Thaqhaña: Un Motor de Búsqueda Inteligente Basado en Recursos Semánticos

Felipe Cujar-Rosero, Estudiante Investigador[1], David Pinchao Ortiz, Estudiante Investigador[1], Ricardo Timarán Pereira, Ph.D[1], y Mateo Guerrero Restrepo, Msc[1].
[1]Universidad de Nariño, Colombia, felipecujar@udenar.edu.co, santipinchi@udenar.edu.co, ritimar@udenar.edu.co jimaguere@udenar.edu.co

Abstract- This paper presents the final results of the research project that aimed to compare the performance of a Semantic Search Engine using an Ontology and a model trained with Machine Learning algorithms that are: Word2vec and Doc2vec to support the semantic search of the research projects of the Research System of the Universidad de Nariño, with respect to a manual search engine. For the construction of THAQHAÑA, as this engine is called, a methodology was used that includes the following stages: appropriation of knowledge; installation and configuration of tools, libraries and technologies; collection, extraction and preparation of research projects; design and development of the Semantic Search Engine; and comparison and final testing between THAQHAÑA and the manual search engine. The main results of the work were the following: a) the complete construction of the Ontology with classes, object properties (predicates), data properties (attributes) and individuals (instances) in Protegé, SPARQL queries with Apache Jena Fuseki and the respective coding with Owlready2 using Jupyter Notebook with Python within the virtual environment of anaconda; b) the successful training of the model for which Machine Learning and specifically Natural Language Processing algorithms such as: SpaCy, NLTK, Word2vec and Doc2vec were used, this was also done in Jupyter Notebook with Python within the virtual environment of anaconda and with Elasticsearch; c) the creation of THAQHAÑA managing, unifying and integrating the queries for the Ontology and for the Machine Learning model with the Word2vec and Doc2vec algorithms; and d) the tests showed that THAQHAÑA was successful in all the searches performed as its results were satisfactory compared to the manual search engine.

Keywords-- Semantic Search Engine, Manual Search Engine, Ontology, Machine Learning, Word2vec, Doc2vec, Research Projects.

I. INTRODUCCIÓN

Los motores de búsqueda semánticos le facilitan el trabajo al usuario, son eficientes en la búsqueda ya que encuentran resultados en función del contexto, proporcionando así información más exacta acerca de lo que se busca, ofreciendo una cantidad de resultados más sesgada, facilitando la labor de filtrar los resultados por parte del usuario. Es de esta forma como estos motores de búsqueda interpretan las búsquedas de los usuarios haciendo uso de algoritmos que simbolizan comprensión, entendimiento e inteligibilidad, ofreciendo resultados precisos de manera rápida y de este modo reconociendo el contexto correcto para las palabras o sentencias de búsqueda. No es más que un motor de búsqueda semántico aquel que realiza la búsqueda fijándose en el significado del grupo de palabras que están escritas [1].

En la Web Semántica la información se ofrece con un significado bien definido, permitiendo a ordenadores y personas trabajar de forma cooperativa. La idea que existe detrás de la Web Semántica es tener datos en la Web definidos y enlazados de manera que puedan ser usados de forma más efectiva para un descubrimiento, una automatización, una integración y una reutilización entre diferentes aplicaciones. El reto de la Web Semántica es ofrecer el lenguaje que exprese tanto datos como reglas para razonar sobre los datos y además permita que las reglas sobre cualquier sistema de representación del conocimiento sean exportadas a la Web, aportando un importante grado de flexibilidad y "frescura" a los sistemas de representación de conocimiento centralizados tradicionales, que se vuelven sumamente agobiantes, crecen rápidamente de tamaño y se vuelven inmanejables. Diferentes sistemas web pueden utilizar diferentes identificadores para un mismo concepto; así, un programa que quiera comparar o combinar información entre dichos sistemas tiene que conocer qué términos significan lo mismo; idealmente, el programa debería tener una forma de descubrir los significados comunes de cualquier base de datos que encuentre. Una solución a este problema es usar un elemento vital en la Web Semántica; colecciones de información denominadas Ontologías [2].

De igual forma se sabe que la gran cantidad de información textual disponible en la red, junto con el aumento de la demanda por parte de los usuarios, hace necesaria la existencia de sistemas que permitan un acceso a aquella información de interés de una forma eficiente y efectiva, ahorrando así tiempo en su búsqueda y consulta [3]. Entre las técnicas existentes para lograr esa eficiencia y efectividad, y a su vez para proporcionar acceso o

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2021.1.1.596
ISBN: 978-958-52071-8-9 ISSN: 2414-6390

facilitar la gestión de información de documentos de texto se encuentran técnicas de Machine Learning, usarlas es altamente conveniente, esto se puede evidenciar en un elevado número de aplicaciones en diferentes ámbitos [3].

Entre algunos trabajos estudiantiles a nivel internacional, relacionados con el tema de esta investigación están: el de Camacho Rodríguez [4] en su trabajo de grado para optar el título de Ingeniería de Telemática propone incorporar una Ontología semántica en la plataforma LdShake para la selección de patrones educativos. Este trabajo fue desarrollado en la Universidad Pompeu Fabra- UPF de Barcelona, España en el año 2013. En este trabajo se analiza la eficiencia de utilizar Ontologías para mejorar considerablemente los resultados y a la vez ganar velocidad en la búsqueda [4]. Amaral [5] presenta un buscador semántico para el idioma portugués donde se hace uso de herramientas de Procesamiento de Lenguaje Natural y un corpus léxico multilingüe donde se evalúan las consultas del usuario, para la desambiguación de palabras polisémicas se hace uso de pivotes mostrados en la pantalla con los diferentes significados de la palabra donde el usuario escoge el sentido con el que desea realizar la consulta. Aucapiña y Plaza [6] en su tesis para la obtención del Título de Ingeniero en Sistemas proponen un buscador semántico universitario para la Universidad de Cuenca en Cuenca, Ecuador en el año 2018, donde se describe de forma detallada como son llevadas a cabo diversas etapas para la consecución del prototipo del buscador semántico siguiendo metodologías probadas, y en ciertos casos siendo soportadas por procesos automatizados [6]. Umpiérrez Rodríguez [7] en su trabajo final de grado en Ingeniería Informática denominado "SPARQL Interpreter" en la Universidad de Las Palmas de Gran Canaria, desarrollado en el año de 2014, expone como SPARQL Interpreter afronta el problema de la comunicación entre un lenguaje de consulta y una base de datos específica [7].

Entre algunos proyectos estudiantiles sobre búsquedas semánticas desarrollados en Colombia están: el de Bustos Quiroga [8] quien desarrolló su tesis de grado en la Maestría en Ingeniería de Sistemas y Computación denominada: "Prototipo de un sistema de integración de recursos científicos, diseñado para su funcionamiento en el espacio de los datos abiertos enlazados para mejorar la colaboración, la eficiencia y promover la innovación en Colombia". En esta tesis del año 2015 en la Universidad Nacional de Colombia se trabaja con Web Semántica en datos enlazados, para mejorar la integración en intemporalidad entre aplicaciones y facilitar el acceso a la información a través de modelos unificados y formatos de datos compartidos [8]. Moreno y Sánchez [9] en su trabajo de grado para optar el título de Ingeniero de Sistemas y Computación proponen un prototipo de motores de búsqueda semántico aplicado a la búsqueda de libros de Ingeniería de Sistemas y Computación en la biblioteca Jorge Roa Martínez de la Universidad Tecnológica de Pereira. Este trabajo fue desarrollado en el año 2012. Este prototipo se desarrolló partiendo de los fundamentos teóricos existentes y, del análisis que se llevó a cabo acerca de las tecnologías que se involucran, como son los agentes inteligentes de software, las Ontologías implementadas en lenguajes como RDF y XML, y demás herramientas de desarrollo [9].

A nivel regional, en la Universidad de Nariño, Benavides y Guerrero [10] desarrollaron el proyecto de trabajo de grado para optar el título de Ingeniero de Sistemas, en el año 2013, denominado "UMAYUX: un modelo de software de gestión de conocimiento soportado en una Ontología dinámica débilmente acoplada con un gestor de base de datos para la Universidad de Nariño" cuyo objetivo fue convertir el conocimiento que en ese entonces era tácito, dentro de los procesos académicos y administrativos de la Universidad de Nariño, en conocimiento explícito, de este modo se permite recopilar, estructurar, almacenar información y transformarla en conocimiento mediante el uso de Ontologías de dominio específico que a cada unidad académica o dependencia administrativa se le permitió construir y acoplar al modelo [10].

Actualmente en el Sistema de Investigaciones de la Universidad de Nariño en la VIIS, el ingeniero encargado, al momento de realizar búsquedas para estas investigaciones de tipo: trabajo de grado, proyectos estudiantiles y proyectos docentes, comenta que existe demora en los procesos, él narra que dichos procesos no son óptimos, que a veces es difícil encontrar lo que se quiere, en muchas ocasiones no ha logrado encontrar lo requerido. Esto indica que no existe ni eficiencia ni efectividad garantizada al momento de realizar búsquedas sobre las investigaciones, puesto que la búsqueda se está realizando de forma manual. Es decir que al tener un sistema de búsqueda manual de la información que ni siquiera tiene el calificativo de motor de búsqueda sintáctico, se deduce que la información no tiene una estructura clara para ser presentada y que los procesos son ineficientes en las búsquedas. Esto lleva a la conclusión de que es totalmente necesario construir el motor de búsqueda semántico inteligente, dado que si el problema persiste, a medida que la información vaya aumentando, las búsquedas serán más tediosas y dispendiosas, adicionalmente con el motor de búsqueda semántico inteligente la cantidad de población de usuarios que realicen búsquedas sobre determinado tema será mayor y estará satisfecha por encontrar los resultados deseados.

II. METODOLOGÍA

La metodología empleada para el trabajo comprende las siguientes etapas: apropiación del conocimiento; instalación y configuración de herramientas, bibliotecas y tecnologías; recolección, extracción y preparación de los

proyectos de investigación; diseño y desarrollo del motor de búsqueda semántico; comparación y pruebas finales entre THAQHAÑA y el motor de búsqueda manual.

III. Arquitectura

La arquitectura de THAQHAÑA se puede evidenciar en Fig. 1.

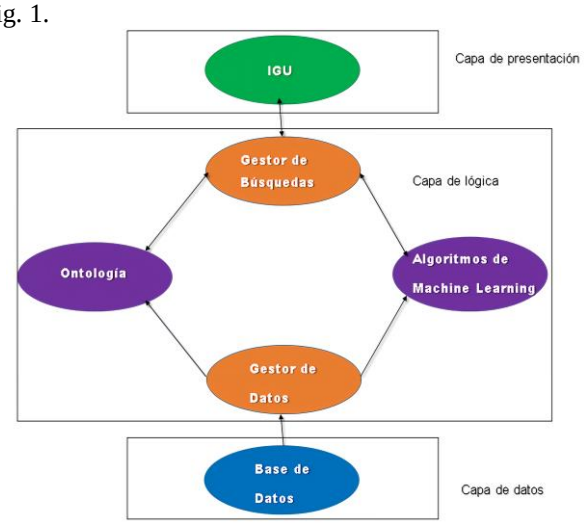

Fig. 1 Arquitectura de THAQHAÑA

A. Capas de la arquitectura del motor de búsqueda

Las capas de la arquitectura de THAQHAÑA, son las siguientes:

1) Capa de Presentación

En esta capa se encuentra el módulo IGU (Interfaz Gráfica de Usuario), el cual presenta la interfaz del motor de búsqueda semántico e interactúa con el usuario final para realizar consultas y mostrar resultados. Se conecta con la capa de lógica.

2) Capa de Lógica

En esta capa se encuentran los módulos: Gestor de Búsquedas, Ontología, Algoritmos de Machine Learning y Gestor de Datos. Estos módulos se encargan de procesar todos los datos de los proyectos de investigación para el motor de búsqueda semántico. Esta capa se conecta con las capas de presentación y de datos.

3) Capa de Datos

En esta capa se encuentra el módulo de Base de Datos, el cual almacena todos los datos de los proyectos de investigación y permite la conexión con la capa de lógica.

B. Módulos de la arquitectura del motor de búsqueda

Los módulos de la arquitectura de THAQHAÑA, son los siguientes:

1) Módulo IGU

Este módulo presenta la Interfaz Gráfica de Usuario del motor de búsqueda semántico, este módulo interactúa con el usuario final presentándole los elementos necesarios para realizar la búsqueda y satisfacer sus necesidades y experiencia de usuario. Este módulo se conecta con el módulo de Gestor de Búsquedas de manera bidireccional para el evento en el que se realice la búsqueda o en el que se muestre el resultado.

2) Módulo Gestor de Búsquedas

Este módulo permite realizar consultas a los módulos de Ontología y Algoritmos de Machine Learning así como retornar el resultado de esas búsquedas al módulo de IGU, para esta labor se instruye de SPARQL, Elasticsearch y scripts que gestionan las búsquedas y las retornan exitosamente.

3) Módulo Ontología

Este módulo almacena todo el conocimiento semántico de la Ontología recorriendo clases, atributos, relaciones e instancias. Se conecta con el módulo Gestor de Búsquedas en el evento en el que se realiza una búsqueda o en el momento en el que se retorna el resultado de esa búsqueda, es por eso que esta conexión se realiza de manera bidireccional. Se conecta con el módulo Gestor de Datos en las etapas iniciales o cuando se alimenta la base de datos, esto es porque este Gestor de Datos realiza las fases de preparación, exploración y análisis de datos y estos datos son insumo para el módulo de Ontología.

4) Módulo Algoritmos de Machine Learning

Este módulo almacena todo el conocimiento semántico de los algoritmos de Machine Learning como: Word2Vec y Doc2Vec. Se conecta con el módulo Gestor de Búsquedas en el evento en el que se realiza una búsqueda o en el momento en el que se retorna el resultado de esa búsqueda, es por eso que esta conexión se realiza de manera bidireccional. Se conecta con el módulo Gestor de Datos en las etapas iniciales o cuando se alimenta la base de datos, esto es porque este Gestor de Datos realiza las fases de preparación, exploración y análisis de datos y estos datos son insumo para el módulo de Algoritmos de Machine Learning.

5) Módulo Gestor de Datos

Este módulo se encarga de gestionar todos los datos que se encuentran en el módulo de Base de Datos con el cual está conectado, esto se logra con la preparación, exploración y análisis de los proyectos de investigación, en donde se destacan los algoritmos de Procesamiento de Lenguaje Natural como: SpaCy, NLTK, BOW y TF-IDF. A su vez se conecta con los módulos de Ontología y Algoritmos de Machine Learning por motivo de suministrar a modo de insumo todos los datos que fueron transformados en este módulo y que posteriormente estos puedan ser fuente para la semántica.

6) Módulo Base de Datos

En este módulo se almacenan los datos de los proyectos de investigación que incluyen: proyectos docentes, proyectos estudiantiles y trabajos de grado. Esto se logra mediante el sistema de gestión de bases de datos PostgreSQL. Este módulo se conecta con el módulo de Gestor de Datos para enviarle datos y que este último pueda procesar los datos y enviarle a los demás módulos respectivos.

IV. RESULTADOS

Se presentan los resultados obtenidos para cada etapa de la metodología aplicada:

A. Apropiación del conocimiento

Se destaca el resultado del conocimiento adquirido de todas las temáticas que abarcan el proyecto, así como también de las diversas herramientas y lenguajes usados. Se obtuvo el aprendizaje de temáticas tales como: semántica, Web Semántica, Ontologías, Motores de Búsqueda, Machine Learning, Procesamiento de Lenguaje Natural y Methontology. De igual forma se apropió y reforzó aprendizaje en lenguajes como: Python, XML, RDF, OWL y SPARQL.

B. Instalación y configuración de herramientas, bibliotecas y tecnologías

Se destaca el resultado de la instalación y configuración de: Jupyter notebook, Protégé, Owlready2, Apache Jena Fuseki, Elasticsearch, Visual Studio Code, Anaconda, Gensim con Word2Vec y Doc2Vec, Pandas, Numpy, NLTK, SpaCy, etc.

C. Recolección y extracción de proyectos de investigación

Se destaca el resultado de recolectar y extraer información de los proyectos de investigación de proyectos docentes, proyectos estudiantiles y trabajos de grado que se encuentran almacenados en el sistema de investigaciones de la Universidad de Nariño.

D. Preparación de proyectos de investigación

Se destaca el resultado de preparar los proyectos de investigación, de modo tal que esto permitió navegar por las siguientes etapas, anticipando y evitando inconvenientes, errores o problemas con respecto a la calidad de los datos.

En este orden de ideas, se resalta la realización de las siguientes fases:

1) Fase de organización de datos

En esta fase se aplicaron algoritmos de ordenamiento (creados por los autores de este trabajo) para los proyectos de investigación, esto dado que los proyectos en la fase de recolección y extracción se encontraban desordenados y en condiciones no aptas para ser tratados, manejados y trabajados. Se utilizó Jupyter Notebook con scripts de Python y Pandas para facilitar el manejo de los datos en las series y data frames.

2) Fase de creación de corpus

En esta fase se creó el corpus para los proyectos de investigación, el cual fue el insumo más potente de la semántica como se podrá observar en las etapas posteriores. Dicho corpus resultó de unificar todos los datos de los proyectos de investigación (ya organizados en la fase previa). En esta fase al igual que la anterior, también se se utilizó Jupyter Notebook, Python y Pandas para facilitar el manejo de los datos en las series y data frames.

3) Fase de preprocesamiento de datos

En esta fase se utilizaron las bibliotecas de NLTK y SpaCy para realizar un preprocesamiento de los datos obtenidos en la anterior fase. Para ello se trabajó con las siguientes subfases:

- *Subfase de tokenización de datos*: en esta subfase se ejecutaron algoritmos de la biblioteca NLTK para separar todas las palabras y poder trabajar con ellas de manera individual.

- *Subfase de normalización de datos*: para esta subfase se aplicaron algoritmos para que todos los datos queden bajo un mismo estándar.

- *Subfase de limpieza de datos*: en esta subfase se aplicaron algoritmos de NLTK y SpaCy junto con expresiones regulares para que los datos queden totalmente limpios, esto con la eliminación de datos nulos, signos de puntuación, caracteres "no ascii" y stopwords.

- *Subfase de lematización de datos*: finalmente en esta subfase se lematizaron los datos a los cuales ya se le realizó la etapa de limpieza.

E. Diseño del Motor de Búsqueda Semántico

Una vez culminada la etapa anterior de preparación de los proyectos de investigación, se procedió a diseñar THAQHAÑA. Este diseño se realizó teniendo en cuenta la fase de conceptualización de la metodología Methontology, en donde se destacan los siguientes resultados:

1) *Fase de conceptualización*

Dentro de la fase de conceptualización se desarrollaron once tareas puntuales que permitieron conceptualizar exitosamente: clases, atributos, relaciones e instancias de la Ontología. Entre estas tareas se destacan tres, que son las siguientes:

Tarea 1. Construir el glosario de términos:

En esta tarea se listaron todos los términos importantes seleccionados después de analizar la anterior fase de especificación con su proceso de adquisición de conocimiento, así mismo se presenta una breve descripción de cada uno como se indica en Tabla I.

TABLA I
GLOSARIO DE TÉRMINOS DE LA ONTOLOGÍA

Término	Descripción
Universidad	Es la entidad orientada a la educación que contiene facultades.
Facultad	Es la entidad que contiene departamentos académicos.
VIIS	Vicerrectoría de Investigaciones e Interacción Social, es la entidad encargada del aspecto investigativo a lo largo de la Universidad, es quien gestiona los recursos económicos para proyectos de investigación.
Departamento	Es la entidad que contiene programas académicos.
Convocatoria	Este término refiere al llamado por parte de la VIIS para que los investigadores acudan a esta convocatoria y presenten proyectos con el fin de que estos sean financiados.
Programa	Es el programa académico dentro del cual se encuentran docentes y estudiantes.
Grupo de investigación	Es el grupo conformado por docentes y/o estudiantes investigadores para presentar proyectos a la convocatoria de la VIIS.
Docente	Es un investigador que pertenece a la Universidad, que realiza proyectos de tipo docente.
Estudiante	Es un investigador que pertenece a la Universidad que realiza proyectos estudiantiles y/o trabajos de grado.
Investigador externo	Es un investigador que es externo a la Universidad pero que se presenta a la convocatoria de VIIS.
Línea de investigación	Es una rama que el grupo de investigación maneja, enfocada a un ámbito específico del conocimiento.
Investigador	Es quien elabora proyectos de investigación y los presenta a la convocatoria de VIIS. Este investigador, puede ser docente, estudiante o investigador externo.
Proyecto de investigación	Es quizá la entidad más importante dentro del dominio investigativo que contiene todo lo referente a un proyecto de investigación.
Palabra	Esta entidad hace referencia a cada una de las palabras que conforma el proyecto de investigación con las cuales se generó gran parte de la semántica y se armó el tesauro.

Tarea 2. Construir taxonomía de conceptos:

En esta tarea se definió la taxonomía o jerarquía de los conceptos o clases de la Ontología que se obtuvo a partir del glosario de términos de la tarea 1, esta taxonomía se muestra en Fig. 2.

Tarea 3. Construir un diagrama de relaciones binarias:

En esta tarea se elaboró el diagrama de relaciones binarias que son los predicados de la Ontología. Se visualizan en Fig. 3 las relaciones de la clase más importante de la Ontología que es: Proyecto de investigación.

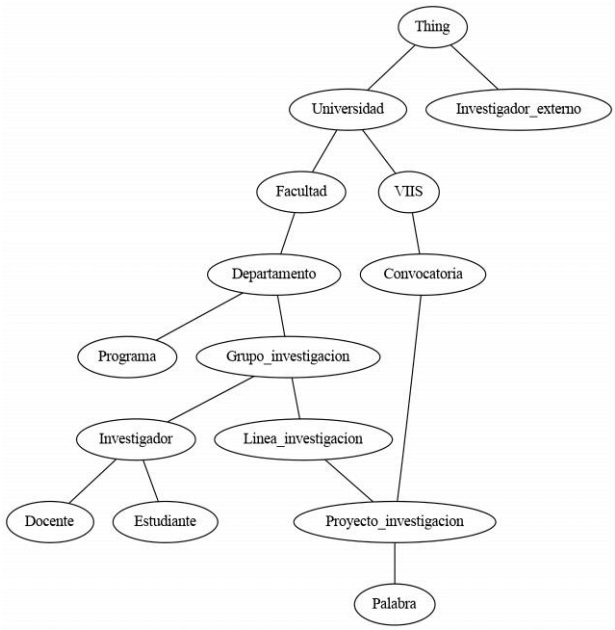

Fig. 2 Taxonomía de conceptos de la Ontología

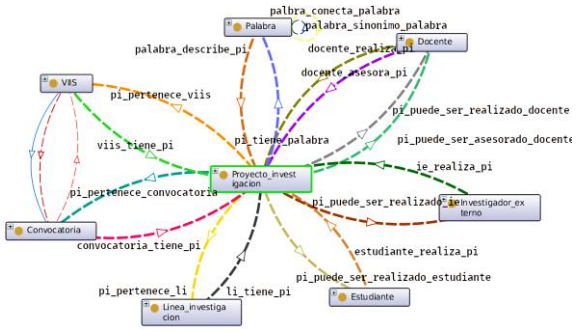

Fig. 3 Diagrama de relaciones binarias de la clase Proyecto de Investigación

F. *Desarrollo del Motor de Búsqueda Semántico*

Se desarrolló THAQHAÑA basado en tres fases en las cuales se destacan los siguientes resultados:

1) *Desarrollo de Ontología*

Para esta fase se aplicaron las tres subfases de Methontology que son: formalización, implementación y evaluación.

- *Fase de formalización*: en esta fase se destacan los resultados obtenidos después usar la herramienta de Protégé para la construcción de la Ontología en términos semi-computables.
- *Fase de implementación:* en esta fase se destacan los resultados de haber usado la biblioteca Owlready2 para codificar una versión computable de la Ontología. Se crearon scripts y se realizó la codificación para el manejo de la Ontología con Python donde se realizó todo un proceso de instanciar objetos de todas las clases.

  En síntesis, todas las clases, atributos y relaciones fueron instanciados dentro de la Ontología.

- *Fase de evaluación:* en esta fase se destacan resultados después de haber realizado pruebas funcionales de manera local y haber recuperado los datos y demás componentes de la Ontología de manera exitosa con el uso de SPARQL y del servidor de Apache Jena Fuseki mediante el manejo de triples de RDF (sujeto predicado objeto).

2) *Desarrollo con Word2vec y Doc2vec*

En esta fase se destacan los resultados producto del entrenamiento con el algoritmo de Machine Learning con Procesamiento de Lenguaje Natural como es Word2Vec, el cual ayudó a encontrar el contexto que una palabra tiene, además se entrenó un modelo con el algoritmo de Doc2Vec, el cual se basa en Word2Vec para encontrar documentos que se relacionan entre sí, estos modelos hacen uso de redes neuronales. En este caso se entrenó el modelo con los algoritmos mencionados previamente en base al modelo Skip-Gram, el cual intenta predecir las palabras o documentos en contexto dada una palabra o un conjunto de palabras base a buscar.

Cabe aclarar que la salida que retornó Word2Vec fue la entrada para el proceso realizado con Doc2Vec, esto es posible dado que ambos algoritmos trabajan de la mano para lograr descubrir relaciones semánticas y recuperar semánticamente la información de manera efectiva.

Para realizar la búsqueda de similitud entre palabras o documentos, de un conjunto de palabras dadas, se utilizó la biblioteca Gensim, la cual hace uso de la normalización de los vectores obtenidos a partir de la palabras a buscar y el cálculo del producto punto entre el vector normalizado y cada uno de los vectores correspondiente a cada palabra o documento entrenado.

Se creó el modelo con los datos de la etapa de preparación de proyectos de investigación, se asignaron los respectivos hiperparámetros, se entrenó dicho modelo, se evaluaron los resultados y se retroalimentó ajustando los hiperparámetros hasta obtener resultados satisfactorios, como se puede evidenciar en Tabla II.

TABLA II
HIPERPARÁMETROS PARA LOS MODELOS WORD2VEC Y DOC2VEC

Nombre	Valor	Descripción
vector_size	300	Dimensión del vector de cada una de las palabras del corpus.
window	5	Refiere al contexto donde se elige la distancia entre palabras predichas.
min_count	1	Mínimo de palabras a buscar.

dm	0	0 indica que se usa PV-DBOW de Doc2Vec que es análogo al modelo de Skip-Gram usado en Word2vec. 1 indica que se usa PV-DM de Doc2Vec que es análogo al modelo de CBOW usado en Word2Vec.
dbow_words	1	0 indica que se va a entrenar con Doc2Vec. 1 indica que se va a entrenar con Doc2Vec teniendo de insumo a Word2Vec.
hs	0	Es el valor con el que se va a castigar a la neurona en caso de que la tarea efectuada no sea la requerida.
negative	20	Número de palabras irrelevantes para el muestreo negativo.
ns_exponent	-0.5	Indica que se van a muestrear frecuencias por igual.
alpha	0.015	Tasa de aprendizaje de la red neuronal.
min_alpha	0.0001	Tasa que se reducirá durante el entrenamiento.
seed	25	Semilla para generar hash para palabras.
sample	5	Número de reducción para palabras con alta frecuencia
epochs	150	Épocas, número de iteraciones para el entrenamiento.

En Fig. 4, Fig. 5 y Fig. 6 se puede observar el resultado de ejecutar la orden de encontrar 10 palabras más similares y relacionadas (según la similitud coseno del algoritmo ordenadas porcentualmente de mayor a menor) a otra palabra que se especifica dentro de todo el corpus investigativo con un método del algoritmo Word2vec.

Fig. 4 indica las 10 palabras más similares y relacionadas a la palabra "medicos".

```
modelo_cargado.wv.most_similar(
    positive=['medicos'], topn=10)
```

```
[('remision', 0.8112903237342834),
 ('reumatologos', 0.775713324546814),
 ('subdiagnostico', 0.7463810443878174),
 ('precoz', 0.7410680055618286),
 ('especialista', 0.7363733649253845),
 ('reumaticas', 0.7290210127830505),
 ('generales', 0.724345862865448),
 ('comprobacion', 0.6957311630249023),
 ('spa', 0.6793832778930664),
 ('nula', 0.6671884655952454)]
```

Fig. 4 Resultado de método con Word2vec para palabra "medicos"

Fig. 5 indica las 10 palabras más similares y relacionadas a la palabra "sentimientos".

```
modelo_cargado.wv.most_similar(
    positive=['sentimientos'], topn=10)
```

```
[('temor', 0.9108718633651733),
 ('frustracion', 0.8689973950386047),
 ('voluntad', 0.8588676452636719),
 ('elegir', 0.8260773420333862),
 ('evocar', 0.7610046863555908),
 ('autoridad', 0.7394462823867798),
 ('emociones', 0.7158118486404419),
 ('filiaciones', 0.7155368328094482),
 ('sonoros', 0.7140575051307678),
 ('sostienen', 0.7129440307617188)]
```

Fig. 5 Resultado de método con Word2vec para palabra "sentimientos"

Fig. 6 indica las 10 palabras más similares y relacionadas a la palabra "redes".

```
modelo_cargado.wv.most_similar(
    positive=['redes'], topn=10)
```

```
[('inalambricas', 0.7252240180969238),
 ('niebla', 0.6841256618499756),
 ('neuronales', 0.660290002822876),
 ('configuraban', 0.658789873123169),
 ('clientelares', 0.6412293314933777),
 ('plausible', 0.6383322477340698),
 ('justo', 0.6135219931602478),
 ('parece', 0.6127933859825134),
 ('potencializan', 0.603103756904602),
 ('bancos', 0.5799424052238464)]
```

Fig. 6 Resultado de método con Word2vec para palabra "redes"

3) Desarrollo integrado de Ontología, Word2vec y Doc2vec

Para esta fase se integran: los algoritmos producto de la ontología junto con Word2vec y Doc2vec brindando así potencia, efectividad y poder semántico para optimizar tiempos, recursos, y para tener mayores posibilidades de encontrar resultados exitosos y satisfactorios a determinadas búsquedas en THAQHAÑA, los resultados se observan en: Fig. 7, Fig. 8, Fig. 9 y Fig. 10.

Esto se logró llevando los vectores que Doc2Vec generó a Elasticsearch; este último ayudó en la etapa de rankeo al tener velocidad, escalabilidad y al ser un motor de análisis distribuido que favorece en la búsqueda y en la indexación de los proyectos de investigación.

Después se crearon scripts para gestionar las consultas de los proyectos de investigación para la Ontología con SPARQL la cual se apoya del modelo Word2Vec entrenado para agregar a la búsqueda palabras adicionales que se relacionan a las solicitadas y así encontrar investigaciones relacionadas a determinada consulta. De

igual forma con Doc2Vec se logró inferir vectores a partir de un conjunto de palabras suministradas, luego como resultado parcial se presentan las investigaciones que se relacionan a dichos vectores inferidos. Finalmente se hace la unión de los resultados obtenidos en la consulta SPARQL y el algoritmo Doc2Vec, así el ranking final de una búsqueda mostrará resultados acordes, coherentes, exitosos y satisfactorios a lo solicitado con la capacidad adicional de recomendar documentos que pueden ser de utilidad e interés para el usuario.

Fig. **7** Implementación de THAQHAÑA: motor de búsqueda semántico

Fig. 8 Primeros resultados para la búsqueda “proyectos de investigacion sobre fisica” en THAQHAÑA

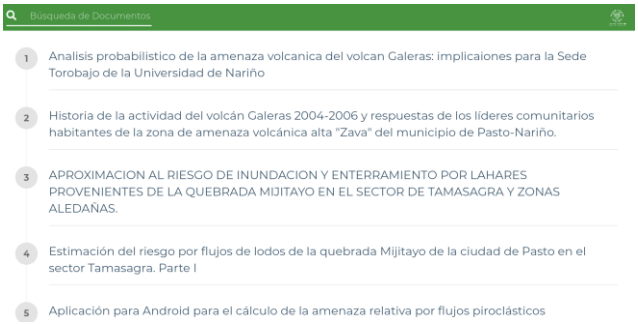

Fig. 9 Primeros resultados para la búsqueda “proyectos de investigacion sobre amenaza volcanica” en THAQHAÑA

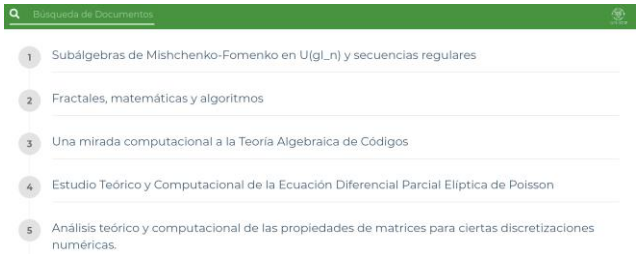

Fig. 10 Primeros resultados para la búsqueda “proyectos de investigacion sobre algebra” en THAQHAÑA

G. Comparación y pruebas finales entre THAQHAÑA y motor de búsqueda manual

Finalmente en esta fase se demuestra la comparación entre THAQHAÑA y el motor de búsqueda manual. Por ejemplo en Fig. 11 no se encontraron resultados para la búsqueda sobre proyectos de investigación de “fisica”, mientras que en Fig. 8 sí. En Fig. 12 no se encontraron resultados para la búsqueda sobre proyectos de investigación de “amenaza volcanica” y por otro lado en Fig. 9 sí. En Fig. 13 no se encontraron resultados para la búsqueda sobre proyectos de investigación de “algebra” y efectivamente en Fig. 10 sí.

Fig. 11 Resultados para la búsqueda “física” en el motor de búsqueda manual

Fig. 12 Resultados para la búsqueda “amenaza volcanica” en el motor de búsqueda manual

Fig. 13. Resultados para la búsqueda “algebra” en el motor de búsqueda manual

La Tabla III indica la comparación final entre THAQHAÑA y el motor de búsqueda manual mediante variables que funcionan como indicadores clave a modo de pregunta. Se respondió a las preguntas mediante respuesta binaria de SI/NO, donde los resultados arrojan que THAQHAÑA tiene un rendimiento general del 100% mientras que el motor de búsqueda manual del 18.75%

TABLA III
PRUEBAS DE COMPARACIÓN ENTRE MOTOR DE BÚSQUEDA SEMÁNTICO Y BUSCADOR MANUAL

Variable a medir	THAQHAÑA: Motor de búsqueda semántico	Motor de búsqueda manual
¿El número de resultados es pertinente y apropiado?	SI	NO
¿Los resultados son de calidad?	SI	NO
¿Se encuentra la búsqueda solicitada y cuando no exista se muestra error?	SI	NO
¿El manejo de la herramienta es fácil?	SI	SI
¿El manejo de la herramienta es intuitivo?	SI	SI
¿El manejo de la herramienta es interactivo?	SI	NO
¿Las vistas de los diseños de la herramienta son agradables al usuario?	SI	NO
¿La herramienta es eficiente y efectiva?	SI	NO
¿Los resultados en respuesta a las búsquedas realizadas en la herramienta son satisfactorios?	SI	NO
¿Se permite la búsqueda múltiple en la herramienta?	SI	SI
¿No existen muchas categorías para realizar las búsquedas lo que vuelve tedioso y dispendioso para el usuario?	SI	NO
¿Funcionan todos los botones en la interacción y dinamismo de la herramienta?	SI	NO
¿La herramienta no se demora actualizando o refrescando?	SI	NO
¿La herramienta no se demora en la búsqueda?	SI	NO
¿Los espacios en el servidor para la herramienta están actualizados y funcionan adecuadamente?	SI	NO
¿La forma de realizar la búsqueda en dicha herramienta es la más óptima?	SI	NO

V. CONCLUSIONES

√ Con la culminación de este trabajo de investigación se obtiene THAQHAÑA: Un Motor de Búsqueda Semántico basado en una Ontología y en un modelo de Machine Learning para proyectos de investigación de la Universidad de Nariño. Mediante el desarrollo exitoso de las etapas del proyecto se soluciona el problema formulado, se cumplen los objetivos planteados y se obtienen resultados satisfactorios. De este modo, esta herramienta facilita la búsqueda exitosa de proyectos de investigación para proyectos docentes, proyectos estudiantiles y trabajos de grado en la Universidad de Nariño.

√ En las etapas de apropiación del conocimiento e instalación y configuración de las herramientas se adquirió un dominio de las diversas temáticas manejadas y eso aparte de contribuir con el desarrollo del trabajo, propendió a la formación personal de los investigadores así como también realizó destacados aportes para el grupo de investigación GRIAS (Grupo de Investigación Aplicado en Sistemas) y para la Universidad de Nariño en general.

√ Las etapas de recolección, extracción y preparación de proyectos de investigación fueron etapas sumamente importantes que actuaron como etapas previas y de antesala a modo de insumo para THAQHAÑA. En este orden de ideas es correcto afirmar que sin estas etapas no se podría haber logrado un buen desarrollo de THAQHAÑA.

√ Methontology fue una metodología que se acopló perfectamente al proyecto y permitió construir la Ontología siguiendo unas fases y tareas puntuales con un orden, entendimiento y exactitud en los procesos.

√ La Ontología integrada con Machine Learning demostraron un gran poder, potencia semántica y una efectividad en los procesos para obtener resultados concretos y acordes a las búsquedas realizadas. Esto es porque los algoritmos de Machine Learning, específicamente de Procesamiento de Lenguaje Natural como Word2vec y Doc2vec trabajan con redes neuronales, las cuales fueron entrenadas con las palabras del corpus de proyectos de investigación, adaptándolas al contexto y encontrando las diversas relaciones semánticas entre las mismas. Así mismo la Ontología actuó como una gran red semántica cuyas instancias, de la mano de clases, relaciones y atributos, interactuaron bajo el esquema de triples que maneja RDF y que consulta SPARQL para extraer todo el conocimiento del dominio de los proyectos de investigación.

AGRADECIMIENTOS

Al Sistema de Investigaciones de la Universidad de Nariño (Colombia), por financiar este proyecto de investigación.

REFERENCIAS

[1] A. De Pedro, "Buscadores Semánticos, para qué sirven". Jun. 10, 2009. [Online]. Available: http://www.alejandrodepedro.es/buscadores-semanticos-el-paso-al-30

[2] F. J. García Peñalvo, "Web Semántica y Ontologías". [Online]. Available: https://www.researchgate.net/publication/267222548_Web_Semantica_y_Ontologias

[3] M. Mouriño García, "Clasificación multilingüe de documentos utilizando machine learning y la wikipedia". 2018. [Online]. Available: https://dialnet.unirioja.es/servlet/tesis?codigo=150295

[4] C. Rodríguez, "Incorporación de un buscador semántico en la plataforma LdShake para la selección de patrones educativos". 2013. [Online]. Available: http://repositori.upf.edu/handle/10230/22172

[5] C. Amaral, D. Laurent, A. Martins, A. Mendes y C. Pinto, "Design and Implementation of a Semantic Search Engine for Portuguese". [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.4090&rep=rep1&type=pdf

[6] Y. del C. Aucapiña Pineda y C. O. Plaza Villa, "Buscador semántico universitario: Caso de estudio Universidad de Cuenca". 2018. [Online]. Available: http://dspace.ucuenca.edu.ec/handle/123456789/30291

[7] F. Umpiérrez, "SPARQL Interpreter". 2014. [Online]. Available: https://nanopdf.com/download/0701044000000000pdf_pdf

[8] G. B. Bustos Quiroga, "Prototipo de un sistema de integración de recursos científicos, diseñado para su funcionamiento en el espacio de los datos abiertos enlazados para mejorar la colaboración, la eficiencia y promover la innovación en Colombia". Aug. 2015. [Online]. Available: https://repositorio.unal.edu.co/handle/unal/55245

[9] C. A. Moreno Agudelo y Y. Sánchez Reyes, "PROTOTIPO DE BUSCADOR SEMÁNTICO APLICADO A LA BÚSQUEDA DE LIBROS DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN EN LA BIBLIOTECA JORGE ROA MARTÍNEZ DE LA UNIVERSIDAD TECNOLÓGICA DE PEREIRA". Jan. 2012. [Online]. Available: repositorio.utp.edu.co/dspace/bitstream/handle/11059/2671/0057565M843.pdf

[10] M. Benavides y J. Guerrero, "UMAYUX: un modelo de gestor de conocimiento soportado en una ontología dinámica débilmente acoplado con un gestor de base de datos". 2014. [Online]. Available: http://sired.udenar.edu.co/2030

[11] A. Andreoni et al., "The ERCIM technical reference digital library". D-Lib Magazine, vol. 5, no. 12. pp. 35–55, Dic. 1999, doi: 10.1045/december99-peters.

[12] J. Baculima and M. Cajamarca, "DISEÑO E IMPLEMENTACIÓN DE UN REPOSITORIO ECUATORIANO DE DATOS ENLAZADOS GEOESPACIALES". Jun. 2014. Available: http://dspace.ucuenca.edu.ec/bitstream/123456789/19876/1/tesis.pdf.

[13] A. C. Barberá et al. "Estudio del buscador semántico Swoogle". 2005. [Online]. Available: https://www.uv.es/etomar/trabajos/swoogle/swoogle.pdf.

[14] A. Budhiraja, "A simple explanation of document embeddings generated using Doc2Vec". May 14, 2018. Available: https://medium.com/@amarbudhiraja/understanding-document-embeddings-of-doc2vec-bfe7237a26da.

[15] R. Pedraza-Jiménez, L. Codina, and C. Rovira, "Semantic web and ontologies in document information processing". Prof. la Inf., vol. 16, no. 6, pp. 569–578, 2007, doi: 10.3145/epi.2007.nov.04.

[16] G. Edwards, "Machine Learning An Introduction". Nov. 17, 2018. Available: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0.

[17] M. Gallo Pérez De Tudela, E. Fabre, and M. Gallo, "¿Qué es un Buscador?". [Online]. Available: http://media.axon.es/pdf/98234_1.pdf.

[18] J. Martín, "Swotti buscador de opiniones". Mar. 06, 2008. Available: https://loogic.com/swotti-buscador-de-opiniones.

[19] B. Shetty, "Natural Language Processing (NLP) for Machine Learning". Nov. 24, 2018. Available: https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b.

[20] G. Shperber, "A gentle introduction to Doc2Vec". Jul. 26, 2017. Available: https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e.

[21] "Sistema de Información de Investigaciones". Available: http://sisinfoviis.udenar.edu.co

[22] "spaCy 101: todo lo que necesita saber". Available: https://spacy.io/usage/spacy-101.

[23] "Wolfram Alpha: Computational Intelligence." Available: https://www.wolframalpha.com.

[24] J. Rocca, "A simple introduction to Machine Learning". Dic. 23, 2019. Available: https://towardsdatascience.com/introduction-to-machine-learning-f41aabc55264.