UACJ INVESTIGACIÓN

Título del Proyecto de Investigación
al que corresponde el Reporte Técnico:

Modelo de distribución de conocimiento tácito a través de redes sociales como apoyo en la gestión de conocimiento en proyectos de innovación.

Tipo de financiamiento

Sin financiamiento

Fecha de Inicio: 08/08/2016
Fecha de Término: 22/11/2021

Tipo de Reporte

[ ] Parcial

[X] Final

Autor (es) del reporte técnico:

José de Jesús Martínez Silva
Dr. Jorge Rodas Osollo
Dra. Karla Olmos Sánchez

Modelo de distribución de conocimiento tácito a través de redes sociales como apoyo en la gestión de conocimiento en proyectos de innovación.

Resumen del reporte técnico en español:

Las empresas o entidades dedicadas al desarrollo tecnológico que atienden las necesidades de desarrollar productos innovadores, se encuentran con ciertos retos en donde el conocimiento necesario para el desarrollo de estos productos generalmente está en dominios de estructura informal, en áreas multidisciplinarias y sin dejar de lado que las cantidades de conocimiento cada día son más amplias, superando las capacidades de las empresas para la realización de un producto satisfactorio en tiempo y forma. Es importante que las entidades encargadas del desarrollo tecnológico identifiquen las fuentes de donde proviene el conocimiento necesario para la elaboración de sus proyectos de investigación y gestión de conocimiento, dichas fuentes suelen contener el conocimiento disperso geográficamente y en distintas partes de internet a lo largo de distintos repositorios. Estas características del conocimiento se convierten en dificultades cuando la cantidad de información a analizar superan los recursos y límites de las entidades desarrolladoras de proyectos innovadores, demandando una gran destreza e inversión de tiempo para la correcta localización del conocimiento necesario. Por estas razones el uso de una herramienta cognitiva como apoyo en la gestión del conocimiento mediante el uso de un modelo de distribución del conocimiento, puede ser una gran ventaja que nos ayude a minimizar las dificultades descritas.

Resumen del reporte técnico en inglés:

Companies or entities dedicated to technological development that meet the needs of developing innovative products, encounter certain challenges where the knowledge necessary for the development of these products is generally in domains of informal structure, in multidisciplinary areas and without neglecting that the amounts of knowledge are growing every day, exceeding the capacities of companies to produce a satisfactory product in a timely manner. It is important that the entities in charge of technological development identify the sources from which the necessary knowledge comes from for the development of their research projects and knowledge management, these sources usually contain the knowledge dispersed geographically and in different parts of the internet throughout different repositories. These characteristics of knowledge become difficulties when the amount of information to be analyzed exceeds the resources and limits of the entities developing innovative projects, demanding great skill and investment of time for the correct location of the necessary knowledge. For these reasons, the use of a cognitive tool to support knowledge management using a knowledge distribution model can be a great advantage that helps us to minimize the difficulties described.

Palabras clave:

Proceso Cognitivo, Gestión del Conocimiento, innovación, Modelo de Distribución, Redes de Conocimiento.

Usuarios potenciales (del proyecto de investigación):

Entidades dedicadas al desarrollo de proyectos de investigación.
Entidades dedicadas al desarrollo de productos innovativos.

Reconocimientos

A mis tutores por ser la guía y fuente de conocimiento a lo largo de este camino, por su interminable paciencia y comprensión.

A mis padres por su incondicional apoyo y eterna confianza, que a pesar de cualquier dificultad siempre se mantuvieron firmes dándome la motivación para seguir superando todos los problemas.

A mi familia y amigos por su alegría que hacen de los días difíciles algo fácil de vencer.

A Dios por permitirme disfrutar de este efímero momento.

Contenido

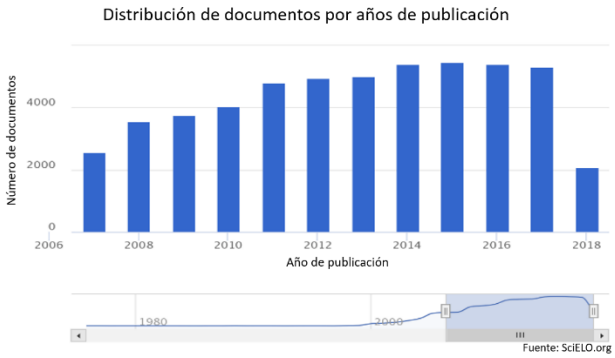
1. Introducción

En la actualidad, las empresas de desarrollo tecnológico que atienden las necesidades de productos innovadores que demanda el mercado, trabajan con Dominios de Estructura Informal (DEI); ya que generalmente desarrollan proyectos multidisciplinarios en los que participan un grupo de especialistas que requieren grandes cantidades de conocimiento, que generalmente es informal, incompleto, parcial, tácito o poco estructurado [1].

Para que el desarrollo de proyectos en los DEI alcance los objetivos planteados es necesario obtener el conocimiento de las entidades proveedoras de este. En particular, el conocimiento heterogéneo y explícito puede encontrarse distribuido en las distintas fuentes externas de conocimiento que un gran número de investigadores distribuyen a través de publicaciones en o en distintos medios de difusión, generalmente de manera internacional. Estas entidades forman redes de trabajo explícitas e implícitas. Las redes de trabajo explícitas son fáciles de identificar, por ejemplo, todos los autores dentro de una investigación o publicaciones expedidas en nombre de una organización. Por otro lado, las redes de trabajo implícitas se forman a través de las múltiples relaciones explícitas entre los distintos autores. Con el tiempo esta red implícita puede ir mutando y evolucionando con el incremento de autores que participan con aportaciones en el mismo dominio. Por lo tanto, la complejidad de interacción de estas redes aumenta debido al crecimiento exponencial de conocimiento disponible y a la integración de distintas tecnologías. Por lo que la compleja tarea de identificar las fuentes de información que contienen el conocimiento necesario para el desarrollo de soluciones de innovación exige del análisis, procesamiento y modelación de estas redes de conocimiento.

En la actualidad, la cantidad de conocimiento formal va en aumento de una manera casi exponencial gracias a un acelerado desarrollo tecnológico, la globalización y otros factores que han impulsado el desarrollo de investigaciones y publicaciones. Dichas publicaciones formales se encuentran distribuidas en distintos repositorios en Internet. Analizar una fuente de conocimiento muy basta y tan distribuida como es el Internet, para el desarrollo de investigaciones o resolución de problemas puede acarrear costos elevados de tiempo, recurso humano y monetarios. Sin embargo, de no considerar este conocimiento dentro del proyecto de innovación o investigación, pondrían en riesgo la calidad del resultado obtenido.

Una de las principales fuentes de conocimiento son los artículos científicos, medio que cuenta con una estructura formal de información que puede ser fácilmente consultada: título, autores, referencias, resúmenes, palabras clave, por mencionar los más importantes. De acuerdo con datos obtenidos del portal SciELO.org (Scientific Electronic Library Online por sus siglas en inglés) [2] [3], una de las hemerotecas virtuales más importantes en américa latina, los accesos a archivos ya sean en *html, pdf* o abstractos, ha ido incrementando tal y como se muestra en la Figura 1.



2 *Figura 1 Archivos registrados en SciELO.org por año.*

Esta alta oferta de conocimiento distribuido demanda el uso de herramientas que apoyen a la búsqueda, identificación y análisis de las fuentes de conocimiento con las que se trabaja. Una herramienta que ofrece este tipo de apoyo son las soluciones cognitivas, las cuales cuentan con la capacidad de realizar dicho análisis ofreciendo resultados procesados gráficos y estadísticos que muestren conocimiento de valor que potencien la innovación en el desarrollo de proyectos o resolución de problemas.

Una solución cognitiva consiste en el uso de conocimiento por medio de un proceso informático determinado para resolver problemas de un dominio en específico y se pueden conformar de varias tecnologías cognitivas. La elección por una solución cognitiva se presenta cuando un problema o necesidad exige una gran cantidad de conocimiento o conocimiento muy especializado, que generalmente está distribuido y que requiere de una gran cantidad de recursos para procesar dicho conocimiento, estas condiciones identifican al problema o necesidad como pertenecientes a los dominios de estructura informal. Para realizar un análisis completo de toda la información encontrada en un dominio de estructura informal de forma convencional, podría resultar en un proceso que demoraría días, meses o años pues debería delimitar todo un estado del arte

referente al dominio del problema. Por otra parte, un proceso que fuera mejorado mediante tecnologías cognitivas podría seleccionar muestras de las fuentes de conocimiento disponible, obtener listados de artículos, realizar clasificaciones e ir integrando una base de conocimiento. Además de ayudar en la identificación de conocimientos o entidades poseedoras de conocimiento confiable.

En conjunto con la utilización de un sistema cognitivo, el manejo de grandes conjuntos de datos requiere de herramientas que permitan adecuada representación del conocimiento, como lo son los sistemas de cómputo basados en visualización el cual ayuda a resaltar puntos clave o a quitarle importancia a datos o información que causan ruido en el conocimiento resultante. Por lo discutido anteriormente, la motivación del siguiente trabajo es brindar soporte al desarrollo de proyectos de innovación mediante el uso de una solución cognitiva para el manejo de la convergencia del conocimiento en dominios de estructura informal utilizando artículos científicos.

2. Planteamiento

El proceso de generación de productos innovadores involucra todo un ejercicio de resolver problemas, satisfacer necesidades o atender dificultades que exigen utilizar conocimiento nuevo o altamente especializado que puede ser tácito o explícito. Además, dicho conocimiento suele encontrarse disperso: el tácito geográficamente y el explícito en distintas fuentes del Internet por medio de publicaciones y repositorios. Para potenciar el proceso de generación de productos innovadores es conveniente identificar la red o conjunto de redes de trabajo que poseen el conocimiento, por medio de visualizaciones que faciliten la comprensión de los datos para atender las dificultades implícitas en el mismo proceso; lo cual, es una ardua tarea que demanda una gran destreza y una buena inversión de tiempo por parte de analistas de información para la localización del conocimiento especializado necesario para proveer el producto pretendido. En la Figura 2 se presenta un modelo conceptual que representa el proceso al trabajar en proyectos de desarrollos que tengan como resultado un producto de innovación en donde podemos identificar algunas áreas donde se encuentra distribuido el conocimiento en un dominio de estructura informal.

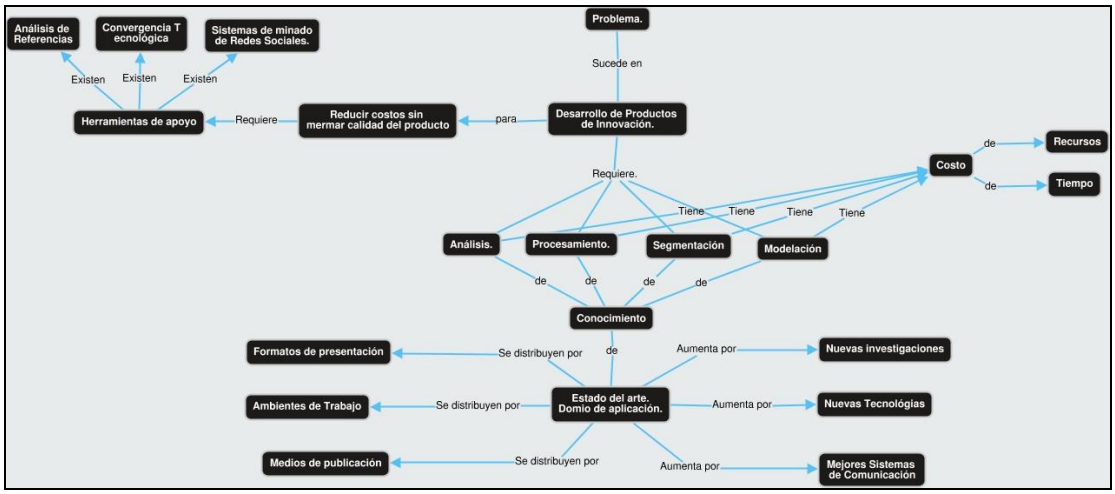


Figura 2 Modelo de problemática para el desarrollo de productos de innovación.

3. Marco teórico

El marco teórico presentado en este apartado describe los conceptos de innovación, convergencia tecnológica y redes colaborativas; las técnicas de análisis de redes de referencia, sistema minado de redes sociales, clusterización y clusterización jerárquica. Por último, se describe la clusterización de redes sociales. Todos estos temas fueron tomados en cuenta para darle forma a la solución propuesta.

3.1.1. Innovación

Mantener el desarrollo de proyectos innovadores les dan valor agregado a las investigaciones ya que aseguran la búsqueda de un nuevo caso de estudio o solución implementando nuevas tecnologías o metodologías. La "innovación" es definida por Garcia, R., y Calantone, R. como un proceso iterativo, iniciado por la percepción de un nuevo mercado y / o una nueva oportunidad de servicio para una invención basada en tecnología que conduce a tareas de desarrollo, producción y marketing que luchan por el éxito comercial de la invención [4]. Este proceso de 'innovación' comprende el desarrollo de una nueva tecnología combinado con la introducción en el mercado de dicha tecnología a los usuarios finales a través de su adopción y difusión.

En las empresas, la adaptación constante al entorno tecnológico cambiante y las necesidades del cliente determina la capacidad de una empresa para mantener una ventaja competitiva. La innovación en las empresas es una función clave que toma ciertos riesgos que conduce a la creatividad, la gestión del conocimiento, especialmente su generación y, finalmente, la realización de la idea creativa a través del desarrollo de competencias y su explotación [5].

La innovación a pesar de ser un proceso de riesgo y que no siempre se obtienen los resultados esperados, ha tenido un incremento en el interés por investigaciones de innovación de acuerdo con datos obtenidos en la Web [6] como se muestra en la Figura 3 con la búsqueda de ‘innovation’ y revisando la cantidad de artículos publicados cada año.

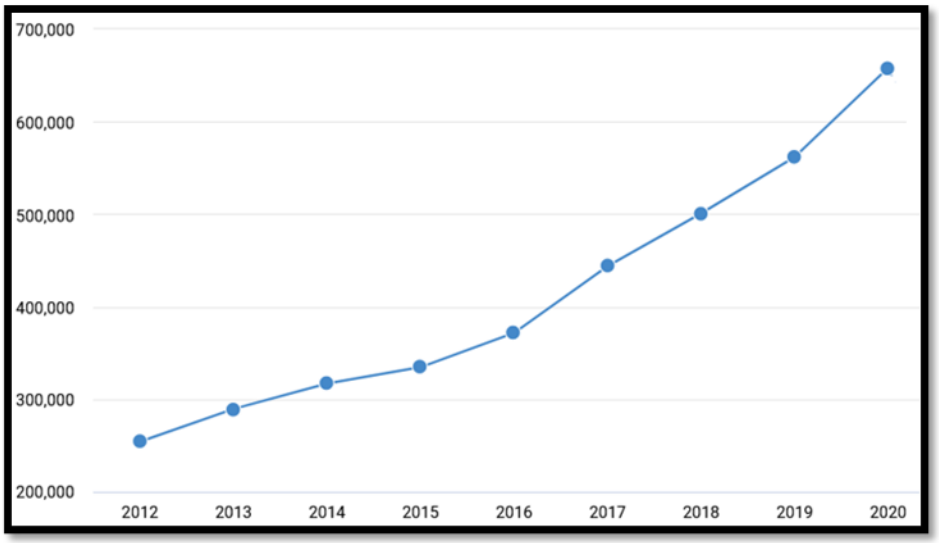


Figura 3. Cantidad de artículos encontrados por la palabra clave 'innovation'.

3.1.2. Convergencia tecnológica

La convergencia tecnológica ayuda entender cómo es que se va dando el nacimiento de nuevos campos de estudio tecnológico gracias a que elementos discretos o heterogéneos, ya sean tecnologías o distintas industrias se fusionan dando lugar a un nuevo campo de estudio, por ejemplo: La robótica. Esta rama tecnológica emerge a través de la unión de otras tecnologías como la ingeniería mecánica, sistemas de control, electrónica y software [7]. Al mismo tiempo esta tecnología podría formar parte de las ramas que convergen en otra tecnología.

Un método reciente para la medición de la relación tecnológica es el método de "Module-based mining methodology" presentado por Kose, T., & Sakata, I. [7] aplicado en el campo de la robótica. Este método busca identificar la convergencia tecnológica de una manera más precisa utilizando un método de modulación de Newman y aplicando un análisis a las redes de referencias.

Las recientes tendencias de transformación industrial pueden caracterizarse por dominios tecnológicos cada vez más convergentes, lo que permite nuevas funcionalidades y oportunidades combinadas para la creación de nuevas tecnologías. La convergencia tecnológica que logra una innovación incremental también puede resultar en la generación de tecnologías disruptivas tal y como Hacklin, F, Raurich, V., & Marxt, C. mencionan en su investigación [8].

3.1.3. Redes colaborativas

Las redes colaborativas están constituidas por una variedad de entidades (por ejemplo, organizaciones y personas) que son en gran mayoría autónomas, geográficamente distribuidas y heterogéneas en términos de su entorno operativo, cultura, capital social y objetivos. [9]. Este tipo de redes tienen una propiedad intencional que se deriva de la creencia compartida de que, juntos, los miembros de la red pueden alcanzar metas que no serían posibles o que tendrían un costo mayor si las intentaran individualmente. Existen muchos tipos de redes colaborativas de acuerdo con las necesidades del objetivo por el cual colaboran, pero todas ellas se encuentran explícitamente formadas.

La cantidad de estas redes han ido creciendo gracias a los avances de las tecnologías de información, crecimiento de exigencias del mercado, y al progreso logrado gracias a la gran cantidad de proyectos internacionales. Es por ello por lo que identificarlas, clasificarlas y entenderlas es de vital importancia para su correcta implementación.

3.1.4 Análisis de redes de referencias

El análisis de redes de referencias es una forma de minado de tecnología también llamado "Tech mining" el cual está compuesto por una combinación de metodologías de investigación científica. Tech mining se ha expandido en los campos de:

- Análisis de futuros tecnológicos.
- Evaluación de investigaciones.
- Inteligencia competitiva.

Esta técnica también ha sido utilizada para determinar o detectar nuevas tecnologías y para rastrear tecnologías emergentes [10] incluyendo el análisis de redes de referencias. El análisis de redes de referencias también ha sido utilizado para:

- Identificar tópicos emergentes.
- Encontrar relación entre tecnología y problemas sociales.
- Medición de la difusión del conocimiento entre los diferentes campos de investigación.
- Detectar frentes de investigación emergentes en campos específicos.
- Ciencia sustentable.

Toshiro Kose y Ichiro Sakata en [7] usan el análisis de redes de referencias para categorizar las publicaciones, factorizar los sectores superpuestos en un simple sector y calcular y evaluar la convergencia de tecnologías.

3.1.5. Sistemas de minado de redes sociales

Existen plataformas las cuales se han encargado de trabajar con las redes sociales académicas y el minado de artículos científicos, por ejemplo, ver [11]. Con el sistema de ArnetMiner realizan una extracción y minado de redes académicas obteniendo perfiles de investigadores automáticamente de la Web y con métodos probabilísticos modelan simultáneamente autores, tópicos de artículos y lugares de publicación. Esta plataforma también ofrece un servicio de búsqueda de experiencia y de asociación. Los métodos probabilísticos, la extracción y minado de datos están enfocados solamente a redes sociales académicas.

Un caso más reciente de minado de redes sociales es el estudio efectuado por Nasution, M. K., & Noah, S. A. [12] en donde se utilizan reglas de asociación para mejorar los métodos superficiales de extracción de redes sociales de bases de datos en Web usualmente basados en las co-ocurencias.

3.1.6. Clusterización

La clusterización en [13] es una de las manifestaciones bien establecidas que ayuda a entender mejor los datos y tomar decisiones. Con su objetivo de revisar en espacios de datos y descubrir su estructura (clústeres de datos), la clusterización es un vehículo ideal para la exploración de vastos territorios de espacios de datos. La clusterización es una metodología general y un marco conceptual y algorítmico notablemente rico para el análisis y la interpretación de datos.

3.1.7. Clusterización jerárquica

La clusterización jerárquica de conocimiento se implementó por primera vez de forma prototípica en Prolog, en un sistema llamado SIPP [14]. La experiencia con SIPP condujo a refinamientos de la idea, lo que resultó en una segunda implementación, esta vez en Lisp. La implementación Lisp, denominada SIPS (Selector de proceso semi-inteligente).

Los resultados de este tipo de clusterización como en la Figura 4 son usualmente representados en forma de dendrogramas.

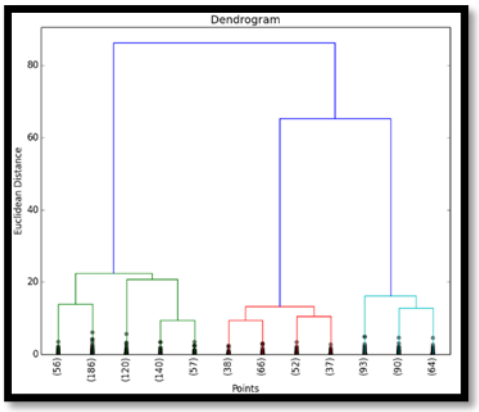


Figura 4 Ejemplo de dendrograma.

Se pueden encontrar en dos tipos [15]:

- Clusterización jerárquica divisoria: En este método, todos los documentos inicialmente son particionados en dos clústeres. Después cada uno de los clústeres que contengan más de un documento es seleccionado y seccionado en dos, repitiendo el proceso *n* veces hasta que ninguno de los clústeres se pueda particionar.
- Clusterización jerárquica aglomerativa: Este tipo de algoritmo inicialmente asigna un clúster a cada uno de los documentos. Después va seleccionando y uniendo en un nuevo clúster pares de clústeres hasta obtener un solo clúster con la totalidad de documentos.

3.1.8. Clusterización de redes sociales

La Clusterización utilizando el algoritmo de caminata aleatoria es un proceso que puede encontrar comunidades en una red, en otras palabras, cuando se usa un algoritmo de caminata aleatoria, escanea los nodos en algunos pasos; comienza con un nodo inicial y, en función de un proceso aleatorio, progresa a los nodos vecinos.

Una de las características más comunes de estas redes se llama estructura de la comunidad, que representa grupos conectados (clústeres) donde deben existir muchos enlaces dentro de cada grupo y pocos entre los grupos. Los grupos resultantes son una fracción de individuos que tienen características similares o están conectados a través de

relaciones. Los grupos en las redes sociales se corresponden con las relaciones sociales y se utilizan para comprender la estructura de datos, como las estructuras de la organización, la colaboración científica y las relaciones en las redes de telecomunicaciones.

El objetivo de la detección de la comunidad de grafos es la identificación de módulos y su estructura jerárquica mediante el uso de la información codificada en la topología de grafos [16].

3.2. Marco tecnológico

3.2.1. Node.js

De acuerdo con [17], Node.js es un manejador de eventos asíncronos en tiempo de ejecución de JavaScript. Node.js está diseñado para crear aplicaciones de red estables. HTTP es manejado de excelente manera en Node.js, diseñado pensando en la transmisión y la baja latencia. Esto hace que Node.js sea adecuado para la base de una biblioteca web o área de trabajo.

3.2.2. React.js

React es una librería javascript enfocada en el desarrollo de interfaces de usuario incluyendo aplicaciones web, es una librería de código libre desarrollada por Facebook y actualmente sigue siendo actualizada por la misma compañía. También react permite el isomorfismo lo cual significa que utilizando el mismo código se podrá renderizar tanto del lado del cliente como del servidor lo cual nos entrega ventajas extra al momento de ser consultado por un buscador.

3.2.3. Archivos JSON

JSON de acuerdo con [18], es un formato ligero de intercambio de datos. Es fácil para los humanos leer y escribir. Es fácil para las máquinas analizar y generar. Se basa en un subconjunto del estándar de lenguaje de programación JavaScript ECMA-262 3.ª edición - diciembre de 1999. JSON es un formato de texto que es completamente independiente del lenguaje, pero utiliza convenciones que son familiares para los programadores de la

familia de lenguajes C, incluido C, C ++, C #, Java, JavaScript, Perl, Python y muchos otros. Estas propiedades hacen de JSON un lenguaje de intercambio de datos ideal.

3.2.4. Visual Studio y C#.NET

El entorno de desarrollo integrado de Visual Studio [19] es una plataforma de lanzamiento creativa que puede usar para editar, depurar y compilar código, y luego publicar una aplicación. Un entorno de desarrollo integrado (IDE) es un programa rico en funciones que se puede utilizar para muchos aspectos del desarrollo de software. Además del editor y depurador estándar que ofrecen la mayoría de los IDE, Visual Studio incluye compiladores, herramientas de finalización de código, diseñadores gráficos y muchas más funciones para facilitar el proceso de desarrollo de software. Cuenta con varios lenguajes para desarrollar, entre ellos C# que es un lenguaje de programación simple, moderno, orientado a objetos y con seguridad de tipos.

3.2.5. Neo4j Base de datos de grafos

Una base de datos de grafos es una base de datos diseñada para tratar las relaciones entre los datos como igual de importantes para los datos como los estos mismos. Está destinado a contener datos sin restringirlos a un modelo predefinido. En cambio, los datos se almacenan como si se dibujaran primero, mostrando cómo cada entidad individual se conecta o se relaciona con otras. Neo4j [20], es la única base de datos de gráficos de nivel empresarial que combina el almacenamiento de gráficos nativo, la arquitectura escalable optimizada para la velocidad. La arquitectura de clúster distribuido de alto rendimiento de Neo4j permite las cargas de trabajo de ciencia de datos más desafiantes. Con Neo4j, puede elegir entre múltiples opciones en la nube: híbrido, multi nube o el servicio en la nube totalmente administrado, Neo4j Aura.

3.2.6. D3.js

Es una biblioteca de JavaScript para manipular documentos basados en datos. D3 le ayuda a dar vida a los datos mediante HTML, SVG y CSS. El énfasis de D3 en los estándares web le brinda todas las capacidades de los navegadores modernos sin atarse a un marco propietario, combinando poderosos componentes de visualización y un enfoque basado en datos para la manipulación del DOM [21].

3.2.7. React Force Graph

Este módulo [22] exporta 4 componentes de React con interfaces idénticas: ForceGraph2D, ForceGraph3D, ForceGraphVR y ForceGraphAR. Cada uno puede usarse para representar una estructura de datos de grafos en un espacio bidimensional o tridimensional utilizando un diseño iterativo dirigido por la fuerza.

3.2.8. Google Cloud

De acuerdo con [23] Google Cloud consiste en un conjunto de recursos físicos, como computadoras y unidades de disco duro, y recursos virtuales, como máquinas virtuales (VM), que se encuentran en los centros de datos de Google en todo el mundo. Cada centro de datos está ubicado en una región, específicamente en Asia, Australia, Europa, América del Norte y América del Sur. Cada región es una colección de zonas aisladas entre sí dentro de cada región.

4. Objetivos

4.1. Objetivo general

Desarrollar una aplicación basada en un modelo de distribución de conocimiento tácito a través de redes sociales como apoyo en la gestión de conocimiento en proyectos de innovación.

4.2. Objetivos específicos

- Conceptualizar y definir los requisitos de la aplicación cognitiva.
- Desarrollar una arquitectura para el proceso de administración de datos y consumo de una solución cognitiva web.

- Verificar el beneficio y viabilidad del uso de una herramienta de apoyo en un ambiente productivo.

5. Metodología

Para poder apoyar a las empresas generadoras de productos de innovación se desarrolló una solución cognitiva para apoyar en la identificación de redes de conocimiento, realizar los clústeres en base a la convergencia del conocimiento y al mismo tiempo poder ofrecer una utilería de búsqueda de artículos científicos usando las redes de conocimiento dentro de la base de conocimiento para brindar soporte durante la toma de decisiones. La herramienta cognitiva que se ha desarrollado usa la base de datos de ami-ner.org en su versión V4 [24].

5.1. Límites y alcances

Se desarrolló una herramienta web ("https://www.eradedatos.info/Inicio") la cual cuenta con una base de conocimiento extensa y las herramientas de visualización necesarias para realizar un análisis y procesamiento de dicho conocimiento dando como resultado modelos que asistirán en el proceso de desarrollo de innovación.

Los usuarios, investigadores o empresas desarrolladoras de innovación a los cuales está dirigido está herramienta tienen a su disposición la posibilidad de subir su propio conjunto de datos para realizar un análisis encapsulado en el entorno al cual se trabaja. A continuación, se enlistan los límites y alcances de la herramienta propuesta:

5.1.1 Limites

- Archivos entregables y reportes que se puedan extraer de la herramienta web.
- Automatización de procesos para la actualización de datos provenientes de repositorios.

5.1.2 Alcances

- Se desarrolló una solución cognitiva para el manejo de convergencia de conocimiento por medio de la visualización de relaciones enfocado para resolver el problema del procesamiento de conocimiento en dominios de estructura informal, utilizando artículos científicos.
- Se realizó la visualización de los clústeres de conocimiento y convergencia de conocimiento utilizando la teoría de redes sociales para la creación de conexiones

entre los datos de las fuentes de información y usando un método para su clusterización aglomerativo.

- Se realizó un sistema de búsqueda parametrizada por título, autor, palabras clave, cantidad de resultados y tipo de repositorio, además de una representación gráfica en donde se muestre la distribución de conocimiento mediante artículos científicos.
- Se realizó una herramienta de análisis de fuentes de conocimiento que ayuda a mantener actualizada la fuente de conocimiento que se utilizará para resolver la problemática, en donde se pueden ejecutar tareas de actualización de manera manual.
- Se realizaron tareas asincrónicas periódicas para la actualización de la base de datos que contenga metadatos de artículos científicos contenidos en repositorios.

5.2. Herramientas tecnológicas

La herramienta está desarrollada en el marco de trabajo de javascript React.js y las bases de conocimientos están almacenadas en Neo4j, por ser una base de datos orientada a grafos, la cual facilita el proceso de generación de modelos resultantes en la herramienta web. Para la realización de los análisis y procesamientos de la fuente de conocimiento se utilizó un método de clusterización jerárquico.

El desarrollo de este proyecto fue enfocado en una plataforma web para que pueda ser accedida por todo tipo de usuarios y en cualquier dispositivo. React siendo una librería javascript enfocada en el desarrollo de interfaces de usuario incluyendo aplicaciones web, es una librería de código libre desarrollada por Facebook y actualmente sigue siendo actualizada por la misma compañía. Esta librería permitió desarrollar la herramienta de una manera más ordenada y con menos código. También, React.js permite el isomorfismo, lo cual significa que utilizando el mismo código se puede renderizar tanto del lado del cliente como del servidor lo cual nos entrega ventajas extra al momento de ser consultado por un buscador. La versión utilizada es la 16.14.0 a la cual se le han integrado una serie de librerías auxiliares necesarias para los distintos procesos del proyecto, como:

- npm install d3 –save: Librería para visualización de datos con la utilización de SVG (Scalable Vector Graphics),
- npm i react-force-graph: Compilado de alto nivel de la librería D3.js usado para la generación de grafos 2d y 3D en base a archivos JSON.
- Canvas y HTML, utilizada para mostrar dendrogramas y diagramas de flujo.
- npm install --save neovis.js: Librería utilizada para la visualización de grafos, específicamente con datos provenientes de la base de datos Neo4j
- npm install neo4j-driver: Librería con driver de conexión para la base de datos Neo4j. Usada para realizar obtención, análisis y actualización de datos.

5.3. Redes de conocimiento

Con el alto crecimiento de publicaciones científicas y su correcta documentación, en la actualidad encontramos millones de artículos disponibles en librerías electrónicas con acceso al público en general, los cuales individualmente cada uno referencian una lista de artículos que representan el conocimiento base usado durante su desarrollo. De esa manera se han ido creando enlaces entre todas las publicaciones realizadas, algunos autores ganando muchas más conexiones gracias a la importancia o novedad del tema y algunos otros quedando una cantidad baja de enlaces como se puede ver en la Figura 5, con estas características podemos representar el conocimiento utilizando nodos y enlaces que formarán una red, esto es a lo que llamamos redes de conocimiento.

Título	#Referencias
C4.5: Programs for Machine Learning	360
Modern Information Retrieval	230

Figura 5 Artículos y referencias.

Pero si visualizáramos esta red de conocimiento resultante, no solo obtendríamos un ranking ordenado por número de referencias, si no que podríamos ver qué artículos se encuentran cerca o más lejos de la nube con mayor interés o donde se está desarrollando con mayor fuerza el tema y así poder discriminar aquellos artículos que a pesar de su número de relaciones no se encuentra tan cerca de aquellos que si son parte primordial de la red.

5.4. Bases y fuentes de conocimiento

Existen múltiples fuentes de bibliotecas virtuales de acceso abierto en web, de los cuales se puede acceder a los artículos en distintos formatos descargándolos directamente de la web del proveedor del servicio, incluso existe una iniciativa: OAI-MPH *Open archives Initiative Protocol for Metadata Harvesting*, el cual es un protocolo usado para la extracción de metadatos de repositorios públicos de información de artículos científicos. Estos metadatos son compartidos en archivos XML usando este servicio. De esta forma se puede acceder de una manera más amplia a las bases de datos públicas sin descargar los documentos enteros de las publicaciones.

Para la realización de pruebas que demuestren la utilidad, se usó la base de datos de la web de dblp.uni-trier.de Computer *science bibliography*, ver [25]. Esta web publica conjuntos de datos de documentos científicos relacionados con la ciencia de la computación, además, también se integró el repositorio de El Colegio de Chihuahua.

5.5. Nodos y relaciones

Para identificar los nodos y relaciones existentes entre artículos científicos utilizamos la definición de tipos de citas tal y como las nombra [7] en donde las "citas directas" son las que se encuentran dentro de las referencias de cada artículo, las "uniones bibliográficas" son aquellos documentos X, Y que tienen una cita directa a un documento A, entre ellos existe una unión bibliográfica, por último tenemos las "Co-Referencias", así llamamos la relación que existe entre los listados de documentos referenciados S y T en un documento inicial A tal y como se muestra en la Figura 6.

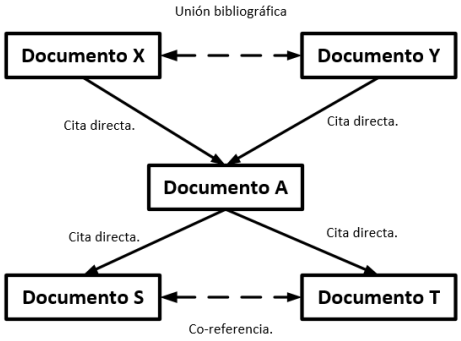


Figura 6 Definición de tipos de citas y relaciones.

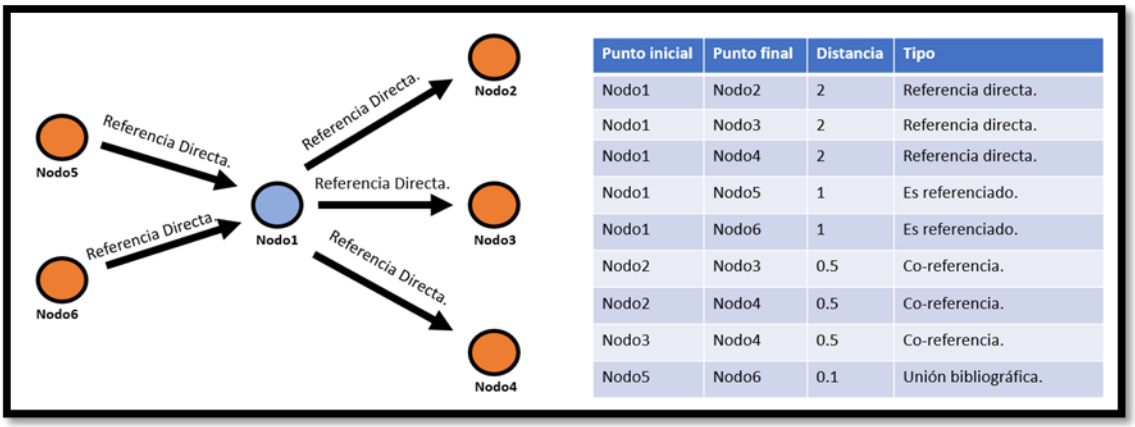
Los valores utilizados para identificar cada uno de los tipos de enlace fueron asignados usando números enteros y consecutivos, siendo los valores más altos asignados a los enlaces que generan una relación más cercana entre los nodos. Las referencias directas son los enlaces principales para las uniones de este proyecto, y contienen un valor y etiqueta “2”. Los enlaces por “Correferencia” son los enlaces de segunda importancia los cuales contienen un valor y etiqueta “0.5”. Por último, usamos los enlaces tipo “Unión bibliográfica” con un valor y etiqueta “0.1”. Estas valoraciones son utilizadas para la representación gráfica mostrando solo las uniones tipo “2” y “0.5”. Los enlaces de tipo “0.1” son mostrados en la representación gráfica solamente cuando un nodo o artículo científico aparece en los resultados, pero sin ningún enlace tipo “2” o “0.5” externos a los resultados iniciales. Esta categorización de enlaces entre artículos es de vital importancia para la solución cognitiva.

5.6. Clusterización de redes de nodos

Para la realización del algoritmo de clusterización aglomerativo se usaron métricos y métodos ajustados al dominio en el cual se trabajó con las relaciones dadas a través de relaciones por referencias y/o autorías de cada instancia de conocimiento. Las distancias entre cada nodo basados en el nivel de importancia e influencia están especificadas de la siguiente manera:

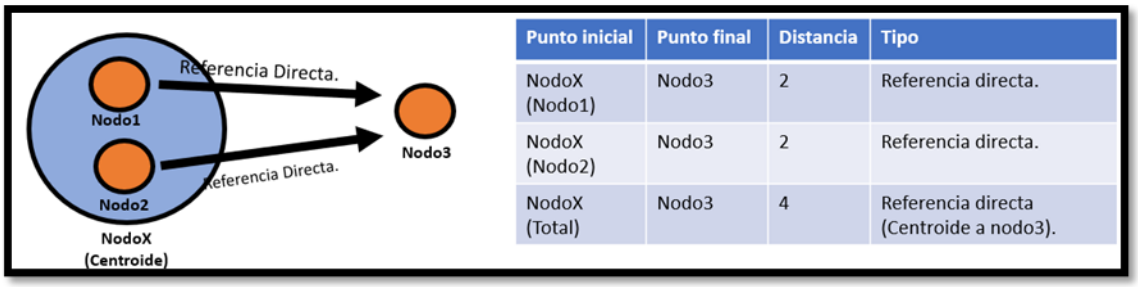
- Referencias directas de nodo de conocimiento a nodo de conocimiento = **2 puntos a cada nodo.**
- Ser referenciado por otro nodo de conocimiento **= 1 punto a cada nodo.**
- Cada referencia es co-referente entre sí, y cada uno de ellos gana un valor de distancia entre ellos de: **0.5 puntos.**
- Los nodos que tengan relación unión bibliográfica con otro nodo de tipo conocimiento obtendrán: **0.1 puntos.**
- Un autor puede ser autor de 1 o más nodos de conocimiento al mismo tiempo, estos nodos de conocimiento relacionados por su autor ganan entre ellos una segunda distancia de: **1 puntos.**
- Relación entre artículos por temática: **0.1 puntos**.
- Relación entre artículos por sub-temática: **0.5 puntos**.

Estas reglas se aplican por cada nodo y a todas sus relaciones existentes han sido ajustadas de tal manera que sus valores les den mayor prioridad a las relaciones más fuertes, sin embargo, estas son configurables. Inicialmente su valor individual no permite aumentar más de 1 vez sus valores en caso de que los nodos vengan repetidos sus enlaces al momento de leer los datos, pero al ir creando centroides sus valores por cada regla si pueden aumentar a X cantidad de nodos dentro de ese centroide.
En la Figura 7 y 8 se muestra un ejemplo más gráfico de estas reglas y sus distancias:



Punto inicial	Punto final	Distancia	Tipo
Nodo1	Nodo2	2	Referencia directa.
Nodo1	Nodo3	2	Referencia directa.
Nodo1	Nodo4	2	Referencia directa.
Nodo1	Nodo5	1	Es referenciado.
Nodo1	Nodo6	1	Es referenciado.
Nodo2	Nodo3	0.5	Co-referencia.
Nodo2	Nodo4	0.5	Co-referencia.
Nodo3	Nodo4	0.5	Co-referencia.
Nodo5	Nodo6	0.1	Unión bibliográfica.

Figura 7 Ejemplo de reglas de distancia aplicadas.



Punto inicial	Punto final	Distancia	Tipo
NodoX (Nodo1)	Nodo3	2	Referencia directa.
NodoX (Nodo2)	Nodo3	2	Referencia directa.
NodoX (Total)	Nodo3	4	Referencia directa (Centroide a nodo3).

Figura 8 Ejemplo de reglas de distancia aplicadas (Centroide).

Para encontrar la distancia más corta todas las distancias se multiplican por -1 y se obtiene el mínimo, esto ya que la relación o distancia entre los nodos entre más fuerte sea, su distancia de acuerdo con las reglas que especificamos aumenta su valor.

$$\boldsymbol{Distancia\ más\ corta = MIN(Distancia * -1)}$$

5.7. Arquitectura general

El proyecto está formado de 4 secciones en general:

- Interfaz Desktop.
- Página Web.
- Servidor en la nube.
- Base de datos Neo4j.

Cada una de estas secciones como se muestra en la Figura 9 fue desarrollada en una plataforma diferente y en conjunto nos ayudan a controlar el flujo y gestión de datos.

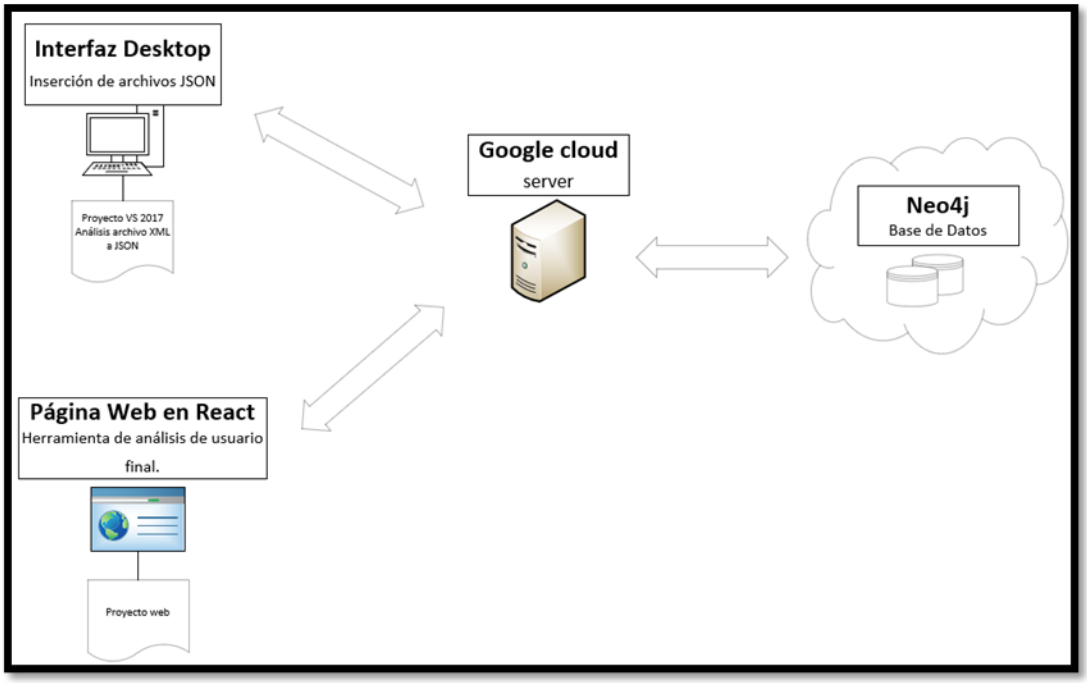


Figura 9 Arquitectura de la aplicación.

5.8. Base de Datos

La representación de la base de datos es simple por estar en una base de datos de grafos, como podemos ver en la figura 10 su estructura principal esta especificada por los nodos: “Artículo” y “People”, y sus relaciones por: “Referencia” y “Author”.

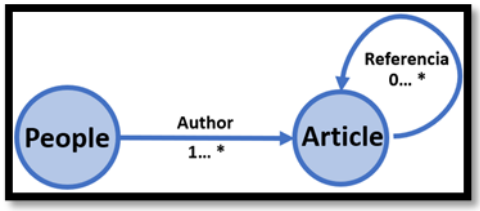


Figura 10 Definición de base de datos personas- artículos.

La estructura para la base de datos personal del usuario se maneja de la misma estructura con la excepción de que lleva la nomenclatura y nombres de nodos distintos como se muestra en la Figura 11.

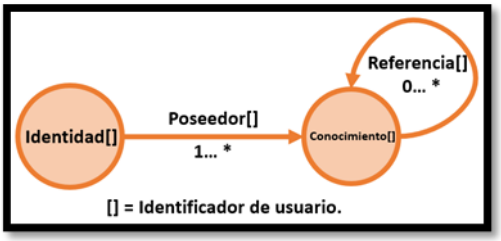


Figura 11 Definición de base de datos de Identidad - Conocimiento.

Los campos de cada uno de los nodos son dinámicos con la posibilidad de que cada nodo tenga más o menos información en sus columnas. Cuanta más información se obtenga de cada uno de los nodos o identidades de conocimiento, los algoritmos de Clusterización y análisis podrán realizar una mejor tarea. Con los datos que se trabajaron en este proyecto en los dos tipos de nodos son los siguientes que se muestran en la Figura 12. Estos datos son usados durante los filtros parametrizados y como información detallada al momento de seleccionar los nodos en la aplicación web.

Personas
Nombre
Segundo Nombre
Apellido
Facha de Nacimiento
Nacionalidad
Orcid
Id Único

Artículo	
Id Único	Mes
Título de Libro	Llave
Referencia cruzada	Titulo
Fecha	URL
Editor	Tipo
Editorial	Editor
ISBN	Volumen
Revista	Páginas
Lenguaje	Numero

Figura 12 Campos generales de los nodos de personas y artículos.

Para guardar los historiales se tiene un tipo de nodo “Historial” y se relacionan con el nodo de usuario en caso de que sean historiales de usuarios autentificados tal y como se ve en la Figura 13.

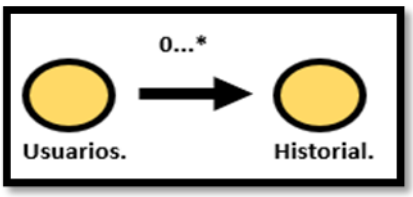


Figura 13 Definición de base de datos de Historial - Usuarios.

5.9. Obtención de datos (Interfaz de escritorio)

La información o metadatos de los artículos científicos fueron obtenidos de la plataforma dblp *Computer science bibliography* [25] siendo la biblioteca online con mayor número de publicaciones en ciencias de la computación. Usando el archivo XML versión 2019-08-08 22:31, se desarrolló una aplicación en lenguaje C# en Visual Studio la cual es la interfaz encargada de actualizar la base de datos de nuestro proyecto en Neo4j por medio de procesos que se tienen que ejecutar manual y periódicamente cada vez que sea necesario. Esta interfaz que se muestra en la Figura 14 está aislada al proyecto web ya que solo es la encargada de subir datos a la base de datos y solo los administradores del proyecto tienen acceso a ello.

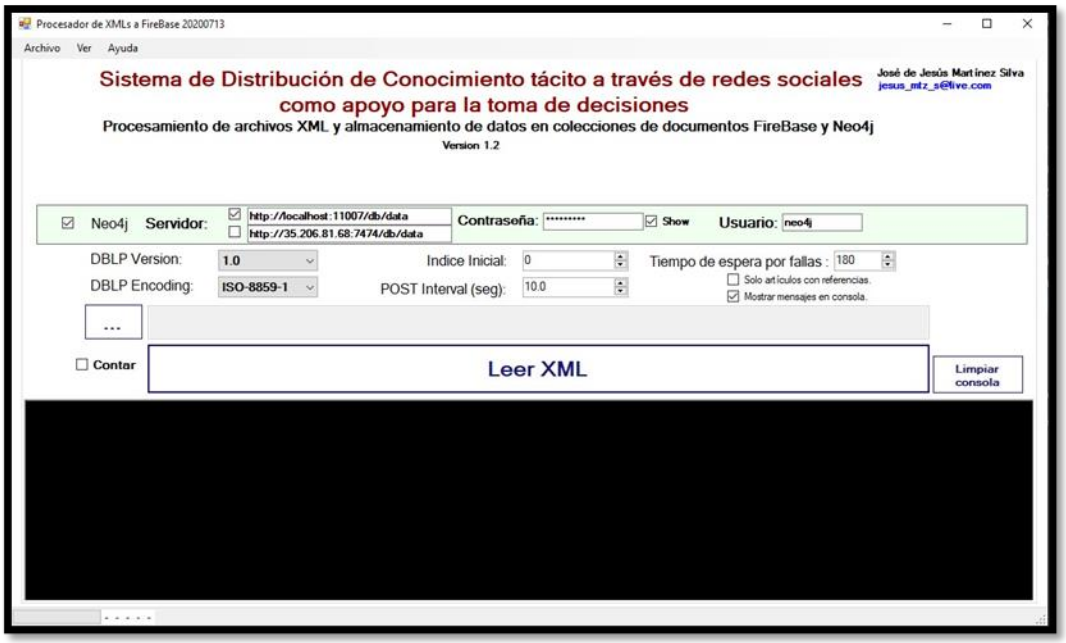


Figura 14 Interfaz de escritorio para la recolección de datos de repositorios públicos.

5.9.1. Módulo de conexión a Neo4j

La interfaz cuenta con una sección que después de realizar la lectura y proceso del archivo XML crea objetos tipo nodo para insertarlos en la base de datos Neo4j del servidor web. Para este módulo se utilizan los siguientes paquetes:

- Neo4j.Driver. Signed.
- Neo4jClient.
- Newtonsoft.Json.

Utilizados principalmente para la conexión a la base de datos Neo4j y para el manejo de objetos de tipo nodo. Para obtenerlos se pueden instalar directamente desde la sección de Nuget Packages de Visual Studio.

5.10. Interfaz WEB

La página web desarrollada en React.js que se muestra en la Figura 15, es la interfaz principal encargada del procesamiento y Clusterización jerárquica necesaria para la obtención de los resultados propuestos por el proyecto. La dirección pública es la siguiente: https://www.eradedatos.info/Inicio

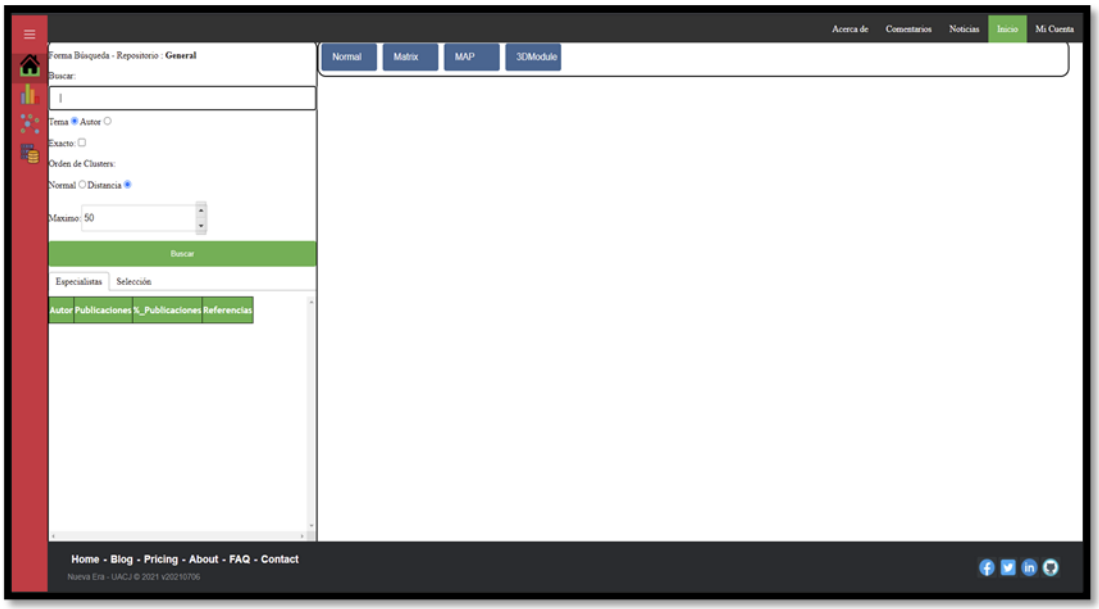


Figura 15 Interfaz web de acceso público para la realización del análisis cognitivo.

Esta interfaz cuenta con autentificación por seguridad para evitar que usuarios suban muchas bases de datos propias y saturen el servidor, también para poder mantener una

relación con las bases de datos personales de cada usuario. Aun así, la aplicación se puede usar sin necesidad de autentificarse.

El análisis principal es realizado por la aplicación web y mostrado con un grafo utilizando una librería llamada “React Force Graph” como se puede ver en la Figura 16, en donde se puede identificar fácilmente las uniones y la cohesión entre distintos nodos, entre mayor cohesión entre nodos de conocimiento podremos encontrar grupos de trabajo implícitos en el área o tema que estemos investigando. Al mismo tiempo se realiza un algoritmo de Clusterización, son los que estarán disponibles en el sistema para mostrar en cualquier nivel del proceso. Además de esto, también se podrá realizar un filtro por fecha para mostrar solo los nodos hasta cierto año y seguir respetando los clústeres generados.

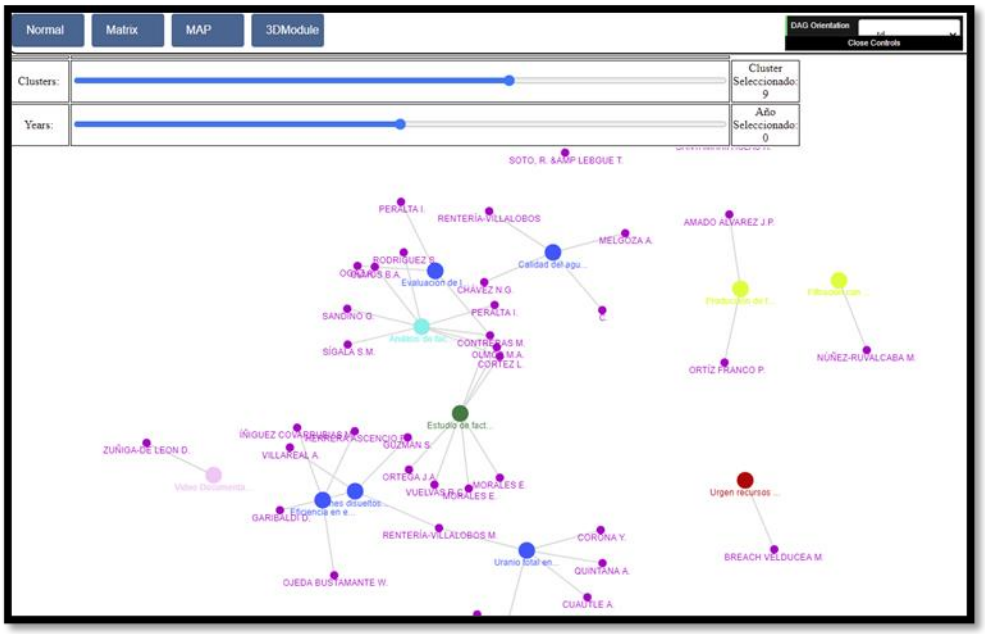


Figura 16 Grafo resultante del análisis cognitivo, representación de clústeres y controles de filtro visual para año de publicación y posición de proceso de clusterización.

Como información principal se muestra una tabla como la de la Figura 17, en donde se ranquea todos los autores disponibles en el análisis basado en el porcentaje de artículos en los cuales participa, también se especifica cuál tiene más probabilidades de ser el especialista del dominio de acuerdo con el proceso realizado.

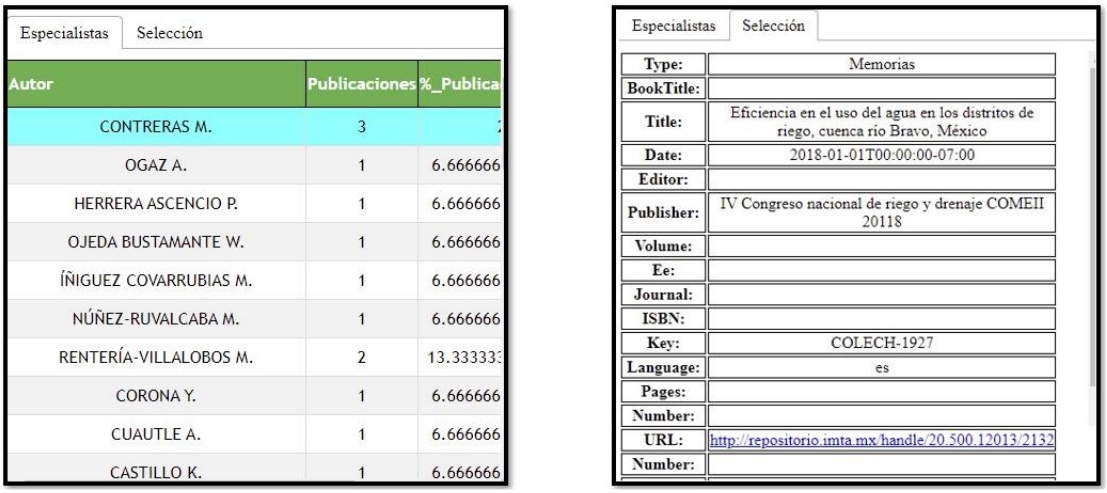


Figura 17 (Izquierda) Especialistas del dominio con ranking de participación. (Derecha) Metadatos de selección en el grafo resultante.

Los datos del análisis por Clusterización realizado también se pueden obtener en forma de grafo 3D como se pueden ver en la Figura 18.

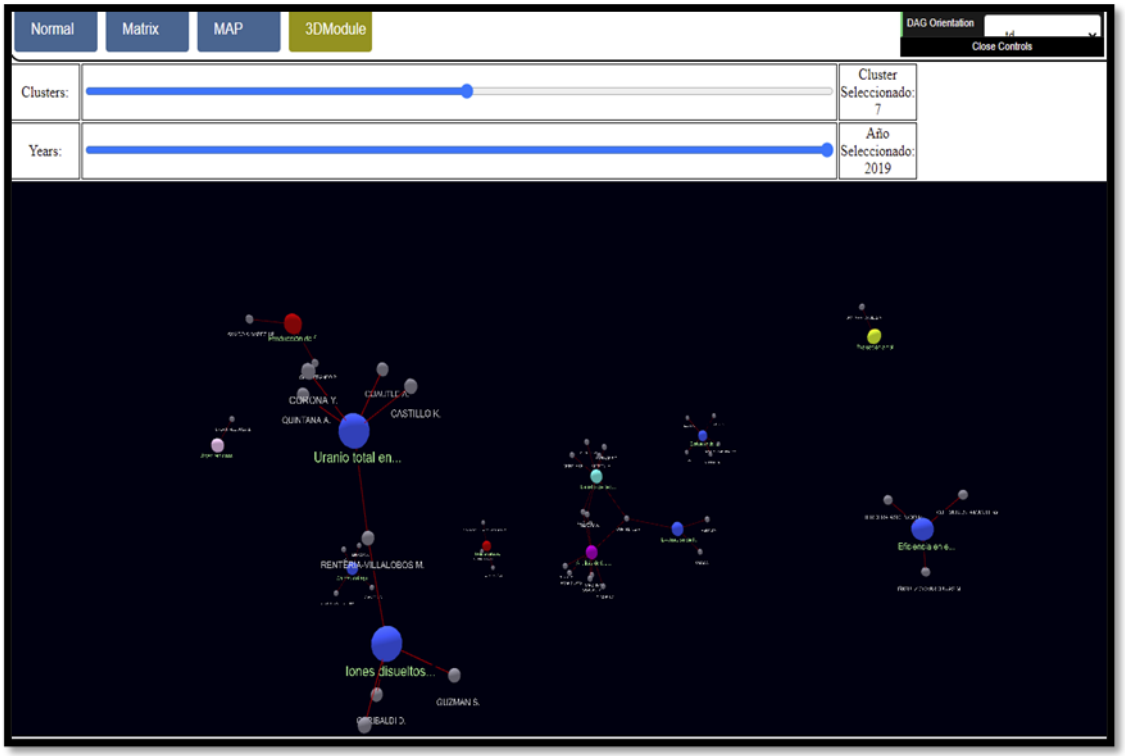


Figura 18 Representación de grafo resultante en 3D para la correcta visualización en resultados con grandes cantidades de relaciones.

5.11. Aplicaciones

Se trabajó con El Colegio de Chihuahua (COLECH), la cual es una institución pública de investigación y posgrado con autonomía e independencia académica, reconocida a nivel nacional e internacional por sus aportaciones a las ciencias sociales y humanidades, con funciones desde el año 2005. Esta institución cuenta con un repositorio público, la Biblioteca Virtual Ambiental del estado de Chihuahua (BVA), la cual es una colección de referencias y documentos digitales que abordan las problemáticas ambientales del estado de Chihuahua México en donde la información se encuentra organizada y clasificada.

5.11.1. Nuevo repositorio – COLECH

Se realizaron algunos scripts para leer la información directamente de los repositorios web, en este caso se ha utilizado el repositorio de El Colegio de Chihuahua (COLECH) y guardar su información en la base de datos de Neo4j. Los scripts corren en la interfaz desktop utilizando la opción que se muestra en la Figura 19 y 20.

Figura 19 Menú de acceso desde la interfaz de escritorio.

Figura 20 Interfaz de extracción de datos especializada en analizar el repositorio público web del COLECH.

Dentro de la aplicación WEB se agrega un menú como el de la Figura 21, para cambiar de repositorio y en donde se listarán los repositorios disponibles, de esta manera los procesos de análisis y filtrado se realizarán solo en un especifico conjunto de datos.

Figura 21 Menú de cambio de repositorio al del COLECH dentro de la aplicación web.

No se pueden generar relaciones a través de referencias usando el repositorio del COLECH ya que esta información no se encuentra disponible, además se contenían nuevas propiedades. Se modificó el sistema para incluir nuevas propiedades de los artículos:

- Palabras clave
- temática
- Sub temática
- Abstracto
- Repositorio (Esta es solo para poder filtrar la información)

Ya que en este repositorio no manejan en los metadatos de los artículos las referencias de cada publicación, fue entonces que se decidió utilizar temática y sub-temática como un enlace extra que nos ayudará a realizar las mediciones de distancias entre nodos.

Aprovechando que ahora se tienen varias reglas en la medición de distancias, se agregó un menú como se muestra en la Figura 22, para editar los valores que tenía especificados por default y así se pueda cambiar la relevancia en los tipos de enlaces entre los nodos.

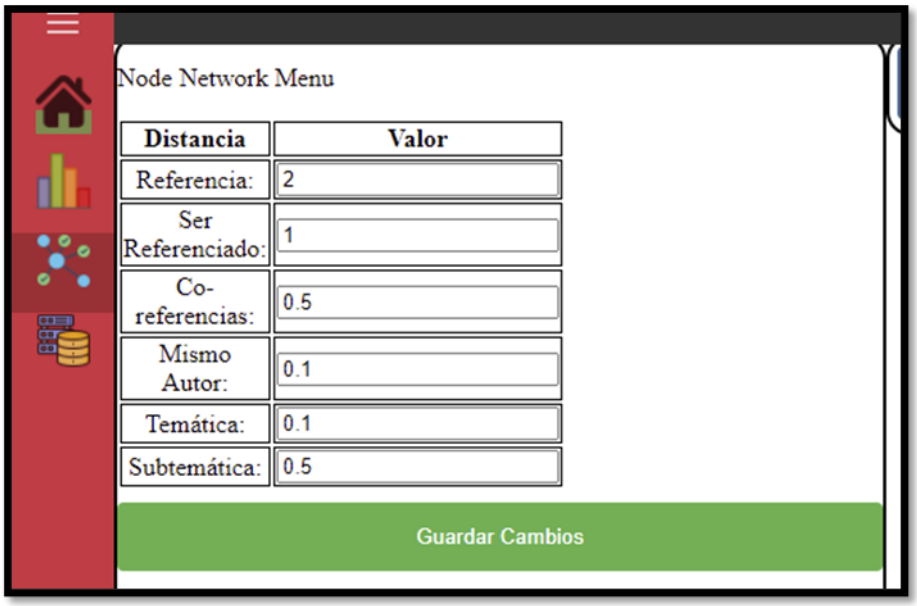


Figura 22 Menú en aplicación web para el ajuste de pesos en el tipo de relaciones usadas en el análisis cognitivo.

Por último, se agrega un acceso a la página para intentar redireccionar los usuarios desde el repositorio original a la página del proyecto con la selección previa al repositorio del COLECH usando el enlace: "https://www.eradedatos.info/COLECH".

5.11.1.1 Plan de actualización

Actualmente la manera de importar los datos de cualquier repositorio externo es realizando o ejecutando rutinas desde la interfaz web de esta aplicación, por lo tanto, cualquier plan para mantener la información actualizada se debe llevar a cabo en tareas periódicas controladas por el administrador.

Con el objetivo de mantener la información mostrada en la aplicación lo más acertada posible, se realizó un plan de actualización, el cual consta de correr las rutinas de lectura

del repositorio del COLECH cada mes al menos para así tener los nuevos artículos que este repositorio pueda tener.

5.11.2. Campaña pública

Para poder recabar más datos y llegar a tener un medible de los resultados, se adquirieron los dominios WEB de “Eradedatos.info” y “nuevaera.biz” uno para la aplicación web y otro para el servidor de base de datos respectivamente, con el fin de darle más confianza al uso de la aplicación además de obtener los certificados SSL.

Con los dominios y el host público se inició una campaña en redes sociales promoviendo el uso de la aplicación y pidiendo retroalimentación sobre la experiencia de uso en la sección de comentarios, teniendo como objetivo los países de habla hispana. Por este motivo se agregó una opción de comentarios como la de la Figura 23.

Figura 23 Menú de captura de comentarios en aplicación web.

En caso de que no se pueda obtener la información detallada a través de la sección de comentarios también se ha agregado un proceso en el cual se guardan datos históricos de uso de la aplicación para las siguientes actividades:

- Inicio de sesión.
- Registro de nuevo usuario.
- Búsquedas (proceso de análisis).
- Cambio de Repositorio.

- Accesos a enlaces externos de artículos científicos.

En cada uno de los casos, se obtiene información del usuario, su nombre en caso de tener sesión iniciada, en caso de ser usuario genérico solo se guarda su IP, fecha y proceso realizado.

6. Resultados

6.1 Resultados del sistema

Se analizó el repositorio general de la aplicación, el cual fue alimentado con la información del conjunto de datos de dblp.org [25] en su versión de noviembre 22 del 2019. Delimitamos el proceso de búsqueda usando de parámetro el tema “lenguaje natural” y con un límite de 1000 nodos. Los resultados se obtuvieron a partir de 48 segundos dándonos como resultado una Clusterización de 51 pasos (niveles de segmentos en el dendrograma) y artículos que varían desde el año 2002 al 2019 tal y como se puede ver en la Figura 24.

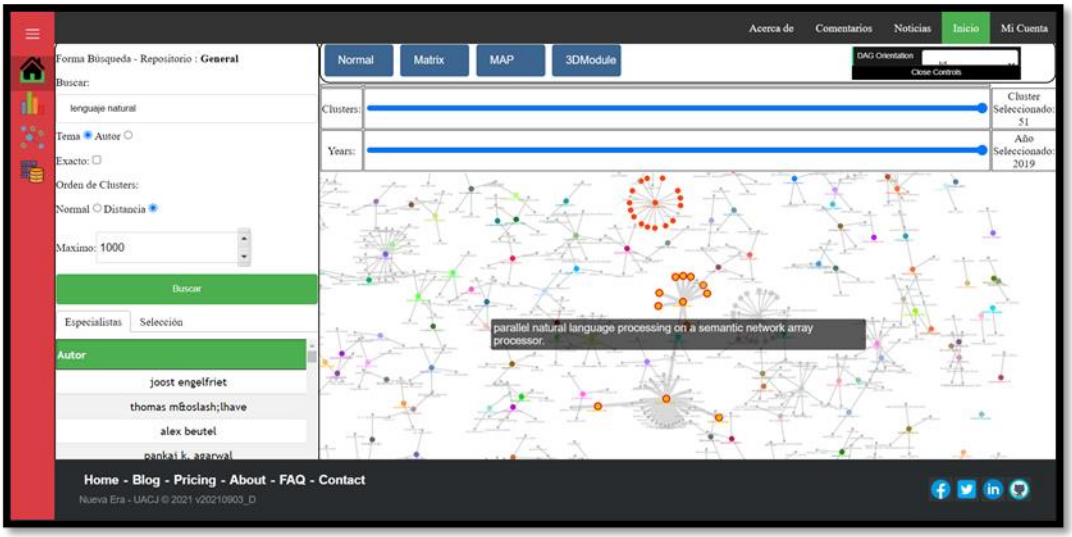


Figura 24 Resultados de análisis de repositorio general con parámetros de "lenguaje natural".

Se obtuvieron 881 autores de los cuales por la evaluación que les da el sistema podemos identificar a 13 autores como posibles especialistas del dominio mostrados en la Tabla 1.

Autor	Publicaciones	%_Publicaciones
alessandro fantechi	2	0.583090379
stefania gnesi	2	0.583090379
rui liu 0003	2	0.583090379
jinlong huang	2	0.583090379
bruce w. ballard	2	0.583090379
ralph m. weischedel	2	0.583090379
keith price	2	0.583090379
david m. w. powers	2	0.583090379
changhoon lee	2	0.583090379
xiaomin pan	2	0.583090379
simone deparis	2	0.583090379
nicholas zabaras	2	0.583090379
jan von plato	2	0.583090379
sergei nirenburg	2	0.583090379
aravind k. joshi	2	0.583090379
dan i. moldovan	2	0.583090379

Tabla 1 Autores con mayor rango en resultados obtenidos del ejercicio.

Usando como referencia los posibles autores como especialistas del dominio obtenidos en base a sus publicaciones y participación dentro de los resultados, se puede observar que el grupo más grande de nodos contiene 3 de los autores identificados mostrados en la Figura 25. Por lo tanto, este es el grupo de trabajo más grande encontrado en los resultados lo cual nos indicaría que gran parte del conocimiento para este tema lo estudian estos autores.

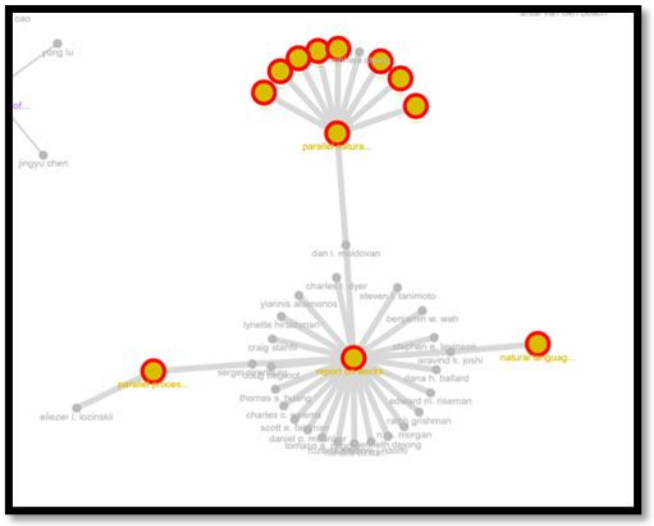


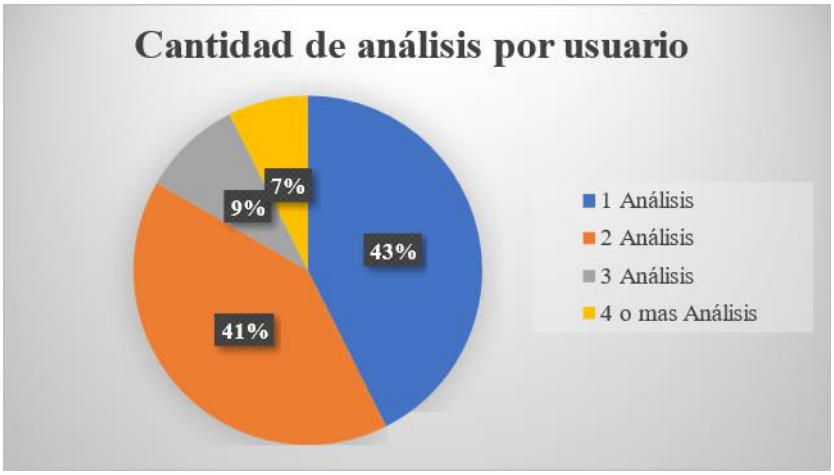
Figura 25 Grupo de trabajo con más cohesión en los resultados.

En cuanto a datos de uso externo por usuarios que han entrado a la aplicación gracias a que se encuentra en una web publica, los resultados recabados hasta el momento gracias a la información de seguimiento de procesos realizados del usuario en la aplicación WEB son los siguientes de la Tabla 2:

Tipo	QTY
Inicio de sesión	0
Cambio de repositorio.	141
Búsqueda	112
Enlaces externos	2
Comentarios	0

Tabla 2 Cantidad de respuestas obtenidas a través de comentarios y acciones de los usuarios.

En la Gráfica 1 tenemos la cantidad de análisis o procesos (búsquedas que desencadenan la ejecución del algoritmo de clusterizacion y visualización de datos) realizados por cada usuario de un total de 54 usuarios de acuerdo con los datos obtenidos:



Gráfica 1 Cantidad de análisis realizado por usuario.

6.2 Resultados del caso de aplicación

Se analizó todo el repositorio del COLECH que se migró a la base de datos de este proyecto (1070 artículos) y se realizó el análisis con el filtro de "Agua" como se puede ver en la Figura 25, el cual arrojó como resultado todos los artículos con tema, subtema y títulos relacionados con el "Agua".

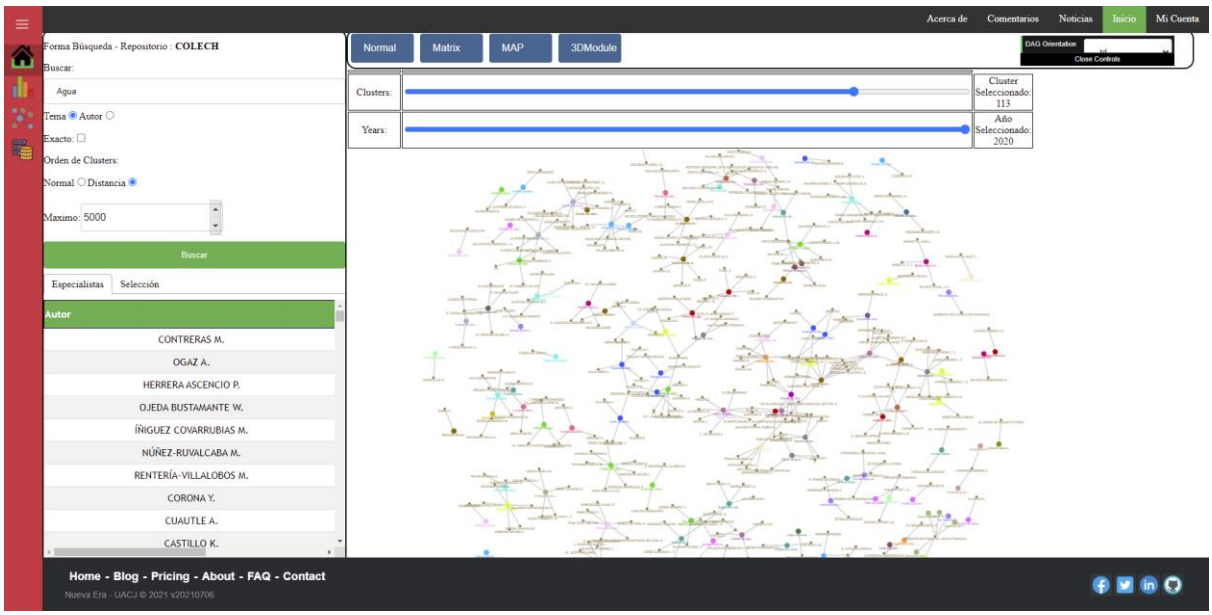


Figura 26 Grafo resultante del análisis del repositorio completo del COLECH con filtro "Agua".

En el análisis se obtuvo 363 distintos autores, que por la limpieza de datos en el repositorio no todos cuentan con una identificación clara. En la Tabla 3 se muestran los mejores 10 autores ranqueados por su porcentaje de participación con publicaciones y referencias en el conjunto de datos resultante.

Autor	Publicaciones	%_Publicaciones
COMISIÓN NACIONAL DEL AGUA (CONAGUA)	6	4.109589041
CONTRERAS M.	4	2.739726027
RENTERÍA-VILLALOBOS M.	3	2.054794521
G.	3	2.054794521
RUBIO-ARIAS	3	2.054794521
CORTEZ L.	2	1.369863014
OLMOS M.A.	2	1.369863014
AMADO ALVAREZ J.P.	2	1.369863014
I.	2	1.369863014
CHÁVEZ	2	1.369863014

Tabla 3 Mejores 10 autores con base a su porcentaje de participación en el análisis realizado.

El proceso de clusterización termina en 141 pasos en donde en el último paso resultan solo 4 clústeres y los artículos se encuentran en un lapso de años entre 1987 y 2020.

Hemos decidido utilizar el paso 130 de la clusterización para este ejercicio ya que la cantidad de clústeres no se encuentra muy extensa y ya podemos identificar algunos grupos fáciles de analizar cómo se puede ver en la Figura 26. El grupo más grande cuenta con 33 artículos en su red.

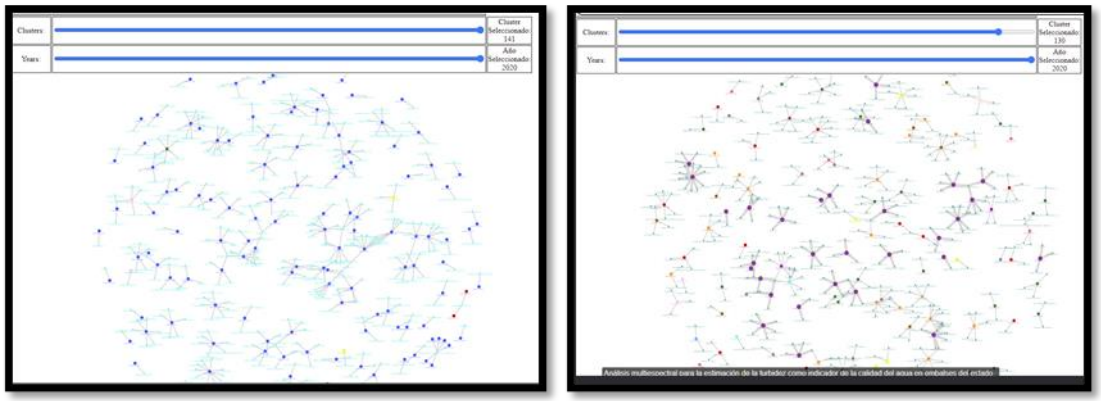


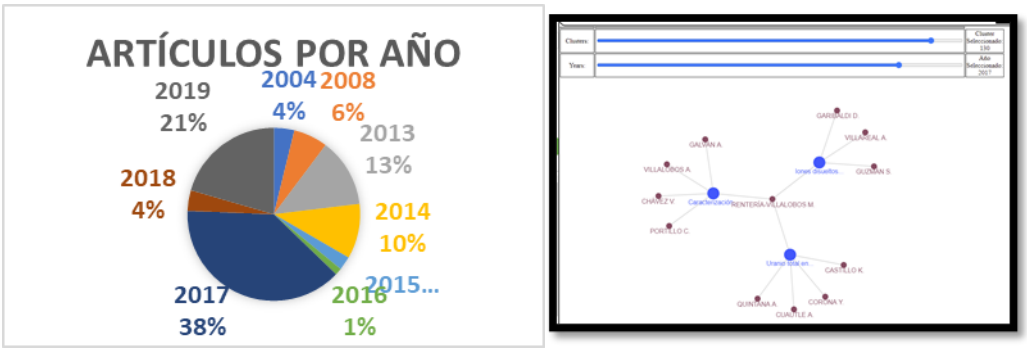
Figura 27 (Izquierda) Grafo con 141 pasos realizados durante la clusterización (Derecha) selección de red de trabajo más grande en clusterización 130

Dentro de este clúster al cual podemos denominar “red de trabajo” encontramos que hay 3 de los autores que se encuentran en los mejores 10 autores como se puede ver en la Tabla 4.

Autor	Publicaciones	%_Publicaciones
CONTRERAS M.	4	2.739726027
RUBIO-ARIAS	3	2.054794521
AMADO-ALVAREZ	1	0.684931507
CONTRERAS M.	1	0.684931507

Tabla 4 Autores con más participación en el dominio y que también aparecen en el clúster seleccionado.

En la Gráfica 2 podemos identificar que, de acuerdo con la fecha de publicación de cada artículo, el año 2017 es donde se realiza la mayor aportación en elementos al dominio, y es también en el 2017 en donde aparece la agrupación más grande de nodos y autores teniendo en cuenta que este repositorio no cuenta con relaciones de referencias entre los artículos.



Gráfica 2 (Izquierda) Porcentajes de elementos publicados por año. (Derecha) Grupo de nodos más grande creado en el año 2017.

7. Conclusiones

7.1. Conclusiones generales

Utilizando los datos obtenidos en las pruebas realizadas se puede definir de grupos de trabajo tal y como se presentan en la propuesta de objetivos, y da una pauta para que el campo de investigación del usuario tenga un orden a seguir con base a datos que la herramienta proporciona ayudando a tomar decisiones que durante etapas tempranas de la investigación pueden ser decisivas al momento de invertir recursos para este proceso.

Al ser una herramienta web esta aplicación cuenta con una utilidad abierta a todos los investigadores, además de que de esta manera tiene la posibilidad de mantener un control y un proceso para la actualización y uso de repositorios públicos que se pueden integrar al proyecto como un área de oportunidad. Una de estas áreas de oportunidad identificadas es la utilización del protocolo de iniciativa de archivos abiertos para la recolección de metadatos (OAI-MPH por sus siglas en inglés) el cual es un protocolo con estándares definidos que se empieza a utilizar cada vez más por repositorios de artículos científicos de distintas instituciones alrededor del mundo.

Se identifican algunas propiedades no incluidas en el proyecto que facilitarían al usuario al momento de extraer los datos, por ejemplo, algunos formatos para descargar la información o importarla para que así se pueda almacenar durante las tomas de decisiones en todo el proceso de la investigación realizada.

7.2. Conclusiones del caso de aplicación

De acuerdo con los datos presentados en este documento referentes a la aplicación del proyecto con el repositorio del COLECH, podemos extraer información importante como:

- El investigador cuenta con herramientas que ayudan a realizar un análisis que facilitan decidir por que grupos de artículos comenzar y dirigir la investigación.
- Utilizando la tabla resultante de autores con su desempeño en el dominio pudimos encontrar los mejores 10 autores que más participación han tenido, pero aun así es necesario identificar que autores de esa selección son los adecuados para poder denominarlos especialistas de nuestro dominio, ya que el área de estudio y análisis fue un tema muy general (Agua). Por lo tanto, podríamos decidir usar o no esta información en un primer plano dependiendo de qué tan especifica sea el área de análisis que se realiza.
- La selección del mejor corte en la clusterización queda a criterio del usuario y de sus recursos para realizar la investigación, pero gracias a la representación por grafos, pudimos identificarla y cambiarla de inmediato.
- A través del uso del filtro de artículos publicados hasta cierto año en combinación con la selección de un clúster, pudimos definir el orden en que los elementos han

enlazado la red de trabajo más grande dentro del clúster en el cual se haya decidido trabajar.

- Utilizando un corte en la clusterización, el grupo con más elementos y los 10 mejores autores en la tabla definimos los autores que forman parte de los especialistas del dominio.

8. Referencias

[1] Olmos, K. ,Rodas, J., "Requirements enngineering process model for informal structural domains.," *International Journal of Computer and Communication Engineering, 2:1,* pp. 75-77, 2013.

[2] SciELO, «Hemeroteca virtual conformada por una red de colecciones de revistas cientificas en texto completo y de accesso abierto y gratuito,» [En línea]. Available: http://www.scielo.org.mx/scielo.php#about.

[3] «Análisis de cantidad de publicaciones realizadas en los ultimos 28 años,» [En línea]. Available: https://analytics.scielo.org/w/publication/article?py_range=1990-2018.

[4] Garcia, R., & Calantone, R. , «A critical look at technological innovation typology and innovativeness terminology: a literature review.,» *Journal of Product Innovation Management: An international publication of the product development & management association, 19(2),,* pp. 110-132., 2002.

[5] B. V. Narayana, «Cognitive architecture: Orchestrator of innovation in organizations. Available at SSRN,» p. 272/522, 2016.

[6] «Dimensions,» 2021. [En línea]. Available: https://dimensions.ai.

[7] Kose, T., & Sakata, I., "Identifying technology convergence in the field of robotics research. Technological Forecasting and Social Change," *Technological Forecasting & Social Change,* pp. 146, 751-766, 2019.

[8] Hacklin, F., Raurich, V., & Marxt, C., «How incremental innovation becomes disruptive: the case of technology convergence.,» *IEEE International Engineering Management Conference (IEEE Cat. No.04CH37574),* vol. 1, pp. 32-36, October 2004.

[9] Camarinha-Matos, L. M., & Afsarmanesh, H., «Collaborative networks: a new scientific discipline.,» *Journal of intelligent manufacturing, 16(4),,* pp. 439-452, 2005.

[10] Tu, Y. N., & Seng, J. L., «Indices of novelty for emerging topic detection. Information processing & management 42(2),» pp. 303-325, 2012.

[11] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z., "Arnetminer: extraction and mining of academic social networks.," *In Proceedings of the 14th ACM SGKDD international conference on Knowledge discovery and data mining,* pp. 990-998, Agust 2008.

[12] Nasution, M. K., & Noah, S. A., «Extraction of academic social network from online database,» *International Conference on Semantic Technology and Information Retrieval,* pp. 64-69, June 2011.

[13] W. Pedrycz, «Knowledge-based clustering: from data to information granules.,» *John Wiley*

& Sons., 2005.

[14] Nau, D. S., & Chang, T. C., «ierarchical representation of problem-solving knowledge in a frame-based process planning system.,» *International Journal of Intelligent Systems, 1(1),,* pp. 29-44, 1986.

[15] Zhao, Y., Karypis, G., & Fayyad, U., «Hierarchical clustering algorithms for document datasets.,» *Data mining and knowledge discovery, 10(2), ,* pp. 141-168, 2005.

[16] Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A., "Comparing community structure identification, 2005(09), P09008.," *Journal of Statistical Mechanics: Theory and Experiment,* 2005.

[17] «Node.js,» 2021. [En línea]. Available: https://nodejs.org/.

[18] «JSON. json.org.,» [En línea]. Available: http://www.json.org.

[19] Microsoft. [En línea]. Available: https://docs.microsoft.com/en-us/visualstudio/get-started/visual-studio-ide?view=vs-2019.

[20] Neo4j, [En línea]. Available: https://neo4j.com/developer/graph-database/.

[21] d3js.org, [En línea]. Available: https://d3js.org/.

[22] npm.io, [En línea]. Available: https://npm.io/package/react-force-graph-vr.

[23] Google. [En línea]. Available: https://cloud.google.com/docs/overview/?hl=es-419.

[24] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su , «AmetMiner: Extraction and Mining of Academic Social Networks,» *In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008),* pp. (pp.990-998).

[25] «Computing science bibliography,» [En línea]. Available: https://dblp.uni-trier.de/xml/.

[26] Allen, G. N., & March, S. T., «A research note on representing part-whole relations in conceptual modeling.,» *MIS Quarterly,* pp. 945-964, 2012.

[27] EI-Aziz, A. A., & Kannan, A., «JSON encryption,» *International Conference on Computer Communication and Informatics,* pp. 1-6, January 2014.

[28] Tang, J., Zhang, J., Yao, L., & Li, J., «Extraction and mining of an academic social network,» *In Proceedings of the 17th international conference on World Wide Web,* pp. 1193-1194, April 2008.

[29] Newman, M. E., & Girvan, M., «Finding and evaluating community structure in networks,» *Physical review E, 62(2), 026113.,* 2004.

[30] Azizifard, N., Mahdavi, M., & Nasersharif, B., «Modularity Optimization for Clustering in Social Networks.,» *In International Conference on Emerging Trends in Computer and Image Processing.,* pp. (pp. 52-55), 2011.

[31] J. Duch, A. Arenas, «Community detection in complex networks using extremal optimization,» *Physical review E72, 027104,2005,* 2005.

[32] I, Czarnowski, P. J. edrzejowicz, «Agent-Based Non-distributed and Distributed Clustering,» *Department of Information Systems, Gdynia Maritime University,* 2009.

[33] J. Jimenez, «Descubre React,» *Recuperado en Octubre, 2 2018,* 2015.

[34] M. Bostock, «D3.js-data-driven documents. Online,» *Disponible: http://d3js.org,* 2015.

[35] Liu, W., Sidhu, A. Beacom, A. M., & Valente, T. W., «Social Network Theory,» *The International Encyclopedia of Media Effects.,* 2017.

[36] O. Borial, «Tacit knowledge and enviromental management.,» *Long range planning, 35(3),* pp. 291,317, 2002.

[37] J. Howells, «Tacit knoledge. Technology analysis & strategic management,» *8(2),* pp. 91-106, 1996.

[38] T. Honderich, «The Oxford companion of philosophy. OUP Oxford,» 2005.

[39] R. A. E. e. A. d. A. d. l. L. <. Española., «Diccionario de la lengua española (23a edición). Madrid: Espasa. ISBN 978-84-670-4189-7,» 2014. [En línea].

[40] Olmos-Sanchez, K. & Rodas-Osollo, J., «Knoledge Management for Informally Structured Domains: Challenges and Proposals Knowledge Management Strategies and Applications,» p. 85, 2017.

[41] Choi, J., Jeong, S., & Kim, K., «A study on diffusion pattem of technology convergence: Patent analysis for Korea. Sustainability, 7(9),» pp. 11546-11569, 2015.

[42] Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., ... & Taylor, A., «Cypher: An evolving query language for property graphs.,» *In Proceedings of the 2018 International Conference on Management of Data,* pp. (pp. 1433-1445), May 2008.

[43] Vargas, G., & Calvo, G., «Seis modelos alternativos de investigación documental para el desarrollo de la práctica universitaria en educación.,» *Educación Superior y Desarrollo, 5(3),,* pp. 7-37, 1987.

[44] Hacklin, F., Raurich, V., & Marxt, C., «How incremental innovation becomes disruptive: the case of technology convergence.,» *IEEE International Engineering Management Conference (IEEE Cat. No. 04CH37574),* vol. 1, pp. 32-36, 2004.

9. Anexos

A. Manual de Usuario

Acceso

La página web que se muestra en la Figura 28, es la interfaz principal encargada del procesamiento y clusterización necesaria para la obtención de los resultados propuestos por el proyecto. La dirección pública es la siguiente: https://www.eradedatos.info/Inicio

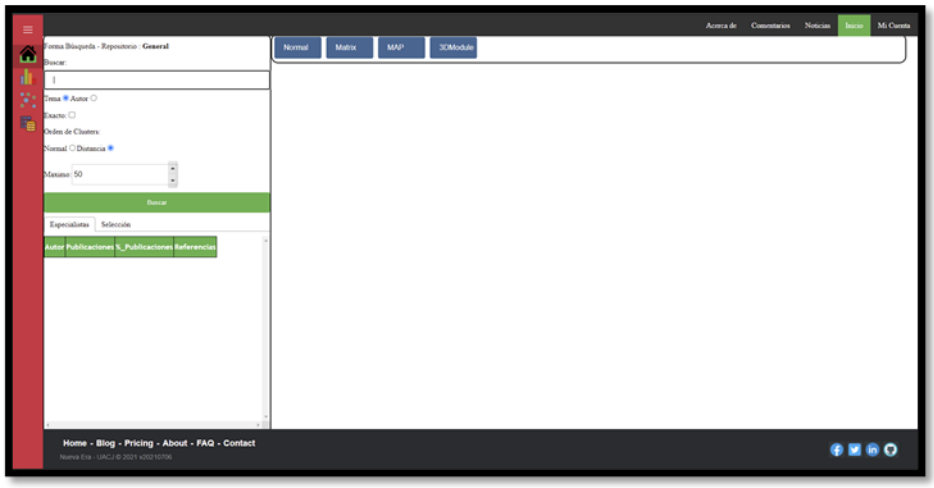


Figura 28 Interfaz web de acceso público.

Pantalla de análisis

La aplicación por default nos redirige a la pantalla principal de la aplicación en donde se pueden realizar los análisis y búsquedas de artículos científicos. Si es necesario volver a esta sección al estar navegando se pude hacer presionando el icono de inicio Figura 29.

Figura 29 Icono de inicio.

En la pantalla principal que se puede observar en la Figura 30, se puede encontrar 2 secciones:

1. Sección de menú de análisis.
2. Sección de gráficos.

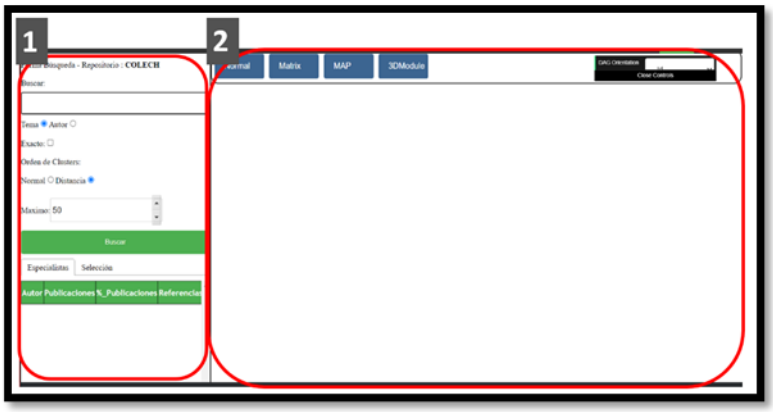


Figura 30 Sección 1: menú de análisis, sección 2: menú de gráficos.

Menú de análisis

En esta pantalla de menú de análisis como se muestra en la Figura 31, el funcionamiento de cada opción es:

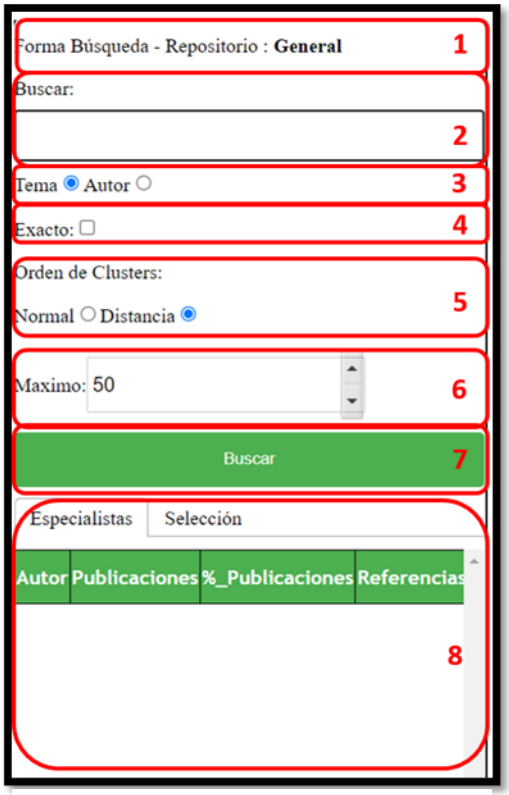


Figura 31 Menú de análisis

1. Título en donde se muestra también el repositorio actual donde se está trabajando.
2. Campo de búsqueda, aquí se insertan los parámetros de búsqueda.
3. Los filtros de búsqueda y análisis se realizarán con base en tema o autores según indique esta selección.
4. Este campo hace que la búsqueda sea por el parámetro exacto que se le especifica, de lo contrario el sistema separará las palabras para realizar la búsqueda de nodos.
5. Esta opción ordenara los clústeres resultantes en: "Normal" el orden es natural como se fueron descubriendo, y "Distancia" usara los valores de distancia entre los nodos para ordenar primero los de mayor distancia.
6. Cantidad de nodos resultantes en la búsqueda.
7. Acción que ejecuta búsqueda y análisis de resultados.
8. Sección en donde se muestran los especialistas del dominio y un ranking de todos los autores en base a los resultados obtenidos. En esta sección también se pueden ver los datos del nodo seleccionado.

Menú de gráficos

En el menú de gráficos como podemos ver en la Figura 32, se tienen las secciones:

- Selección de tipo de grafo.
    - Normal: grafo 2D.
    - Matrix: matriz de clústeres.
    - MAP: Dendrograma.
    - 3D Module: Grafo 3D.
- Selección de clústeres y selección de fecha o filtro de tiempo.
- Área del grafo, en esta área se puede manipular con el ratón de la computadora para manipular el grafo.

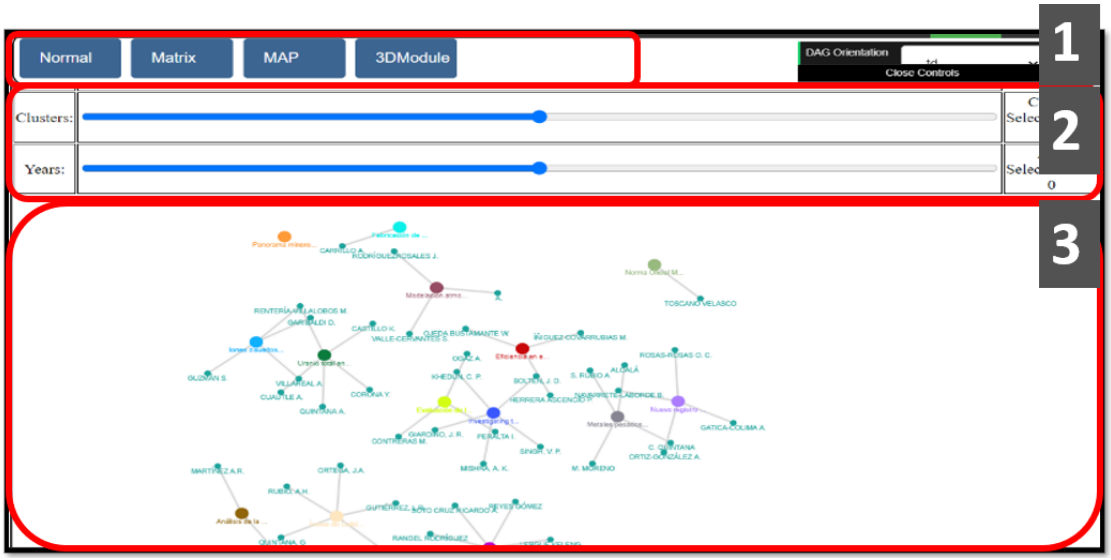


Figura 32 Secciones de menú de gráficos.

Repositorios

El sistema tiene la capacidad de realizar los análisis presentados delimitando el repositorio de búsqueda en específico. Para cambiar de repositorio es necesario seleccionar del menú principal la opción mostrada en la Figura 33.

Figura 33 Menú para cambiar de repositorios.

Al seleccionar esta opción se despliegan las opciones de repositorios disponibles, solo es necesario dar clic sobre la opción y el repositorio de búsqueda será cambiado. Los repositorios disponibles se muestran en la Figura 34.

Figura 34 Menú de selección de repositorios.

Administración de Pesos

Los pesos para la medición de distancias utilizados en el proceso de análisis están especificados por defecto, pero es posible ajustarlos a los valores que mejor se adapten a las necesidades del usuario. Para editar estos valores es necesario seleccionar el menú que se muestra en la Figura 35.

Figura 35 Menú de edición de pesos.

Al seleccionar esta opción se despliega un menú como en la Figura 36 en donde podremos editar cada uno de los valores utilizados por el sistema. Los valores deben ser mayores a 0 y para guardar solo es necesario presionar el botón de “Guardar Cambios”.

- Referencia: distancia con el artículo al cual se tiene la referencia.
- Ser referenciado: Distancia con el nodo que está referenciando al artículo actual.
- Co referencias: Distancia entre las referencias del artículo actual.
- Mismo autor: Distancia entre los artículos que comparten el mismo autor.
- Temática: Distancia entre artículos que comparten la misma temática.
- Sub temática: Distancia entre artículos que comparten la misma sub temática.

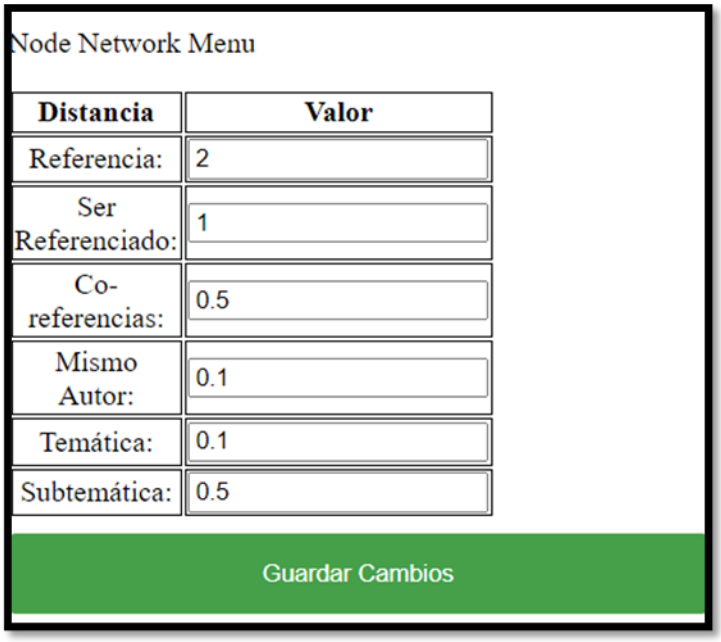


Figura 36 Formulario de edición de pesos.

Análisis de base de datos propia

En el momento que cambiamos de repositorio a "repositorio propio" el sistema busca realizar los análisis usando la base de datos propia del usuario quien utiliza el sistema, pero es necesario que previamente el usuario suba una base de datos propia para poder realizar este proceso. Para subir esta información es necesario seleccionar la opción que se muestra en la Figura 37.

Figura 37 Menú de base de datos propia.

Al seleccionar esta opción se mostrará un nuevo menú como en la Figura 38 en donde se podrá subir un archivo que contenga la base de datos propia.

Figura 38 Formulario para subir base de datos propia.

Para realizar este proceso debemos seleccionar un archivo seleccionándolo con la opción "Choose File" y buscar su localización en la computadora. El archivo seleccionado debe estar en formato JSON y debe estar en un formato especifico.

Formato de archivo JSON

El formato de los archivos JSON para que pueda ser procesado por el sistema debe ser tal y como se muestra en la Figura 39. El formato se divide en dos secciones: nodos y enlaces.

- Nodos. Los nodos son cada uno de los artículos científicos y cada uno de los autores de los artículos, las propiedades requeridas son las que se encuentran en la Figura 39, pero de ser necesario se pueden agregar más propiedades siguiendo el mismo formato.
- Enlaces. Los enlaces son las relaciones entre los artículos y/o autores, los cuales pueden ser de tipo "Referencia" o "Autoría".

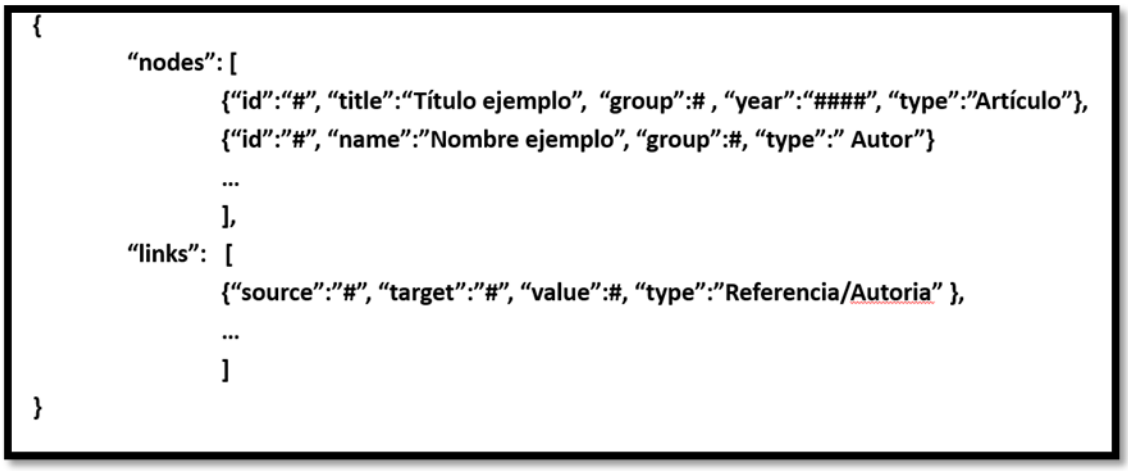


Figura 39 Formato de archivos JSON.

El nombre de las propiedades se escribe primero encerradas en doble comilla, después separado por dos puntos escribimos el valor de la propiedad también encerrada en doble comilla, a excepción de valores numéricos como en la propiedad "group", todos los demás valores de las propiedades deben estar entre comillas dobles.

B. Taxonomía de los Roles de Colaborador (con las actividades logradas)

Roles	Definición de los roles	Nombre de él(la) investigador(a)	Figura	Grado de contribución	Actividades logradas durante el proyecto	Tiempo promedio semanal (en horas) dedicado al proyecto
Autor	-Desarrollador del proyecto.	José de Jesús Martínez Silva	Desarrollador del proyecto.	-Principal	Desarrollo del proyecto.	10
Director	Coordinar la planificación y ejecución de la actividad de investigación.	Dr. Jorge Rodas Osollo	Director del proyecto	-Principal	-Coordinación del proyecto	4
Codirector	Coordinar la planificación y ejecución de la actividad de investigación.	Dra. Karla Olmos Sánchez	Codirectora del proyecto	-Principal	-Coordinación del proyecto	3
. Responsabilidad de supervisión	-Revisar progreso y avances semestralmente	Dra. Julia Patricia Sánchez Solís	Supervisora del proyecto	- De apoyo	-revisión de progreso semestral y recomendación de cambios.	0.5
Responsabilidad de supervisión	-Revisar progreso y avances semestralmente	Dr. Rogelio Florencia Juárez	Supervisor del proyecto	- De apoyo	-revisión de progreso semestral y recomendación de cambios.	0.5
Supervisora de caso de estudio	-Poner a disposición un caso de estudio real y recomendar cambios al proyecto.	Dra. Esmeralda Cervantes Rendón	Directora de caso de estudio	- De apoyo	-Poner a disposición un caso de estudio real en la biblioteca virtual de la UACH. -recomendar cambios al proyecto que se adapten al caso de estudio.	1

C. Estudiantes participantes en el proyecto

Nombre de estudiante(s)	Matrícula	Tiempo promedio semanal (en horas) dedicado al proyecto	Actividades logradas en la ejecución del proyecto
José de Jesús Martínez Silva	160519	10	Desarrollo