Universidad de las Ciencias Informáticas

Facultad 3

Título: Extracción de relaciones jerárquicas entre conceptos en corpus documentales legales

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas

Autor: Harold Lugo Rodríguez

Tutores: Ing. Guillermo Manuel Negrín Ortiz

MSc. Julio César Díaz Vera

La Habana, 6 de junio de 2019

"Año 60 de la Revolución"

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

Harold Lugo Rodríguez	**Ing. Guillermo Manuel Negrín Ortiz**	**MSc. Julio César Díaz Vera**
Firma del Autor	Firma del Tutor	Firma del Tutor

Datos de Contacto

MSc. Julio Cesar Díaz Vera

Universidad Martha Abreu, Villa Clara, Cuba.

Correo electrónico: jcdiaz@uci.cu

Ing. Guillermo Manuel Negrín Ortiz

Universidad de las Ciencias Informáticas, La Habana, Cuba.

Correo electrónico: gmnegrin@uci.cu

Dedicatoria

A mis padres,

Por su ejemplo de perseverancia y dedicación,

por su estimulante optimismo y confianza,

por su amor sin límites

Agradecimientos

A mi mamá y mi papá, por ser el motor impulsor de cada proyecto trazado en mi vida, por nunca decirme que no, por ese apoyo incondicional que me han mostrado siempre. Gracias por su sacrificio e inagotable cariño, por su confianza, por ser mí meta, por guiarme siempre por un camino de bien y apoyarme siempre que lo he necesitado. A los dos, una y mil veces. Muchas Gracias.

A mis familiares en especial a mi tía Fina, Negre, y a mi abuela Isabel. Gracias por esta ahí siempre que lo necesité y hacerme sentir especial.

A mi hermana Cinthya por el apoyo que me dio para poder terminar mi tesis.

A mis tutores Julio y Guillermo, por tantos consejos y atenciones inmerecidas.

A mis amigos, Claudia, José Pablo, Luis Alberto, Raykof, Alexis Roberto, Christopher, Ariandi, Iván, Robinson, con los cuales he compartido los años más productivos de mi vida, por tanto apoyo, consejos y buenos momentos.

A todos mis compañeros de grupo, de todos me llevo lo mejor, gracias por dejarme compartir junto a ustedes esta etapa inolvidable.

Al tribunal de manera general por sus correcciones y apoyo en cada momento

A los profesores, Cordoví, Dariela y Elizabeth por su ayuda en cada momento que necesité.

A todos los que han contribuido a mi desarrollo profesional y al resultado de este trabajo. En general, a todas las personas que me preguntaron alguna vez: “Como va la tesis”, que de una forma u otra me dieron fuerza para seguir adelante, sin ustedes no podría haberlo conseguido.

A todos, muchas gracias.

Resumen

La extracción de relaciones entre conceptos en un corpus documental es un área de minería de texto que ha alcanzado mucho interés en tareas de análisis y recuperación de información. En el contexto del gobierno electrónico, esta área de investigación ha tomado auge a partir de la necesidad del conocimiento de las leyes por la ciudadanía en general, y no solo por el personal especializado en la rama del derecho. Sin embargo, para el lenguaje español no existen técnicas ni estándares que ayuden a llevar a cabo el proceso automatizado de extracción de relaciones jerárquicas entre conceptos. En este trabajo se presenta un procedimiento que utiliza patrones sintácticos para abordar el problema de la extracción de relaciones jerárquicas en un corpus documental como un problema de reconocimiento de patrones. La presente investigación valida el procedimiento propuesto utilizando métricas de Sensibilidad y Precisión para determinar su eficiencia en la clasificación y extracción de relaciones jerárquicas entre conceptos.

PALABRAS CLAVE

Relaciones jerárquicas, conceptos, corpus documental

ABSTRACT

The extraction of relationships between concepts in document corpuses is a text mining area that has gained great interest in the tasks of analyzing and retrieving information. In the context of e-government, this area of research has become a priority based on the need of citizenship in general to know the law, and not only by specialized personnel in the field of law. However, for the Spanish language there are no techniques or standards that help to carry out an automated process of extraction of hierarchical relationships between concepts. This research presents a procedure that uses syntactic patterns to address the problem of extracting hierarchical relationships in a documentary corpus as a problem of pattern recognition. The present investigation validates the procedure by using Sensitivity and Precision benchmarks to determine its efficiency in the classification and extraction of hierarchical relationships between concepts

KEYWORDS

Hierarchical relations, concepts, document corpus

Índice

Índice de tablas

Índice de figuras

Introducción

El avance de las tecnologías y de las comunicaciones a nivel mundial, la velocidad con la que se realizan distintos tipos de operaciones a través de la utilización de Internet y las ventajas que todo esto implica provocan un cambio fundamental en la manera de relacionarse. Este cambio tecnológico afecta todas las áreas de la sociedad y la forma de establecer relaciones entre ellas. El Estado, en su papel regulador de la sociedad también ante este fenómeno debe adaptarse a las nuevas tendencias, implementando de esta manera el concepto de Gobierno Electrónico (GE).

El GE ha tomado auge en la últimas dos décadas, facilitando la interacción entre el gobierno y los ciudadanos. Se concibe como una nueva forma de gobierno que utiliza las TIC con modalidades para la gestión, planificación y administración a través de portales en Internet con información referente a las dependencias de la administración pública, órganos de gobierno, poderes, servicios y trámites (Pérez Zúñiga et al. 2016). De este modo se permite a la sociedad lograr un mayor contacto con el Estado.

Internet puede aportar mucho para acercar los ciudadanos a la gestión gubernamental. De este modo, la población demanda servicios públicos en línea que además de amenizar trámites y procedimientos, ahorran tiempo y dinero. Además, están operativos las 24 horas del día, los 7 días de la semana, durante todo el año (Naser y Concha 2011).

Los servicios públicos en línea implementados como parte de la estrategia de GE traen consigo una mejora en la interacción entre el par gobierno sociedad y por lo tanto permite que los ciudadanos lleguen a entender mejor la manera en que funcionan los servicios que se les ofrecen. Desde el punto de vista de los procesos legales, en alguna manera estos se agilizan y se extiende su disponibilidad a través de la tecnología y la infraestructura de redes.

La disponibilidad de información y los servicios al alcance de la mano han dado pie a una sociedad del conocimiento en la que el entendimiento de la ley juega un papel relevante. A pesar del desarrollo y el fácil acceso a la información que brinda un gobierno electrónico a la sociedad, las personas en muchas ocasiones no cuentan con conocimiento suficiente para entender los servicios y leyes que ponen a su disposición.

La poca comprensión o entendimiento incompleto de la ley tiene consecuencias perjudiciales; el correcto entendimiento de los derechos y deberes que deben cumplir los ciudadanos es un factor clave para mejorar la sociedad. El Centro de Estudios Jurídicos (CEJ) reconoce que las leyes son difíciles de comprender, en buena medida por sus connotaciones técnicas, pero también por el escaso conocimiento que de ella tienen los ciudadanos (Ballesteros Muñoz y Ruiz Jiménez 2011).

El lenguaje jurídico se ha caracterizado tradicionalmente por ser un lenguaje especializado y complejo, lo que lo hace en algunos casos ininteligible para el ciudadano. Sin embargo, las especificidades del lenguaje jurídico deberían ser compatibles con la claridad, pues no tiene como destinatarios únicos a los profesionales del Derecho y de la Administración, sino también a la ciudadanía en general (Relinque Barranca 2017).

La comprensión de la ley por los ciudadanos se dificulta por dos cuestiones fundamentales: el dominio jurídico complejo y la ambigüedad del lenguaje común. La primera mantiene palabras del latín, emplea un lenguaje protocolario y utiliza un léxico culto. Mientras que la segunda afecta la comprensión ya que no siempre existe claridad en el significado de un término o frase. Es por ello que en muchos países se planteó la necesidad de modernizar este lenguaje con el fin de conseguir que la Administración y la Justicia se acerquen al ciudadano (Relinque Barranca 2017).

De esta forma, han surgido movimientos en prácticamente todos los países occidentales para conseguir la simplificación del lenguaje jurídico. Los países de habla inglesa son los que más han avanzado en este campo, especialmente los Estados Unidos (Relinque Barranca 2017). En España se ha avanzado muy poco en este tema, ya que a pesar de que se han puesto en marcha varias iniciativas desde el Gobierno, ninguna se ha visto reflejada en la práctica (Relinque Barranca 2017).

Hasta ahora se ha hecho evidente la necesidad de buscar un mecanismo para mejorar el entendimiento de los textos jurídicos. En este aspecto entra a jugar un papel importante el idioma en que esté escrito el texto, si el caso es el español, hay que tener en cuenta su complejidad en la construcción de frases por el gran número de palabras y las numerosas formas verbales irregulares que posee. Independientemente de la manera en que estén redactados los documentos jurídicos y el idioma en que estén escritos, su comprensión, en su forma más simple, se trata de encontrar palabras y establecer relaciones entre ellas de una manera en que puedan ser interpretados por los ciudadanos. En la literatura se han

definido estas palabras importantes como conceptos y se han establecido dos tipos de relaciones: las jerárquicas y las no jerárquicas.

Las relaciones jerárquicas están basadas en niveles jerárquicos de superioridad o subordinación entre conceptos. El concepto superior constituye una clase, mientras que los conceptos subordinados representan elementos o partes de esa clase.

La extracción manual de relaciones entre conceptos es una tarea que consume gran cantidad de recursos y la participación de muchos expertos. En contraste la detección automatizada de relaciones entre conceptos podría ser de utilidad. Para ello es imprescindible que una vez extraídos los conceptos desde un corpus documental legal se incorporen las relaciones entre ellos.

A partir de la situación descrita anteriormente y para guiar la investigación se plantea como problema a resolver: ¿Cómo extraer relaciones jerárquicas entre conceptos desde corpus documentales legales?

Para resolver el problema planteado se define como objetivo general de la investigación: Obtener las relaciones jerárquicas entre conceptos desde corpus documentales legales.

Como objeto de estudio de la presente investigación se define: Procesamiento de lenguaje natural.

Enmarcándose en el campo de acción: Técnicas de extracción de relaciones jerárquicas entre conceptos para el idioma español.

Para dar cumplimiento al objetivo general se desglosan los siguientes objetivos específicos:

- Establecer el marco conceptual para el desarrollo de la investigación
- Definir un mecanismo para la extracción de relaciones jerárquicas entre conceptos en el proyecto de constitución de la República de Cuba
- Validar la propuesta.

La tesis está estructurada en tres capítulos cuyos contenidos son:

Capítulo 1. “Fundamentación Teórica”. En el capítulo se describen los conceptos asociados a la investigación para el desarrollo de la propuesta de solución. Se realiza la revisión

sistemática de la bibliografía la cual permite llevar a cabo un estudio de las soluciones existentes y las investigaciones abordadas en la tesis. Se describe un procedimiento para extracción de relaciones jerárquicas entre conceptos.

Capítulo 2. “Diseño e Implementación”. Se presenta el procedimiento a seguir para la extracción de relaciones entre conceptos. El procedimiento se divide en tres etapas: creación del corpus documental legal, pre procesamiento y extracción de características.

Capítulo 3. “Validación”. En el capítulo se valida la solución mediante una metodología de experimentación para verificar la calidad y el funcionamiento del procedimiento del capítulo anterior. Se evalúa mediante las métricas precisión y sensibilidad.

Capítulo 1: Fundamentación Teórica

Introducción del capítulo

En este capítulo se presentan los fundamentos teóricos de la investigación, se emplea el método de revisión sistemática de la literatura relacionada con el campo de acción. Se realiza el análisis de la literatura, la planificación de la investigación y la elección de los artículos mediante criterios de inclusión y exclusión. Además del análisis de las bases de datos, todo esto se toma en consideración para el estudio del estado del arte de las investigaciones relacionadas con el tema del presente trabajo. Por último, se propone la estrategia para orientar la investigación.

1.1 Marco conceptual

1.1.1 Concepto

En la comunicación, no todos los objetos individuales en el mundo son diferenciados y nombrados. En cambio, a través de la observación y un proceso de abstracción llamado conceptualización, los objetos se clasifican en clases, que corresponden a unidades de conocimiento llamadas conceptos.

Para efectuar o acumular experiencias, es decir, para integrarlas en la vida, se necesitan conceptos, pues los conceptos permiten guardar y retener las experiencias incluso cuando éstas ya se han desvanecido. Además se necesitan conceptos para saber lo que sucedió, para almacenar el pasado en el lenguaje y para integrar las experiencias vividas en sus capacidades lingüísticas y en su comportamiento (Koselleck 2004).

Según la vigésimo tercera edición del diccionario de la Real Academia de la Lengua Española la palabra concepto posee 6 acepciones («Real Academia Española» 2018):

1. Idea que concibe o forma el entendimiento.
2. Sentencia, agudeza, dicho ingenioso.
3. Opinión, juicio.
4. Crédito en que se tiene a alguien o algo.
5. Aspecto, calidad, título.
6. Representación mental asociada a un significante lingüístico.

La norma ISO 704 de 1987 define el término concepto como: constructos mentales o abstracciones que se pueden emplear para clasificar los distintos objetos del mundo exterior e interior. Engloban y construyen los conocimientos básicos, complejos o determinados del ser humano. Se compone por un conjunto de proposiciones o pensamientos que se formulan en base a un conocimiento previo, es decir: de una o varias ideas se concibe o se da forma a un entendimiento (Hernández 2002).

La norma ISO 704 de 2009 define que, en lenguaje natural, los conceptos pueden representarse por términos, denominaciones, definiciones, otras formas lingüísticas o también pueden estar representados por símbolos. En lenguaje artificial, pueden representarse mediante códigos o fórmulas, mientras que en multimedia pueden representarse mediante íconos, imágenes, diagramas, gráficos, clips de sonido, videos u otras representaciones multimedia. Los conceptos también pueden representarse con el cuerpo humano como lo son en lenguaje de señas, expresiones faciales o movimientos corporales.

La formación del concepto está estrechamente ligada a la experiencia propia, la cultura, la sociedad y el lenguaje, proporcionando habilidades como la expresión oral, la escritura y la comunicación. Surge como necesidad para generalizar y clasificar las experiencias y conocimientos de la vida.

De acuerdo al contexto de la presente investigación se asume la definición abordada por Chantal Pérez Hernández en (Hernández 2002) basado en la ISO 704 de 1987.

1.1.2 Relación

El origen de la palabra relación en el vocablo latino “relatio”, integrada por el prefijo de reiteración “re”, por “lat” en el sentido de llevar y el sufijo de efecto “tio”, o sea su significado sería “volver a llevar alguna cosa”.

Según la vigésimo tercera edición del diccionario de la Real Academia de la Lengua Española la palabra relación posee 8 acepciones («Real Academia Española» 2018):

1. Exposición que se hace de un hecho.
2. Conexión, correspondencia de algo con otra cosa.
3. Conexión, correspondencia, trato, comunicación de alguien con otra persona.
4. Trato de carácter amoroso.
5. Lista de nombres o elementos de cualquier clase.

6. Informe que generalmente se hace por escrito, y se presenta ante una autoridad.
7. Conexión o enlace entre dos términos de una misma oración.
8. Resultado de comparar dos cantidades expresadas en números.

La relación alude a la existencia de una conexión entre dos o más seres vivos, cosas, hechos o circunstancias.

En el marco del trabajo de investigación, se toma como principal definición de la palabra relación, la segunda propuesta realizada por la vigésimo tercera edición del diccionario de la Real Academia de la Lengua Española, ya que en la investigación se relacionan conceptos que se corresponden

1.1.3 Relación Jerárquica

Las relaciones jerárquicas son consideradas relaciones taxonómicas que asocian una entidad de un tipo específico (hipónimo) a otra entidad de un tipo más general (llamado hiperónimo). Las relaciones de hiperónimo son conocidas como relaciones "*is-a*", "*class-inclusion*" o subsunción. Esta relación jerárquica se representa mediante el enlace (ES-UN), es decir, es el vínculo del concepto *X* con el *Y*, al mismo tiempo que unen el *Y* con el *X* (Tovar et al. 2015) (García de Quesada 2001).

En una relación jerárquica, los conceptos se organizan en niveles de conceptos superiores y subordinados. Para que haya una jerarquía, debe haber al menos un concepto subordinado debajo de un concepto superior. Como se muestra en **¡Error! No se e ncuentra el origen de la referencia.**.

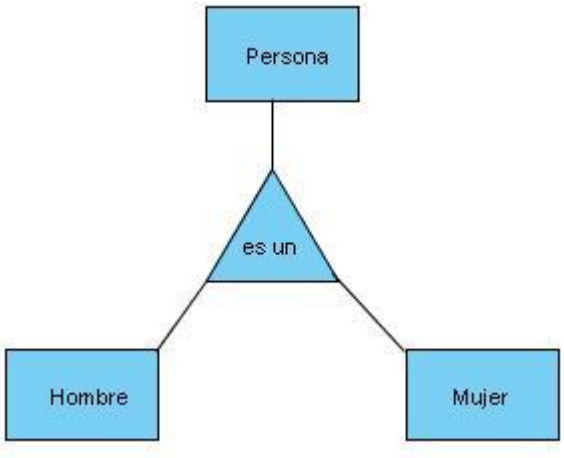

Ilustración 1 Relación jerárquica entre tres conceptos

Los conceptos superiores pueden subdividirse de acuerdo con más de un criterio de subdivisión (es decir, pueden verse desde más de una dimensión). Los conceptos subordinados en el mismo nivel y que resultan de la aplicación del mismo criterio de subdivisión se denominan conceptos de coordenadas.

La norma ISO 704, reconocen dos tipos de relaciones jerárquicas:

- relaciones genéricas
- relaciones parciales

Existe una relación genérica entre dos conceptos cuando la intensión del concepto subordinado incluye la intensión del concepto superordinado más al menos una característica de delimitación adicional. El concepto superior en una relación genérica se llama concepto genérico y el concepto subordinado se llama concepto específico.

Se dice que una relación partitiva existe cuando el concepto superior representa un todo, mientras que los conceptos subordinados representan partes de ese todo. Las partes se juntan para formar el todo. El concepto superior en una relación partitiva se llama el concepto integral y el concepto subordinado se llama el concepto partitivo. Los conceptos subordinados en el mismo nivel y compartir la misma dimensión también se denominan conceptos de coordenadas («International Standar ISO 704» 2009).

1.1.4 Procesamiento del lenguaje natural

El lenguaje es una de las herramientas centrales en la vida social y profesional. Los seres humanos son muy buenos a la hora de articular y utilizar el lenguaje, pero no tanto a la hora de entender y describir formalmente las reglas que lo gobiernan. Por este motivo, entender y producir el lenguaje por medio de una computadora es un problema muy difícil de resolver. Este problema, es el campo de estudio de lo que en la inteligencia artificial se conoce como Procesamiento del Lenguaje Natural (PLN).

El PLN es un conjunto de técnicas y procesos computacionales que trabajan sobre textos escritos en lenguaje natural para, con base en análisis lingüísticos, resolver tareas útiles que requieren de dicho conocimiento lingüístico (Liddy 2001).

El lenguaje natural se entiende como el lenguaje hablado y escrito con el propósito que exista comunicación entre una o varias personas, es más directo para expresar lo que se

quiere comunicar. El PLN es un área de investigación de constante desarrollo, se aplica en la actualidad en muchas actividades como son los sistemas de recuperación de información, la traducción automática, e interfaces en lenguaje natural.

El procesamiento del lenguaje natural es una aproximación válida para la extracción de relaciones jerárquicas y que a continuación se realiza un estudio del arte para formalizar un procedimiento y para la extracción de relaciones jerárquicas.

1.2 Método de revisión sistemática

Los métodos de revisión sistemática o también llamados: revisión integrada, meta-análisis meta-resumen, meta-síntesis o teoría fundada, se utilizan para identificar lo medular de una revisión bibliográfica, realizando la búsqueda y extracción de lo más relevante acorde a criterios que han sido evaluados y respetados por otros autores (Urra Medina y Barría Pailaquilén 2010).

Es un diseño de investigación eficiente que permite incrementar el poder y la precisión de la estimación, así como la consistencia y generalización de los resultados; y hacer además una evaluación estricta de la información publicada (Manterola et al. 2011).

Las revisiones sistemáticas son el mejor esfuerzo por recopilar y sintetizar evidencia científica sobre un tema. Este tipo de investigación establece un resumen de evidencias a través de métodos de búsqueda sistemática y síntesis de información seleccionada. Estas presentan una serie de pasos, los cuales incluyen: formular las preguntas de investigación, definir criterios de selección, la búsqueda en la literatura, evaluar los datos, analizar los datos y, presentar los resultados.

Existen dos tipos de revisiones sistemáticas (cualitativas o cuantitativas/meta-análisis). Las revisiones cualitativas presentan la evidencia en forma "descriptiva" y sin análisis estadístico, también conocidas como revisiones sistemáticas (revisiones sistemáticas sin meta-análisis). Las revisiones cuantitativas también pueden presentar la evidencia de forma descriptiva, pero la gran diferencia con la revisión cualitativa radica principalmente en el uso de técnicas estadísticas para combinar "numéricamente" los resultados frente a un estimador puntual, también denominado "meta-análisis" (Aguilera Eguía 2014).

En la presente investigación se utiliza el método de revisión sistemática de la literatura para la construcción del estado del arte y la recomendación de trabajos previos. El tipo de revisión sistemática utilizada fue la cuantitativa, dividiéndose el proceso en tres etapas:

planeación de la investigación, ejecución y análisis de resultados. Las siguientes secciones describen cómo se ejecutaron cada una de estas etapas.

1.3 Planeación de la investigación

En la etapa de planeación se aborda la definición de las preguntas científicas, definición de los temas de interés, selección de las fuentes de datos, definición de palabras claves, estrategias de búsqueda y criterios para inclusión y exclusión. La formulación de las preguntas inician la revisión sistemática y se establece como preguntas conocidas y delimitadas que sean accesibles e identificables en la literatura (Urra Medina y Barría Pailaquilén 2010b). Por lo tanto, se definen las siguientes preguntas de investigación:

1. ¿Cuál es el enfoque de los investigadores?
2. ¿Cuáles dataset fueron utilizados?
3. ¿Cuáles fueron los algoritmos y métodos aplicados?
4. ¿Cuál es la distribución temporal de los artículos?

Una vez formuladas las preguntas de investigación, se determinan los criterios de inclusión y excusión para elegir los artículos de investigación. La Tabla 1 que se presenta a continuación se establecen los criterios utilizados en este estudio:

Tabla 1:Criterios de selección de exclusión e inclusión

Criterios	Descripción
Inclusión	• Estudios realizados en idioma inglés • Documentos que constituyan artículos científicos. • Estudios relevantes sobre extracción de relaciones jerárquicas entre conceptos en texto.
Exclusión	• Estudios fuera del contexto de trabajo. • Estudios consignados en idioma diferente al inglés. • Estudios publicados antes del 2010

Una vez definidos los criterios, se verifica en las bases de datos y otras bases electrónicas si hay revisiones ya publicadas del tema seleccionado. La búsqueda de estudios en la

literatura científica se realiza a través de una estrategia de búsqueda que cumpla con los criterios propuestos. La lectura del título, el resumen y la revisión completa de los artículos, seleccionando aquellos que reúnen los criterios de selección. Estos estudios constituyen la revisión, de ellos se extraen los datos requeridos. En la presente investigación se utilizan las siguientes bases de datos electrónicas para encontrar artículos:

- ACM Digital Library: www.dl.acm.org
- IEEE Xplore Digital Library: www.ieeexplore.ieee.org
- Science Direct: www.sciencedirect.com
- Semantic Scholar: www.semanticscholar.org

El método seleccionado para buscar en estas bases de datos fue la recuperación booleana. Esencialmente, divide un espacio de búsqueda, identificando un subconjunto de documentos en una colección, de acuerdo con los criterios de consulta (Karimi et al. 2010). En este caso, la clave es la siguiente cadena: *(("relation" OR "relationship" OR " relationships between concepts ") AND ("concept" O "taxonomy" O " hierarchical relations)).*

1.4 Plan de ejecución

El plan de ejecución consta de cinco pasos:

1. Realizar la búsqueda de en las bases de datos seleccionadas
2. Comparación de resultados de búsquedas para excluir artículos repetidos
3. Aplicación de criterios de inclusión, exclusión y calidad;
4. Evaluación de todos los estudios que pasaron la revisión inicial;
5. Síntesis de datos

La búsqueda en las fuentes bibliográficas se ejecutó a partir de una cadena de búsqueda en las bases de datos mencionadas en el epígrafe 0. En la cual se obtuvo como resultado un total de 570 artículos para la revisión.

Para ayudar a la revisión y lograr una mayor precisión y confiabilidad, se usó la herramienta StArt (estado del arte a través de revisiones sistemáticas). Esta herramienta tiene el propósito de apoyar a los investigadores en su análisis sistemático.

Utiliza las extensiones BibTeX (archivo de formato bibliográfico utilizado en los documentos de LaTeX) para realizar estos análisis. Por lo que las extensiones también se extrajeron de las bases de datos mencionadas anteriormente. Es importante tener en cuenta que los

archivos BibTeX se exportaron sin ningún filtro, lo que explica el número de investigaciones devueltas.

De este total se eliminaron 400 artículos identificando aquellos correspondientes a estudios realizados antes del 2010, obteniendo una diferencia de 170 artículos. Después, se realizó un análisis por título, palabras claves y resúmenes, con el objetivo de eliminar aquellos artículos cuyo contenido no era coherente con la temática de investigación. Obteniéndose 13 artículos con información sobre extracción de relaciones jerárquicas entre conceptos como se muestra en la Ilustración 2.

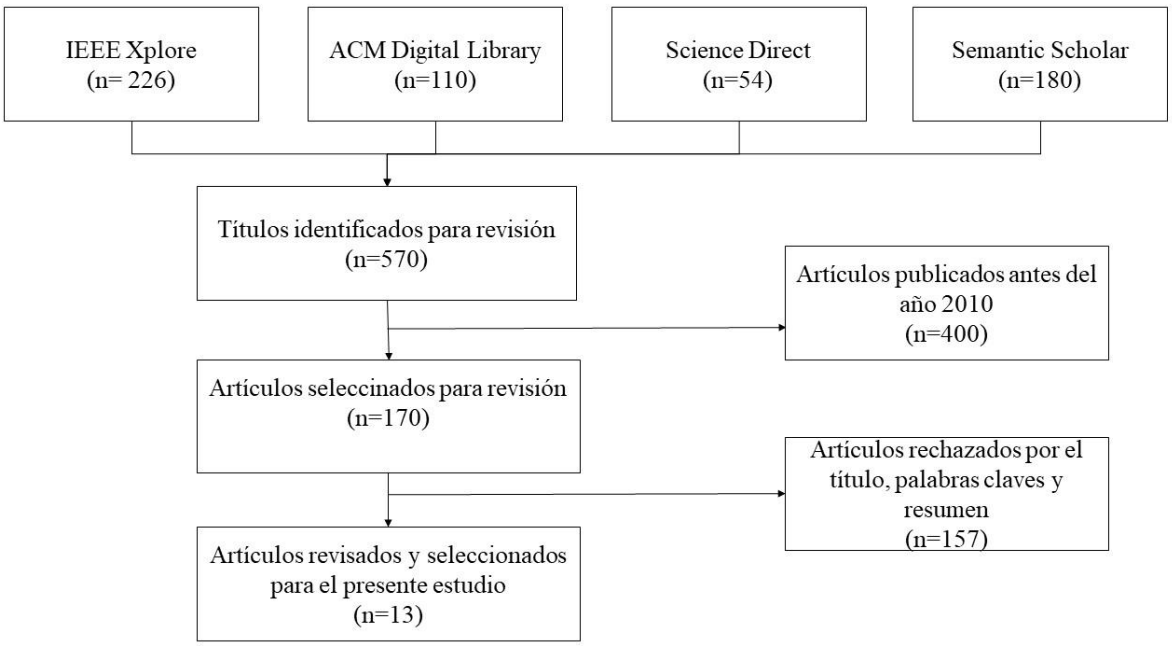

Ilustración 2 Flujo de investigación de la revisión sistemática

A continuación, se muestra un resumen de los 13 artículos escogidos:

No. 1	**Título:** Un enfoque semántico para extraer taxonomías de dominio de texto
Autores: Kevin Meijer, Flavius Frasincar, Frederik Hogenboom **Breve descripción**: En este artículo se presenta un marco para la construcción automática de una taxonomía de dominio a partir de cuerpos de texto, denominada Construcción de taxonomía automática a partir de texto (ATCT). Este marco consta de	

cuatro pasos. Primero, los términos se extraen de un corpus de documentos. De estos términos extraídos, los que son más relevantes para un dominio específico se seleccionan utilizando un enfoque del filtrado en el segundo paso. En tercer lugar, los términos seleccionados se desambiguan mediante una técnica de desambiguación con sentido de la palabra y se generan conceptos. En el paso final, las relaciones más amplias y más estrechas entre conceptos se determinan utilizando una técnica de subsunción. Para la evaluación, se evalúa el rendimiento del marco ATCT utilizando la precisión semántica, el recuerdo semántico y la medida F taxonómica que tienen en cuenta el concepto de semántica. El marco propuesto se evalúa en el campo de la economía y la gestión, así como el dominio médico.

No. 2	**Título:** Aprendizaje automatizado de taxonomías de dominio a partir de texto utilizando antecedentes de crecimiento
Autores: Julia Hoxha, Guoqian Jiang, Chunhua Weng **Breve descripción**: En este artículo se presenta un método automatizado para el aprendizaje de taxonomía, centrado en la formación de conceptos y el aprendizaje de relaciones jerárquicas. Para inferir tales relaciones, se divide los conceptos extraídos y se agrupan en grupos estrechamente relacionados mediante el uso de Agrupación jerárquica aglomerada, informados por funciones de comparación sintáctica y relación semántica. Presentamos un método novedoso, sin supervisión, para la detección de clústeres basado en la eliminación automática de dendrogramas, que es dinámico para cada partición. Se evalúa con dos tipos diferentes de cuerpos textuales, la descripción de ensayos clínicos y resúmenes de publicaciones de MEDLINE	

No. 3	**Título:** Identificación automática de relaciones jerárquicas entre palabras basadas en agrupación
Autores: Wei Zhou, Yuncheng Du, Hongwei Wang, Xueqiang Lv	

Breve descripción: En este documento, se describieron varios métodos comunes para identificar el concepto de relación jerárquica: el método de identificación basado en el análisis de ocurrencia concurrente, la distribución del método de cálculo de similitud y el método basado en el emparejamiento sintáctico de patrones, propuso una combinación de métodos de estadística de co-ocurrencia y distribución. del cálculo de similitud para identificar el concepto de relaciones jerárquicas entre palabras y se comprobó mediante experimentos que el método es factible y efectivo.

No. 4	**Título:** Generando automáticamente una jerarquía de conceptos con gráficos.
Autores: Pucktada Treeratpituk, Madian Khabsa, C. Lee Giles **Breve descripción**: En este artículo se propone un nuevo enfoque basado en gráficos para construir una jerarquía de conceptos a partir de un gran cuerpo de texto. El algoritmo incorpora tanto co-ocurrencias estadísticas como similitud léxica en la optimización de la estructura de la taxonomía. Para generar automáticamente taxonomías dependientes del tema a partir de un gran corpus de texto, primero se extraen términos tópicos y sus relaciones del corpus. Después el algoritmo construye un gráfico ponderado que representa los temas y sus asociaciones. Luego se usa un algoritmo de partición de gráficos, este divide recursivamente el gráfico de temas en una taxonomía.	

No. 5	**Título:** Aprendizaje de taxonomía de dominios a partir de texto: el método de subsunción versus agrupamiento jerárquico.
Autores: Jeroen de Knijff, Flavius Frasincar, Frederik Hogenboom **Breve descripción**: Este documento se propone un marco para construir automáticamente taxonomías a partir de un corpus de documentos de texto. Este marco extrae primero los términos de los documentos utilizando un analizador de parte del discurso. Estos términos se filtran utilizando la pertinencia del dominio, el consenso del	

dominio, la cohesión léxica y la relevancia estructural. Los términos restantes representan conceptos en la taxonomía. Estos conceptos se organizan en una jerarquía con el método de subsunción extendida que da cuenta de los antepasados del concepto al determinar el elemento primario de un concepto o un algoritmo de agrupación jerárquica que utiliza varios ámbitos de documentos y ventanas basados en texto para las co-ocurrencias del concepto. El método de subsunción es preferible para taxonomías poco profundas, mientras que el algoritmo de agrupamiento jerárquico se recomienda para taxonomías profundas.

No. 6	**Título:** Relación de aprendizaje de la taxonomía a largo plazo de una gran colección de texto sin formato.

Autores: Shang-En Yang, Vorakit Vorakitphan, Hung-Yuan Chen, Yao-Chung Fan

Breve descripción: En este documento se presenta una heurística para etiquetar un término determinado como una etiqueta de taxonomía. Específicamente, para determinar una relación "is-a" entre el término dado y un concepto inferido. Dicho problema de etiquetado de término no es nuevo, pero las soluciones existentes requieren un procesamiento de entrenamiento semi-supervisado, por ejemplo, LDA supervisado, o dependen de lexicógrafos, por ejemplo, wordnet. El costo de construcción del modelo se convierte en una carga para emplear dicha capacidad de comprensión semántica en varias aplicaciones emergentes. Con el objetivo de abordar estos problemas, en este estudio, presentamos un enfoque ligero con las siguientes características. Primero, el enfoque propuesto no está supervisado y solo toma el texto principal como entradas. Segundo, el enfoque propuesto permite la construcción del modelo incremental. Tercero, el enfoque propuesto es simple pero efectivo y computacionalmente eficiente en comparación con las soluciones existentes. Demostramos estos resultados a través de experimentos comparando nuestro enfoque con DBpedia y empleamos los términos de búsqueda populares como conjunto de pruebas. A partir de los resultados del experimento, observamos que el enfoque propuesto puede lograr una mejora del 30 por ciento en la precisión.

No. 7	**Título:** Aprendizaje temporal de relaciones semánticas entre conceptos mediante el repositorio web

Autores: Zheng Xua, Junyu Xuan

Breve descripción: En este trabajo, se propone el estudio sobre la generación de aprendizaje temporal de las relaciones entre conceptos. El propósito del estudio propuesto es anotar una relación entre conceptos con información semántica, temporal, concisa y estructurada, que puede liberar la carga cognitiva de las relaciones de aprendizaje entre conceptos para los usuarios. Las anotaciones temporales pueden ayudar a los usuarios a aprender y comprender las relaciones desconocidas o nuevas surgidas entre los conceptos. Se propone un método general para generar el aprendizaje temporal de una relación entre conceptos mediante la construcción de sus palabras de relación, oraciones de relación, gráfico de relación y factor de relación. Los experimentos empíricos en el conjunto de datos de estrellas de cine muestran que el algoritmo propuesto es efectivo y preciso.

No. 8	**Título:** Desambiguación del sentido de la palabra para la construcción automática de taxonomía a partir de cuerpos web basados en texto

Autores: Jeroen de Knijff, Kevin Meijer, Flavius Frasincar, Frederik Hogenboom

Breve descripción: En este documento, proponemos el marco de Construcción de Taxonomía Automática a partir de Texto (ATCT) para la construcción de taxonomías a partir de web basados en texto. El marco se compone de múltiples pasos de procesamiento. En primer lugar, los términos de dominio se extraen utilizando un método de filtrado. Posteriormente, la Desambiguación de los sentidos de las palabras (WSD) se aplica opcionalmente para determinar los sentidos de estos términos. Luego, mediante una técnica de subsunción, los conceptos resultantes se ordenan en una jerarquía. Construimos taxonomías con y sin WSD e investigamos el efecto de WSD en la calidad de las relaciones de tipo de concepto utilizando un marco de evaluación que usa una

taxonomía de oro. Encontramos que WSD mejora la calidad de la taxonomía construida en términos de la F-Measure taxonómica.

No. 9	**Título:** TaxoLearn: un enfoque semántico para el aprendizaje de la taxonomía de dominios

Autores: Emmanuelle-Anna Dietz, Damir Vandic, Flavius Frasincar

Breve descripción: Este documento propone TaxoLearn, un enfoque para la construcción automática de taxonomías de dominio. TaxoLearn es una nueva metodología que combina aspectos de los enfoques existentes, pero también contiene nuevos pasos para mejorar la calidad de la taxonomía de dominio resultante. La contribución de este trabajo es triple. Primero, empleamos un paso de desambiguación del sentido de las palabras al detectar conceptos en el texto. En segundo lugar, mostramos el uso de la agrupación jerárquica basada en la semántica con el fin de aprender taxonomía. En tercer lugar, proponemos un nuevo procedimiento de etiquetado dinámico para los grupos de conceptos. Evaluamos nuestro enfoque comparando la taxonomía generada por la máquina con una taxonomía de oro construida manualmente. Basado en un corpus de documentos en el campo de la economía financiera, TaxoLearn muestra una alta precisión para las relaciones aprendidas del concepto taxonómico de las relaciones de tipo de concepto utilizando un marco de evaluación que usa una taxonomía de oro. Encontramos que WSD mejora la calidad de la taxonomía construida en términos de la F-Measure taxonómica.

No.10	**Título:** Inducción de taxonomía usando sub secuencias hipernímicas

Autores: Amit Gupta, Hamza Harkous, Rémi Lebret, Karl Aberer

Breve descripción: Se propone un enfoque novedoso, semi-supervisado hacia la inducción de la taxonomía de dominio a partir de un vocabulario de entrada de términos semilla. A diferencia de todos los enfoques anteriores, que normalmente extraen bordes de hipernimias directos para los términos, este utiliza un nuevo marco probabilístico para extraer sub secuencias hipernímicas. La inducción de taxonomía a partir de sub secuencias extraídas se emite como una instancia del problema de flujo de costo mínimo

en un gráfico dirigido cuidadosamente diseñado. A través de experimentos, se demuestra que el enfoque propuesto supera los enfoques de inducción de taxonomía de vanguardia en cuatro idiomas. Es importante destacar que también mostramos que el enfoque es robusto a la presencia de ruido en el vocabulario de entrada.

No.11	**Título:** Aprendizaje de jerarquías de conceptos a partir de recursos textuales para la construcción de ontologías.

Autores: Ana B. Rios-Alvarado, Ivan Lopez-Arevalo, Victor J. Sosa-Sosa

Breve descripción: Las ontologías desempeñan un papel muy importante en la gestión del conocimiento y en la Web Semántica, su uso ha sido explotado en muchas aplicaciones actuales. Las ontologías son especialmente útiles porque apoyan el intercambio y el intercambio de información. El aprendizaje ontológico del texto es el proceso de derivar conceptos de alto nivel y sus relaciones. Una tarea importante en el aprendizaje de ontologías del texto es obtener un conjunto de conceptos representativos para modelar un dominio y organizarlos en una estructura jerárquica (taxonomía) a partir de información no estructurada. En el proceso de construcción de una taxonomía, la identificación de las relaciones de hiperónimos / hipónimo entre los términos es esencial. Cómo construir automáticamente la estructura apropiada para representar la información contenida en textos no estructurados es una tarea difícil. Este artículo presenta un método novedoso para obtener, a partir de textos no estructurados, conceptos representativos y sus relaciones taxonómicas en un dominio de conocimiento específico. Este enfoque construye una jerarquía de conceptos a partir de un corpus de dominio específico mediante el uso de un algoritmo de agrupamiento, un conjunto de patrones lingüísticos e información contextual adicional extraída de la Web que mejora el descubrimiento de las relaciones hipernímicas / hiponímicas más representativas. Un conjunto de experimentos se llevó a cabo utilizando cuatro cuerpos diferentes. Evaluamos la calidad de las taxonomías construidas frente a las ontologías estándar de oro, los experimentos muestran resultados prometedores.

No.12	**Título:** Inducción de ontología de dominio usando incrustaciones de palabras

Autores: Niharika Gupta, Sanjay Podder, Annervaz K M, Shubhashis Sengupta

Breve descripción: Se ha demostrado que la ontología, la conceptualización formal compartida de la información de dominio, tiene múltiples aplicaciones para modelar, procesar y comprender texto en lenguaje natural. En este trabajo, se usó vectores de palabras distribuidos de varios modelos de lenguaje recientes de *Deep Learning* para la creación de ontologías de dominios semi-automatizadas para dominios cerrados. Se cubrieron todos los aspectos principales de la Inducción de Ontología de Dominio o el aprendizaje como la identificación de conceptos, la identificación de atributos, la identificación de relaciones taxonómicas y no taxonómicas utilizando los vectores de palabras distribuidos. Los resultados preliminares mostraron que los métodos simples basados en agrupamiento que utilizan vectores de palabras distribuidos de estos modelos de lenguaje superan a los métodos que utilizan modelos como LSI en el aprendizaje de ontología para dominios cerrados.

No.13	**Título:** Generación automática de taxonomías jerárquicas a partir de texto libre utilizando algoritmos lingüísticos

Autores: Juan Lloréns, Hernán Astudillo

Breve descripción: Este artículo presenta una técnica, basada en algoritmos lingüísticos, para construir taxonomías jerárquicas a partir de texto libre. Estas jerarquías, así como otras relaciones, se extraen del texto libre mediante la identificación de estructuras verbales con significado semántico. Las taxonomías creadas se pueden usar como núcleo para la representación completa del dominio de un área de conocimiento particular. La representación del dominio podría permitir a los arquitectos de software reutilizar la información. Para probar los algoritmos, se presenta un ejemplo de dominio para ejemplificar el uso de la técnica.

1.4 Análisis de Resultados

En este epígrafe, se presentan los resultados una vez terminado el plan de ejecución, se plantea el enfoque utilizado por los investigadores, se analizan los conjuntos de datos utilizados, los algoritmos, y el resultado de cada artículo científico analizado. En los siguientes sub-epígrafes se le dará respuesta a cada una de las preguntas de investigación.

1.4.1 Enfoque de los investigadores

Para proporcionar una visión general de las temáticas que se han propuesto en los artículos, estos fueron categorizados en tres clases:

1. Construcción automática de una taxonomía de dominio a partir de los cuerpos de texto
2. Aprendizaje automatizado de taxonomías de dominio a partir de texto
3. Aprendizaje de jerarquías de conceptos para la construcción de ontologías

Se presenta un resumen de los artículos seleccionados en el epígrafe anterior de acuerdo a su clase temática. Seguido de su identificador con el prefijo **N_o**.

Para una mejor comprensión de estas categorías y los artículos seleccionados, se presenta la Tabla 2.

Tabla 2 Clase temática de los artículos

Clase Temática	Artículos relacionados
Construcción automática de una taxonomía de dominio a partir de cuerpos de texto	• No.1 No.3 No.4 No.8 • No.10 No.13
Aprendizaje automatizado de taxonomías de dominio a partir de texto	• No.2 No.5 No.6 No.7 No.9
Aprendizaje de jerarquías de conceptos para la construcción de ontologías	• No.11 No.12

Como se puede observar la clase temática más utilizada en los artículos es la “Construcción automática de taxonomía de dominio partir de cuerpos de texto” con una cantidad de 6 artículos. Como segunda clase temática más utilizada es el “Aprendizaje automático de taxonomías de domino a partir de texto” con una cantidad de 5 artículos. Las dos clases

temáticas tienen relación ya que se enfocan en formar una estructura de organización de la información que está formada por un conjunto de categorías y subcategorías, las cuales unen entidades que comparten alguna característica común.

1.4.2 Dataset utilizados

Para dar respuesta a la pregunta de investigación, se estudian los conjuntos de datos de los artículos revisados en el proceso de extracción de relaciones jerárquicas. Véase en la Tabla 3.

Tabla 3 Dataset más utilizados en los artículos

Artículo	Conjunto de datos
Repositorio RePub	N_o.1
Repositorio RePEc	N_o.1
Texto de noticia (Diario del Pueblo)	N_o.3
Biblioteca digital CiteSeerX	N_o.4
DBpedia	N_o.6
MEDLINE	N_o.2
Colección de películas de Yahoo!	N_o.7
Corpus económico	N_o.9
Lonely planet data set	N_o.12

Como se muestra en la Tabla 3 se puede llegar a la conclusión que los dataset más utilizados son Repositorio RePub y Repositorio RePEc, destacándose en el área de la salud y economía principalmente. Se destacan también DBpedia y la biblioteca digital CiteSeerX en diferentes artículos.

1.4.3 Algoritmos y métodos aplicados

Para dar respuesta a la pregunta de investigación, se analizan los artículos para extraer los métodos y los algoritmos que en estos se utilizan. Véase la siguiente Tabla *4*

Tabla 4 Algoritmos utilizados en los artículos

Algoritmos	Artículos
Algoritmo de subsunción	N_o.1 N_o.5 N_o.8
Algoritmo de agrupamiento jerárquico	N_o.5 N_o.9 N_o.11 N_o.12
Algoritmo de co-ocurrencia	N_o.3 N_o.4 N_o.6

Agrupación jerárquica aglomerada	N_o.2
Algoritmo de Similitud léxica	N_o.4
Método Propuesto	N_o.7 N_o.13

- **Algoritmo de subsunción:** construye el concepto, relaciones más amplio y más estrecho basado en la co-ocurrencia de conceptos**,** si un concepto ocurre con frecuencia al mismo tiempo que otro concepto, se crea una relación padre-hijo entre los conceptos. Puede categorizar conceptos en un tiempo relativamente corto debido a su simplicidad y, por lo tanto, es un buen método para aplicar en grandes conjuntos de datos.
- **Algoritmo de co-ocurrencia:** se basa en la frecuencia de co-ocurrencia de dos términos para determinar las dependencias entre palabras. Primero, divide las palabras en varios niveles según su frecuencia, alta frecuencia palabras en alto nivel, baja frecuencia en bajo nivel; entonces calcula la similitud entre dos palabras en niveles adyacentes, toma las dos palabras que tienen la mayor similitud para constituir una relación jerárquica: la palabra en alto nivel como término superior, el otro como término inferior.
- **Algoritmo de agrupamiento jerárquico:** todos los términos al inicio son individuales. Se van agrupan los términos de acuerdo a las distancias entre ellos creando grupos y se vuelve a calcular la distancia entre cada grupo. Los grupos más cercanos entre sí son fusionados este proceso continúa hasta que permanece un grupo.
- **La agrupación aglomerada:** es el tipo más común de agrupación jerárquica utilizada obtener agrupaciones de objetos en función de su similitud. También es conocido como AGNES (Aglomerative Nesting). El algoritmo comienza tratando cada objeto como un único grupo. A continuación, los pares de grupos se fusionan sucesivamente hasta que todos los grupos se han fusionado en un gran grupo que contiene todos los objetos (Kassambara 2018).

  Funciona de abajo hacia arriba. Es decir, cada objeto se considera inicialmente como un grupo de un solo elemento (hoja). En cada paso del algoritmo, los dos grupos que son más similares se combinan en un nuevo grupo más grande (nodos). Este procedimiento se repite hasta que todos los puntos son miembros de un solo gran clúster (raíz) (Kassambara 2018).

- **La similitud léxica:** se encarga de comparar textos para conocer el parecido entre ellos, incluso si se encuentran en diferentes idiomas, puede ayudar a las personas a encontrar información relevante. En operaciones de minería de texto tales como: la búsqueda y recuperación de información, clasificación de texto, extracción de información, agrupación de documentos, análisis de sentimientos, traducción automática, resumen de texto y procesamiento de lenguaje natural (Dwi Prasetya, Prasetya Wibawa y Hirashima 2018).

1.4.4 Distribución temporal de los artículos

Al analizar la distribución temporal de los artículos incluidos, se observó que el año 2016 se reportó la mayor cantidad de publicaciones.

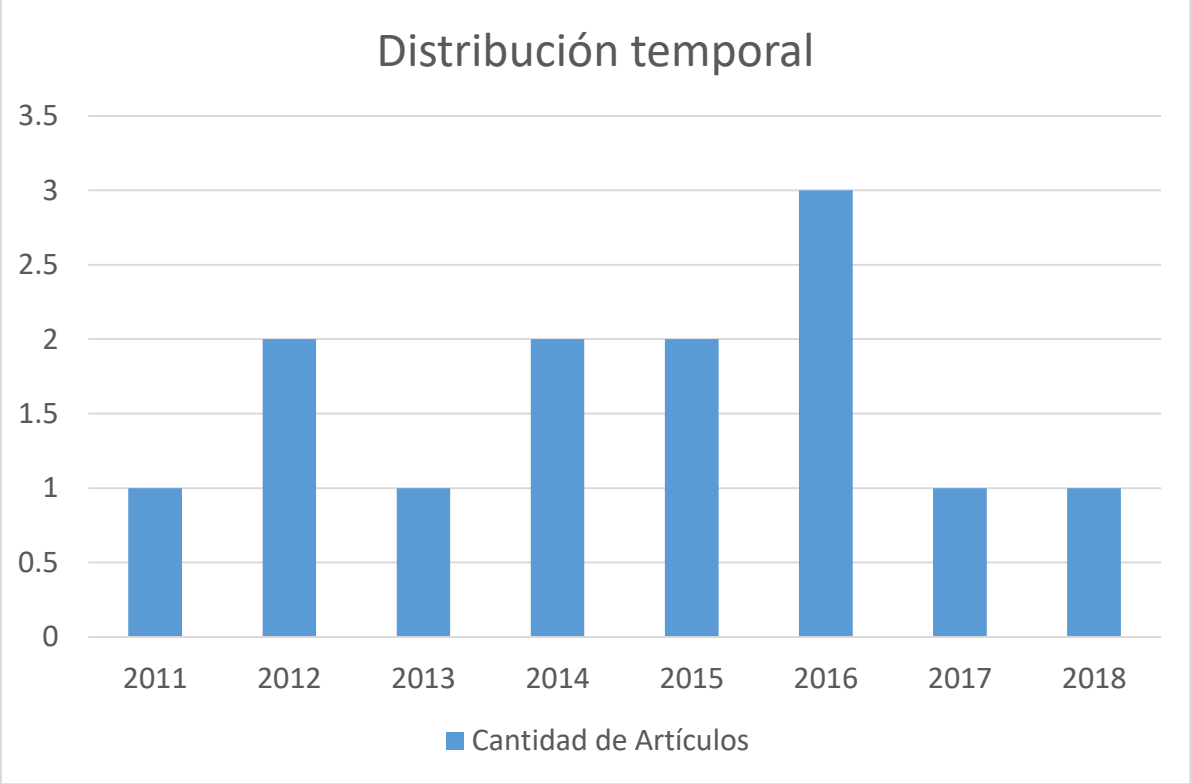

Ilustración 3 Distribución temporal de los artículos

1.5 Procedimiento para extracción de relaciones jerárquicas entre conceptos

Para continuar con la presente investigación es preciso definir un procedimiento que orientará las siguientes tres etapas:

1. Creación del corpus documental: los corpus documentales son una estructura lógica que constituye la unidad estructural en la que puedan agruparse los

diferentes textos sujetos a las tareas de minería. Contiene el conjunto de materiales lingüísticos, que conforman el objeto de estudio de la investigación. Está compuesto por uno o varios ficheros de texto, es decir, libros, revistas, artículos o cualquier otra forma con texto. Esta fase se dedica a la creación del corpus documental.

2. Pre procesamiento: en la rama de la minería de texto el pre procesamiento es una de las tareas que se llevan a cabo antes del análisis directamente con el texto para agrupar todos los documentos que serán procesados. Luego de agrupar todos los documentos en un corpus, se realiza el pre procesamiento del documento que consiste en la limpieza y trasformación de un texto para la siguiente fase de minería. De manera general se modificará el corpus a texto minúscula y se realizará el proceso de tokenización.
3. Extracción de características: es la fase en la que se aplica el procedimiento propuesto, el cual se centra en encontrar patrones lingüísticos en texto que corresponden a relaciones jerárquicas.

1.6 Conclusiones parciales

A partir del estado del arte se determina que la mayoría de las investigaciones se centran en algoritmos de co-ocurrencia y agrupamiento jerárquico. Se emplean diferentes Dataset para la realización de pruebas y experimentos. A partir del análisis de resultados en el método de revisión sistémica de la bibliografía se logra proponer un procedimiento con tres etapas para la extracción de relaciones jerárquicas entre conceptos.

Capítulo 2: Diseño e Implementación

Introducción del capítulo

En el presente capítulo se desarrolla la extracción de relaciones jerárquicas entre conceptos en texto legal, a partir de encontrar patrones sintácticos en un corpus documental. Se desarrollan los dos primeros pasos definidos en el capítulo anterior, implementando la transformación del corpus documental legal a letra minúscula, la tokenización y la extracción de características. Para la implementación se utilizó en lenguaje de programación Python y el entorno de desarrollo integrado PyCharm.

2.1 Patrones lingüísticos en un corpus

Las relaciones jerárquicas están basadas en conceptos de subordinación o superordinación entre dos conceptos, esto implica la existencia de un concepto genérico, lo que significa que un concepto hereda características de un concepto superior. Esta relación se representa mediante el enlace: *es un*, donde un concepto X es un concepto Y (García de Quesada 2001).

La identificación de relaciones jerárquicas en corpus es un trabajo que permite el entendimiento de estos. Para su identificación en texto es necesario encontrar patrones lingüísticos que reflejen la relación.

Según (Rodilla, 2019), se determinó que los patrones lingüísticos que reflejan la relación jerárquica son la unión de un concepto con las formas verbales *es* y *son*, los artículos determinados y los indeterminados. Teniendo en cuenta estos elementos, se pueden construir un grupo de sentencias que recogen las relaciones de jerarquía en un texto escrito en idioma español de la siguiente forma: concepto + forma verbal del verbo ser + artículo + concepto, lo que se puede concretar en las expresiones listadas a continuación:

1. Concepto + *es un* + concepto
2. Concepto + *es una* + concepto
3. Concepto + *son unos* + concepto
4. Concepto + *son unas* + concepto
5. Concepto + *es el* + concepto
6. Concepto + *es la* + concepto
7. Concepto *son los* + concepto
8. Concepto + *son las* + concepto

Establecer las relaciones jerárquicas como patrones sintácticos permite abordar el problema de la extracción de relaciones jerárquicas en un corpus como un problema de reconocimiento de patrones, donde el corpus corresponde al sensor y los patrones lingüísticos corresponden a las características que necesitan ser extraídas en el texto y los resultados arrojarían las relaciones encontradas.

2.2 Creación del corpus documental

El reconocimiento de patrones lingüísticos es una tarea de la minería de texto. El sensor debe ser simulado sobre la base de un texto que cumpla con las propiedades necesarias para aplicar técnicas de minería de texto. Tiene que estar limpio, estructurado y debe permitir agrupar diferentes documentos. Las estructuras más comunes dentro de este tipo de tareas son: en oraciones o en palabras. Para el caso de los patrones lingüísticos se requiere trabajar por oraciones ya que no tiene sentido detectar un patrón que se forme entre el final de una oración y el inicio de la otra si estuviese estructurado en palabras.

Los corpus documentales son una estructura lógica que constituye la unidad estructural en la que puedan agruparse los diferentes textos sujetos a las tareas de minería. Contiene el conjunto de materiales lingüísticos, que conforman el objeto de estudio de la investigación. Está compuesto por uno o varios ficheros de texto, es decir, libros, revistas, artículos o cualquier otra forma con texto (Sierra Martínez 2015).

En el marco de esta tesis el interés está en el minado de texto legal (texto que contenga normas legales), por ello, la entrada a los algoritmos de minería de texto serán corpus documentales legales. Un corpus documental legal es un corpus documental en el que cada uno de los documentos que lo integran contiene una norma jurídica o un parte de una norma jurídica.

A modo de caso de estudio, para ejecutar una prueba de conceptos que permita validar la propuesta se va a utilizar la Constitución de la República de Cuba, aprobada en referéndum popular el 24 de febrero del 2019. El corpus documental construido a partir de la Constitución contiene un único fichero con los 229 artículos presentes.

2.2 Pre procesamiento

En la última década el crecimiento de los datos en inmensas cantidades está siendo elemento clave en el actual escenario de pre procesamiento de datos. Tecnologías como

internet generan grandes cantidades de datos (texto, video, imágenes) gracias al desarrollo del almacenamiento y los recursos de red. El pre procesamiento de los datos es una tarea necesaria para el análisis de los datos.

La etapa de pre procesamiento presenta un conjunto de tareas o técnicas las cuales mejoran la calidad de texto haciendo uso de técnicas que tienen el objetivo de inicializar correctamente la entrada para los algoritmos de minería. Este tipo de técnicas son de uso obligatorio, ya que sin ellas los algoritmos de extracción de conocimiento ofrecerían resultados erróneos. En esta área se incluye la transformación de la fuente tipográfica a minúscula, limpieza de texto, lematización y tokenización.

Transformación de la fuente tipográfica a minúscula, se define como el proceso de convertir o transformar toda letra del texto a minúscula (Pino Mejías 2018).

Limpieza de texto, consiste en la eliminación de las palabras que no tienen interés para la minería de textos, como son artículos, pronombres, preposiciones, conjunciones, e incluso palabras que se usan muy frecuentemente y no ayudan a la diferenciación de un documento a otro (Pino Mejías 2018).

Tokenización es el proceso de dividir un documento de texto (o una colección de ellos) en la lista de palabras que lo conforman, mediante la identificación espacios en blanco o los signos de puntuación (Kannan y Gurusamy 2014).

Lematización consiste en la eliminación de plurales, pasados, futuros, entre otras cosas, esto con la finalidad de dejar solamente la raíz de la palabra, ésta será el lema de la palabra (Pilar Socha Díaz, Martínez Serna y Medina Mosquera 2017).

En el marco de esta tesis se propone refinar la información contenida en el corpus documental y potenciar la accesibilidad para facilitar la aplicación del algoritmo. La transformación de la fuente tipográfica en minúscula y la tokenización del corpus son las técnicas utilizadas.

La primera actividad que se realiza, es la transformación del corpus documental legal a letra minúscula. De este modo se facilita la identificación de conceptos y se evitan errores de comparación entre el corpus documental legal y la lista de conceptos. Para esta actividad se tiene el siguiente código en lenguaje de programación Python:

```
55  concepts = open('concept.txt').read()
56  corpus = open('corpus.txt').read()
57  corpus = corpus.lower()
```

Ilustración 4 Transformación de fuente tipográfica en minúscula

El algoritmo de transformación inicia con la creación de dos variables: *concept* y *corpus*. A través de la función *open()* se abren los ficheros *concept.txt* y *corpus.txt*. Luego se utiliza el método *read()* que se encarga de guardar la información contenida en cada fichero en las variables *concept* y *corpus*. De este modo se tiene guardada la lista de conceptos y el corpus documental en dos variables. Después se aplica la función *lower()* a la variable corpus, la cual transforma la tipografía del texto a minúscula.

La próxima actividad es la tokenización del corpus documental legal. Entendiéndose por tokenización al proceso de segmentación de una sentencia en unidades más simples denominadas *tokens* (Vallejo Huanga 2016). En esta acción los segmentos de texto más grandes pueden ser convertidos en oraciones y las oraciones pueden ser segmentadas en palabras.

El presente trabajo de diploma, se propone la tokenización de un corpus en una lista de palabras. Con este objetivo se utiliza la función *word_tokenize()* de la biblioteca *Natural Language Toolkit* (NLTK).

```
58  concept_tk = nltk.word_tokenize(concepts)
59  corpus_tk=nltk.word_tokenize(corpus)
```

Ilustración 5 Tokenización

Utilizando la función *word_tokenize()* se realiza la tokenización de las variables *concepts* y *corpus*. De este modo se obtienen los *tokens* de cada sentencia, los resultados se almacenan en las variables *concept_tk* y *corpus_tk*. La tokenización implica además la clasificación por parte del discurso (pos), de modo que cada palabra es un *token* y una etiqueta de parte del discurso. Así se garantiza que se tendrá una palabra y además su clasificación gramatical. Con esta finalidad se define la clase *Word* cuya implementación se muestra a continuación:

```
7   class Word(object):
8       def __init__(self, token, pos):
9           self.token = token
10          self.pos = pos
11      def __repr__(self):
12          return self.token+'('+self.pos+')'
```

Ilustración 6 Clasificación gramatical

2.3 Extracción de características

Esta etapa se centra en encontrar patrones lingüísticos, que corresponden a la característica del problema de reconocimiento de patrones. Debido a que se ha logrado establecer una estructura fija para estos patrones (concepto + forma verbal es o son + artículo + concepto) es relativamente sencillo determinar los patrones en el texto a través del uso de expresiones regulares.

Las expresiones regulares son una serie de caracteres que forman un patrón, normalmente representativo de otro grupo de caracteres mayor, de tal forma que se puede comparar el patrón con otro conjunto de caracteres para ver las coincidencias («The Python Standard Library» 2018).

Comúnmente la expresión regular no trabaja con objetos y por ende no es capaz de manipular *tokens* que estén categorizados como ocurre en el caso de los sustantivos que son de interés para encontrar patrones lingüísticos que representan jerarquías.

En este trabajo se asume como conocido el listado de conceptos (sustantivos) presente en el corpus y se requiere modificar la estructura de los *token* para añadir a cada uno de ellos la parte del habla (pos por sus siglas en inglés) a la que pertenece. Debido a que solo se requiere el conocimiento de qué palabra es un sustantivo para encontrar los patrones lingüísticos relevantes, solo se clasificarán los sustantivos y se dejarán nulas las clasificaciones del resto de las palabras en lo adelante.

```
list_tk=[]
for concept in corpus_tk:
    if concept in concept_tk:
        list_tk.append(Word(concept,'sust'))
    else:
        list_tk.append(Word(concept,'no_sust'))
```

Ilustración 7 Clasificación gramatical del corpus

En la Ilustración 7*Ilustración 7* se asume que todos los sustantivos relevantes están almacenados en la lista *concept_tk*. Se recorren todas las palabras presentes en el corpus y se les añaden la clasificación de "*sust"* si aparece en *concept_tk* y por tanto es un sustantivo, o la clasificación "*no_sust*" en el caso contrario.

Las relaciones jerárquicas entre conceptos se extraen a través de patrones lingüísticos que se implementan en el procedimiento propuesto. Para ello se realizan una serie de reglas capaces de encontrar relaciones jerárquicas. En la Ilustración 8 se definen un conjunto de reglas utilizadas en la implementación.

```
rules = [

    Rule(condition=W(pos="sust")+ W(token="es")+ W(token="un")+ W(pos="sust"),action=result),

    Rule(condition=W(pos="sust")+ W(token="es")+ W(token="una")+ W(pos="sust"),action=result),

    Rule(condition=W(pos="sust")+ W(token="es")+ W(token="la")+ W(pos="sust"),action=result),

    Rule(condition=W(pos="sust")+ W(token="son")+W(token="unos")+ W(pos="sust"),action=result),

    Rule(condition=W(pos="sust")+ W(token="son")+W(token="unas")+ W(pos="sust"),action=result),

    Rule(condition=W(pos="sust")+ W(token="es")+ W(token="el")+ W(pos="sust"),action=result),

    Rule(condition=W(pos="sust")+ W(token="son")+W(token="los")+ W(pos="sust"),action=result),

    Rule(condition=W(pos="sust")+ W(token="son")+W(token="las")+ W(pos="sust"),action=result),
]
```

Ilustración 8 Reglas gramaticales

En la variable *rules* se guardan una lista de reglas o patrones lingüísticos a identificar en el corpus. Las reglas se componen por una condición. La condición a cumplir está sujeta a un *W(pos="sust")* lo que significa que se identificará como primer elemento un sustantivo. Seguido, como segundo elemento un *W(token="es")* + *W(token="un")* lo que significa que se identifica una secuencia de *tokens* de tipo jerárquico. Como último se identifica nuevamente un *W(pos="sust")* completando el tipo de condición a identificar.

Para hacer coincidir los patrones lingüísticos en el corpus se utiliza la librería de Python re. Este módulo proporciona operaciones de coincidencias de expresiones regulares similares. Las funciones en este módulo le permiten verificar si una cadena en particular coincide con una expresión regular dada (o si una expresión regular dada coincide con una cadena en particular, que se reduce a lo mismo)(«The Python Standard Library» 2018).

Los métodos empleados para hacer posible coincidir los patrones en la lista se encuentran dentro de la clase W como se muestra en la Ilustración 9:

```
14  class W(Predicate):
15      def __init__(self, token=".*", pos=".*"):
16          self.token = re.compile(token + "$")
17          self.pos = re.compile(pos + "$")
18          super(W, self).__init__(self.match)
19
20      def match(self, word):
21          m1 = self.token.match(word.token)
22          m2 = self.pos.match(word.pos)
23          return m1 and m2
```

Ilustración 9 Método match()

Como método de la clase W se tiene al constructor __init__, el cual hace uso de la función *compile()* . El uso de *re.compile()* define un patrón de expresión regular que se usa para hacer coincidencias, en el caso del código se hace coincidir *tokens* y *pos*.

Para realizar las comparaciones, *compile*, hace uso de los métodos de comparación *match()* y *search().* En la implementación se utiliza el método *match()* y no el método *search()* ya que la comparación de los patrones se hace desde el principio y no desde cualquier posición de la cadena. De esta manera se comprueba si se encuentra el *token* y el pos que se encuentra en las reglas.

El algoritmo termina una vez que se identificaron todas las reglas en el corpus. Para ello se implementó el método *result()* el cual imprime aquellas reglas que son identificadas. El formato de salida de es de hijo – padre. Véase en la Ilustración 10

```
37  def result(x):
38      print(x[0].token+"---"+x[len(x)-1].token)
```

Ilustración 10 Método result()

2.4 Conclusiones parciales

Se creó un corpus documental legal con los 229 artículos presentes. Se realizó el pre procesamiento al copus legal, transformando su tipografía a minúscula y realizando la tokenización de cada palabra en él. A cada palabra se le asignó su clasificación como parte del discurso, y con el empleo de expresiones regulares se definieron una serie de reglas para encontrar patrones sintácticos que identifiquen en el texto relaciones jerárquicas entre conceptos.

Capítulo 3: Validación

Introducción del capítulo

En este capítulo se muestra la metodología de un experimento y los resultados alcanzados. Se realiza la matriz de confusión y se evalúan los resultados mediante las métricas de precisión y sensibilidad.

3.1 Metodología de experimentación

Para validar la propuesta de solución se diseñó un experimento con el objetivo de verificar la efectividad del procedimiento a partir de las métricas Precisión y Sensibilidad.

Sensibilidad

La sensibilidad es el número de verdaderos positivos dividido por el número de verdaderos positivos más el número de falsos negativos. Los verdaderos positivos son puntos de datos clasificados como positivos por el modelo que en realidad son positivos (lo que significa que son correctos), y los falsos negativos son puntos de datos que el modelo identifica como negativos que en realidad son positivos (incorrectos).

La sensibilidad se puede considerar a partir de la capacidad de un modelo para encontrar todos los puntos de datos de interés en un conjunto de datos. Su valor oscila entre cero y uno, mientras más cercano a uno mejor será la capacidad del modelo de encontrar puntos de interés en los datos; es decir menos sensible es el modelo (Jun Lee et al. 2019).

Precisión

La precisión se define como el número de verdaderos positivos dividido por el número de verdaderos positivos más el número de falsos positivos. Los falsos positivos son casos que el modelo señala incorrectamente como positivos que son realmente negativos. Su valor oscila entre cero y uno, mientras más cercano a uno más preciso es el modelo (Jun Lee et al. 2019).

Mientras que sensibilidad expresa la capacidad de encontrar todas las instancias relevantes en un conjunto de datos, la precisión expresa la proporción de los puntos de datos que nuestro modelo dice que eran relevantes y en realidad eran relevantes (Jun Lee et al. 2019).

Matriz de confusión

La matriz de confusión (Visa et al. 2011) es un mecanismo a partir el cual se pueden calcular de manera directa las métricas de precisión y sensibilidad. La matriz asume que en cada fila se almacena la información real del dataset y en cada columna la información obtenida aplicando los procedimientos automáticos. La estructura de matriz de dos clases, en este caso + y -, se representa de la siguiente forma:

Tabla 5: Matriz de confusión

		Propuesta	
		+	**-**
Real	**+**	**TP**	**FN**
	-	**FP**	**TN**

Cada columna de la matriz representa el número de propuestas para cada clase realizadas por el modelo, y cada fila los valores reales por cada clase. Con lo cual los conteos quedan divididos en cuatro clases, VP, FN, FP, VN, que significan lo siguiente:

TP – Verdaderos positivos: es el número de predicciones correctas para la clase +.

FN – Falsos negativos: es el número de predicciones negativa cuando realmente el valor tendría que ser positivo.

FP – Falsos positivos: es el número de predicciones positivas cuando realmente el valor tendría que ser negativo.

TN – Verdaderos negativos: es el número de predicciones correctas para la clase -.

Las métricas de Precisión y Sensibilidad se calculan de la siguiente manera:

$$Sensibilidad = \frac{TP}{(TP + FN)} \qquad Precision = \frac{TP}{(TP + FP)}$$

Para el desarrollo del experimento se utilizó una muestra de seis corpus documentales diferentes para realizar las pruebas. Estos corresponden al texto de la Constitución, la gaceta N_o.29, la gaceta N_o.30, la gaceta N_o.31 la gaceta N_o.32 y la gaceta N_o.33 respectivamente. En la Tabla 12 se recogen los corpus utilizados para cada en la columna N_o 1.

La metodología a seguir consta de las siguientes tareas:

1. Extraer las relaciones jerárquicas reales de los corpus documentales: el desarrollo del experimento requiere conocer las relaciones jerárquicas que están presentes en el corpus documental. Para lograr este resultado se realizó la extracción manual de todas las relaciones jerárquicas reales en cada uno de los corpus documentales utilizados en el experimento. En la Tabla *12* se recoge la cantidad de relaciones presentes en cada corpus en la columna N_o 2.

2. Extraer las relaciones jerárquicas generadas por el procedimiento de los corpus documentales: para cada corpus se aplica el procedimiento de extracción de relaciones jerárquicas y se registra la cantidad de relaciones encontradas. En la Tabla 12 se recoge la cantidad de relaciones presentes en cada corpus en la columna N_o 3.

3. Construir la matriz de confusión para cada corpus documental: para construir la matriz de confusión es necesario describir las clasificaciones realizadas por el procedimiento en verdaderas positivas, falsas positivas y falsos negativas. Entre las Tabla 6 y Tabla 11 se muestran las matrices de confusión para cada corpus.

4. Calcular el valor de sensibilidad y precisión para cada corpus documental: el cálculo de las métricas sensibilidad y precisión se realizará apoyándose en la matriz de confusión. Los resultados se registran en la columna N_o 4 y N_o 5 de la Tabla 12.

Construir la matriz de confusión para cada corpus documental

Tabla 6: Matriz de confusión. Corpus Constitución

		Propuesta	
		+	**-**
Real	**+**	**8**	**1**
	-	**2**	**--**

Tabla 7: Matriz de Confusión Gaceta Oficial 29

		Propuesta	
		+	**-**
Real	**+**	**9**	**2**
	-	**1**	**--**

Tabla 8: Matriz de Confusión Gaceta Oficial 30

		Propuesta	
		+	-
Real	+	10	2
	-	2	--

Tabla 9 Matriz de Confusión Gaceta Oficial 31

		Propuesta	
		+	-
Real	+	8	1
	-	2	--

Tabla 10 Matriz de Confusión Gaceta Oficial 32

		Propuesta	
		+	-
Real	+	9	1
	-	1	--

Tabla 11 Matriz de Confusión Gaceta Oficial 33

		Propuesta	
		+	-
Real	+	9	2
	-	2	--

Tabla 12: Resultados del experimento. Precisión y sensibilidad

Corpus	Relaciones reales	Relaciones procedimiento	Sensibilidad	Precisión
Constitución	10	8	0.88	0.8
Gaceta 29	11	9	0.81	0.9
Gaceta 30	12	10	0.83	0.83
Gaceta 31	10	8	0.88	0.8
Gaceta 32	10	9	0.9	0.9
Gaceta 33	11	9	0.81	0.81

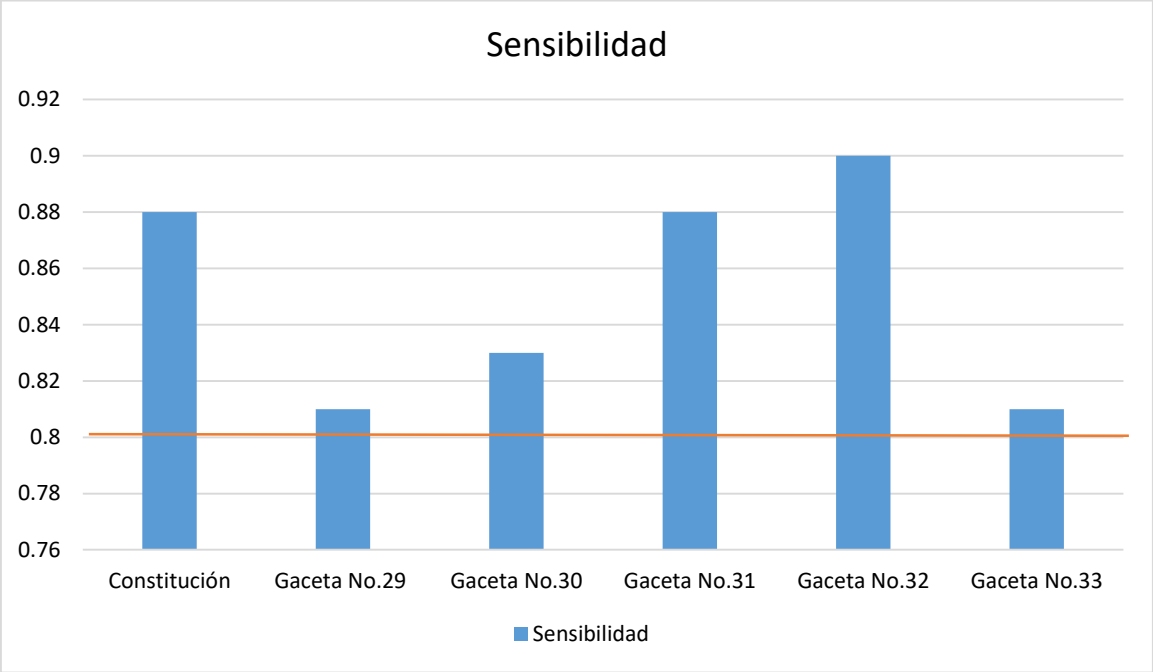

Ilustración 11 Comparación de la métrica Sensibilidad para cada uno de los corpus

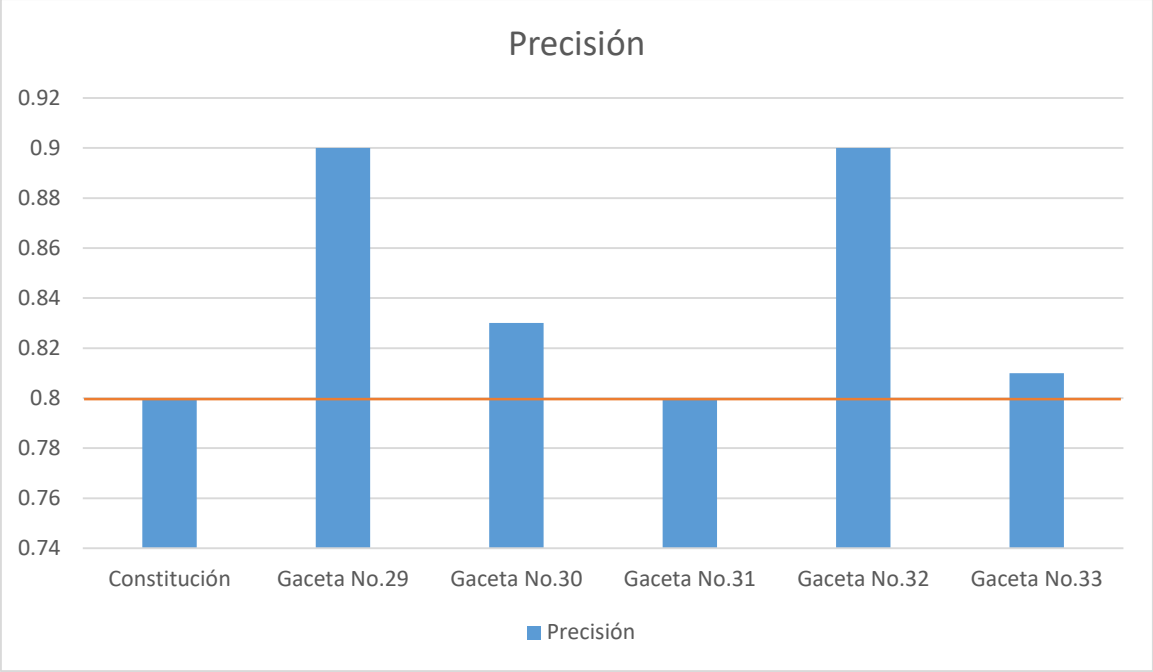

Ilustración 12 Comparación de la métrica Precisión para cada uno de los corpus

En los corpus Constitución y Gaceta N_o 31 se puede apreciar el procedimiento tuvo mejor sensibilidad que precisión, por lo que en estos casos el procedimiento tuvo una tendencia a encontrar los casos positivos.

En el corpus Gaceta N_o 29 se puede apreciar que el procedimiento tuvo mejor precisión que sensibilidad, lo que quiere decir que el procedimiento tuvo una tendencia a realizar la mayor cantidad de clasificaciones correctas.

En los corpus Gaceta N_o 30, N_o 32 y N_o 33 se puede apreciar que el procedimiento tuvo valores similares de sensibilidad y precisión, por lo que se puede decir que existe un equilibrio entre su capacidad de encontrar los casos positivos y la capacidad de realizar clasificaciones correctas.

A partir de los resultados alcanzados en los experimentos se puede observar en la Ilustración 11 y la Ilustración 12 que las métricas de sensibilidad y precisión alcanzaron valores por encima de 0.8, lo que son considerados valores adecuados en comparación con las propuestas en el estado del arte.

3.3 Conclusiones parciales

Se realizó la validación de la propuesta a través de las métricas de precisión y sensibilidad en seis corpus documentales diferentes. Los resultados alcanzados permiten asegurar que el procedimiento tiene un buen desempeño en comparación con los resultados en el estado del arte.

Conclusiones

1. A partir del estado del arte se determina que la mayoría de las investigaciones se centran en métodos o algoritmos de co-ocurrencia (subsunción). Se emplean diferentes Dataset para la realización de pruebas y experimentos en algunas de las investigaciones. A partir del análisis de resultados en el método de revisión sistémica de la bibliografía se logra proponer un procedimiento con tres etapas para la extracción de relaciones entre conceptos.
2. Se implemetó en el lenguaje Python el procedimiento formalizado en el capítulo anterior para extraer relaciones jerárquicas utilizando patrones liguísticos que reflejen la relación. Se determina que las expresiones regulares son una aproximación válida para establecer relaciones jerárquicas como patrones sintácticos demostrando que se puede abordar esta tarea como un problema de reconocimiento de patrones.
3. Las comparaciones realizadas entre los resultados de las métricas de Sensibilidad y Precisión y los corpus documentales arrojaron valores superiores a 0.80. Estos valores son consistentes a los alcanzados en el estado del arte lo que permite demostrar la validez del método.

Recomendaciones

- Realizar un trabajo regular de soporte y mantenimiento de la solución obtenida, para mejorar su funcionamiento.
- Construir una interfaz para visualizar los resultados del modelo de relación jerárquica extraída

Referencias

AGUILERA EGUÍA, R., 2014. ¿Revisión sistemática, revisión narrativa o metaanálisis? *Revista de la Sociedad Española del Dolor*, vol. 21, no. 6, pp. 359-360. ISSN 1134-8046. DOI 10.4321/S1134-80462014000600010.

BALLESTEROS MUÑOZ, M.M. y RUIZ JIMÉNEZ, A., 2011. *Informe de la Comisión de modernizacion del lenguaje jurídico*. 2011. S.l.: s.n. Gobierno de España

DWI PRASETYA, D., PRASETYA WIBAWA, A. y HIRASHIMA, T., 2018. The performance of text similarity algorithms. En: Department of Electrical Engineering, State University of Malang, Indonesia, *International Journal of Advances in Intelligent Informatics*, vol. 4, pp. 63-69. ISSN ISSN 2442-6571.

GARCÍA DE QUESADA, M., 2001. Estudios de Lingüística del Español. En: Universidad de GranadaFacultad de Traducción e InterpretaciónDepartamento de Traducción e Interpretación, ISSN 1139-8736.

HERNÁNDEZ, C., 2002. Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. *Estudios de Lingüística del Español*, no. 18. ISSN 1139-8736.

International Standar ISO 704. , 2009.

JUN LEE, T., TATBUL, N., GOTTSCHLICH, J., METCALF, E. y ZDONIK, S., 2019. Precision and Recall for Time Series. ,

KANNAN, S. y GURUSAMY, V., 2014. *Preprocessing Techniques for Text Mining*. S.l.: s.n.

KARIMI, S., POHL, S., SCHOLER, F., CAVEDON, L. y ZOBEL, J., 2010. Boolean versus ranked querying for biomedical systematic reviews. *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, pp. 58. ISSN 1472-6947. DOI 10.1186/1472-6947-10-58.

KASSAMBARA, A., 2018. Hierarchical Clustering in R: The Essentials. .

KOSELLECK, R., 2004. Historia de los conceptos y conceptos de historia. [en línea], no. 53. [Consulta: 2 mayo 2019]. ISSN 1134-2277,. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=1034819.

MANTEROLA, C., ASTUDILLO, P., ARIAS, E. y CLAROS, N., 2011. Revisiones sistemáticas de la literatura. ¿Qué se debe saber acerca de ellas? En: Universidad de La Frontera, Temuco, Chile, vol. 3, pp. 149-155.

MERCEDES GARCÍA DE QUESADA, 2001. ESTRUCTURA DEFINICIONAL TERMINOGRÁFICA EN EL SUBDOMINIO DE LA ONCOLOGÍA CLÍNICA. [en línea]. [Consulta: 30 noviembre 2018]. Disponible en: http://elies.rediris.es/elies14/index.html#indice.

NASER, A. y CONCHA, G., 2011. El gobierno electrónico en la gestión pública. *Naciones Unidas*, vol. 73. ISSN 1680-8827.

PÉREZ ZÚÑIGA, R., CAMACHO CASTILLO, O., MENA HERNÁNDEZ, E. y ARROYO CERVANTES, G., 2016. Análisis general del gobierno electrónico en México. En: Paakat: Revista de Tecnología y Sociedad, ISSN 2007-3607.

PILAR SOCHA DÍAZ, D.M., MARTÍNEZ SERNA, J.S. y MEDINA MOSQUERA, C.C., 2017. *Minería de texto histórica -colaboración al proyecto 'RevealingCooperation and Conflict Project'*. S.l.: Facultad Ingeniería de SistemasEscuela Colombiana de Ingeniería Julio Garavito.

PINO MEJÍAS, J.L., 2018. *Técnicas Estadísticas en Minería de Textos*. S.l.: Universidad de Sevilla.

Real Academia Española. [en línea], 2018. España: Disponible en: www.rae.es.

RELINQUE BARRANCA, M., 2017. El proceso de modernización del lenguaje jurídico en el Reino Unido, los Estados Unidos y España, y su reflejo en el lenguaje utilizado por los jueces. , vol. 4. ISSN 2341-3778. FITISPos international journal: public service interpreting and translation

SIERRA MARTÍNEZ, G.E., 2015. Introducción a los corpus linguisticos. *Universidad Nacional Autónoma de México*, pp. 214 pp. ISSN 978-607-029-898-1.

The Python Standard Library. [en línea], 2018. Disponible en: https://docs.python.org/2/library/re.htm.

TOVAR, M., PINTO, D., MONTES, A., GONZÁLEZ-SERNA, G. y VILARIÑO, D., 2015. Evaluación de relaciones ontológicas en corpora de dominio restringido. *Computación y Sistemas*, vol. 19, no. 1, pp. 135-149. ISSN 1405-5546. DOI 10.13053/CyS-19-1-1954.

URRA MEDINA, E. y BARRÍA PAILAQUILÉN, R.M., 2010a. La revisión sistematica y su relación con la práctica basada en la evidencia en salud. *Latino-Am. Enfermagem*,

URRA MEDINA, E. y BARRÍA PAILAQUILÉN, R.M., 2010b. Systematic Review and its Relationship with Evidence-Based Practice in Health. *Revista Latino-Americana de Enfermagem*, vol. 18, no. 4, pp. 824-831. ISSN 0104-1169. DOI 10.1590/S0104-11692010000400023.

VALLEJO HUANGA, D.F., 2016. *CLUSTERING DE DOCUMENTOS CON RESTRICCIONES DE TAMAÑO*. Trabajo Fin de Máster. S.l.: Escuela Técnica Superior de Ingeniería Informática Universidad Politécnica de Valencia.

VISA, S., RAMSAY, B., RALESCU, A. y KNAAP, E., 2011. Confusion Matrix-based Feature Selection. *CEUR Workshop Proceedings*. S.l.: s.n., pp. 120-127.