de términos relacionados con las superclases, integran sinónimos y cuenta con dos versiones: conceptos compuestos y simples	295 elementos compuestos y 264 simples
Corpus Wiki	Artículos aleatorios de Wikipedia lematizados, los cuales fueron extraídos con python	174,605 instancias
Corpus Libros	Libros, compilaciones, antologías, diccionarios y enciclopedias de ediversos temas de pedagogía (corpus lematizado)	113 libros
Gold de artículos	Artículos por superclase etiquetados manualmente con “SI” o “NO”, dependiendo se pertenecen o no al enfoque teórico del corpus inicial	166 instancias
Corpus extendido	Artículos del corpus inicial, unidos con los artículos etiquetados como “SI” en los experimentos	Tres versiones: 125, 362 y 207 instancias
Ontologías manuales	Ontologías sobre tipos de inteligencias, estrategias de enseñanza y estilos de aprendizaje	3 ontologías

La tabla muestra un *gold*, ontologías manuales y varias versiones de corpus. Los recursos básicos fueron el corpus extendido y el *gold* de términos, con los cuales se realizaron los experimentos y su respectiva evaluación. Los restantes corpus son auxiliares para la obtención de métricas en las que se requiere una gran cantidad de texto (por ejemplo PMI). Estos recursos

pueden ser auxiliares en otras investigaciones de PLN que tomen como enfoque el dominio pedagógico.

Se aportan distintas visiones de métricas y procesos para la detección de conceptos principales. Esta actividad no solo forma parte de la construcción de ontologías, sino en otras tareas importantes en el área, como la creación de resúmenes y filtrado de documentos. Otra aportación importante es que al ser un método semiautomático, la metodología se puede ver como un proceso iterativo, que se puede aplicar a otros subdominios del área, cambiando una o varias de las clases principales.
Como trabajo en progreso, se está trabajando en las siguientes actividades:

- Experimentar con otros subdominios de la pedagogía, a fin de seguir analizando elementos que pueden tener impacto en el aprendizaje significativo.
- Expandir las técnicas utilizadas para la detección de relaciones, a fin de formalizar los experimentos ya realizados.
- Enriquecer las ontologías creadas con un proceso de poblado, aplicado a un caso específico.
- Aplicar las ontologías creadas a investigaciones pedagógicas enfocadas en el aprendizaje significativo.

Apéndices

Apéndice A

Artículos del corpus inicial

A continuación se listan los 51 artículos que componen la primera versión del corpus:

1. Macías, María (2002). Las múltiples inteligencias. Psicología desde el Caribe.
2. López, Consuelo (2003). Evaluación de los estilos de aprendizaje en estudiantes de enfermería mediante el cuestionario CHAEA. Enfermería global: Revista electrónica semestral de enfermería.
3. Vargas, Ana (2004). Antes y después de las inteligencias múltiples. Revista electrónica Educare.
4. Guzman, Belkys (2005). Las inteligencias múltiples en el aula de clases. Revista de investigación.
5. García, Hécmy (2007). Variables académicas y estilos de aprendizaje en estudiantes del ciclo de iniciación universitaria. Revista de educación.
6. Enríquez, Álvaro (2007). estrategias de aprendizaje para la empleabilidad en el mercado del trabajo de profesionales recién egresados. Universitas Psychologica.
7. Santiago, Álvaro (2007). Estrategias y enseñanza-aprendizaje de la lectura. Folios.
8. García, José (2008). Análisis de datos obtenidos a través del cuestionario CHAEA en línea de la página web www.estilosdeaprendizaje.es. Revista de estilos de aprendizaje.
9. Lamas, Héctor (2008). Aprendizaje autorregulado, motivación y rendimiento académico. Liberabit.
10. Paniagua, Liziano (2008). La teoría de las inteligencias múltiples en la práctica docente en educación preescolar. Revista Electrónica Educare. "
11. Herrera, Lucía (2009). Estrategias de aprendizaje en estudiantes universitarios. Un aporte a la construcción del Espacio Europeo de Educación Superior. Educación y Educadores."
12. Klimenko, Olena (2009). Aprender cómo aprendo: la enseñanza de estrategias metacognitivas. Educación y Educadores.
13. Esguerra, Gustavo (2010). Estilos de aprendizaje y rendimiento académico en estudiantes de Psicología. Diversitas: Perspectivas en Psicología.
14. Juárez, Jaqueline (2010). Inteligencias Múltiples: Una innovación pedagógica para potenciar el proceso enseñanza aprendizaje. Investigación y Posgrado.
15. Escurra, Luis (2011). Análisis psicométrico del Cuestionario de Honey y Alonso de Estilos de Aprendizaje (Chaea) con los modelos de la Teoría Clásica de los Tests y de Rasch. Persona: Revista de la Facultad de Psicología,.
16. Villamizar, Gustavo (2011). Estilos de aprendizaje y rendimiento académico en estudiantes de ingeniería civil. Informes Psicológicos.
17. Juárez, Carlos (2012). El cuestionario de estilos de aprendizaje CHAEA y la escala de estrategias de aprendizaje acra como herramienta potencial para la tutoría académica. Revista de estilos de aprendizaje.

18. Morales, Alejandra (2012). estilos de aprendizaje en estudiantes universitarios de ingeniería en computación e informática administrativa. Revista de estilos de aprendizaje.
19. Valencia, Nilson (2012). Procesos cognitivos y metacognitivos en la solución de problemas de movimiento de figuras en el plano a través de ambientes computacionales. Tecné, Episteme y Didaxis: TED.
20. Inciarte, Nerylena (2012). Inteligencias múltiples en la formación de investigadores. Multiciencias.
21. Freiberg, Agustín (2013). Cuestionario Honey-Alonso de estilos de aprendizaje: Análisis de sus propiedades Psicométricas en Estudiantes Universitarios. SUMMA psicolÓgica UST.
22. Muñetón, Bahamón (2013). Estilos y estrategias de aprendizaje relacionadas con el logro académico en estudiantes universitarios. Pensamiento Psicológico.
23. Mayora, Isamar (2013). Estrategias Metacognitivas aplicadas en la comprensión de la lectura por estudiantes de Inglés I. Caso Vice. Revista de Investigación.
24. Cala, Ramón (2014). determinación de los estilos de aprendizaje de estudiantes de 1er curso de ing. industrial y electrónica de la universidad técnica del norte. ibarra. ecuador. Revista de estilos de aprendizaje.
25. Juarez, Carlos (2014). Propiedades psicométricas del cuestionario Honey - Alonso de estilos de aprendizaje (CHAEA) en una muestra mexicana. Revista de estilos de aprendizaje.
26. López, Adriana (2014). Estilos de aprendizaje y su transformación a los largo de la trayectoria escolar. Esseñanza e Investigación en Psicología.
27. Salas, Jorge (2014). Estilos de aprendizaje en estudiantes de la Escuela de ciencias del Movimiento Humano y Calidad de Vida, Universidad Nacional. Costa Rica. Revista Electrónica Educare.
28. Sotillo, Juan (2014). El cuestionario CHAEA-junior o cómo diagnosticar el estilo de aprendizaje en alumnos de primaria y secundaria. Revista de estilos de aprendizaje.
29. Campos, Karolina (2014). Actividades de aprendizaje y TIC: Usos entre docentes de la Educación General Básica costarricense. Aproximación diagnóstica. Revista Electrónica Educare.
30. Carrillo, Maria (2014). La teoria de las inteligencias multiples en la enseñanza de las lenguas. Contextos educativos.
31. García, María (2015). Estrategias utilizadas por estudiantes universitarios en el aprendizaje de la lengua extranjera según el género y nivel de competencia. Docencia e Investigación.
32. Sánchez, Iván (2015). Estrategias cognitivas de aprendizaje significativo en estudiantes de tres titulaciones de Ingeniería Civil. Paradigma.
33. Vázquez, Ana (2015). La metacognición: Una herramienta para promover un ambiente áulico inclusivo para estudiantes con discapacidad. Revista Electrónica Educare.
34. Malnieri, Aida (2015). Conocimientos teóricos y estrategias mtodológicas que emplean docentes de primer ciclo en la estimulación de las inteligencias múltiples. Actualidades investigativas en Educación.
35. Nadal, Blanca (2015). Las inteligencias múltiples como una estrategia didáctica para atender a la diversidad y aprovechar el potencial de todos los alumnos.. Revista nacional e internacional de educación inclusiva.
36. Ruíz, Daniela (2015). Inteligencias múltiples en alumnos de la Universidad Americana de Asunción. ACADEMO Revista de Investigación en Ciencias Sociales y Humanidades.
37. Sandoval, Aida (2015). Estimación de la inteligencia lingüística-verbal y lógico-matemática según el género y la ubicación geográfica. TELOS. Revista de Estudios Interdisciplinarios en Ciencias Sociales.
38. Campo, Kiara (2016). Metacognición, escritura y rendimiento académico en universitarios de Colombia y Francia. Avances en Psicología Latinoamericana.
39. Hernández, Jaqueline (2016). Metacognición y comprensión oral en L2. Observación de la práctica docente en nivel universitario. Revista electrónica de investigación educativa.
40. Zambrano, Carolina (2016). Autoeficacia, Prácticas de Aprendizaje Autorregulado y Docencia para fomentar el Aprendizaje Autorregulado en un Curso de Ingeniería de Software. Formación universitaria.
41. Álvarez, David (2016). Una mira al futuro ante la relación de las inteligencias múltiples y el rendimiento escolar. Una apuesta hacia nuevas metodologías docentes en la escuela del siglo XXI. Aula encuentro: Revista de investigación y comunicación de experiencias educativas.
42. Barraza, René (2016). Rendimiento académico y autopercepción de inteligencias múltiples e inteligencia emocional en universitarios de primera generación. Actualidades investigativas en Educación.
43. Cordeiro, Dayane (2016). Múltiples puertas de entrada a la mente de nuestros alumnos: las inteligencias múltiples en el aula de E/LE. Foro de Profesores de E/LE.
44. Garzón, Ana (2016). La i ntegración TIC-Inteligencias múltiples (IM): Una oportunidad de cambio en el proceso educativo. Revista de pedagogía.

45. Perozo, Carmen (2016). Teoría de inteligencias múltiples una alternativa en la didáctica de la química. Aula de encuentro: Revista de investigación y comunicación de experiencias educativas.
46. Alducin, JuanManuel (2017). Estilos de aprendizaje, variables sociodemográficas y rendimiento académico en estudiantes de Ingeniería de Edificación. Revista Electrónica Educare.
47. Gómez, Edna (2017). Estilos de aprendizaje en universitarios, modalidad educación a distancia. Revista virtual Universidad Católica de Chile.
48. Díaz, Alejandro (2017). Impacto de un entrenamiento en aprendizaje autorregulado en estudiantes universitarios. Perfiles educativos.
49. Freiberg, Agustín (2017). Estilos, Estrategias y Enfoques de Aprendizaje en Estudiantes Universitarios de Buenos Aires. Psicodebate.
50. Ventura, Ana (2017). Aprendizaje autorregulado en el nivel universitario: Un estudio situado con estudiantes de psicopedagogía de diferentes ciclos académicos. Revista Electrónica Educare.
51. Etchegaray, Maria (2017). Diseño de un recurso multimedia on line basado en Inteligencias Múltiples. Campus virtutales.

Apéndice B

Cuestionario Honey - Alonso

Instrrucciones para responder al cuestionario.

- ➥ Este cuestionario ha sido diseñado para identificar tu estilo preferido de aprender. **No** es un test de inteligencia, ni de personalidad.
- ➥ No hay límite de tiempo para contestar el cuestionario.
- ➥ No hay respuestas correctas o erróneas. Será útil en la medida que seas sincero (a) en tus respuestas.
- ➥ Si estás más de acuerdo que en desacuerdo con la sentencia, pon un signo más (+). Si por el contrario, estás más en desacuerdo que de acuerdo, pon un signo menos ().
- ➥ Por favor contesta todas las sentencias.

1. ( ) Tengo fama de decir lo que pienso claramente y sin rodeos.
2. ( ) Estoy seguro(a) de lo que es bueno y lo que es malo, lo que está bien y lo que está mal.
3. ( ) Muchas veces actúo sin mirar las consecuencias.
4. ( ) Normalmente trato de resolver los problemas metódicamente y paso a paso.
5. ( ) Creo que los formalismos coartan y limitan la actuación libre de las personas.
6. ( ) Me interesa saber cuáles son los sistemas de valores de los demás y con qué criterios actúan.
7. ( ) Pienso que el actuar intuitivamente puede ser siempre tan válido como actuar reflexivamente.
8. ( ) Creo que lo más importante es que las cosas funcionen.
9. ( ) Procuro estar al tanto de lo que ocurre aquí y ahora.

10. ( ) Disfruto cuando tengo tiempo para preparar mi trabajo y realizarlo a conciencia.
11. ( ) Estoy a gusto siguiendo un orden, en las comidas, en el estudio, haciendo ejercicio regularmente.
12. ( ) Cuando escucho una nueva idea enseguida comienzo a pensar como ponerla en práctica.
13. ( ) Prefiero las ideas originales y novedosas aunque no sean prácticas.
14. ( ) Admito y me ajusto a las normas solo si me sirven para lograr mis objetivos.
15. ( ) Normalmente encajo bien con personas reflexivas, y me cuesta sintonizar con personas demasiado espontáneas, imprevisibles.
16. ( ) Escucho con más frecuencia que lo que hablo.
17. ( ) Prefiero las cosas estructuradas a las desordenadas.
18. ( ) Cuando poseo cualquier información, trato de interpretarla bien antes de manifestar alguna conclusión.
19. ( ) Antes de hacer algo estudio con cuidado sus ventajas e inconvenientes.
20. ( ) Me crezco con el reto de hacer algo nuevo y diferente.
21. ( ) Casi siempre procuro ser coherente con mis criterios y sistemas de valores. Tengo principios y los sigo.
22. ( ) Cuando hay una discusión no me gusta ir con rodeos.
23. ( ) Me disgusta implicarme afectivamente en mi ambiente de trabajo. Prefiero mantener relaciones distantes.
24. ( ) Me gustan más las personas realistas y concretas que las teóricas.
25. ( ) Me cuesta ser creativo(a), romper estructuras.
26. ( ) Me siento a gusto con personas espontaneas y divertidas.
27. ( ) La mayoría de las veces expreso abiertamente cómo me siento.
28. ( ) Me gusta analizar y dar vueltas a las cosas.
29. ( ) Me molesta que la gente no se tome en serio las cosas.
30. ( ) Me atrae experimentar y practicar las últimas técnicas y novedades.
31. ( ) Soy cauteloso(a) a la hora de sacar conclusiones.
32. ( ) Prefiero contar con el mayor número de fuentes de información. Cuantos más datos reúna para reflexionar, mejor.
33. ( ) Tiendo a ser perfeccionista.
34. ( ) Prefiero oír las opiniones de los demás antes de exponer la mía.
35. ( ) Me gusta afrontar la vida espontáneamente y no tener que planificar todo previamente.
36. ( ) En las discusiones me gusta observar cómo actúan los demás participantes.
37. ( ) Me siento incómodo(a) con las personas calladas y demasiado analíticas.
38. ( ) Juzgo con frecuencia las ideas de los demás por su valor práctico.
39. ( ) Me agobio si me obligan a acelerar mucho el trabajo para cumplir un plazo.
40. ( ) En las reuniones apoyo las ideas prácticas y realistas.

41. ( ) Es mejor gozar del momento presente que deleitarse pensando en el paso o en el futuro.
42. ( ) Me molestan las personas que siempre desean apresurar las cosas.
43. ( ) Aporto ideas nuevas y espontáneas en los grupos de discusión.
44. ( ) Pienso que son más consistentes las decisiones fundamentadas en un minucioso análisis que las basadas en la intuición.
45. ( ) Detecto frecuentemente la inconsistencia y puntos débiles en las argumentaciones de los demás.
46. ( ) Creo que es preciso saltarse las normas muchas más veces que cumplirlas.
47. ( ) A menudo caigo en la cuenta de otras formas mejores y más prácticas de hacer las cosas.
48. ( ) En conjunto hablo más que escucho.
49. ( ) Prefiero distanciarme de los hechos y observarlos desde otras perspectivas.
50. ( ) Estoy convencido(a) que debe imponerse la lógica y el razonamiento.
51. ( ) Me gusta buscar nuevas experiencias.
52. ( ) Me gusta experimentar y aplicar las cosas.
53. ( ) Pienso que debemos llegar pronto al grano, al meollo de los temas.
54. ( ) Siempre trato de conseguir conclusiones e ideas claras.
55. ( ) Prefiero discutir cuestiones concretas y no perder el tiempo con charlas vacías.
56. ( ) Me impaciento cuando me dan explicaciones irrelevantes e incoherentes.
57. ( ) Compruebo antes si las cosas funcionan realmente.
58. ( ) Hago varios borradores antes de la redacción definitiva de un trabajo.
59. ( ) Soy consciente de que en las discusiones ayudo a mantener a los demás centrados en el tema, evitando divagaciones.
60. ( ) Observo que, con frecuencia, soy uno(a) de los(as) más objetivos(as) y desapasionados(as) en las discusiones.
61. ( ) Cuando algo va mal, le quito importancia y trato de hacerlo mejor.
62. ( ) Rechazo ideas originales y espontáneas si no las veo prácticas.
63. ( ) Me gusta sopesar diversas alternativas antes de tomar una decisión.
64. ( ) Con frecuencia miro hacia adelante para prever el futuro.
65. ( ) En los debates y discusiones prefiero desempeñar un papel secundario antes que ser el(la) líder o el(la) que más participa.
66. ( ) Me molestan las personas que no actúan con lógica.
67. ( ) Me resulta incómodo tener que planificar y prever las cosas.
68. ( ) Creo que el fin justifica los medios en muchos casos.
69. ( ) Suelo reflexionar sobre los asuntos y problemas.
70. ( ) El trabajar a conciencia me llena de satisfacción y orgullo.
71. ( ) Ante los acontecimientos trato de descubrir los principios y teorías en que se basa.
72. ( ) Con tal de conseguir el objetivo que pretendo soy capaz de herir sentimientos ajenos.

73. ( ) No me importa hacer todo lo necesario para que sea efectivo mi trabajo.
74. ( ) Con frecuencia soy una de las personas más anima las fiestas.
75. ( ) Me aburro enseguida con el trabajo metódico y minucioso.
76. ( ) La gente con frecuencia cree que soy poco sensible a sus sentimientos.
77. ( ) Suelo dejarme llevar por mis intuiciones.
78. ( ) Si trabajo en grupo procuro que se siga un método y un orden.
79. ( ) Con frecuencia me interesa averiguar lo que piensa la gente.
80. ( ) Esquivo los temas subjetivos, ambiguos y poco claros.

Para evaluación:

1. Rodea con un círculo cada uno de los números que has señalado con un signo más (+).
2. Suma el número de círculos que hay en cada columna.
3. Coloca estos totales en la gráfica (Figura B.1). Une los cuatro para formar una figura, así comprobarás cuál es tu estilo do estilos de aprendizaje preferentes.

Activo	Reflexivo	Teórico	Pragmático
3	10	2	1
5	16	4	8
7	18	6	12
9	19	11	14
13	28	15	22
20	31	17	24
26	32	21	30
27	34	23	38
35	36	25	40
37	39	29	47
41	42	33	52
43	44	45	53
46	49	50	56
48	55	54	57
51	58	60	59
61	63	64	62
67	65	66	68
74	69	71	72
75	70	78	73
77	79	80	76

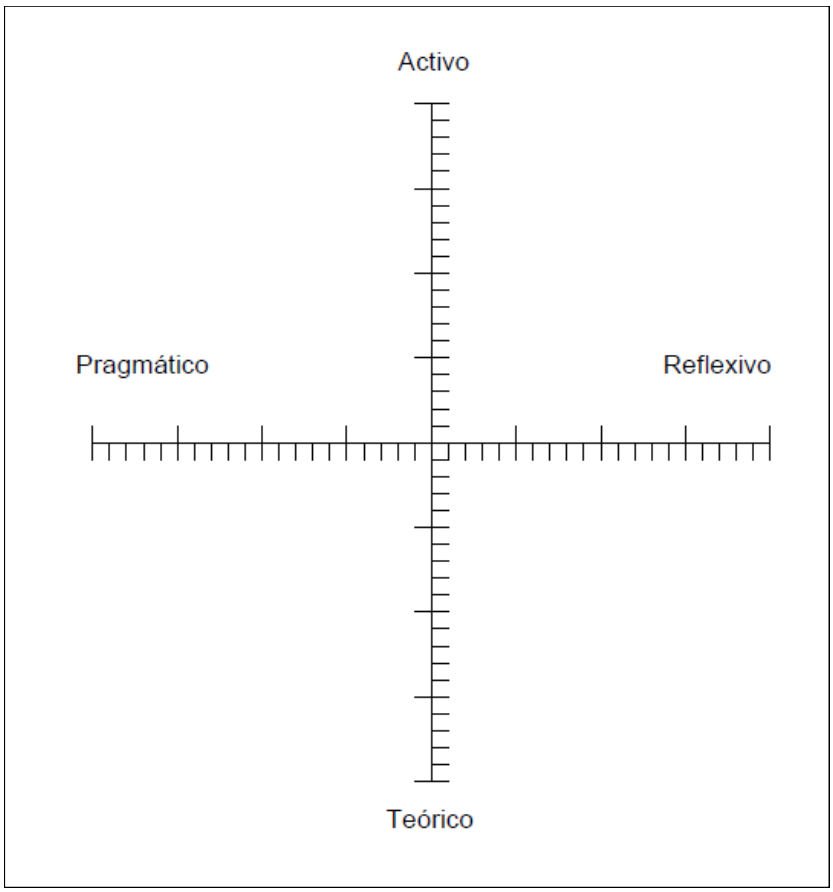


Figura B.1: Gráfica para Estilos de Aprendizaje

Apéndice C

Test de inteligencias múltiples

INSTRUCCIONES: lee cada una de las afirmaciones. Si expresan características fuertes en tu persona y te parece que la afirmación es veraz entonces coloca una V (en una hoja junto al número de la pregunta) y si no lo es, coloca una F.

1. Prefiero hacer un mapa que explicarle a alguien como tiene que llegar.
2. Si estoy enojado(a) o contento (a) generalmente sé exactamente por qué.
3. Sé tocar (o antes sabía tocar) un instrumento musical.
4. Asocio la música con mis estados de ánimo.
5. Puedo sumar o multiplicar mentalmente con mucha rapidez
6. Puedo ayudar a un amigo a manejar sus sentimientos porque yo lo pude hacer antes en relación a sentimientos parecidos.
7. Me gusta trabajar con calculadoras y computadores.
8. Aprendo rápido a bailar un ritmo nuevo.
9. No me es difícil decir lo que pienso en el curso de una discusión o debate.
10. Disfruto de una buena charla, discurso o sermón.
11. Siempre distingo el norte del sur, esté donde esté.
12. Me gusta reunir grupos de personas en una fiesta o en un evento especial.
13. La vida me parece vacía sin música.
14. Siempre entiendo los gráficos que vienen en las instrucciones de equipos o instrumentos.
15. Me gusta hacer rompecabezas y entretenerme con juegos electrónicos
16. Me fue fácil aprender a andar en bicicleta. (o patines)

17. Me enojo cuando oigo una discusión o una afirmación que parece ilógica.
18. Soy capaz de convencer a otros que sigan mis planes.
19. Tengo buen sentido de equilibrio y coordinación.
20. Con frecuencia veo configuraciones y relaciones entre números con más rapidez y facilidad que otros.
21. Me gusta construir modelos (o hacer esculturas)
22. Tengo agudeza para encontrar el significado de las palabras.
23. Puedo mirar un objeto de una manera y con la misma facilidad verlo.
24. Con frecuencia hago la conexión entre una pieza de música y algún evento de mi vida.
25. Me gusta trabajar con números y figuras
26. Me gusta sentarme silenciosamente y reflexionar sobre mis sentimientos íntimos.
27. Con solo mirar la forma de construcciones y estructuras me siento a gusto.
28. Me gusta tararear, silbar y cantar en la ducha o cuando estoy sola.
29. Soy bueno(a) para el atletismo.
30. Me gusta escribir cartas detalladas a mis amigos.
31. Generalmente me doy cuenta de la expresión que tengo en la cara
32. Me doy cuenta de las expresiones en la cara de otras personas.
33. Me mantengo .^{en} contactoçon mis estados de ánimo. No me cuesta identificarlos.
34. Me doy cuenta de los estados de ánimo de otros.
35. Me doy cuenta bastante bien de lo que otros piensan de mí.

Ahora revisa las siguientes preguntas en el orden dado: si pusiste V asignales un punto a cada una y suma los puntos

Verbal	9 -10-17-22-30
Lógico matemática	5-7-15-20-25
Visual espacial	1-11-14-23-27
Kinesésica corporal	8-16-19-21-29
Musical	3-4-13-24-28
Intrapersonal	2-6-26-31-33
Interpersonal	12-18-32-34-35

En las filas que obtengas 4, tienes la habilidad marcada y si obtienes 5 eres sobresaliente.

Apéndice D

Corpus auxiliares

Para complementar los experimentos, se crearon dos corpus externos al compuesto por artículos. El corpus $Wiki$ está constituido por entradas de Wikipedia, mientras que el corpus $Libros$ es una colección de diversos recursos como libros, antologías y diccionarios temáticos.

El corpus $Wiki$ se construyó con la ayuda de la librería $urllib$[1], la cual contiene métodos para extraer entradas aleatorias de Wikipedia. Se ejecutó dos veces el código diseñadom obteniendo 100,000 entradas por ejecución, como ámbas ejecuciones fueron independientes, se verificó que no se tuvieran entradas repetidas,eliminando una cuando se dió el caso. En promedio, cada instancia tiene alrededor de 300 palabras, pero hay algunas con solo 30. El vocabulario aunque es extenso, no se relaciona del todo con el dominio pedagógico. La siguiente lista muestra algunas de las palabras con mayor frecuencia:

- número: 191,387
- primero: 124,934
- familia: 108,728
- científico: 87954

Sin embargo, conceptos relacionados a la investigación tienen pocas apariciones en el mismo. Al momento de analizar automáticamente este corpus, estas palabras importantes para la investigación no aparecen en las salidas de los algoritmos implementados. Algunos de estos conceptos son:

[1] `https://docs.python.org/3/library/urllib.html`

- aprendizaje: 1,016
- estrategia: 1,376
- inteligencia: 1,512

El corpus denominado $Libros$ contiene recursos de acceso libre descargados de foros de instituciones educativas y repositorios de pedagogía. Dichos recursos fueron descargados en formato PDF y posteriormente pasados a texto libre para ser procesados. Al ser libros o diccionarios, cada instancia tiene más palabras, aproximadamente 9650 palabras en promedio. Además, aunque no son hablan específicamente de las clases analizadas en la investigación, tratan temas relacionados con el dominio pedagógico, por lo que los conceptos que forman el vocabulario van más relacionados a conceptos como $escuela$, $estudiante$, $aprendizaje$, $inteligencias$, entre otros. La siguiente lista contiene los nombres de los recursos que integran este corpus.

Acceso abierto al conocimiento científico
Aprendiendo a transformar el entorno de la Educación Superior
Aprendizaje activo en ambientes enriquecidos con tecnología
Autoconcepto y rendimiento escolar
Avances y desafíos en la evaluación educativa
Calidad Equidad y Reformas en la enseñanza
Ciencia Tecnología y Sociedad en Iberoamericana
Cómo podemos fomentar participación 1
Cómo podemos fomentar la participación 2
Compendio de pedagogía teórico practica
Competencias profesionales para la enseñanza aprendizaje
Década de la Educación para la Sostenibilidad
Diccionario enciclopédico de ciencias de la educación
Diseño y análisis de un sistema web educativo
Diversidad cultural e igualdad escolar
Docencia constructivista en la universidad
Educación Ciencia Tecnología y Sociedad
Educación no formal y educación pupular
Educación para el desarrollo sostenible
Educaciones y pedagogías criticas desde el sur
El aprendizaje significativo en la practica
El conocimiento libre y los recursos educativos abiertos
Entornos de Aprendizaje Claves para el ecosistema educativo en red
Estilos de aprendizaje de docentes y alumnos
Estrategias de enseñanza aprendizaje en distintos niveles
Estrategias de enseñanza aprendizaje y su importancia en el entorno educativo
Estrategias docentes para un aprendizaje significativo
Fundamentos de educación a distancia
Guía para la integración alumnado con TEA
Hacia la sociedad del conocimiento
Inteligencias múltiples: la teoría en la practica
Innovación Docente Universitaria en entornos de aprendizaje
Innovaciones y educación para la paz
Inteligencias múltiples de Gadner
Introducción a la educación interactiva
Investigación de los saberes pedagógicos
La enseñanza de las ciencias naturales
Las TICs y la Crisis de la Educación
Lineas generales de pedagogía comparada
Los desafíos de las TICS para el cambio educativo
Los estilos de aprendizaje en la enseñanza
Manual de teorías emocionales y motivacionales
Manual Formación Competencias informáticas e informacionales
Nuevos retos de la profesión docente
Pedagogía ambiental para el planeta
A educación o educación
Acceso al conocimiento
Animación a la lectura y TICs
Aprender la condición humana
Aprendizaje activo
Aprendizaje y conducta
Cartas a quien pretende enseñar
Contribuciones para la pedagogía
Diccionario Akal de Filosofía
Diccionario de la filosofía tomo1
Diccionario de la filosofía tomo2
Diccionario pedagógico
Didáctica de la pedagogía
Educación 2.0
Educación alternativa
Educación expandida
Educación y democracia
Educación y sociedad volumen1
Educación y sociedad volumen2
Educación y tecnologías
El museo y la escuela
El valor de educar
Enseñanza y aprendizaje
Enseñanza y educación
Estilos de aprendizaje
Estructuras de la mente
Experiencias educativas en el sigo XXI
Generaciones y tecnologías
Guía de apoyo en las escuelas
Investigación educativa
La educación ayer Hoy y Mañana
La escuela y el maestro
La vida en las escuelas
Las tic del aula agenda política
Las TIC en Educación CLM
Las TIC en la formación docente
Las TICs en la educación
Literatura Argentina para niños
Manual de estilos de aprendizaje
Manual de estrategias didácticas
Manual de pedagogía teatral
Modelo andragógico fundamentos
Organización de centro escolar
Pedagogía aplicada a la conducción
Pedagogía de la esperanza

Pedagogía hospitalaria y de la salud
Pedagogía tradicional y pedagogía critica
Pedagogía universitaria en América Latina
M-learning en España Portugal y América Latina
Tecnologías de la información en educación superior
Tendencias emergentes en educación con TIC
Universidad e investigación científica
Uso Inteligente de las Nuevas Tecnologías
Uso Inteligente de las Nuevas Tecnologías para Alumnos 10 12
Uso Inteligente de las Nuevas Tecnologías para Alumnos 12 14
Uso Inteligente de las Nuevas Tecnologías para Alumnos 14 16
Uso Inteligente de las Nuevas Tecnologías para Alumnos 8 10
Pedagogía de la praxis
Pedagogía del oprimido
Pedagogía didáctica y autismo
Pedagogía Kant
Profesores excelentes
Psicología y pedagogía
Tecnología y escuela
TIC para la inclusión social
Una pedagogía de la comunicación
Una pedagogía praxeológica
Vida y profesión del pedagogo

Apéndice E

Publicaciones realizadas

En las siguientes listas se muestran los artículos publicados a lo largo de los estudios doctorales:

2020:

- Alemán, Y., Somodevilla, M., & Vilariño, D. (2020). *An Analysis of Variance Method for Detection of Collocations in a Pedagogical Domain Corpus.* Computación y Sistemas, 24(2).
- Yuridiana Alemán, María J. Somodevilla, Darnes Vilariño. *Ontología de estilos de aprendizaje para asistir en el aprendizaje significativo*. Research in Computing Science (En proceso de publicación). ISSN: 1870-4069

2019:

- Yuridiana Alemán, María J. Somodevilla, Darnes Vilariño. *Similarity metrics analysis for principal concepts detection in ontology creation*. Special Section: Intelligent and Fuzzy Systems applied to Language & Knowledge Engineering. Journal of Intelligent & Fuzzy Systems. Vol. 36 No. 5, pp. 4753-4764. ISSN: 1875-8967
- Yuridiana Alemán, MJ Somodevilla, D Vilariño. *Extracción de conceptos y relaciones para la creación de una ontología en el dominio pedagógico*. Avances en Tecnologías del lenguaje y del conocimiento. BUAP ISBN:978-607-7512-91-2

2018:

- Yuridiana Alemán, María J. Somodevilla, Darnes Vilariño. *A class validation proposal of a pedagogic domain ontology based on clustering analysis*. Computer and Information Science. Vol. 11 No 1, Febrero 2018 ISSN 1913-8989 (Print) ISSN 1913-8997 (Online).

- Helena Gómez-Adorno, Carolina Martín-del-Campo-Rodríguez, Grigori Sidorov, Yuridiana Aleman, Darnes Vilariño and David Pinto. *Hierarchical Clustering Analysis: The bestperforming approach at PAN 2017 author clustering task. CLEF*
- Alemán, Y., Somodevilla, M., & Vilariño, D. *Identificación de relaciones taxonómicas de dominio usando métricas textuales*. Research in Computing Science, 147, 71-84.
- Yuridiana Alemán, María J. Somodevilla, Darnes Vilariño. *Una metodología para el aprendizaje ontológico semiautomático de dominio pedagógico*. Tópicos actuales en la ingeniería del lenguaje y del conocimiento Marzo 2019 ISBN 978-607-97282-7-4 pp. 67 - 76.

2017:

- Helena Gómez-Adorno, Yuridiana Alemán, Darnes Vilariño, Miguel Sanchez-Perez, David Pinto, Grigori Sidorov. *Author clustering using Hierarchical Clustering Analysis*. CLEF (Working Notes).
- Yuridiana Alemán, María J. Somodevilla, Darnes Vilariño. *A semantic proposal for semiautomatic corpus creation in the pedagogic domain.* Research in Computing Science Journal Vol. 145. ISSN: 1870-4069
- Yuridiana Alemán, María J. Somodevilla, Darnes Vilariño. *A proposal for Domain Ontological Learning*. Research in Computing Science (Advances in Pattern Recognition). Vol 133, pp. 63-70. ISSN: 1870-4069

2016:

- Yuridiana Alemán, Darnes Vilariño, David Pinto. *Análisis de selección de atributos con ganancia de información y* X_2. Tendencias en la Ingeniería del Lenguaje y del Conocimiento 2015-2016. ISBN: 978-607-525-149-3 pp. 54-63
- Yuridiana Alemán, Darnes Vilariño, Josefa Somodevilla. *Clasificación de polaridad con un diccionario mediante el algoritmo de Bayes*. Special issue of the Research in Computing Science Journal, ISSN 1870-4069

Referencias

Aguilar, C., Acosta, O., Sierra, G., Juárez, S., & Infante, T. (2016). Extracción de contextos definitorios en el área de biomedicina. *Procesamiento del Lenguaje Natural*, *57*(0), 167–170.

Aguirre, É. B. (2008). Inteligencias múltiples y estilos de aprendizaje en los estudiantes de primer semestre de contaduría pública en la Universidad de la Salle. *Psicogente*, *11*(20).

Al-Shamri, M. Y. H. (2014). Power coefficient as a similarity measure for memory-based collaborative recommender systems. *Expert Systems with Applications*, *41*(13), 5680 – 5688.

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., & Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, *18*(1), 14–21.

Alducin-Ochoa, J. & Vázquez, A. (2016). Estilos de aprendizaje, variables sociodemográficas y rendimiento académico en estudiantes de ingeniería de edificación. *Revista Electrónica Educare*, *21*, 1.

Alonso, C., Gallego, D., & Honey, P. (2007). *Los Estilos de Aprendizaje: Procedimientos de diagnóstico y mejora*.

Ameen, A., Khan, K. U. R., & Rani, B. P. (2012). Creation of ontology in education domain. In *2012 IEEE Fourth International Conference on Technology for Education*, (pp. 237–238).

Aminah, S., Afriyanti, I., & Krisnadhi, A. (2017). Ontology-based approach for academic evaluation system. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, (pp. 1569–1574).

Arias, W. L. (2014). Estilos de aprendizaje e inteligencia en estudiantes universitarios de Arequipa, Perú. *Journal of Learning Styles*, *7*(14).

Arthur, D. & Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA 07, (pp. 1027–1035)., Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Ausubel, D. & Novak, J. (1983). *Psicología educativa, un punto de vista cognoscitivo*. México: Trillas.

Bagiampou, M. & Kameas, A. (2012). A use case diagrams ontology that can be used as common reference for software engineering education. In *2012 6th IEEE International Conference Intelligent Systems*, (pp. 035–040).

Barciela, B. C. & Padilla, A. P. (2012). Representación gráfica de documentos para extracción automática de relaciones. *Procesamiento del Lenguaje Natural*, *49*(0), 57–64.

Barriga, F. & Hernández, G. (2004). *Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista*. México: McGraw Hill.

Bergasa-Suso, J., Sanders, D. A., & Tewkesbury, G. (2005). Intelligent browser-based systems to assist internet users. *IEEE Trans. Education*, *48*(4), 580–585.

Bouamrane, M. M., Luz, S., & Masoodian, M. (2008). Ontologies in Interactive Systems. In *2008 First International Workshop on Ontologies in Interactive Systems*, (pp. 3–6).

Bucos, M., Dragulescu, B., & Veltan, M. (2010). Designing a semantic web ontology for E-learning in higher education. In *2010 9th International Symposium on Electronics and Telecommunications*, (pp. 415–418). IEEE.

Cakula, S. & Sedleniece, M. (2013). Development of a personalized e-learning model using methods of ontology. *Procedia Computer Science*, *26*, 113 – 120. ICTE in Regional Development, December 2013, Valmiera, Latvia.

Cano García, F. (2000). Diferencias de género en estrategias y estilos de aprendizaje. *Psicothema*, *12*(3).

Chance, P. (2001). *Aprendizaje y conducta*. Manual moderno.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51–89.

Cimiano, P. (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Cimiano, P. & Völker, J. (2005). Text2onto. In Montoyo, A., Muńoz, R., & Métais, E. (Eds.), *Natural Language Processing and Information Systems*, (pp. 227–238)., Berlin, Heidelberg. Springer Berlin Heidelberg.

Corda, I., Bennett, B., & Dimitrova, V. (2008). Interacting with an ontology to explore historical domains. In *2008 First International Workshop on Ontologies in Interactive Systems*, (pp. 65–74).

Cullen, J. & Bryman, A. (1988). The knowledge acquisition bottleneck: Time for reassessment? *Expert Systems*, *5*(3), 216–225.

Dai, X. & Li, X. (2010). Study of learning source ontology modeling in remote education. In *2010 International Conference on Multimedia Technology*, (pp. 1–4).

Das, P., Das, A., Nayak, J., Pelusi, D., & Ding, W. (2019). A graph based clustering approach for relation extraction from crime data. *IEEE Access*, *PP*, 1–1.

Díaz Barriga, F. (2008). Educación y nuevas tecnologías de la información: ¿hacia un paradigma educativo innovador? *Sinéctica, Revista Electrónica de Educación*.

De la Villa Moreno, M. A. (2016). *Método para la Construcción Automática de Ontologías Basadas en Patrones Lingüisticos*. PhD thesis, Universidad Politécnica de Madrid.

De Luca, S. L. (2000). El docente y las inteligencias múltiples. *Revista Iberoamericana de la educación*, *11*.

Diaz-Barriga, F. & Hernández-Rojas, G. (2010). *Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista*. McGraw-Hill Interamericana.

Dinu, A., Dinu, L., & Sorodoc, I. (2014). Aggregation methods for efficient collocation detection. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, (pp. 4041–4045)., Reykjavik, Iceland. European Language Resources Association (ELRA).

Dorantes, M. A., Pimentel, A., Sierra, G., Bel-Enguix, G., & Molina, C. (2017). Extracción automática de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con un sistema basado en patrones lingüísticos. *Linguamática*, *9*(2), 33–44.

Du, L., Zheng, G., You, B., Bai, L., & Zhang, X. (2012). Research of online education ontology model. In *2012 Fourth International Conference on Computational and Information Sciences*, (pp. 780–783).

Esguerra, G. & Guerrero, P. (2010). Estilos de aprendizaje y rendimiento académico en estudiantes de psicología. *Diversitas: Perspectivas en Psicología*, *6*, 97–109.

Fan, Z., Apple, F., Horace, I., & Jiaheng, C. (2008). Engonto: Integrated Multiple English Learning Ontology for Personalized Education. In *2008 International Conference on Computer Science and Software Engineering*, volume 5, (pp. 210–213). IEEE.

Faria, C., , & Girardi, R. (2014). A domain-independent process for automatic ontology population from text. *Science of Computer Programming*, *95, Part 1*, 26 – 43. Special Issue on Systems Development by Means of Semantic Technologies.

Ferreira, A. & Atkinson Abutridy, J. A. (1998). Un modelo de agente de búsqueda y filtrado de información inteligente apoyado por interacciones en lenguaje natural. *Revista Facultad de Ingenier'ıa*, (5).

Ferreira, H. N. M., Brant-Ribeiro, T., Araújo, R. D., Dorça, F. A., & Cattelan, R. G. (2016). An automatic and dynamic student modeling approach for adaptive and intelligent educational systems using ontologies and bayesian networks. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, (pp. 738–745).

Fu, J., Jia, K., & Xu, J. (2008). Domain ontology learning for question answering system in network education. In *2008 The 9th International Conference for Young Computer Scientists*, (pp. 2647–2652).

García-Miguel, J. M., Vaamonde, G., & Domínguez, F. G. (2010). 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 2010)*, Valletta, Malta. European Language Resources Association (ELRA).

Gardner, H. (2001). *Estructuras de la Mente*.

Genovard, C. & Gotzens, C. (1990). *Psicología de la instrucción*. Santillana.

Gomaa, W. & Fahmy, A. (2013). A survey of text similarity approaches. *68*.

Gong, S. & Gao, W. (2016). Ontology learning algorithm via wmw optimization model. In *2016 12th International Conference on Computational Intelligence and Security (CIS)*, (pp. 431–434).

González, M. & Tourón, J. (1992). *Autoconcepto y rendimiento escolar: sus implicaciones en la motivación y en la autorregulación del aprendizaje*. Eunsa.

Grandbastien, M., Azouaou, F., Desmoulins, C., Faerber, R., Leclet, D., & Quenu-Joiron, C. (2007). 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).

Grljevic, O. & Bosnjak, Z. (2015). Development of serbian higher education corpus. In *2015 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*, (pp. 177–181).

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, *43*, 907–928.

Guan, H., Zhou, J., & Guo, M. (2009). A class-feature-centroid classifier for text categorization. In *Proceedings of the 18th International Conference on World Wide Web*, WWW 2009, (pp. 201–210)., New York, NY, USA. ACM.

Hajiabadi, H. (2014). Ontology based data mining approach on web documents. *Computer and Information Science*, *7*(4), 123.

Han, J., Kambre, M., & Pei, J. (2000). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Hanani, U., Shapira, B., & Shoval, P. (2001). Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, *11*(3), 203–259.

Hssina, B., Bouikhalene, B., & Merbouha, A. (2017). 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.

Hu, J., Li, Z., & Xu, B. (2016). 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).

Huang, C.-H., Yin, J., & Hou, F. (2011). A text similarity measurement combining word semantic information with tf-idf method. *34*, 856–864.

Jing, D., Yang, H., & Tian, Y. (2013). Abstraction based domain ontology extraction for idea creation. In *2013 13th International Conference on Quality Software*, (pp. 341–348).

Kang, Y.-B., Haghighi, P. D., & Burstein, F. (2014). Cfinder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications*, *41*(9), 4494 – 4504.

Kaushik, N. & Chatterjee, N. (2018). Automatic relationship extraction from agricultural text for ontology construction. *Information Processing in Agriculture*, *5*(1), 60 – 73.

Khan, S. S. & Madden, M. G. (2010). A survey of recent trends in one class classification. In Coyle, L. & Freyne, J. (Eds.), *Artificial Intelligence and Cognitive Science*, (pp. 188–197)., Berlin, Heidelberg. Springer Berlin Heidelberg.

Kolb, D. (1976). *Learning style inventory*. Boston USA: MA: Hay Group, Hay Resources Direct.

Lee, C.-S., Kao, Y.-F., Kuo, Y.-H., & Wang, M.-H. (2007). Automated ontology construction for unstructured text documents. *Data & Knowledge Engineering*, *60*(3), 547–566.

Li, C., Zhou, W., Ji, F., Duan, Y., & Chen, H. (2018). A deep relevance model for zero-shot document filtering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 2300–2310). Association for Computational Linguistics.

Li, M., Lang, B., & Wang, J. (2015). Compound concept semantic similarity calculation based on ontology and concept constitution features. (pp. 226–233).

Li, R. (2018). An information filtering model based on neural network. In Li, K., Li, W., Chen, Z., & Liu, Y. (Eds.), *Computational Intelligence and Intelligent Systems*, (pp. 217–227)., Singapore. Springer Singapore.

Lima, R., Espinasse, B., & Freitas, F. (2019). A logic-based relational learning approach to relation extraction: The ontoilper system. *Engineering Applications of Artificial Intelligence*, *78*, 142 – 157.

López, F. B. & Castillo, J. N. P. (2013). Esquema metodológico para la construcción automática de ontolog'ıas. *Revista V'ınculos*, *10*(1), 20–30.

Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416.

Álvarez Carmona, M. n. (2014). *Detección de similitud semántica en textos cortos*. PhD thesis, Instituto Nacional de Astrofísica Óptica y Electrónica.

Maedche, A. & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, *16*(2), 72–79.

Mahesh, K. (1996). Ontology development for machine translation: Ideology and methodology.

Mala, V. & Lobiyal, D. K. (2015). Concepts extraction for medical documents using ontology. In *2015 International Conference on Advances in Computer Engineering and Applications*, (pp. 773–777).

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.

Méndez, N. D. D., Carranza, D. A. O., & Ocampo, M. G. (2015). Representación ontológica de perfiles de estudiantes para la personalización del aprendizaje. *Revista Educación en Ingeniería*, *10*(19), 105–115.

Noguera Robles, E., Llopis, F., & Ferrández, A. (2006). Filtrado de información para la búsqueda de respuestas. *Procesamiento del lenguaje natural, nº 37 (sept. 2006), pp. 145-152*.

Noy, N. & Mcguinness, D. (2001). Ontology development 101: A guide to creating your first ontology. *Knowledge Systems Laboratory*, *32*.

Ochoa, J. L., Hernández-Alcaraz, M. L., Valencia-García, R., & Martínez-Béjar, R. (2011). *A Semantic Role-Based Approach for Ontology Learning from Spanish Texts*, (pp. 273–280). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ochoa, J. L., Hernández-Alcaraz, M. L., Almela, A., & Valencia-García, R. (2011). 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).

Ochoa, J. L., Hernández-Alcaraz, M. L., Valencia-Garcia, R., & Martinez-Bejar, R. (2011). A semantic role-based methodology for knowledge acquisition from Spanish documents. *International Journal of Physical Sciences*, *6*(7), 1755–1765.

Olivos, P., Santos, A., Martín, S., Cañas, M., Gómez, E., & Maya, Y. (2016). The relationship between learning styles and motivation to transfer of learning in a vocational training programme. *Suma Psicológica*, *23*(1), 25–32.

Orta Palacios, C. P. (2008). *Métodos Basados en Patrones Léxicos para la Extracción de Información*. PhD thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica.

Ortega-Mendoza, R., Aguilar, C., Villaseñor-Pineda, L., Montes, M., & Sierra, G. (2011). Hacia la identificación de relaciones de hiponimia/hiperonimia en internet. *Revista signos*, *44*, 68–84.

Pazos, J.-M. & Pamies, A. (2006). Detecci/'on automatizada de colocaciones y otras unidades fraseológicas en un corpus electrónico. *Letras de Hoje*, *41*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Raad, J. & Cruz, C. (2015). A survey on ontology evaluation methods. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, IC3K 2015, (pp. 179–186)., Portugal. SCITEPRESS - Science and Technology Publications, Lda.

Ramón, F. (2006). *Estrategias didácticas del aprendizaje cooperativo*. Trillas.

Rettinger, A., Lösch, U., Tresp, V., D Amato, C., & Fanizzi, N. (2012). Mining the semantic web. *Data Mining Knowledge Discovery*, *24*(3), 613–662.

Rodríguez, A. & Simón, A. (2013). Método para la extracción de información estructurada desde textos. *Revista Cubana de Ciencias Informáticas*, *7*, 55 – 67.

Rodríguez, J., Bravo, M., & Guzmán, R. (2012). Multi-dimensional ontology model to support context-aware systems. In *The Seventh International Conference on Internet and Web Applications and Services*.

Rodriguez, M. (2011). La teoría del aprendizaje significativo: una revisión aplicable a la escuela actual. *Revista Electrònica d?Investigació i Innovació Educativa i Socioeducativa*, *3*, 29–50.

Rojas, C., Díaz, C., Vergara, J., Alarcón, P., & Ortiz, M. (2016). Estilos de enseñanza y estilos de aprendizaje en educación superior: Análisis de las preferencias de estudiantes de pedagogía en inglés en tres universidades chilenas. *Revista Electrónica Educare*, *20*, 1–29.

Senso, J. A., Leiva-Mederos, A., & Dominguez-Velasco, S. (2011). Modelo para la evaluacion de ontologÍas. aplicacion en Onto-Satcol. *Revista EspaÑola de Documentación Científica*, *34*(3), 334–356.

Silva Sprock, A. & Ponce, J. (2013). Reingeniería de una ontología de estilos de aprendizaje para la creación de objetos de aprendizaje. *Eduweb*, *7*, 49–64.

Smith, B. (2004). *Ontology and Information Systems*. Stanford.

Sánchez López, S. E. (2007). *Modelo de indexación de formas en sistemas VIR basado en ontologías*. PhD thesis, Escuela de ingeniería y Ciencias. Universidad de las Américas Puebla.

Somodevilla, M. J., Mena, I., Pineda, I. H., & d. Celis, M. C. P. (2015). Deducting lifestyle patterns by ontologies' SWRL rules. In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, (pp. 9–13).

Staab, S. & Studer, R. (2009). *Handbook on Ontologies* (2nd ed.). Springer Publishing Company, Incorporated.

Suárez, J., Maíz, F., & Meza, M. (2010). Inteligencias múltiples: una innovación pedagógica para potenciar el proceso de enseñanza aprendizaje. *Investigación y Posgrado*, *25*(1), 81–94.

Tapias, M. (2018). Estilos de aprendizaje, estrategias para enseñar. su relación con el desarrollo emocional y aprender a aprender. *Tendencias Pedagógicas*, *31*.

Teixeira, J., Sarmento, L., & Oliveira, E. (2011). 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).

Tovar, M., Pinto, D., Rendón, A. M., Serna, J. G. G., & Ayala, D. V. (2014). Identification of ontological relations using formal concept analysis. In *LANMR*.

Uskov, V., Pandey, A., Bakken, J. P., & Margapuri, V. S. (2016). Smart engineering education: The ontology of internet-of-things applications. In *2016 IEEE Global Engineering Education Conference (EDUCON)*, (pp. 476–481).

Valencia, R. (2005). *Un entorno para la extracción incremental de conocimiento desde texto en lenguaje natural*. PhD thesis, Departamento de Ingeniería de la información y las comunicaciones. Universidad de Murcia.

Valle, A., González, R., Cuevas, L., & Fernández, A. (1998). Las estrategias de aprendizaje: características básicas y su relevancia en el contexto escolar. *Revista de Psicodidáctica*, 53–68.

Vargas-Vera, M. & Celjuska, D. (2004). Event recognition on News Stories and semi-automatic population of an ontology. In *Web Intelligence, 2004. Proceedings*, (pp. 615–618).

Vasilateanu, A., Goga, N., Tanase, E.-A., & Marin, I. (2015). Enterprise domain ontology learning from web-based corpus.

Vázquez Cuchillo, J. (2008). *Recuperación de Información utilizando Frecuencias Frecuentes Maximales*. PhD thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).

Weigand, H. (1997). A multilingual ontology-based lexicon for news filtering-the TREVI project. In *Proceedings of the IJCAI Workshop on Multilingual Ontologies-Nagoya*.

Weinstein, C. E. & Mayer, R. E. (1986). The teaching of learning strategies. In *Innovation abstracts*, volume 5. ERIC.

Wong, W., Liu, W., & Bennamoun, M. (2011). Ontology learning from text: A look back and into the future. *44*, 1–36.

Wu, H. (2008). Research of internet education system based on ontology. In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, volume 4, (pp. 602–605).

Yu, H., Zhai, C., & Han, J. (2003). Text classification from positive and unlabeled documents. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM 03*, (pp. 232–239)., New York, NY, USA. ACM.

Zhang, J. & Hu, J. (2008). Image segmentation based on 2d otsu method with histogram analysis. In *CSSE (6)*, (pp. 105–108). IEEE Computer Society. 978-0-7695-3336-0.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). Birch: an efficient data clustering method for very large databases. In *In Proc. of the ACM SIGMOD Intl. Conference on Management of Data (SIGMOD)*, (pp. 103–114).

Zhu, F. & Yao, N. (2009). Ontology-based learning activity sequencing in personalized education system. In *2009 International Conference on Information Technology and Computer Science*, volume 1, (pp. 285–288). IEEE.

Zorić, B., Bajer, D., & Martinović, G. (2018). Utilising filter inferred information in nature-inspired hybrid feature selection. In *2018 International Conference on Smart Systems and Technologies (SST)*, (pp. 117–123). IEEE.