INGENIERÍA Y DESARROLLO

ARTÍCULO DE REVISIÓN / REVIEW ARTICLE
http://dx.doi.org/10.14482/inde.36.1.10949

Una revisión de la literatura sobre población de ontologías

A State-of-the-art Review About Ontology Population

Juan Carlos Blandón Andrade*
Universidad Católica de Pereira

Carlos Mario Zapata Jaramillo**
Universidad Nacional de Colombia - Sede Medellín

* Ingeniero de Sistemas, Especialista en Docencia Universitaria en la Universidad Piloto de Colombia, Magíster en Ingeniería énfasis Sistemas y Computación, Pontificia Universidad Javeriana, sede Cali, Candidato a Doctor en Ingeniería - Sistemas e Informática de la Universidad Nacional de Colombia, sede Medellín. Docente Tiempo Completo, Facultad de Ciencias Básicas e Ingeniería, Programa de Ingeniería de Sistemas y Telecomunicaciones, Universidad Católica de Pereira. *juanc.blandon@ucp.edu.co*

** Ingeniero Civil, Especialista en Gerencia de Sistemas Informáticos, Magíster en Sistemas y Doctor en Ingeniería con énfasis en Sistemas de la Universidad Nacional de Colombia. Presidente del Comité Ejecutivo del Capítulo Latinoamericano de Semat, uno de los Traductores Oficiales del libro "La Esencia de la Ingeniería de Software: aplicando el núcleo de Semat". Profesor Titular, Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión, Universidad Nacional de Colombia, Sede Medellín. *cmzapata@unal.edu.co*

Correspondencia: Juan Carlos Blandón Andrade. Universidad Católica de Pereira, Carrera 21 N.o 49-95 Bloque Aletheia. Piso 2, Oficina 31, Av. de las Américas, Pereira, Colombia. Tel. móvil. 3174591969.

Volumen 36, n.º 1
Enero-junio, 2018
ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

Resumen

El principal objetivo de las ontologías en computación es la definición de un vocabulario común para describir conceptos básicos y sus relaciones en un dominio específico. Los principales componentes de las ontologías son clases (conceptos), instancias, propiedades, relaciones y axiomas, entre otros elementos. El proceso de población de ontologías se refiere a la recepción de una ontología como entrada, para luego extraer y relacionar las instancias a cada clase de la ontología desde fuentes de información heterogéneas. En este artículo se realiza una revisión sistemática de literatura sobre la población de ontologías. Se seleccionan artículos de bases de datos especializadas y se crea una pregunta de investigación que permita dirigir la búsqueda de los artículos. Los resultados de la revisión apuntan a que la población de ontologías es un tema de interés para los investigadores. A pesar de que existen muchas técnicas para realizar el proceso, hace falta crear herramientas automáticas y con altos niveles de *precision* y *recall*.

Palabras clave: Extracción de información, máquinas de aprendizaje, ontologías, población de ontologías, procesamiento de lenguaje natural, revisión sistemática de la literatura.

Fecha de recepción: 5 de julio de 2017
Fecha de aceptación: 10 de septiembre de 2017

Abstract

The main goal of ontologies in computing is related to the definition of a common vocabulary for describing basic concepts and relationships on a specific domain. Main components of ontologies are classes—concepts—, instances, properties, relations, and axioms, among other elements. The ontology population process is intended to receive an ontology as input in order to extract and relate the instances of each ontology class from heterogenous information sources. In this paper we perform a systematic state-of-the-art review about ontology population. We select papers from specialized databases and we create a research question for driving paper search. The results of our review points out ontology population as an interesting topic for researchers. Even though we have several techniques for driving the process, fully automated tools are still missing and we also miss high levels of precision and recall.

Keywords: Information extraction, machine learning, natural language processing, ontologies, ontology population, systematic state-of-the-art review.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

I. INTRODUCCIÓN

En inteligencia artificial, los esfuerzos se enfocan en comprender cómo funciona la mente humana y, con base en ello, se intenta construir entidades inteligentes a fin de sintetizar y automatizar tareas intelectuales [1], [2]. En este entorno, las ontologías se pueden usar con el propósito de mejorar la búsqueda de información, realizar inferencia computacional, etc. [3]. Una definición de Borst señala: "las ontologías son especificaciones formales de una conceptualización compartida" [4]. Las ontologías son medios para modelar formalmente la estructura de un sistema, es decir, las entidades y relaciones relevantes que surgen desde su observación y son útiles para un propósito particular [5]. Las ontologías se pueden conceptualizar, es decir, determinar el conjunto de clases, objetos, relaciones y restricciones que caracterizan un dominio; por ejemplo, el dominio *viaje*, puede contener clases como locación, transporte y avión, también instancias como la ciudad de Buenos Aires y un avión de Matricula AA7462. La clase, entonces, hace referencia a un conjunto de objetos, los cuales son instancias de esta; las instancias se definen como los objetos del dominio de interés; las relaciones se refieren a relaciones binarias entre individuos; y las restricciones se expresan por medio de axiomas, esto es, condiciones que se cumplen siempre y permiten realizar las inferencias [6]. El proceso de población de ontologías consiste en insertar instancias de conceptos y relaciones dentro de una ontología existente y se puede realizar desde fuentes de información estructuradas, semiestructuradas y libres [7]. Para realizar el proceso de población existen muchas herramientas manuales y semiautomáticas, pero no se detectan en la revisión herramientas completamente automáticas.

El método de revisión sistemática de literatura constituye una manera de evaluar e interpretar toda la información disponible que sea relevante respecto de un interrogante de investigación particular, en un área temática o fenómeno de interés [8]. En este artículo se realiza una revisión sistemática de la literatura especializada en la población de ontologías, con el método de Kitchenham [8]. Los resultados muestran que el tema de población de ontologías es un tema de actualidad y de interés para la comunidad científica porque existe el interés de crear nuevas herramientas automáticas y así realizar la población de ontologías [9].

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

Este artículo se ordena de la siguiente forma: en la Sección II se presenta información para comprender el proceso de población de ontologías. En la III, se presenta la metodología utilizada para la revisión, y en la IV se desarrolla el proceso de revisión sistemática de literatura. En la V se realiza una síntesis de datos y análisis de resultados, y en la VI se presentan las conclusiones del tema revisado.

II. POBLACIÓN DE ONTOLOGÍAS

El proceso de población de ontologías se puede realizar de forma semiautomática o automática. El proceso se debe centrar en enriquecer, por medio de instancias de clases o relaciones, una ontología ya existente, la cual se encuentra vacía. La importancia de poblar ontologías se debe a que las ontologías que existen en los sistemas informáticos se deben actualizar de manera constante y porque, al existir muchas ontologías, se abarcan muchos dominios específicos. Los objetos o instancias de las ontologías contribuyen en muchas tareas como, por ejemplo, la realización de búsquedas de información en Internet [10]. El proceso a seguir a fin de poblar una ontología requiere un corpus, es decir, un conjunto de textos y un motor de extracción de instancias que se encarga de localizar las instancias de clases y relaciones en el corpus. Luego, se debe procesar el corpus utilizando el motor de extracción, de forma que se puedan localizar conceptos dentro del texto, e inmediatamente después se crea una lista con posibles instancias de conceptos y relaciones, las cuales, después, en un proceso adicional, se utilizan para poblar la ontología [11].

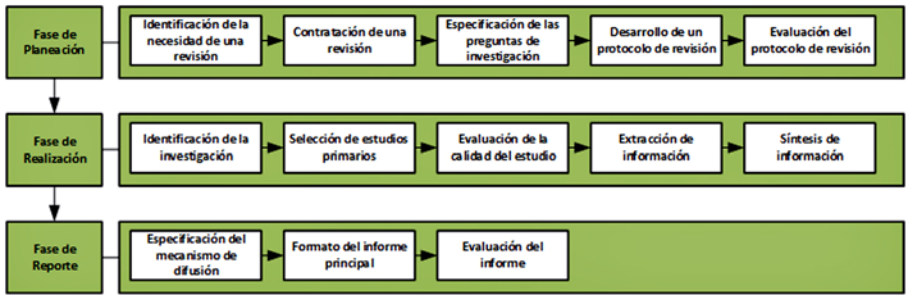
Existen métodos que permiten poblar ontologías. Los métodos estadísticos, basados en la distribución de las palabras en el corpus, constituyen una primera aproximación. Usualmente, se basan en métodos estocásticos y probabilísticos que permiten resolver ambigüedad en frases largas y procesar gramáticas que pueden generar muchos análisis posibles [12], [13], [14]. Los métodos de extracción de información se basan en el análisis del lenguaje natural para luego extraer información de interés de forma automática. Entre las técnicas más representativas se encuentran el reconocimiento de entidades nombradas y la resolución de correferencia [15]. Los métodos de procesamiento de lenguaje natural (NLP, por sus siglas en inglés) tienen el propósito de lograr el análisis, la representación y la generación de textos, para lo cual se utilizan una serie de herramientas computacionales que buscan el procesamiento lingüístico a nivel morfológico, sintáctico y semántico [16],

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

[17], [18]. Los métodos basados en aprendizaje de máquinas (ML, *machine learning*) se refieren a la creación de algoritmos que sean capaces de generalizar comportamientos y reconocer patrones con información suministrada en forma de ejemplos [19], [20]. Los métodos basados en reglas utilizan un conjunto de reglas manuales que asocian características del texto a entidades. Las reglas se usan para apoyar la toma de decisiones en clasificación, regresión y tareas de asociación [21]. Finalmente, los métodos híbridos [22] realizan combinaciones entre los métodos existentes, de tal manera que se puedan optimizar los recursos de cómputo y aumentar la efectividad.

III. METODOLOGÍA

La metodología que se utiliza en este trabajo (véase la Fig. 1), se basa en la revisión sistemática de literatura [8].



Fuente: propia con base en [8]

Figura 1. Proceso de revisión sistemática de literatura

Se compone de tres fases: 1. Fase de planeación; 2. Fase de realización; y 3. Fase de reporte. En la primera fase se determina la necesidad de la revisión, las preguntas de investigación, el protocolo a seguir y la evaluación del protocolo. En la segunda se debe identificar la investigación mediante la definición de cadenas de carateres, con el fin de llevar a cabo la búsqueda en las bases de datos especializadas; a partir de esto, se deben seleccionar unos estudios primarios con los resultados de la búsqueda y así relizar la evaluación de la calidad del estudio y la extracción de la información relevante de los artículos seleccionados. Finalmente, de acuerdo con unos criterios, se sintetiza la información, lo cual en este artículo se realiza con una Tabla. En la tercera fase se escriben los resultados de la revisión y se comunican a la comunidad científica, en este caso mediante este artículo.

IV. PROCESO DE REVISIÓN

A. Fase de planeación

Identificación de las necesidades de la revisón

Se pretende identificar los aspectos más relevantes en la población de ontologías, lo que incluye técnicas, dominios y niveles de automatización.

Contratación de la revisión

Los investigadores interesados tienen la experiencia necesaria para llevar a cabo el estudio; por lo tanto, no es necesario contratos para desarrollar la revisión.

Especificación de las preguntas de investigación

La pregunta de investigación es: ¿Cuáles son los aspectos más relevantes que guían la población de ontologías, desde el punto de vista de criterios tales como las técnicas a emplear, el tipo de dominio en que se aplica y los niveles de automatización que tienen las diferentes propuestas?

Desarrollo del protocolo

Los métodos de la estrategia para el desarrollo de la revisión se definen y se aplican de acuerdo con el proceso de revisión sistemática de literatura.

Evaluación del protocolo

Para el caso de la población de ontologías, el protocolo se aplica inicialmente a tres artículos [23] - [25]. Esa muestra permite ajustar el protocolo para su aplicación al universo del tema.

B. Fase de realización

Identificación de la investigación

Las cadenas de búsqueda definidas en el caso de la construcción de la revisión sobre población de ontologías son las siguientes:

Cadena 1: "Ontology population".

Cadena 2: "Automatic ontology population".

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

Selección de estudios primarios

Identificación de fuentes de estudio

La búsqueda de la literatura sobre la población de ontologías se hace mediante el uso de fuentes digitales, tales como: ACM Digital Library, Ebsco, IEEE Xplore Digital Lybrary, Science Direct, bdigital Repositorio Institucional UN y Scopus. Asimismo, otros estudios de la comunidad de ontologías y fuentes digitales se incluyen en este proceso de identificación. Los resultados del ejercicio se muestran en la Tabla 1.

Selección de estudios

Los criterios de inclusión de la revisión de población de ontologías se relacionan con la población, el enriquecimiento de las ontologías y el uso de métodos semiautomáticos o automáticos y que realicen el proceso desde diferentes fuentes de información. Los criterios de exclusión de la síntesis de población de ontologías son los estudios relacionados con métodos desarrollados para un idioma diferente al inglés, así como los métodos manuales de población de ontologías. El proceso se realiza en tres iteraciones considerando: 1. El título del estudio; 2. Resúmen y palabras clave; 3. Conclusiones. Los resultados de aplicar los criterios de inclusión y exclusión se muestran en la Tabla 2.

Tabla 1. Resultados de busquedas en fuentes digitales

Fuente de estudio	Estudios seleccionados		
	Cadena 1	Cadena 2	TOTAL
ACM Digital Library	80	45	125
bdigital Repositorio Institucional UN	120	6	126
Ebsco	40	24	64
IEE Xplore Digital Library	180	110	290
ScienceDirect	90	75	165
Scopus	20	14	34
TOTAL	530	274	804

Fuente: propia

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

Tabla 2. Resultados de la aplicación de los criterios de inclusión y exclusión

Fuente de estudio	Estudios seleccionados		
	Cadena 1	Cadena 2	TOTAL
ACM Digital Library	10	7	17
bdigital Repositorio Institucional UN	15	8	23
Ebsco	15	5	20
IEE Xplore Digital Library	40	24	64
ScienceDirect	25	10	35
Scopus	15	6	21
TOTAL	120	60	180

Fuente: propia

Evaluación de la calidad del estudio

En el caso de esta revisión, estos criterios se aplican como filtros adicionales a fin de evitar sesgos y asegurar la inclusión de estudios relevantes.

Extracción de información

Extraer cualquier tipo de información desde el lenguaje natural es una labor muy útil porque, de esta forma, se puede tratar de automatizar procesos y evitar tareas largas y complejas. La extracción de instancias de una clase o población de ontologías se inició aproximadamente en la década de los ochenta y, actualmente, es un tema de interés para la comunidad científica. A continuación, se presenta una serie de métodos que intentan solucionar el problema de población de ontologías. Esos enfoques se encuentran en los artículos científicos seleccionados de acuerdo con el protocolo establecido.

Abbott [23] propone la extracción de sustantivos para definir tipos de datos y presenta una técnica para desarrollar programas informáticos desde descripciones informales, es decir, desde textos con una estructura básica sin reglas sintácticas y semánticas rigurosas. Esas descripciones deben ser precisas y se trabajan para el idioma inglés. La técnica demuestra cómo derivar tipos de datos (categorías de seres o cosas) desde sustantivos comunes, variables desde verbos y atributos y estructuras de control desde sus equivalentes en inglés. La principal contribución de este trabajo es la relación propuesta entre sustantivos comunes y tipos de datos. La idea es capturar estos elementos y transformarlos en un programa escrito en ADA.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

El artículo también presenta una discusión de cómo hacer la transformación entre sintagmas nominales, tipos de datos y objetos.

Contreras [24] propone una arquitectura de adquisición de contenido para la web semántica que provee un marco conceptual, el cual permite desarrollar sistemas de procesamiento de contenido web y mapear contenidos semánticamente anotados y, a su vez, el procesamiento por medio de agentes de software y aplicaciones de web semántica. Asimismo, genera una ontología automáticamente, extrae instancias desde textos, asigna instancias a clases y extrae valores de atributos. La arquitectura acepta como entrada archivos TXT para luego utilizar técnicas de procesamiento de lenguaje natural.

Pasca [25] propone un método para adquirir entidades nombradas en categorías arbitrarias utilizando patrones léxico-sintácticos. También hace referencia al refinamiento de una consulta en una búsqueda web. Las categorías de nombres recogidas se fusionan eficazmente y luego resumen las relaciones semánticas detectadas en los documentos iniciales. Se extraen en pares, por ejemplo: NombreNavegador y Google. Luego, se utilizan los patrones léxico-sintácticos para extraer las instancias. Esos patrones se obtienen de documentos de entrenamiento automáticamente, los cuales constituyen las reglas que se deben cumplir antes de hacer la extracción de las instancias.

Geleijns *et al.* [26] hacen mención al hallazgo de la relación entre dos instancias, por ejemplo "Britney Spears" y "Cristina Aguilera". Se usa la co-ocurrencia para encontrar el vecino más cercano, es decir, las instancias más relacionadas. Las tres clasificaciones presentadas en el documento se pueden combinar. Finalmente, se podrá decir que esas instancias pertenecen a la categoría "artistas pop". El método consiste en que se tienen dos conjuntos dados: uno de instancias y otro de categorías. Luego, se busca en Google cada pareja instancia-categoría. Luego de verificar la co-ocurrencia en la web, se utilizan métodos para establecer si esa instancia pertenece a esa categoría.

Geleijns *et al.* [27] presentan un método que utiliza patrones hechos a mano y los construyen a la medida para las clases y relaciones consideradas. Los patrones se consultan en Google, donde los resultados se utilizan para buscar otras instancias. Las instancias que se encuentran se utilizan dentro de los patrones, de tal forma que el algoritmo puede poblar la ontología, al

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

utilizar unas pocas instancias de una ontología parcial dada. También, se debe construir una ontología parcial en forma de tupla. Luego, se alimenta el sistema con pocas instancias de clases y relaciones escritas a mano, y el sistema busca en la web qué instancias diferentes puede encontrar. Así, se logran poblar las clases y las relaciones.

De Boer *et al.* [28] presentan una propuesta para la extracción de instancias de relaciones, por ejemplo, la relación artista-estilo artista. También, se trabaja un dominio de fútbol. Teniendo como base la ontología, se pueden extraer las instancias de las relaciones desde un corpus, que en este caso es la web. Específicamente, el método necesita dos conjuntos de instancias de clases Ci y Cj. Después, toma una instancia i de Ci, con la cual se eligen los documentos desde la web. Posteriormente, se utilizan todas las instancias de Cj para saber cuántas existen en cada uno de los documentos encontrados. De esta forma, se obtienen las instancias de las relaciones entre Ci y Cj.

Yoon *et al.* [29] proponen un método automático para población de ontologías con datos en formato estructurado. Las instancias se extraen desde páginas web utilizando *wrappers* o sentencias mediante técnicas de procesamiento de lenguaje natural. El método requiere una ontología e instancias semilla y se extraen instancias desde documentos semiestructurados o no estructurados. Su precision es del 98 %.

Talukdar *et al.* [30] presentan un algoritmo de propagación de etiquetas semi-supervisado, el cual utiliza un gráfico denominado "*Adsorption*". Después, utilizan fuentes estructuradas y no estructuradas de información con el fin de adquirir clases etiquetadas y las instancias en un dominio abierto. Así, construyen un grafo donde cada nodo representa una instancia o una clase y existe un puente entre un nodo instancia y un nodo clase, siempre y cuando la instancia pertenezca a esa clase. Esta herramienta requiere una clase y cinco instancias semilla que se utilizan para evaluar el texto y extraer instancias, las cuales permiten construir el grafo, el cual finalmente encuentra otras instancias que ayudan a etiquetar las clases a las que pertenecen.

Manine *et al.* [31] presentan una arquitectura para integrar ontologías en el dominio biomédico. La entrada del sistema debe partir de documentos muy especializados a fin de poderlo entrenar y, después, se logra extraer instancias de la web de forma automática. Se utilizan técnicas de extracción

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

de información y aprendizaje de máquina. Finalmente, se obtienen buenos niveles de *precision* y *recall*.

Ruiz-Martínez *et al.* [32] presentan un *framework* que procesa textos mediante herramientas de procesamiento de lenguaje natural. Realizan las pruebas con la ontología denominada "*Travel.owl*", la cual se descarga desde la página de Protégé y a la cual realizan algunos cambios. Utilizan una página web para extraer instancias pertenecientes a la clase "Hotel".

Danger *et al.* [33] utilizan una ontología de referencia, reconocedores de entidades y desambiguadores de entidades con el propósito de crear y combinar adecuadamente un conjunto inicial de instancias. El análisis exhaustivo y la experimentación de la propuesta se llevan a cabo en una variedad de escenarios de aplicación. En el proceso, se define una ontología en formato OWL (*Web Ontology Language*) que tiene conceptos y relaciones. Existen lexicones que describen las reglas léxicas, para después identificar conceptos y relaciones en el texto. Luego de extraer entidades, se define un conjunto de instancias inicial que usa reglas de inferencia y se genera, finalmente, un conjunto de instancias complejas que definen semánticamente el documento de acuerdo con la ontología dada.

Faria *et al.* [34] presentan una propuesta para semiautomatizar la población de ontologías desde textos. Utilizan técnicas de NLP y extracción de información (EI) para clasificar instancias de ontologías. El proceso tiene dos fases. En la primera fase se realiza la extracción y la clasificación de instancias que, a su vez, incluye tres tareas. La primera tarea es el análisis de corpus, mediante el cual se estructura el corpus y se realizan tres actividades (análisis morfoléxico—que identifica las categorías gramaticales—, reconocimiento de nombre de entidades—que identifica nombre de personas, organizaciones o lugares—, y la identificación de correferencia—que identifica las correferencias de pronombres y correferencias nominales—). La segunda tarea es la especificación de reglas de clasificación y extracción, en las que el usuario se basa en la ontología y los patrones léxico-sintácticos definidos previamente, a fin de generar un conjunto de reglas de extracción. La tercera tarea es la extracción y la clasificación de instancias en las que se utilizan las reglas de la tarea previa. La segunda fase es la representación de instancias, en la cual se realizan dos tareas (el refinamiento de instancias y la población de la ontología). Los autores mencionan que están evaluando las ventajas de combinar técnicas NLP con *soft computing*.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

Rauf *et al.* [35] mencionan la creación de un *framework*, el cual tiene como base que los documentos de requisitos contienen instancias de estructuras lógicas (plantillas) tales como casos de uso, requisitos funcionales, reglas del negocio, etc. El sistema permite hacer dos cosas: ingresar estructuras lógicas al sistema para luego, a partir de documentos RTF, extraer las instancias de esas estructuras.

Schlaf *et al*. [36] presentan un *framework* para aprender categorías y sus instancias mediante características contextuales. Su *framework* se basa en el uso de textos en lenguaje natural como ejemplos de entrenamiento. Consta de tres pasos: 1. Aprendizaje de reglas desde los textos de ejemplo; 2. Selección de las reglas de alta calidad, por medio de dos filtros (el número de ocurrencias de la regla y dos características dependientes); y 3. Identificación de nuevas instancias de la categoría teniendo en cuenta las reglas de filtrado, que se basan en cuatro categorías (nombre, apellido, profesión y ciudad), que permiten ubicar palabras (profesor, ingeniero o abogado) y así detectar automáticamente que son instancias de una clase (en este caso, profesión).

IJntema *et al.* [37] proponen un método que utiliza reglas para aprender instancias de ontologías desde textos, con el fin de contribuir con el proceso de población de ontologías. Las reglas léxico-semánticas explotan las capacidades de inferencia de las ontologías. Este sistema necesita ontologías del dominio a trabajar, para luego definir patrones léxico-semánticos. Con estas herramientas se procesa el documento con el fin de evaluar qué instancias se pueden extraer de páginas web de noticias.

Ruiz-Martinez *et al.* [38] presentan un método para la población de ontologías del dominio biomédico. El sistema se alimenta con una ontología de dominio biológico, enriquecida con instancias de textos en lenguaje natural. El proceso tiene tres capas: 1. Las ontologías de nivel superior, las cuales definen las relaciones semánticas básicas a mapear, dentro de recursos que permiten etiquetar roles semánticos; 2. La ontología del dominio a poblar, que se relaciona con el modelo ontológico; y 3. La ontología del dominio poblada, que se puebla mediante los modelos ontológicos y recursos lingüísticos. Los autores utilizan procesamiento de lenguaje natural y logran extraer instancias en el dominio biomédico, así como asignar instancias a clases automáticamente. Finalmente, obtienen buenos resultados de *recall* y *precision* en ese dominio.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

Faria *et al.* [39] proponen un proceso para la población automática de ontologías desde textos. El proceso aplica procesamiento de lenguaje natural y técnicas de extracción de información para adquirir y clasificar instancias de ontologías. Es un paso inicial hacia la utilización de una ontología que permite generar reglas automáticamente desde ella, extraer instancias desde textos y clasificarlas en las clases de la ontología. Las reglas se generan a partir de ontologías de cualquier dominio a fin de lograr el objetivo de independencia del dominio. El proceso tiene tres fases: 1. Identificación de instancias candidatas; 2. Construcción de un clasificador; y 3. Clasificación de instancias. El sistema se prueba en el dominio legal y el de turismo.

De Araujo *et al.* [40] proponen un método que permite poblar ontologías con instancias de eventos. La principal contribución se relaciona con la exploración de la flexibilidad de reglas lingüísticas y la representación del dominio de conocimiento mediante su manipulación e integración con un sistema de razonamiento. Los documentos a procesar se deben tratar con un programa de análisis lingüístico profundo (PALAVRAS), y luego representar en OWL con el modelo de datos POWLA. Posteriormente, se puedan usar reglas lingüísticas y los conceptos de la ontología del dominio. La gran cantidad de información OWL que se genera sirve para hacer inferencias lógicas y como salida se obtiene la ontología con las reglas lógicas y la ontología del dominio; con todos esos elementos se utiliza un razonador que permite extraer las instancias.

Sadoun *et al.* [41] presentan un enfoque que se centra en la identificación de instancias de propiedades mencionadas en textos, el uso de reglas de extracción que se obtienen desde rutas sintácticas recurrentes y la vinculación de términos que denotan conceptos e instancias de propiedades. El proceso requiere una ontología del dominio, al igual que un corpus de entrenamiento. Con esos elementos se definen reglas de extracción, a fin de extraer instancias de propiedades mencionadas en los textos. Las reglas explotan conocimiento léxico, sintáctico y semántico. Finalmente, los autores demuestran que con esa información pueden extraer instancias de clases implícita o explícitamente.

Ríos [42] presenta la generación de ontologías, lo que incluye axiomas de clases e instancias de manera automática a partir de textos en idioma inglés. Se utilizan técnicas como NLP, algoritmos de agrupamiento y EI. Se obtienen ontologías que incluyen conceptos, relaciones jerárquicas, axio-

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

mas e individuos. Finalmente, se construyen ontologías que se comparan manualmente con la ontología de referencia *goldstandard*. En cuanto a las instancias, se obtienen resultados de *precision* del 56,8 %.

Faria *et al.* [43] presentan un método de dominio independiente y ajustan algunas partes del enfoque con el fin de mejorar los valores de *precision* y *recall*. Como entrada, es necesario un corpus y una ontología vacía para realizar la población. El método cuenta, básicamente, con tres tareas: 1. Identificación de instancias candidatas, mediante técnicas NLP y modelos estadísticos SIR; 2. Construcción de un clasificador, por medio de una herramienta de extracción de información (EI) y otra de ML, en la que la función de la tarea consiste en seleccionar clases, propiedades y relaciones, además de seleccionar los disparadores y generar reglas; y 3. Clasificación de instancias, que necesita como entradas el clasificador y el corpus anotado, para luego utilizar NLP y ML para asignar las instancias a las clases, propiedades y relaciones y, finalmente, obtener como resultado la ontología poblada. Por otra parte, a fin de lograr la independencia del dominio, generan un clasificador desde la ontología procesada, de modo que, sin importar la ontología de entrada, se puebla desde documentos en lenguaje natural. Las pruebas se realizan utilizando una ontología en un dominio de leyes (legal) y otra con un dominio en turismo; al final se obtienen buenos resultados.

Lima *et al.* [44] presentan un sistema que se basa en programación lógica inductiva (PLI) y que, automáticamente, induce reglas de extracción simbólicas que se utilizan para poblar un dominio de ontología con instancias de clases. El método explota la similitud semántica y tiene cuatro fases: 1. La recuperación del corpus, donde se recuperan oraciones desde la web para construir un corpus de trabajo (patrones *Hearst*), además de que el usuario elige una clase desde una ontología de dominio y después el sistema recupera algunos documentos con la elección del usuario; 2. El preprocesamiento del texto, en el que se realiza un análisis léxico-sintáctico por medio del analizador de *Stanford* y se mide semánticamente la distancia entre la clase y las instancias candidatas mediante *Wordnet*; 3. El mejoramiento de las reglas que se deben aplicar y las cuales se encuentran en la base de conocimiento; y 4. La aplicación de las reglas y la extracción de las instancias.

Nederstig *et al.* [45] presentan un proceso semi-automático para poblar ontologías, especialmente valores de atributos relacionados con información

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

de productos desde textos semiestructurados o almacenes web. Inicialmente, se utiliza una ontología predefinida y compatible con la ontología *GoogRelation*, que se utiliza para el dominio de comercio electrónico. Luego, el método contiene un léxico junto con patrones para clasificar productos, mapear propiedades y crear valores de instancias.

Colace *et al.* [46] presentan un sistema para el aprendizaje y población de ontologías que combina metodologías estadísticas (*Latent Dirichlet Analysis*, LDA) y semánticas (*WordNet*). El sistema recibe como entrada un conjunto de documentos desde diferentes fuentes web o desde colecciones relacionadas específicas para un dominio de interés, que se clasifican de acuerdo con temas disjuntos semánticamente y así producir una ontología terminológica. Incluye dos componentes principales: 1. Aprendizaje de ontologías, que usa LDA sobre los documentos de entrada y produce una representación WWP (*Weighted Word Pairs*), la cual contiene los conceptos del dominio más relevantes y sus valores de co-ocurrencia (relaciones) en el conjunto que se analiza; 2. Refinamiento de ontologías, que utiliza propósito general o bases de datos léxicas de dominio específico, refina los conceptos descubiertos previamente, explota sus relaciones léxicas (por ejemplo, relaciones taxonómicas *is_a*), agrega conceptos ocultos y produce el esquema de la ontología final y la población de la misma. Algunos experimentos se llevan a cabo con documentos TREC-8 (*Text RETrieval Conference*), a fin de demostrar la efectividad del enfoque propuesto.

Santos *et al*. [47] presentan *Apponto-Pro*, que es la unificación de varias propuestas. Proponen un proceso incremental para lograr la construcción y posterior población de una ontología de aplicación. El sistema es capaz de generar todos los elementos de la ontología, tales como clases, taxonomía, relaciones no taxonómicas, instancias, propiedades y axiomas en un archivo de extensión OWL. El proceso se compone de seis fases: 1. Recolección de objetivos, en la que se requiere un experto que entregue al sistema un conjunto de objetivos; 2. Construcción de una ontología base, la cual tiene como entrada un conjunto de objetivos que el usuario alimenta manualmente y entrega como salida una ontología base con clases, taxonomía, propiedades y axiomas; 3. Aprendizaje de clases y relaciones taxonómicas, la cual aprende otras clases y relaciones taxonómicas mediante un algoritmo que extrae elementos a partir de un corpus y la ontología base; 4. Aprendizaje de relaciones no taxonómicas, en la cual se aplican técnicas estadísticas y

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

de procesamiento de lenguaje natural y se realizan otras actividades como la anotación del corpus; 5. Población de ontologías, en la que se realiza la identificación, extracción y clasificación de instancias de relaciones no taxonómicas y propiedades de una ontología desde el corpus anotado y se utilizan técnicas NLP y EI a fin de obtener como salida la ontología poblada; y 6. Inserción de axiomas, en la cual se necesita la ontología poblada de la fase anterior y utiliza reglas de inferencia en programación lógica inductiva con el propósito de lograr extraer nuevas reglas en lógica de primer orden para nuevas relaciones e instancias. Finalmente, se entrega una ontología totalmente poblada y con nuevas reglas. Las pruebas se realizan bajo el dominio derecho de familia.

Kordjamshidi *et al.* [19] presentan un *framework* para poblar instancias de relaciones en ontologías desde el lenguaje natural. Se incluye un modelo estructurado de ML, el cual tiene muchas variables y restricciones. Una estrategia que se utiliza es subdividir el problema en subproblemas. A su vez, cada subproblema se resuelve por medio de la programación lineal. Se utilizan conceptos de relaciones espaciales tales como trayectoria, puntos de referencia e indicadores espaciales. Para las pruebas del *framework*, utilizan datos de los métodos de evaluación *SemEval*-2012 y *SemEval*-2013.

Blandón [48] presenta un método computacional automático que utiliza técnicas de extracción de información y procesamiento de lenguaje natural, a fin de extraer instancias de una clase y generar como resultado un archivo con una ontología completa en formato OWL, utilizando la herramienta GATE (*General Architecture for Text Engineering*). La entrada del sistema es un texto escrito en lenguaje natural en formato Word o PDF. Después, realiza varios procesos para separar las palabras en *tokens*, utiliza diccionarios, divide el texto en oraciones, agrega categorías gramaticales a las palabras, etiquetado semántico y resolución de correferencia. Luego, realiza el diseño e implementación de un sistema de reglas con patrones sintácticos genéricos que sirven de apoyo para etiquetar las diferentes entidades ontológicas como clases, instancias, relaciones y atributos. Las reglas se implementan en lenguaje JAPE (*Java Annotation Patterns Engine*) que es propio de GATE. Después, se envía la información sobre etiquetas a un archivo con extensión "XML". Finalmente, implementa un proceso en lenguaje Java, el cual se encarga de recibir todas las etiquetas con entidades ontológicas, para luego generar la ontología en un archivo con extensión OWL que se puede editar en Protégé [49].

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

Síntesis de la información

Con el fin de sintetizar los diferentes enfoques encontrados en la revisión de literatura, se utiliza la Tabla 3, en la cual se tienen en cuenta los criterios más importantes para evaluar cada uno de los métodos, a saber: 1. Nivel de actualidad de los documentos revisados para la población de ontologías; 2. Tipo de documento (estructurado E, semi-estructurado SE o libre L) que los autores utilizan para la población de ontologías; 3. Tipo de dominio (general G o específico Esp) utilizado para realizar la población de ontologías; 4. Tipo de técnicas utilizadas para realizar la población de ontologías (NLP, EI, ML u otras); 5. Nivel de automatización del método desarrollado (automático A o semiautomático SA); 6. Nivel de los criterios *precision* (el número de instancias correctamente extraídas sobre el número instancias extraidas) y *recall* del método evaluado (el número de instancias bien extraídas sobre el número de instancias en el corpus evaluado [34], [50]); y 7. Construcción de la ontología desde cero.

Tabla 3. Síntesis de los trabajos sobre población de ontologías

Propuesta	Criterios de Comparación						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Contreras 2004 [24]	I	g	NLP	A	--	--	Sí
Pasca 2004 [25]	SE, I	g	Patrones Léxico Sintácticos	SA	88,0	--	No
Geleijnse *et al.* 2005 [27]	SE	g	Patrones Hearst	SA	78,0	93,8	No
De Boer *et al.* 2007 [28]	SE	Esp	Co-ocurrencia en la web	SA	--	--	No
Yoon *et al.* 2007 [29]	I	Esp	NLP	SA	95,2	--	No
Talukdar *et al.* 2008 [30]	E, I	g	Algoritmo ADSORPTION	SA	77,4	--	No
Manine *et al.* 2008 [31]	SE, E	Esp	ei, ml	A	89,6	89,3	No
Ruiz-Martínez *et al.* 2008 [32]	SE, I	Esp	NLP	A	93,2	94,9	No
Danger *et al.* 2009 [33]	SE	g	Proceso no Monolítico	A	90,0	90,0	No
Faria *et al.* 2011 [34]	L	Esp	NLP, ei	SA	95,0	75,0	No
Schlaf *et al.* 2012 [36]	L	Esp	ML	A	84,9	45,0	No
IJntema *et al.* 2012 [37]	L	Esp	Patrones Léxico Sintácticos y Léxico Semánticos	SA	80,0	70,0	No
Ruiz-Martinez *et al.* 2012 [38]	L	Esp	NLP	A	79,6	69,0	No
Faria *et al.* 2012 [39]	L	Esp	NLP, EI	A	81,9	82,0	No
De Araujo *et al.* 2013 [40]	L	Esp	EI, NLP	A	98,0	91,5	No

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

Propuesta	Criterios de Comparación						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sadoun *et al.* 2013 [41]	L	Esp	ML y Reglas de Extracción	A	95,0	63,0	No
Ríos 2013 [42]	L	Esp	NLP, algoritmo de agrupamiento, EI	A	56,8	--	Sí
Faria *et al.* 2013 [43]	L	g	NLP, EI	A	87,3	86,5	No
Lima *et al.* 2014 [44]	SE	g	Programación Lógica Inductiva, ML, NLP	A	94,0	59,0	No
Nederstig *et al.* 2014 [45]	SE	Esp	Léxico y Patrones	SA	96,0	89,0	No
Colace *et al.* 2014 [46]	L	g	Metodologías estadísticas y semánticas	A	--	--	Sí
Santos *et al.* 2014 [47]	L	Esp	Ciclo Incremental por Objetivos	SA	--	--	Sí
Kordjamshidi *et al.* 2015 [19]	L	g	ML	A	--	--	Sí
Blandón 2017 [48]	L	g	NLP, EI	A	94	89,56	Sí

(1) Tipo texto procesado, **(2)** Dominio, **(3)** Técnicas usadas, **(4)** Nivel de Automatización, **(5)** % Nivel de *precision*, **(6)** % Nivel de *recall*, **(7)** ¿Genera Ontología?

Fuente: propia

C. Fase de reporte

Especificación de mecanismos de difusión

En este caso, la revisión se escribe como una Sección de una Tesis Doctoral. La difusión se realiza mediante este artículo y mediante la defensa de la Tesis Doctoral.

Formato del informe principal

Corresponde al formato de esta revista y el de la Tesis Doctoral.

Evaluación del informe

Tres jurados con doctorado aprobaron la Tesis Doctoral con su contenido, incluyendo la revisión que se consigna en este artículo. Dos jurados adicionales revisaron el artículo mismo.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

V. RESULTADOS Y DISCUSIÓN

Los resultados de la revisión sistemática de literatura demuestran que el tema de población de ontologías es un tema pertinente y de actualidad para la comunidad científica. También, resaltan que, para realizar un buen proceso de población de ontologías, se debe tener en cuenta que existen diferentes fuentes de información desde donde se extraen elementos. Además, en la creación de nuevos métodos para población de ontologías, estos se deben enfocar en la realización del proceso desde tipos de textos en formato libre.

Muchos de los métodos encontrados permiten realizar el proceso de extracción de instancias desde dominios específicos, tales como dominio biomédico, turismo, arqueología, cine, fútbol, finanzas, legal, *e-commerce*, etc. Otros métodos mencionan que realizan la extracción desde dominios generales, sin embargo, a excepción de Blandón [48] —quien menciona que realizan pruebas con doce dominios diferentes—, en los demás trabajos sólo presentan, a lo sumo, dos dominios diferentes. Según la revisión de literatura, los métodos se deben enfocar en dominios generales, lo que significa que el método debe poblar cualquier ontología sin importar el dominio de aplicación.

Durante la revisión se evidenció que se viene experimentando con diferentes técnicas para resolver el problema de población de ontologías. Entre estas se encontraron patrones léxico-sintácticos, procesos no monolíticos, patrones *Hearst*, algoritmos de agrupamiento, programación lógica inductiva, métodos estadísticos y co-ocurrencia en la web, entre otros. Las técnicas que más se utilizan hasta hoy se enfocan en EI, NLP y ML. También, se encontró que algunos autores prefieren realizar una combinación entre dos o más métodos.

En cuanto al nivel de automatización de los métodos encontrados, se puede mencionar que los primeros métodos de población de ontologías fueron manuales, pero realmente son muy costosos porque requieren un ingeniero de conocimiento que esté actualizando las ontologías constantemente y, con esto, tampoco se garantiza que las instancias sean las actuales, debido a la gran cantidad de documentos existentes. Por consiguiente, aparecen los métodos semiautomáticos que requieren la calibración de algunos parámetros por parte de un humano, pero pueden evaluar muchos más documentos y en menor tiempo. Los métodos automáticos buscan evaluar muchos documentos en poco tiempo y sin la necesidad de intervención de

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

un ser humano; actualmente los desarrollos de métodos de población de ontologías se dirigen a la completa automatización del proceso.

A fin de calcular el nivel del criterio *precision,* se deben contar las instancias encontradas correctamente y se deben dividir entre el número de instancias extraídas; en la mayoría de los métodos este valor se encuentra por encima del 80 %. En el caso del criterio *recall*, se divide el valor de intancias correctamente extraídas sobre las instancias posibles de hallar en el documento; los valores encontrados en los métodos estudiados en esta revisión fueron muy variados, porque algunos trabajos no presentan este valor, otros se encuentran entre 50 % y 90 %, y muy pocos trabajos suben de ese valor.

Finalmente, se evaluó si los autores experimentan sobre la posibilidad de realizar el proceso de aprendizaje de ontologías y, a su vez, la población de ontologías, y todo ese proceso automáticamente, más específicamente generar una ontología desde cero. Un 75 % de los trabajos estudiados en esta revisión no genera la ontología automáticamente y directamente desde el texto. Por otra parte, se encontró que un 25 % de los métodos presentados generan la ontología, pero no muestran sus respectivos valores de los criterios *precision* y *recall* en relación con la población de ontologías. Sólo Ríos [42] presenta un *precision* de 56,8 %, y Blandón [48] presenta *precision* de 94 %, *recall* de 89,56 % y genera la ontología desde cero.

VI. CONCLUSIONES

En este artículo se desarrolló una revisión sistemática de literatura sobre los autores y sus respectivos métodos para realizar el proceso de población de ontologías. Para esto, se aplicó un método de revisión sistemática de la literatura [8], con los productos de trabajo necesarios que permitieran fundamentar la seriedad de la revisión. Los resultados se agruparon en una Tabla que muestra los criterios más importantes de cada propuesta y se discutieron en una Sección del artículo.

La importancia de la población de ontologías radica en que las ontologías se deben actualizar constantemente, porque al existir muchas también existirán muchas específicas, y los objetos que se encuentren pueden ayudar en varias tareas, por ejemplo, en la búsqueda de información en Internet.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

La revisión muestra que, si bien existen métodos semiautomáticos, es necesaria la creación de herramientas automáticas, con las que se puedan extraer instancias u objetos desde textos sin la necesidad de alimentar el sistema mediante parámetros o reglas, es decir, se requiere que el proceso sea automático con buenos niveles de *precision* y *recall*. Algunos autores mencionan que realizan el proceso de aprendizaje y población de ontologías de forma automática, aunque la mayoría de ellos no presentan los resultados de los criterios de *precision* y *recall*. La literatura evidencia la necesidad de crear métodos de población de ontologías que sean de dominio general y no específico.

Los métodos más utilizados para realizar el proceso de población de ontologías son los métodos estadísticos, la extracción de información, el procesamiento de lenguaje natural, el aprendizaje de máquina, los métodos basados en reglas y, finalmente, los métodos híbridos en los que se realizan combinaciones entre ellos.

Se puede concluir que el proceso de población de ontologías es un tema de interés para la comunidad científica, debido a que las ontologías son estructuras de información con las que se pueden realizar inferencias computacionales, las cuales permiten descubrir conocimiento oculto en grandes cantidades de información en la web, lo cual es importante para los sistemas de software en la actualidad. La importancia radica en que se requiere explotar esa información, y es muy costoso que un ingeniero de conocimiento esté alimentado constantemente las ontologías existentes, por lo que se hace necesario crear sistemas que realicen la población de ontologías de forma automática desde diferentes fuentes de información.

REFERENCIAS

[1] S. J. Rusell, P. Norvig, Inteligencia artificial. Un enfoque moderno. 2ª ed. España: Pearson Prentice Hall, 2004.

[2] G. Luger, Artificial intelligence: structures and strategies for complex problem solving, 5ª ed. Edinburgh Gate, Inglaterra: Addison-Wesley, 2005.

[3] M. Uschold, R. Jasper, "A framework for understanding and classifying ontology applications", en Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5), Estocolmo, Suecia, 1999, pp. 12-24.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

[4] W. Borst, Construction of engineering ontologies. Enschede, The Netherlands: Centre for Telematica and Information Technology, University of Tweenty, 1997.

[5] S. Staab, R. Studer, Handbook on ontologies. 2ª ed. Londres: Springer, 2004.

[6] A. Gómez Pérez, O. Corcho, M. Fernández López. Ontological Engineering. 1ª ed. Londres: Springer-Verlag, 2004.

[7] R. Gil, M. J. Martin-Bautista, "SMOL: a systemic methodology for ontology learning from heterogeneous sources," *Journal of Intelligent Information Systems*, vol. 42, n.º 3, pp. 415-455, 2014. doi:10.1007/s10844-013-0296-x

[8] B. Kitchenham, "Procedures for performing systematic reviews", RU, Keele University TR/SE-0401/NICTA Technical Report 0400011T.1, vol. 33, 2004.

[9] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review", *Information and Software Technology*, vol. 51, n.º 1, pp. 7-15, 2009.

[10] J. M. Ruiz Martínez, "Metodología para la población automática de ontologías: aplicación en los dominios de medicina y turismo", *Sociedad Española para el Procesamiento del Lenguaje Natural*, vol. 48, pp. 123-126, 2012.

[11] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, E. Zavitsanos, "Ontology Population and Enrichment: State of the Art", en Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, G. Paliouras, C. D. Spyropoulos, and G. Tsatsaronis Eds., Springer Berlin Heidelberg, 2011, pp. 134-166.

[12] G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, vol. 24, n.º 5, pp. 513-523, 1988. doi: 10.1016/0306-4573(88)90021-0

[13] J. E. Gómez Balderas, J. Á. Vera Félix, O. A. Olivas Zazueta, "Métodos estadísticos en procesamiento de lenguaje natural y su uso en alineación de los corpus paralelos", Instituto Politécnico Nacional. Centro de Investigación en Computación, Ciudad de México, *Tech. Rep.*, 217, 2006.

[14] C. Sammut, G. I. Webb, Eds., "Statistical Natural Language Processing", en Encyclopedia of Machine Learning, Boston, MA: Springer US, 2010, pp. 916-924.

[15] H. Cunningham, "Information Extraction, Automatic", en Encyclopedia of Language & Linguistics. 2ª ed. Oxford: Elsevier, 2006, pp. 665-677.

[16] R. Mitkov, The Oxford Handbook of Computational Linguistics. 1ª ed. Oxford, GB: Oxford University Press, 2003.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

[17] A. Clark, C. Fox, S. Lappin, Eds., The handbook of computational linguistics and natural language processing, vol. 57. Malden, MA, EE. UU.: Wiley-Blackwell, 2010.

[18] S. Linckels, C. Meinel, "Natural Language Processing", en E-Librarian Service, Berlin: Springer Berlin Heidelberg, 2011, pp. 61-79.

[19] P. Kordjamshidi, M.-F. Moens, "Global machine learning for spatial ontology population", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 30, pp. 3-21, 2015. doi: 10.1016/j.websem.2014.06.001

[20] P. Cimiano, Ontology learning and population from text-algorithms, evaluation and applications. Karlsruhe, Alemania: Springer US, 2006.

[21] W. Duch, "Rule-Based Methods", Department of Informatics, Nicolaus Copernicus University, Polonia, 2010.

[22] R. C. Wang, "Language-independent class instance extraction using the web," Tesis doctoral, School of Computer Science Carnegie Mellon University, Pittsburgh, 2009.

[23] R. J. Abbott, "Program Design by Informal English Descriptions", *Communications of the ACM*, vol. 26, n.º 11, pp. 882-894, 1983.

[24] J. Contreras, "Incremento crítico del conocimiento de la web semántica mediante poblado automático de ontologías", Tesis doctoral, Facultad de Informática Universidad Politécnica de Madrid, Madrid, 2004.

[25] M. Pasca, "Acquisition of categorized named entities for web search", en Proceedings of the thirteenth ACM international conference on information and knowledge management, Nueva York, 2004, pp. 137-145. doi: 10.1145/1031171.1031194

[26] G. Geleijnse, J. Korst, V. de Boer, "Instance classification using co-occurrences on the web", en Proceedings of the ISWC 2006 Workshop on Web Content Mining with Human Language Technologies, Atenas, 2006, pp. 3-12.

[27] G. Geleijnse, J. H. Korst, "Automatic Ontology Population by Googling", en Proceedings of the 17th Belgium-Netherlands Conference on Artificial Intelligence, Bruselas, 2005, pp. 120-126.

[28] V. de Boer, M. van Someren, B. J. Wielinga, "A redundancy-based method for the extraction of relation instances from the web", *International Journal of Human Computer Studies,* vol. 65, n.º 9, pp. 816-831, 2007. doi: 10.1016/j.ijhcs.2007.05.002

[29] H.-G. Yoon, Y. J. Han, S.-B. Park, S.-Y. Park, "Ontology Population from Unstructured and Semi-Structured Texts", en Sixth International Conference on Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007, Luoyang, Henan, 2007, pp. 135-139. doi: 10.1109/ALPIT.2007.30

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

[30] P. P. Talukdar, J. Reisinger, M. Paşca, D. Ravichandran, R. Bhagat, F. Pereira, "Weakly-supervised acquisition of labeled class instances using graph random walks", en Proceedings of the Conference on Empirical Methods in Natural Language Processing, Seattle, 2008, pp. 582-590.

[31] A. P. Manine, E. Alphonse, P. Bessieres, "Information extraction as an ontology population task and its application to genic interactions," en 20th IEEE International Conference on Tools with Artificial Intelligence, 2008. ICTAI '08, Dayton, OH, 2008, vol. 2, pp. 74-81. doi: 10.1109/ICTAI.2008.117

[32] J. M. Ruiz-Martinez *et al.*, "Populating ontologies in the etourism domain", en IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08, Sydney, NSW, 2008, vol. 3, pp. 316-319. doi: 10.1109/WIIAT.2008.278

[33] R. Danger, R. Berlanga, "Generating complex ontology instances from documents", *Journal of Algorithms*, vol. 64, n.º 1, pp. 16-30, 2009. doi: 10.1016/j.jalgor.2009.02.006

[34] C. Faria, R. Girardi, "An information extraction process for semi-automatic ontology population", en Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011, Springer Berlin Heidelberg, 2011, pp. 319-328.

[35] R. Rauf, M. Antkiewicz, K. Czarnecki, "Logical structure extraction from software requirements documents", en Requirements Engineering Conference (RE), 2011 19th IEEE International, Trento, 2011, pp. 101-110. doi: 10.1109/RE.2011.6051638

[36] A. Schlaf, R. Remus, "Learning Categories and their Instances by Contextual Features", en Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC, Estambul, 2012, vol. 3, pp. 1235-1239.

[37] W. IJntema, J. Sangers, F. Hogenboom, F. Frasincar, "A lexico-semantic pattern language for learning ontology instances from text", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 15, pp. 37-50, 2012. doi: 10.1016/j.websem.2012.01.002

[38] J. M. Ruiz Martínez, R. Valencia García, R. Martínez Béjar, A. Hoffmann, "BioOntoVerb: A top level ontology based framework to populate biomedical ontologies from texts", *Knowledge-Based Systems*, vol. 36, pp. 68-80, 2012. doi: 10.1016/j.knosys.2012.06.002

[39] C. Faria, R. Girardi, P. Novais, "Using domain specific generated rules for automatic ontology population", en 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), Kochi, 2012, pp. 297-302. doi: 10.1109/ISDA.2012.6416554.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

[40] D. A. De Araujo, S. J. Rigo, C. Muller, R. Chishman, "Automatic Information Extraction from Texts with Inference and Linguistic Knowledge Acquisition Rules", en 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, 2013, vol. 3, pp. 151-154. doi: 10.1109/WI-IAT.2013.171

[41] D. Sadoun, C. Dubois, Y. Ghamri-Doudane, B. Grau, "From Natural Language Requirements to Formal Specification Using an Ontology", en 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI), Herndon, 2013, pp. 755-760. doi: 10.1109/ICTAI.2013.116

[42] A. B. Ríos, "Obtención de axiomas en el aprendizaje de ontologías," Tesis doctoral, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Victoria, Tamaulipas, México, 2013.

[43] C. Faria, I. Serra, R. Girardi, "A domain-independent process for automatic ontology population from text", *Science of Computer Programming*, vol. 95, pp. 37-50, 2013. doi: 10.1016/j.scico.2013.12.005

[44] R. Lima, H. Oliveira, F. Freitas, B. Espinasse, "Ontology population from the web: an inductive logic programming-based approach", in 2014 11th International Conference on Information Technology: New Generations (ITNG), Las Vegas, 2014, pp. 473-478. doi: 10.1109/ITNG.2014.60

[45] L. J. Nederstigt, S. S. Aanen, D. Vandic, F. Frasincar, "FLOPPIES: A Framework for Large-Scale Ontology Population of Product Information from Tabular Data in E-commerce Stores", Decision Support Systems, vol. 59, pp. 296-311, 2014. doi: 10.1016/j.dss.2014.01.001

[46] F. Colace, M. De Santo, L. Greco, F. Amato, V. Moscato, A. Picariello, "Terminological ontology learning and population using latent Dirichlet allocation", *Journal of Visual Languages & Computing*, vol. 25, n.º 6, pp. 818-826, 2014. doi: 10.1016/j.jvlc.2014.11.001

[47] S. Santos, R. Girardi, "Apponto-Pro: An incremental process for ontology learning and population", en 2014 9th Iberian Conference on Information Systems and Technologies (CISTI), Barcelona, 2014, pp. 1-6. doi: 10.1109/CISTI.2014.6876966

[48] J. C. Blandón A., "Extracción de instancias de una clase desde textos en lenguaje natural independientes del dominio de aplicación", Tesis doctoral, Universidad Nacional de Colombia, Facultad de Minas, 2017.

[49] H. Knublauch, R. W. Fergerson, N. F. Noy, and M. A. Musen, "The Protégé OWL plugin: an open development environment for semantic web applications", en The Semantic Web-ISWC 2004, S. A. McIlraith, D. Plexousakis, F. van Harmelen, Eds., Springer Berlin Heidelberg, 2004, pp. 229–243. doi: 10.1007/978-3-540-30475-3_17

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)

[50] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, n.º 1, pp. 37-63, 2011.

ISSN: 0122-3461 (*impreso*)
2145-9371 (*on line*)