



Análisis de literatura científica en el dominio del cáncer usando técnicas de redes complejas

Jaime Andrés Hurtado Giraldo

jaime.hurtado@correounivalle.edu.co

Maestría en ingeniería con énfasis en Ingeniería de Sistemas

Director: Oswaldo Solarte Pabón, Ph.D.

Co Director: Victor Andrés Bucheli, Ph.D.

ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

FACULTAD DE INGENIERÍA

SANTIAGO DE CALI - COLOMBIA

2025

Resumen

El cáncer es una de las principales causas de muerte en la actualidad. Analizar grandes volúmenes de literatura científica es crucial para ayudar a los médicos a mejorar diagnósticos y tratamientos. En este sentido, la presente investigación logra construir un prototipo para encontrar patrones en literatura científica del dominio del cáncer usando técnicas de redes complejas. Para ello, se implementó la metodología DSRM (Design science research methodology) para desarrollar la herramienta I-ADC-Univalle que toma artefactos de software como librerías de ia, agregadores de información científica, y librerías de análisis de redes complejas para estudiar la literatura científica relacionada con el cáncer de mama y presentar patrones en términos de autores, instituciones, países y palabras clave relevantes; que de luces a los médicos en los casos clínicos. A través de la herramienta I-ADC-Univalle se realizó un estudio de caso utilizando 232 cadenas de búsqueda que permitieron construir un corpus de 11.348 documentos científicos relacionados con el cáncer de mama. Posteriormente, se procesó la información para encontrar las relaciones entre los metadatos y lograr la construcción y análisis de redes. El aplicativo permitió extraer los patrones representativos de la literatura consultada ofreciendo información relevante sobre las publicaciones científicas sobre el cáncer de mama.

Palabras Clave: Cáncer de mama, integración continua, análisis de redes, Cienciometría.

Tabla de contenido

1. Introducción	6
2. Marco teórico	8
Descubrimiento basado en la literatura	8
Análisis cuantitativo	9
Análisis de redes complejas	9
Grafo	10
Nodos, ejes y vecinos	10
Modelado de redes	11
Métricas de redes	11
Medidas de centralidad	11
Tipos de redes	12
Etapas en el análisis de redes sociales	15
Herramienta computacional	16
Integración continua	17
Entrega continua	17
Despliegue continuo	18
Servidor de integración continua	18
Contenerización y Docker	20
Kubernetes	20
Google Cloud Platform	21
3. Estado del Arte	21
Representación y resumen de información biomédica	21
Análisis de redes de colaboración científica	22
Cienciometría y visualización científica	23
Construcción de redes biológicas	25
Aplicación de aprendizaje profundo en análisis de imágenes y texto clínico	26
Curación de conocimiento y mapeo en identificadores biomédicos	27
4. Problema de investigación	28
5. Objetivos de investigación	30
Objetivo general	30
Objetivos específicos	30
6. Metodología	31
Búsqueda de bases de datos y agregadores de literatura científica	31
Sistema prototipo para la descarga de artículos científicos	33
Tecnologías para implementación de integración y despliegue continuo	33
Sistema prototipo basado en servicios web I-ADC-UNIVALLE	33

Extracción de información de artículos científicos	33
Revisión de literatura sobre análisis de información científica en el dominio del cáncer	34
Obtención de matrices de adyacencia	34
Sistema prototipo para el análisis de redes complejas	34
Estudio de UX/UI basado en encuestas a personal médico	34
Aplicativo web prototipo I-ADC-UNIVALLE	35
Patrones en literatura científica de cáncer de mama usando técnicas de redes complejas	35
7. Resultados	35
Búsqueda de bases de datos y agregadores de literatura científica	35
APIs para consulta de literatura científica	36
Librerías para interactuar con las APIs	38
PyPI - arXiv	39
Implementación Core	40
PyPI - Metapub	40
Ecuaciones de búsqueda	41
Sistema prototipo para la descarga de artículos científicos	42
Tecnologías para implementación de integración y despliegue continuo	42
Sistema prototipo basado en servicios web I-ADC-UNIVALLE	43
Implementación integración continua	44
Implementación despliegue continuo	46
Arquitectura	47
Extracción de información de artículos científicos	49
Metadatos a extraer	49
Implementación de la base de datos	50
Revisión de literatura sobre análisis de información científica en el dominio del cáncer	53
Obtención de matrices de adyacencia	54
Sistema prototipo para el análisis de redes complejas	56
Estudio de UX/UI basado en encuestas a personal médico	58
Desarrollo de un aplicativo web prototipo basado en microservicios I-ADC-UNIVALLE	59
Características del aplicativo	60
Manejo de usuarios y proyectos	60
Visualización de las redes construidas y sus métricas	62
Consulta de los nodos de interés sus metadatos e información relacionada	63
Patrones en literatura científica de cáncer de mama usando técnicas de redes complejas	64
Autores	65
Organizaciones	67
Países	68

Palabras clave	69
8. Conclusiones	70
9. Trabajo futuro	72
10. Referencias	73

1. Introducción

En 2015 se presentaron alrededor de 8,8 millones de muertes por cáncer en todo el mundo [1]. En esta enfermedad las células del cuerpo se dividen incesantemente y se diseminan a los tejidos circundantes [2]. Los tumores se pueden formar debido a muchos tipos de cáncer y pueden aparecer en varias partes del cuerpo utilizando la sangre o el sistema linfático como medio. Más de 100 tipos de cánceres afectan a los humanos como una enfermedad parcialmente genética y susceptible de heredarse [1]. En los hombres, los pulmones, la próstata, el colon y el estómago son los órganos más predominantes del cáncer. En las mujeres, los cánceres de mama, colorrectal, de pulmón y de cuello uterino son los más comunes. En particular, el cáncer de mama es la segunda causa universal de muerte en mujeres [3].

Por otro lado, se sabe que el proceso de cuidado y atención de pacientes diagnosticados con cáncer genera una gran cantidad de datos [4]. La comunidad médica y científica también genera grandes volúmenes de datos fruto del proceso investigativo, los cuales se reportan en artículos científicos. Analizar estos datos podría ser de utilidad para la comunidad médica. Sin embargo, la mayoría de estos datos son no estructurados, escritos en lenguaje natural, y su tratamiento representa varios retos. El uso de técnicas de inteligencia artificial ofrece una alternativa para abordar estos retos. Específicamente, las redes complejas se podrían usar para el análisis de grandes cantidades de información, la minería de datos y también para encontrar patrones en literatura científica. En cuanto a esto, [5] menciona que el descubrimiento basado en la literatura acelera el proceso de adquisición de conocimientos reflejando así una tendencia a implementar sistemas de búsqueda más inteligentes capaces de encontrar significados y conceptos en los textos publicados proporcionando búsquedas de mayor valor agregado.

Considerando este panorama, en el presente proyecto se propone el desarrollo de un prototipo de sistema informático para buscar y procesar literatura científica relacionada con el dominio del cáncer aplicando técnicas de análisis de redes. El sistema permitirá recopilar documentos, extraer metadatos y luego hacer un análisis de redes mediante la implementación de pipelines basados en microservicios y desplegados en la nube. Para ello, se espera implementar pipelines tecnológicos que permitan la selección y procesamiento de literatura científica relacionada con el cáncer, el análisis de los corpus de literatura obtenidos para revelar información relevante en la toma de decisiones médicas. Se espera que el sistema sea capaz de construir automáticamente redes de autores, organizaciones, países y palabras clave que estén relacionadas con el estudio de la enfermedad de cáncer, así como suministrar acceso inmediato al texto de los artículos y demás información disponible.

La metodología propuesta incluye una revisión sistemática de literatura (RSL) que permitirá encontrar el avance tecnológico relacionado con el análisis de corpus de literatura científica en el dominio del cáncer. Adicionalmente, se analizará y probará algunas tecnologías que muestren potencial para aplicar en este tema de estudio. Posteriormente se seleccionarán varias herramientas para el desarrollo e implementación de una plataforma que contribuya con la finalidad de presentar al personal médico información relevante, de fácil acceso y representativa de los grandes volúmenes de literatura científica disponible para el cáncer. Una vez se implemente el prototipo desarrollado, se evaluará la tecnología mediante un caso de estudio. Al finalizar el proyecto se espera que mediante un aplicativo web, el personal médico pueda recuperar literatura relacionada a criterios de búsqueda sobre la enfermedad del cáncer y obtenga datos que le permitan analizar e identificar los mejores autores, países e instituciones

reconocidas, para así contribuir a la selección de información actualizada y efectiva para el diagnóstico y tratamiento de los casos.

2. Marco teórico

Descubrimiento basado en la literatura

Por definición, el descubrimiento basado en la literatura (LBD) es un enfoque de minería de texto para generar automáticamente hipótesis de investigación. El objetivo de LBD es estimular y apoyar la creatividad humana para encontrar conexiones importantes entre literaturas dispares mediante la identificación de relaciones ocultas y previamente desconocidas del conocimiento existente. Hoy en día, LBD es un campo de investigación interdisciplinario y se considera una rama tanto de las ciencias de la computación como de las ciencias de la información [6].

El paradigma LBD de Swanson se basa en la noción de conceptos relevantes para tres dominios de la literatura: A, B y C como se muestra en la Figura 2.1. Por ejemplo, supongamos que hemos encontrado una relación entre una enfermedad A y un gen B. A continuación, supongamos que otro grupo de investigación ha publicado el efecto de un fármaco C sobre el gen B. El uso de la metodología LBD puede proponer una relación AC, lo que sugiere que el fármaco C puede potencialmente tratar la enfermedad A. Tal vínculo latente puede representar una hipótesis novedosa para una relación potencial, aún no confirmada.

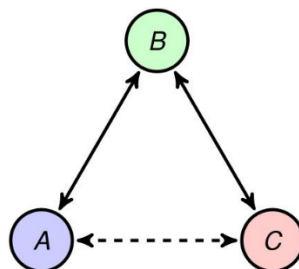


Figura 2.1. Modelo de descubrimiento ABC de Swanson.

Análisis cienciométrico

La revisión manual es un enfoque satisfactorio para presentar una descripción detallada del área de investigación estudiada. Sin embargo, es susceptible a problemas relacionados con las opiniones sesgadas del autor y las interpretaciones subjetivas. Una revisión sistemática de la literatura (SLR) tampoco puede caracterizar suficientemente todo un campo de estudio debido a su amplio rango de aplicación en diferentes dominios de investigación. Por su parte un análisis cienciométrico ayuda a encontrar asociaciones entre conceptos en la literatura a través de numerosas técnicas cienciométricas. Por lo tanto, ayuda a los investigadores a identificar nueva información en la literatura que puede ser ignorada en las revisiones tradicionales realizadas manualmente. El concepto de cienciometría puede verse como la evaluación cuantitativa de la investigación prevista para representar el crecimiento en el campo de interés. También se puede definir este concepto como una técnica para delinear el corpus de conocimiento, mejorar la comprensibilidad de los mecanismos de citación, medir la influencia de la investigación y resaltar las tendencias evolutivas en un dominio basado en datos bibliográficos [7].

Análisis de redes complejas

Una red, en general, es un conjunto de objetos/nodos con interconexiones. Estudiar y analizar las redes es de interés en la comunidad científica porque las redes están en todas partes, por ejemplo, redes sociales, redes de amistad, redes de comunicación por correo electrónico, redes de familiares, redes de transporte y movilidad, redes de información, internet, relaciones con sitios web, redes biológicas, interacciones de proteínas, redes financieras, etc. El análisis de redes ayuda a comprender fenómenos complejos [8].

Grafo

Un grafo es una estructura de datos común que consta de un número finito de nodos (o vértices) y un conjunto conectado de ejes (o aristas) [8]. Una de las aplicaciones de los grafos es representar y modelar las redes.

Nodos, ejes y vecinos

Un grafo se compone de una colección de elementos, conocidos como nodos, que están unidos por vínculos, conocidos como ejes, para formar una red de conexiones. En la Figura 2.2 se muestra un grafo usando una serie de círculos para representar los nodos y una línea para conectar cada par de nodos que están relacionados entre sí que corresponde a un eje. Este grafo consta de cuatro nodos indicados por las letras A, B, C y D, cada uno de los cuales está unido a los otros tres por ejes. Además, los nodos C y D también están unidos por una eje. Se requiere que dos nodos estén unidos por un eje para denominarlos vecinos [8].

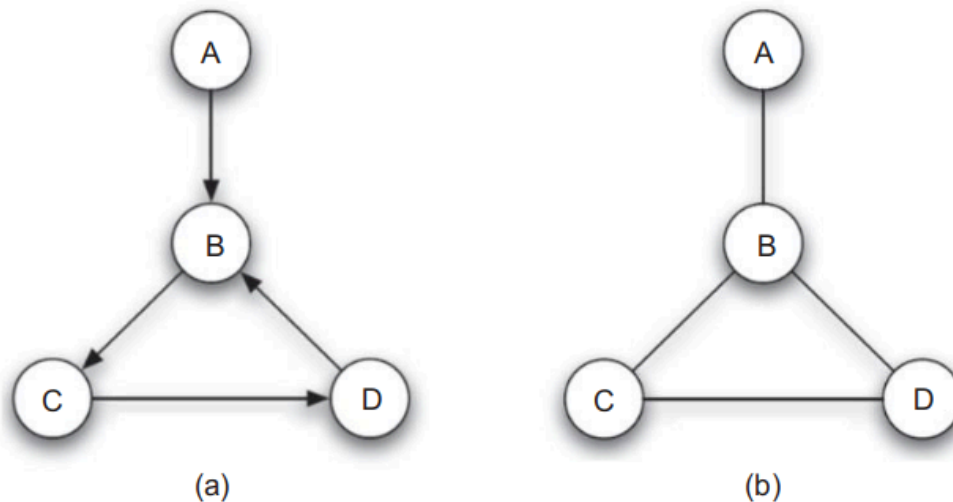


Figura 2.2. (a) Grafo dirigido con 4 nodos. (b) grafo con 4 nodos [8].

Un grafo dirigido (Figura 2.2a) se compone de una colección de nodos y una colección de ejes dirigidos. Los nodos dirigidos son conexiones entre nodos en los que la dirección de la

conexión es esencial. Se puede utilizar un grafo no dirigido (Figura 2.2b) para llamar la atención sobre el hecho de que hay una relación entre los nodos sin ser de interés establecer alguna propiedad adicional que indique la dirección de la relación [8].

Modelado de redes

Un modelo de red es un proceso que produce grafos con propiedades similares al sistema real. Sin embargo, la ventaja de utilizar un modelo es reducir la complejidad del mundo real a un nivel en el que se pueda tratar de una manera más práctica. Por lo tanto, las redes se consideran un medio poderoso para representar patrones de conexiones entre partes de sistemas como Internet, redes eléctricas, redes alimentarias y redes sociales [8].

Métricas de redes

Se ha propuesto diversos parámetros usados frecuentemente para la caracterización de redes y grafos entre ellos las medidas de centralidad, la distancia geodésica y métricas para caracterizar el tipo de red [8].

Medidas de centralidad

Las medidas de centralidad son un indicador importante utilizado en el análisis de redes. Existen diferentes tipos de medidas de centralidad. Algunas medidas destacadas se dan a continuación [8]:

- **Centralidad por intermediación:** Suponiendo que los nodos importantes conectan otros nodos, la centralidad por intermediación se define como el número de veces que un nodo está en el camino más corto entre dos nodos para todas las combinaciones de pares de nodos posible en la red.
- **Centralidad por cercanía:** Suponiendo que el grafo está conectado el nodo con mayor centralidad por cercanía es el que está más cerca de todos los nodos.

- **Centralidad por grado:** Suponiendo que en la red todos los nodos están conectados, puede haber uno o más nodos que tienen un valor alto de esta métrica por el mayor número de conexiones en comparación con otros nodos vecinos.

Tipos de redes

Por un lado, las redes tecnológicas como Internet son redes de infraestructura física que pueden representarse mediante redes estáticas, es decir, los nodos y ejes son fijos o cambian muy lentamente con el tiempo. Por otro lado, las redes que describen alguna forma de interacción social entre personas tienen una naturaleza intrínseca que varía en el tiempo y que debe tenerse en cuenta. Por ejemplo, en una red de colaboración científica, los autores no están en contacto con todos sus colaboradores simultáneamente durante todo el tiempo. Las redes de contacto reales son dinámicas con conexiones que aparecen y desaparecen con diferentes escalas de tiempo características.

Desde el punto de vista de los procesos dinámicos, ambas clases de redes (estáticas y temporales) son importantes. El primero ha sido ampliamente estudiado y es conveniente para la manejabilidad analítica, mientras que el segundo es, en algunos casos, más realista. Dentro de estas dos grandes clases, existen varias clasificaciones que se distinguen según su distribución de grado, elementos de conexiones como la distancia euclidiana u otras métricas [9]:

- **Redes Regulares:** En una red regular (Figura 2.3), cada nodo puede representar, por ejemplo, un individuo ubicado en una cuadrícula regular de puntos. Están conectados sólo con sus vecinos más cercanos. Las redes regulares son homogéneas cuando todos los nodos tienen el mismo número de conexiones.

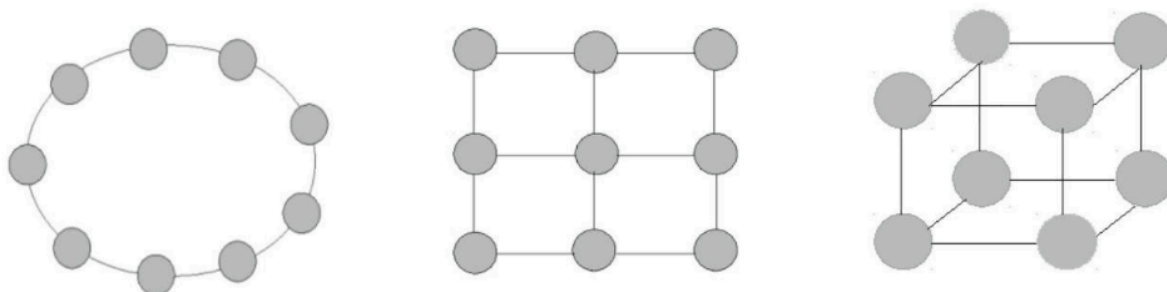


Figura 2.3. Redes regulares de una, dos y tres dimensiones, respectivamente [9].

- **Redes aleatorias:** Es un tipo simple de red que fue investigado por Erdos y Rényi en 1959. En su formulación original, corresponde a un grafo que se construye a partir de un conjunto de N nodos y todos los enlaces entre ellos tienen la misma probabilidad de existir. Esto genera un grafo en el que los nodos tienen un número de vecinos que no difieren mucho del grado promedio, cuando los nodos y ejes adoptan un valor finito, esta red corresponde a una distribución binomial, y cuando el número de nodos y ejes tienden a infinito, adoptan una distribución Poisson.
- **Red de mundo pequeño:** En 1998, Watts y Strogatz (WS) propusieron un modelo más realista inspirado en las redes sociales que pasó a ser conocido como el modelo del mundo pequeño. Las redes WS se construyen a partir de N vértices inicialmente dispuestos en una red unidimensional con condiciones de contorno periódicas, cada vértice tiene m conexiones con los vecinos más cercanos. Una observación característica de esta red es que a pesar de que la mayoría de los nodos no son vecinos entre sí, se puede llegar a cualquier nodo de la red con pocos saltos [9]. De manera similar ocurre en muchas redes del mundo real que se caracterizan por agrupaciones locales y rutas cortas. La estructura del fenómeno del mundo pequeño es un excelente ejemplo de lo que ocurre en las redes sociales, que resulta en la formación de redes masivas que deben ser

administradas por componentes masivos y herramientas de software científicas. Proporcionar modelos que se asemejan a mundo pequeño, o más precisamente, una muestra de esa red masiva en un tamaño más pequeño, es un método para analizar y obtener datos aproximados de la red [8].

- **Redes libres de escala:** La red libre de escala (Figura 2.4) es una alternativa a los modelos de red tradicionales, que brindan la configuración para simular redes grandes. Las principales propiedades que hacen que la red de mundo pequeño sea diferente a la libre de escala son que la red de mundo pequeño crece con el tiempo y los nodos y ejes están dispuestos a unirse a otros por medio de unión preferencial, es decir hay nodos muy populares que atraen las nuevas conexiones. La red libre de escala, Figura 2.4, contiene nodos con un alto grado de vecindad, que son los llamados nodos centrales, así como algunos nodos con un bajo grado de vecindad, donde el mayor número está con los nodos con menor grado o tasa de vecindad [8].



Figura 2.4. Red libre de escala [8].

El modelado de redes de mundo pequeño no podría replicar el comportamiento de la red libre de escala en determinadas circunstancias dado que el comportamiento de la distribución de grado de la red libre de escala sigue la ley de la potencia es decir unos pocos nodos tienen gran número de conexiones y muchos nodos tienen pocos vecinos.

Etapas en el análisis de redes sociales

El análisis de redes sociales (SNA) implica un proceso de tres etapas: captar, comprender y presentar [10]. En el primer paso, se reconoce la información relacionada con el tema de investigación y a través de las plataformas de redes sociales, bases de datos u otras fuentes en internet se obtiene una mayor cantidad de información. Los datos recopilados deben procesarse para pasar al siguiente paso. Los procesos incluyen conectar información de diversas procedencias, guardar los datos obtenidos en un único centro de datos en forma compilada, desarrollar modelos de datos de ciertos tipos, extraer partes importantes de los datos, eliminar datos no deseados y eliminar el ruido de los datos para lograr un resultado significativo en la captura de información [8].

La etapa de comprender es importante en SNA porque permite obtener una idea o identificar aspectos relevantes sobre la información consultada evaluando su significado y generando métricas útiles. Evaluar el significado de los datos obtenidos puede implicar varios métodos estadísticos y otras técnicas derivadas de la minería de textos, procesamiento del lenguaje natural, máquinas de traducción y análisis de redes. Dependiendo de las técnicas utilizadas y de la información que se busca, ciertos análisis pueden procesarse fuera de línea mientras que otros se calculan sobre la marcha utilizando estructuras de datos optimizadas para este fin. La intervención humana es válida para ayudar a la comprensión cuando se utilizan

análisis visuales para permitir ver varios tipos y representaciones de datos a la vez o crear porciones visuales que hacen que los patrones sean más evidentes [10].

La última etapa en el proceso de SNA es la presentación. En este punto los resultados de diferentes análisis se resumirán, evaluarán y mostrarán a los usuarios de una manera fácil de entender. Se pueden utilizar varias técnicas de visualización para presentar información útil. Uno de los diseños de interfaz más utilizados es el panel visual (Dashboard) que agrega y muestra información de diversas fuentes. Los análisis van más allá de simplemente mostrar información. Al admitir vistas personalizadas para diferentes usuarios, estos paneles ayudan a dar sentido a grandes volúmenes de información, incluidos patrones que son más evidentes para las personas que para las máquinas. Los analistas de datos y los estadísticos pueden dar un aporte adicional durante esta etapa [10].

Herramienta computacional

Muchos editores, bases de datos y productos académicos ofrecen herramientas que permiten a los usuarios acceder, analizar, extraer corpus para minería de datos, textos completos, y metadatos mediante programación para una variedad de propósitos de investigación. [11] muestra una lista de recursos académicos con licencia y disponibles para la comunidad educativa. Estos recursos recopilan esfuerzos para permitir que sus trabajos estén disponibles para su uso brindando un acceso programático y herramientas computacionales para usar sus recursos. Una interfaz de programación de aplicaciones (API), es una herramienta que se utiliza para compartir contenido y datos entre aplicaciones de software. Por lo general las API de diversos proyectos de investigación son abiertas y tienen una documentación que permiten utilizar estas herramientas o recursos, obtener una información general y saber cómo acceder a ellos.

Integración continua

En el desarrollo de software, la integración continua (CI) es la práctica de integrar regularmente todas las versiones de código de trabajo en un canal compartido o repositorio varias veces al día [12]. CI reúne todas las alteraciones de código en un solo lugar, prepara su producción, luego organiza y prueba el lanzamiento del código. La integración continua denota principalmente la infraestructura de pruebas y desarrollo de software con la intención de producir software de calidad con el menor esfuerzo y tiempo. CI es la fase inicial para la entrega y el despliegue continuos, ayuda a los desarrolladores a combinar partes de código en una rama de código principal sin interrumpir el código colectivo. Ejecuta la automatización de pruebas para que los errores no se introduzcan en el código y garantiza que el producto final no se rompa cada vez que se incorporan nuevas confirmaciones en la línea principal.

Entrega continua

La entrega continua (CD) es una práctica de desarrollo de software que funciona en conjunto con la CI para automatizar el proceso de aprovisionamiento de infraestructura y lanzamiento de aplicaciones. Una vez que el código se ha probado y creado como parte del proceso de CI, CD se hace cargo durante las etapas finales para garantizar que esté empaquetado con todo lo que necesita para implementarse en cualquier entorno en cualquier momento. El CD puede cubrir todo, desde el aprovisionamiento de la infraestructura hasta la implementación de la aplicación en el entorno de prueba o producción. Con el CD, el software está diseñado para que pueda implementarse en producción en cualquier momento. Luego, puede activar los despliegues manualmente o pasar a un despliegue continuo, donde las implementaciones también están automatizadas [13].

Despliegue continuo

El despliegue continuo o implementación continua es una versión totalmente automatizada y altamente avanzada en la entrega continua. Luego de pasar por los procesos mencionados de producción, los cambios se pueden enviar a sus usuarios a través de versiones automáticas, esto también elimina la necesidad de intervención humana [13]. Los despliegues automáticos permiten al equipo concentrarse en el trabajo sin preocuparse por el despliegue. Pueden tener retroalimentación de los clientes con mucha rapidez y comprobar los resultados a los pocos minutos de completarlo. CD sienta las bases para la entrega continua y el despliegue continuo, son parecidos porque ayudan a liberar cambios; sin embargo, las versiones del despliegue se producen automáticamente. Tanto el despliegue como la entrega se producen cuando las modificaciones han pasado por cada fase de desarrollo y están listas para ser lanzadas. El objetivo del Despliegue Continuo es realizar despliegue a cualquier hora del día sin detener el servicio. De esta manera, se puede decidir el ritmo que mejor se adapte a los requisitos del negocio y publicarlo mensualmente, semanalmente o diariamente, siempre será mejor hacer entregas más rápidas y pequeñas.

Servidor de integración continua

Un servidor de integración continua esencialmente administra el repositorio compartido y actúa como administrador del código que ingresa. Cuando los desarrolladores hacen un aporte al repositorio de código, el servidor de CI inicia una compilación y documenta los resultados de la misma indicando si el código cumple con los requerimientos configurados. De los equipos que practican CI, la mayoría confía en un servidor de CI para automatizar las compilaciones. Para los equipos que no utilizan un servidor de CI, aún pueden lograr una integración continua usando scripts que ellos mismos crean o activar compilaciones manualmente. La integración continua no

es una herramienta en sí misma, es un marco más amplio con un conjunto de prácticas asistidas por ciertas herramientas. La mayoría de las herramientas que consideramos herramientas de integración continua son, de hecho, servidores de CI. Algunos incluyen funciones adicionales, como gestión de código fuente, entrega continua y pruebas. Algunos servidores de CI son: GitLab CI, Jenkins y CircleCI. Los servidores de CI son útiles en el desarrollo de software porque pueden ofrecer ciertas ventajas en el proceso [14]:

- **Pruebas automatizadas:** Cuando el código se confirma mediante un servidor CI, los errores en el código se detectan automáticamente. El servidor prueba el código y proporciona retroalimentación rápidamente, sin que éste tenga que iniciar una compilación manual.
- **Colaboración:** En lugar de depender de los desarrolladores para mantener actualizado el repositorio compartido, el servidor de CI administra todo el código que ingresa y mantiene la integridad del código fuente mediante compilaciones y pruebas automáticas. Esto permite a los desarrolladores centrarse en sus propios proyectos en lugar de preocuparse de que otros proyectos rompan sus pruebas. Los equipos pueden colaborar sin preocuparse por el código.
- **Flujo de trabajo optimizado:** Sin servidores CI, los desarrolladores trabajan en diferentes capas de la aplicación con código guardado en máquinas locales. Si bien los equipos que realizan pruebas manuales o autogestionadas pueden solucionar estos posibles problemas, un servidor de CI elimina esta capa adicional de coordinación del flujo de trabajo.

Contenerización y Docker

La contenerización es una forma de ejecutar múltiples aplicaciones de software en una misma máquina, cada una de las cuales se ejecuta en un entorno aislado llamado contenedor. Un contenedor es un entorno cerrado para el software que incluye todos los archivos y librerías que la aplicación necesita para funcionar correctamente. Se pueden implementar varios contenedores en la misma máquina y compartir los recursos. Docker usa imágenes para generar contenedores las cuales son una copia del sistema completo con su estado en un momento determinado. Se pueden crear varios contenedores utilizando una imagen. Las imágenes se pueden crear en forma local o se pueden publicar y descargar por medio de DockerHub, un repositorio en línea administrado por Docker [15].

Kubernetes

Kubernetes también se escribe como K8, que significa piloto en griego. Entonces, es el capitán del barco con todos los contenedores corriendo por dentro. Sus responsabilidades de gestión de contenedores incluyen el despliegue, escalado, desescalado y balanceo de carga de los contenedores. Kubernetes es una plataforma de código abierto para gestionar cargas de trabajo y servicios en contenedores y para automatizar la implementación, escalado y orquestación de aplicaciones en contenedores. Kubernetes trata de ejecutar múltiples contenedores conectados, todos organizados en pods. Así como Docker es el estándar de facto para contenedores, Kubernetes es el estándar de facto para la orquestación de contenedores. Kubernetes no limita los tipos de aplicaciones admitidas y tiene el poder de admitir una variedad de cargas de trabajo y de procesamiento de datos. Si el contenedor puede ejecutar la aplicación, Kubernetes definitivamente puede ejecutar la aplicación [15].

Google Cloud Platform

Google Cloud Platform (GCP) ofrece una amplia gama de servicios para proteger, almacenar, servir y analizar datos. Estos servicios en la nube forman un perímetro de nube seguro para los datos, donde se pueden realizar diferentes operaciones y transformaciones sobre la información sin que estos abandonen nunca el ecosistema de la nube. Los servicios ofrecidos por Google Cloud incluyen computación, almacenamiento, big data/análisis, inteligencia artificial (IA) y otros servicios de gestión, desarrollo y redes [16]. GCP puede ser utilizado en el proceso de CI y CD, debido a los distintos servicios que ofrece. Un pipeline utilizando un servidor de CI y una app de grandes capacidades es factible de implementar utilizando GCP.

3. Estado del Arte

Numerosos trabajos realizados durante los últimos cinco años emplean técnicas para encontrar patrones en corpus de literatura científica de cáncer [2]. A continuación se realiza una revisión de los trabajos relacionados con el análisis de corpus de literatura científica en el dominio del cáncer que han sido publicados entre los años 2018 y 2023.

Representación y resumen de información biomédica

Representar y resumir información puede contribuir en estudios de cáncer. Casama [17], un sistema de representación y resumen para la literatura biomédica sobre el cáncer de pulmón. Se caracteriza por agregar contexto a la información recopilada tanto a nivel de estudio, como a nivel de paciente y entorno. Múltiples usuarios revisaron y calificaron las anotaciones encontrando mejores capacidades de resumen de Casama comparadas con SemRep [18], una representación similar pero libre de contexto.

Un método [19] de análisis de red de texto y modelado de temas permitió explorar las principales tendencias en la investigación de niños y adolescentes sobrevivientes de cáncer mediante la estructuración de una red de palabras clave (morfemas semánticos) y su coocurrencia en los resúmenes de artículos publicados en 5 bases de datos importantes de 2016 a 2020. Éste estudio es significativo porque confirma la estructura de conocimiento basada en las principales palabras clave y las tendencias transversales en la temática de investigación.

Análisis de redes de colaboración científica

Algunos trabajos se basan principalmente en el análisis de redes y grafos que ayudan a destacar información de literatura científica en el dominio del cáncer. [20] mediante un método automatizado utiliza el conocimiento publicado en la literatura para sugerir la información relevante mediante la implementación de modelos que combinan la lectura de máquina, la métrica basada en la clase de frecuencia y el análisis grafos. Se logra al organizar los eventos extraídos de la literatura un grafo de colaboración que depende de la frecuencia de aparición de eventos en la literatura. Se muestra que las métricas comunes de centralidad de grafos varían en la evaluación de los eventos extraídos y evalúan la fiabilidad del método propuesto utilizando algunos casos de uso.

En el trabajo de [2] se investiga la visibilidad de países e instituciones en términos de su colaboración en investigación en el dominio del cáncer en el contexto de documentos científicos. Para ello examinan los patrones estructurales de los coautores internacionales y las coinstituciones en los artículos considerando las citaciones. El estudio utiliza métricas de análisis de redes sociales, incluida la centralidad de grado, la centralidad de intermediación, la centralidad de vector propio y la eficacia para investigar los efectos de las redes sociales en el área de investigación del cáncer.

En [21] también se usó análisis de redes sociales para evaluar la coautoría de artículos científicos de 2007 a 2014 entre científicos que son miembros de cuatro programas de investigación formales en el Centro de Cáncer de Markey (MCC) de la Universidad de Kentucky. Lograron obtener cambios descriptivos de las redes a lo largo del tiempo e implementaron modelos de grafos STERGM para estimar el efecto de las variables del autor y de la red en la tendencia a formar un vínculo de coautoría. Encontraron que con el tiempo, los miembros de MCC colaboraron más con otros fuera de su programa de investigación y fuera de sus densos grupos iniciales de coautoría. Sin embargo, la formación de vínculos continúa siendo impulsada por la coautoría con personas del mismo programa de investigación y departamento académico.

El estudio de [22] utiliza indicadores bibliométricos y de integración de redes para analizar la actividad científica sobre la neumonía asociada a ventilación mecánica. A diferencia de otros trabajos relacionados, utilizan el concepto de puntos de corte de la teoría de grafos para identificar a los agentes científicos más relevantes. Los puntos de corte son nodos que aseguran la conectividad y la cohesión global de las redes. La eliminación de estos puntos recortará unas zonas de la red de otras, aumentando el número de componentes y reduciendo su tamaño. De esta manera se puede identificar nodos cuya participación es crucial para asegurar tanto la integración en el foco principal de investigación de unas comunidades con alto rendimiento investigador y la cohesión global de una red de coautoría.

Cienciometría y visualización científica

Un trabajo interesante [23] aunque no relacionado específicamente al dominio del cáncer tuvo como objetivo utilizar la técnica de la cienciaometría para investigar, analizar y visualizar los resultados científicos de Eugene Garfield, científico influyente, promotor de esta técnica, con

numerosos aportes en extracción del conocimiento y creador del Science Citation Index (SCI). Para el análisis de los datos se utilizaron técnicas de análisis de coautoría, análisis de cocitación y también historiografía. Referencian el uso de herramientas de software como HistCite, VOSviewer, Ucinet y NetDraw para dibujar los mapas y analizar los datos.

Por su parte, [6] realizaron un análisis cuantitativo con publicaciones recuperadas de Web of Science (WoS), Scopus, PubMed y otras bases de datos bibliográficas relevantes analizando metadatos bibliográficos de índices de citas (títulos, resúmenes, nombres de revistas, nombres de los autores, afiliación de los autores) para inferir la producción, el impacto, los campos de interés y características generales del descubrimiento basado en la literatura (LBD) en los últimos 35 años. Este trabajo permitió crear un perfil cuantitativo en el área de investigación de LBD identificando autores, países, instituciones y principales palabras clave relacionadas. Adicionalmente, dedujeron los temas de investigación emergentes más notables en el campo de LBD y compararon los trabajos más influyentes basándose en estadísticas de citas.

En el estudio de [7] se explora la estructura del conocimiento para crear trayectorias evolutivas a partir de las publicaciones de congruencia socio-técnica (STC). Para lograr este objetivo, se realizó un análisis cuantitativo que combinó una revisión bibliográfica crítica (CLR) de investigaciones publicadas relacionadas con STC en las bases de datos Web of Science y Scopus desde 2000 hasta 2020. El análisis cuantitativo se lleva a cabo a través de cuatro técnicas cuantitativas: análisis de red de copalabras, análisis de red de coautores, análisis de cocitación, agrupamiento de documentos con análisis de línea de tiempo.

Un análisis bibliométrico [24] sobre el tema de aprendizaje profundo en imágenes de histopatología para el cáncer de mama permitió obtener tendencias de publicación anual, redes

de coautoría entre países, autores y revistas científicas. Se obtuvieron obras de literatura relevantes de la base de datos Scopus entre 2014 y 2021 utilizando las herramientas VOSviewer y Bibliometrix para el análisis a través de varias formas de visualización. Se analizó la red de coocurrencia de las palabras clave de los autores para posibles direcciones futuras del campo. Este estudio proporciona una revisión reciente de este campo de rápido crecimiento para resaltar el estado y las tendencias mediante la visualización científica.

Construcción de redes biológicas

El análisis de redes biológicas ha permitido identificar patrones relacionados con el cáncer. En [1] propusieron un enfoque prometedor para encontrar genes relacionados con la enfermedad del cáncer mediante la construcción de una red de interacción gen gen para todo el genoma humano y un análisis de red para identificar genes relacionados con enfermedades de cáncer. Las diferentes medidas de centralidad de red les permitió decidir sobre la importancia de cada gen e identificar nuevos genes candidatos que podrían vincularse directamente con las enfermedades. Otro trabajo [25], propone un nuevo marco de análisis de red basado en la importancia del nodo (NIBNA) para detectar impulsores de cáncer codificantes y no codificantes. NIBNA detecta los impulsores del cáncer mediante la construcción de una red de condición específica incorporando datos de expresión génica, la estimación de las estructuras comunitarias en la red y la aplicación de una métrica basada en la centralidad para calcular la importancia de los nodos.

En aplicaciones sobre morbilidad múltiple, [26] proponen la construcción de redes donde los nodos representan distintas enfermedades o condiciones, y los ejes identifican grupos de enfermedades por pares observados en una cohorte de población. Este enfoque se emplea en un número creciente de artículos que utilizan análisis de red para examinar la morbilidad múltiple.

Estos estudios utilizaron el análisis de redes para investigar cómo variaba la prevalencia de enfermedades en personas de diferentes sexos, personas de 50 años o más y con diversas condiciones índice. Se demostró que el análisis de redes puede ser una herramienta importante en la investigación de múltiples condiciones de salud y cómo afectan los resultados de los pacientes. En este estudio, se considera la eficacia y la viabilidad del uso de hipergráficos en el análisis de datos a gran escala recopilados de forma rutinaria para investigar grupos de morbilidad múltiple. Usando este formalismo de hipergrafía, se puede ver qué enfermedades y conjuntos de enfermedades coexisten más con otras. Esto proporciona ventajas sobre los métodos actuales que pueden descartar información sobre interacciones entre más de dos enfermedades.

Aplicación de aprendizaje profundo en análisis de imágenes y texto clínico

El aprendizaje profundo se ha aplicado para analizar datos de textos clínicos en cáncer, destacándose en tareas como la identificación de patrones en histopatología y la detección de negación en textos. El trabajo de [4] propone un enfoque para la negación y detección de incertidumbre en textos clínicos escritos en español. Este enfoque explora dos métodos de aprendizaje profundo para realizar la negación y la detección de incertidumbre: la Memoria bidireccional a largo y corto plazo con una capa de campo aleatorio condicional (BiLSTMCRF) y la representación de codificador bidireccional para transformadores (BERT). El enfoque propuesto aprovecha las técnicas de aprendizaje por transferencia para realizar detección de incertidumbre y negación en textos clínicos. En este enfoque, se explotan recursos previamente entrenados para realizar la detección de negación e incertidumbre como una tarea de etiquetado de secuencias evitando así el proceso de ingeniería de características que consume mucho tiempo. Los resultados obtenidos han demostrado la capacidad de los métodos de aprendizaje profundo para realizar tareas de identificación de señales y reconocimiento de alcance. Este

enfoque es útil para aplicaciones de minería de textos médicos porque puede reconocer la negación y la incertidumbre en un solo paso. También se llevó a cabo un proceso de validación utilizando datos clínicos de la vida real que mostró resultados prometedores. Los modelos de aprendizaje profundo entrenados en corpus NUBES han demostrado la capacidad de predecir la negación y la incertidumbre en conjuntos de datos que son diferentes de aquellos en los que fueron entrenados.

Curación de conocimiento y mapeo en identificadores biomédicos

Los identificadores biomédicos como ontologías, taxonomías y vocabularios controlados comúnmente tienen entradas equivalentes que pueden ayudar a la integración de conocimientos, sin embargo puede haber lagunas en la definición de entradas relacionadas de estos identificadores. Para este caso, Biomappings [27] implementa un flujo de trabajo de curación para mapeos faltantes que combina la predicción automatizada con la curación humana en el circuito. Admite múltiples enfoques de predicción y proporciona una interfaz de usuario basada en la web para revisar la exactitud de las asignaciones predichas, combinada con una verificación de consistencia automatizada. Las asignaciones previstas y seleccionadas están disponibles en archivos de recursos públicos controlados por versión en GitHub proporcionando mapeos que antes faltaban entre recursos de identificadores ampliamente utilizados que cubren moléculas pequeñas, líneas celulares, enfermedades y otros conceptos. Se ha demostrado el valor de Biomappings en estudios de casos que involucran la predicción y curación de mapeos faltantes entre líneas de células cancerosas, así como pequeñas moléculas probadas en ensayos clínicos.

El trabajo [28] tuvo como objetivo agregar y ampliar el conjunto de bases de conocimiento (KB) clínicamente relevantes para la oncología de precisión que también incluyan KB de literatura médica y registros de ensayos clínicos registrados. Además de las colecciones

integrales existentes de KB, se propuso una categorización de KB a lo largo del proceso de diagnóstico. Además, se intentó guiar a los implementadores de herramientas de interpretación de variantes de cáncer mediante una evaluación de las opciones de acceso programático a las KB, así como una clasificación de relevancia basada en los comentarios de los profesionales clínicos.

Existe un amplio conjunto de datos actualmente inaccesibles pero normalizables en publicaciones de ensayos clínicos basados en biomarcadores. Un enfoque propuesto [29] demuestra que los ensayos clínicos de biomarcadores tienen suficientes datos normalizables. Se analizó 13 publicaciones de ensayos clínicos de mCRPC que informan los resultados de ensayos que miden biomarcadores de recombinación homóloga y de deficiencia de reparación del ADN permitiendo informar la selección de biomarcadores en el diseño de ensayos clínicos, comparar y clasificar las métricas de supervivencia, investigar los efectos de las relaciones biomarcador resultado, evaluar eventos adversos y automatizar el descubrimiento de mecanismos terapéuticos. Los resultados podrían evitar a los oncólogos, las agencias reguladoras e investigadores invertir esfuerzos en buscar, leer y recuperar datos de publicaciones para el descubrimiento de nuevas terapias. Aunque se demostró su utilidad para el cáncer de próstata metastásico, este enfoque es independiente del dominio y puede extenderse fácilmente a otros dominios clínicos y avanzar en la medicina de precisión.

4. Problema de investigación

Dado que la cantidad de literatura biomédica publicada, en particular la relacionada con el cáncer, aumenta a un ritmo sin precedentes, con tratamientos novedosos y actualizados que pueden considerar nuevas variantes de las enfermedades, es un desafío para un médico

mantenerse actualizado. Agregar y resumir el estado actual del conocimiento en el dominio de una enfermedad de cáncer puede ayudar a informar al investigador médico sobre los procesos de la enfermedad y la efectividad de las estrategias de tratamiento [17]. Una revisión sistemática de la literatura y un metanálisis, que son métodos comunes utilizados para analizar una temática, se basan en preguntas de investigación específicas y se analizan para un número limitado de artículos relevantes. Aunque es posible investigar el resultado de un tema de interés de esta manera, existe la limitación de que la estructura de conocimiento general puede variar respecto al contexto en el que se estudia el tema y los conceptos principales de la investigación puedan no ser revelados [19].

El entorno médico reviste diversas complicaciones de tiempo y recursos. Cuando un paciente ingresa puede requerir un diagnóstico y tratamientos urgentes para casos muy particulares por lo cual según el criterio y experiencia del personal a cargo pueden intentar documentarse con recursos de consulta limitados y finalmente pueden decidirse a emplear tratamientos tradicionales. Aunque estos tratamientos pueden contribuir a la mejora del paciente, podrían no estar a la vanguardia de las investigaciones en el campo. Esto brinda una oportunidad para reducir costos al sistema de salud porque al conocer información científica actualizada se puede brindar mejores diagnósticos y tratamientos más efectivos que posiblemente reduzcan la permanencia y recurrencia en la atención de los pacientes. Adicionalmente existirían beneficios del lado del paciente de cáncer porque se aumenta la posibilidad de brindar solución a su caso, mejorar su condición de salud y su calidad de vida.

Esto indica la necesidad que tiene el personal de salud de aumentar la calidad de la información científica disponible así como disminuir el tiempo de búsqueda de literatura biomédica. En la actualidad hay muchos esfuerzos en ese sentido tratando de usar los desarrollos

tecnológicos actuales para apoyar las decisiones médicas. Sin embargo, las implementaciones tecnológicas para esta finalidad podrían requerir grandes recursos de cómputo así como sistemas informáticos distribuidos, tecnologías de big data, inteligencia artificial, aprendizaje de máquina entre otras debido al enorme volumen de información que se requiere procesar. En este sentido, como aporte del presente trabajo en la generación de nuevo conocimiento está el lograr relacionar la terminología médica con información relevante para el diagnóstico y tratamiento del cáncer. Específicamente desarrollar un aplicativo que sea una herramienta práctica, de uso cotidiano, que apoye al investigador médico y pueda considerar grandes volúmenes de información científica extrayendo los trabajos más relevantes para la mejora del criterio médico y toma de decisiones apoyadas en la tecnología.

5. Objetivos de investigación

Objetivo general

Desarrollar un prototipo para encontrar patrones en literatura científica del dominio del cáncer usando técnicas de redes complejas.

Objetivos específicos

Desarrollar un mecanismo para la búsqueda automática de literatura científica a partir de palabras claves asociadas al cáncer.

Extraer automáticamente metadatos a partir del análisis de literatura científica en el dominio del cáncer.

Desarrollar un módulo para análisis de redes complejas usando artículos científicos en el dominio del cáncer.

Implementar un sistema prototipo basado en microservicios para el apoyo a la toma de decisiones médicas.

6. Metodología

En este trabajo, se combinaron actividades de investigación y desarrollo de software que fueron implementadas siguiendo la metodología DSRM [30] para la obtención de cuatro prototipos que contribuyeron a la solución del problema abordado.

En la tabla 6.1, se especifican los prototipos a implementar y las actividades a realizar en cada fase de la metodología. Para cada uno de los prototipos la identificación del problema y la definición de objetivos están relacionados con los formulados en las secciones anteriores. La fase de diseño y desarrollo requirió un primer componente investigativo en el cual se consultaron y abordaron diferentes tecnologías para luego realizar el desarrollo del prototipo, en todos los casos se requirió varias iteraciones que se pueden ver en la historia del repositorio de código [31]. Las fases de demostración y evaluación para cada prototipo están ampliadas en la sección de resultados.

Las diversas actividades de las fases de diseño, desarrollo, demostración y evaluación de la metodología propuesta en la tabla 6.1 se describen a continuación:

Búsqueda de bases de datos y agregadores de literatura científica

Mediante esta actividad se buscará bases de datos o agregadores de literatura científica que puedan ser accedidos a través de API. Adicionalmente se buscará librerías de software relacionadas que permitan la consulta en estos agregadores usando palabras clave relacionadas al cáncer.

Etapa	Prototipo			
	Descarga artículos	Extracción de información	Redes complejas	Aplicativo web
Identificación del problema	Obtener literatura científica	Agregadores no tienen metadatos requeridos	Obtener patrones en corpus de literatura científica	Herramienta para relacionar terminología médica con información científica relevante
Objetivo	Desarrollar mecanismo para la descarga de literatura científica	Extraer metadatos a partir del texto completo	Desarrollar módulo para analizar redes de literatura científica	Desarrollar aplicativo web que integre las tecnologías desarrolladas
Diseño y desarrollo	Búsqueda de agregadores Construcción del prototipo	Búsqueda de tecnologías de integración y despliegue continuos Construcción prototipo	Revisión literatura sobre análisis de información científica Construcción prototipo	Estudio UX/UI Definición arquitectura y requerimientos Desarrollo e implementación aplicativo web
Demostración y evaluación	Uso del prototipo para descargar literatura	despliegue de ambientes locales y en la nube, ejecución del prototipo para extraer información	Uso del prototipo para obtener redes de literatura y métricas de análisis de redes	Búsqueda de patrones de literatura científica de cáncer de mama
Comunicación	Repositorio de código Trabajo de investigación	Repositorio de código Trabajo de investigación	Repositorio de código Trabajo de investigación	Repositorio de código Trabajo de investigación

Tabla 6.1. Propuesta metodológica del tipo DSRM para la definición de actividades a realizar para la implementación de prototipos tecnológicos del proyecto.

Sistema prototipo para la descarga de artículos científicos

Se realizará la implementación de un prototipo de software para la descarga de artículos científicos y metadatos usando conexión vía API con los agregadores seleccionados. Este sistema a partir de palabras clave será capaz de recuperar metadatos y la URL del acceso directo al artículo científico en formato PDF.

Tecnologías para implementación de integración y despliegue continuo

Se realizará la búsqueda y pruebas de tecnologías que permitan la implementación de los microservicios a través de herramientas que permitan integración y despliegue continuos.

Sistema prototipo basado en servicios web I-ADC-UNIVALLE

Considerando las tecnologías abordadas para microservicios, integración y despliegue continuos, se implementará un prototipo inicial que permita ir soportando las diferentes funcionalidades requeridas por el sistema y a su vez pueda ser desplegado de manera continua a partir de la integración con un repositorio de código. Mediante los microservicios se espera separar las diferentes funcionalidades como consulta y descarga de artículos científicos, procesamiento de artículos científicos para obtener metadatos, construcción de redes de autores instituciones países y palabras clave, análisis de redes y la visualización e interfaz de usuario.

Extracción de información de artículos científicos

Se creará un microservicio que permita la extracción de información a partir del texto completo de un artículo científico. Se usarán expresiones regulares para detectar correos electrónicos y modelos entrenados de Inteligencia artificial para detectar el idioma, ubicaciones y palabras clave permitiendo la extracción de metadatos de interés en la investigación.

Revisión de literatura sobre análisis de información científica en el dominio del cáncer

Se realizará una revisión sistemática de literatura para encontrar trabajos relacionados con el análisis de literatura científica en el dominio del cáncer. Los temas relevantes serán enriquecimiento semántico, ciencia de redes y redes de coautoría. Se propone buscar en Scopus y Web of Science.

Obtención de matrices de adyacencia

Se creará un nuevo microservicio que acceda a la base de datos que contiene los metadatos extraídos de los artículos científicos y realice el proceso de determinar las relaciones entre autores, instituciones, países y palabras clave. El microservicio utilizará la librería NetworkX y devolverá una matriz de adyacencia con la información de la red creada.

Sistema prototipo para el análisis de redes complejas

A partir de las matrices de adyacencia obtenidas se probará y utilizará librerías que permitan obtener métricas de redes así como la obtención de gráficos y visualización. A través de la implementación basada en microservicios, se podrá acceder a las diversas características implementadas. De esta manera se espera configurar reportes que permitan identificar información de interés como autores o trabajos principales, instituciones reconocidas, trabajo en los países y palabras clave relevantes.

Estudio de UX/UI basado en encuestas a personal médico

Se realizará un estudio de UX/UI en forma de encuestas dirigidas a personal médico. De esta manera se espera identificar las necesidades específicas que requiere el personal médico para investigaciones relacionadas con el cáncer. Con esta información se procederá a definir la arquitectura y contenido de la interfaz de usuario apropiada para la investigación y consulta de la información brindada por el sistema.

Aplicativo web prototipo I-ADC-UNIVALLE

Finalmente se creará otro microservicio enfocado en proveer una interfaz de usuario tipo aplicativo web. Esta interfaz dispondrá de manejo de cuentas de usuario, configuración de proyectos, interfaz para que los investigadores introduzcan los estudios e interfaz para que los usuarios médicos puedan acceder a la información disponible.

Patrones en literatura científica de cáncer de mama usando técnicas de redes complejas

Usando la tecnología desarrollada se realizará una investigación relacionada con el cáncer de mama y se presentará la información obtenida. Los patrones a identificar en la literatura serán autores, instituciones, países y palabras clave relevantes.

7. Resultados

Búsqueda de bases de datos y agregadores de literatura científica

La primera actividad consistió en identificar bases de datos de literatura científica accesibles a través de APIs que permitan la recuperación eficiente de artículos sobre cáncer. Hay diversas soluciones disponibles en línea, [11] incluye una lista de recursos académicos de uso común, que ponen a disposición sus API, proporcionan acceso programático y herramientas computacionales para utilizar sus recursos o proporcionan corpus disponibles para la minería de texto y datos. Aunque se exploraron diversas fuentes, la elección final dependió de que las APIs fueran de libre acceso y tuvieran una implementación tipo librería que permitiera su uso de manera estructurada. Por consiguiente, se consideró la búsqueda de integraciones disponibles en repositorios públicos y la facilidad para recuperar no solo artículos completos, sino también metadatos relevantes como autores, instituciones y palabras clave. Para cada agregador

seleccionado se desarrolló un script que interactúa con las respectivas APIs, utilizando términos específicos para obtener artículos relacionados con la temática del cáncer.

APIs para consulta de literatura científica

La disponibilidad de las API permite que los investigadores puedan acceder de manera programática a la literatura disponible en agregadores permitiendo ventajas en tiempo y recursos para sus investigaciones. Al explorar las diversas API, se observa que cada una ofrece información o metadatos con formatos específicos. En la tabla 7.1 se presentan algunas tecnologías abordadas en esta fase y se observa que cumplen un propósito general de organizar información sobre publicaciones científicas y construir metadatos de interés, sumado a ello se observa una implementación de diversas métricas que ayudan a complementar la información suministrada.

Cada una de las implementaciones en la tabla 7.1 fueron probadas de manera simple accediendo al endpoint suministrado como se observa en la figura 7.1 donde a través del navegador web fue posible acceder a la API de arXiv. En algunos casos fue necesario disponer de la API Key por medio de registro en el proveedor. Es de notar que el acceso a Scopus y Web of Science fue posible por la suscripción institucional de la Universidad del Valle [39]. Las APIs de PubMed, arXiv y Core destacaron por su accesibilidad gratuita, de ellas Pubmed fue la de mayor interés en nuestra investigación dado que es muy fuerte en el campo de la medicina, arXiv incluso no requería de una API Key y Pubmed ofrece acceso de menor velocidad cuando la consulta se hace sin API Key.

Proveedor	Propósito	Datos proporcionados	Acceso
Scopus [32]	Búsqueda y análisis bibliográfico, incluyendo análisis de redes de citación, productividad científica, colaboraciones	Artículos, citas, afiliaciones, métricas de impacto	Suscripción institucional
CrossRef [33]	Referencias de citación y metadatos para artículos publicados, búsqueda de publicaciones, actualización de datos bibliográficos	DOIs, metadatos de publicaciones, citas cruzadas entre editores	Registro gratuito o pago según uso
Web of Science [34]	Análisis de citas, productividad académica, impacto de la investigación y métricas bibliométricas	Artículos, métricas de citación, datos de patentes	Registro y plan de suscripción
PubMed [35]	Consulta de bases de datos en biología, medicina, química, bioinformática, comparación de secuencias y análisis estructural	Artículos científicos, secuencias biológicas, información química	Acceso HTTP gratuito
CORE [36]	Acceso a artículos de investigación en acceso abierto, integración en herramientas de búsqueda, análisis y clasificación de literatura	Metadatos y texto completo de artículos científicos	API Key (opciones gratuitas y comerciales)
SciELO OAI-PMH [37]	Recolección de metadatos académicos en repositorios, interoperabilidad entre sistemas, aumento de visibilidad de publicaciones académicas	Metadatos de títulos, resúmenes, afiliaciones	URL de protocolo OAI-PMH
arXiv [38]	Acceso a artículos en ciencias y tecnología	Títulos, autores, resúmenes, enlaces a texto completo	Acceso libre y sin autenticación

Tabla 7.1. Proveedores de API para consulta de literatura científica abordados en la investigación.

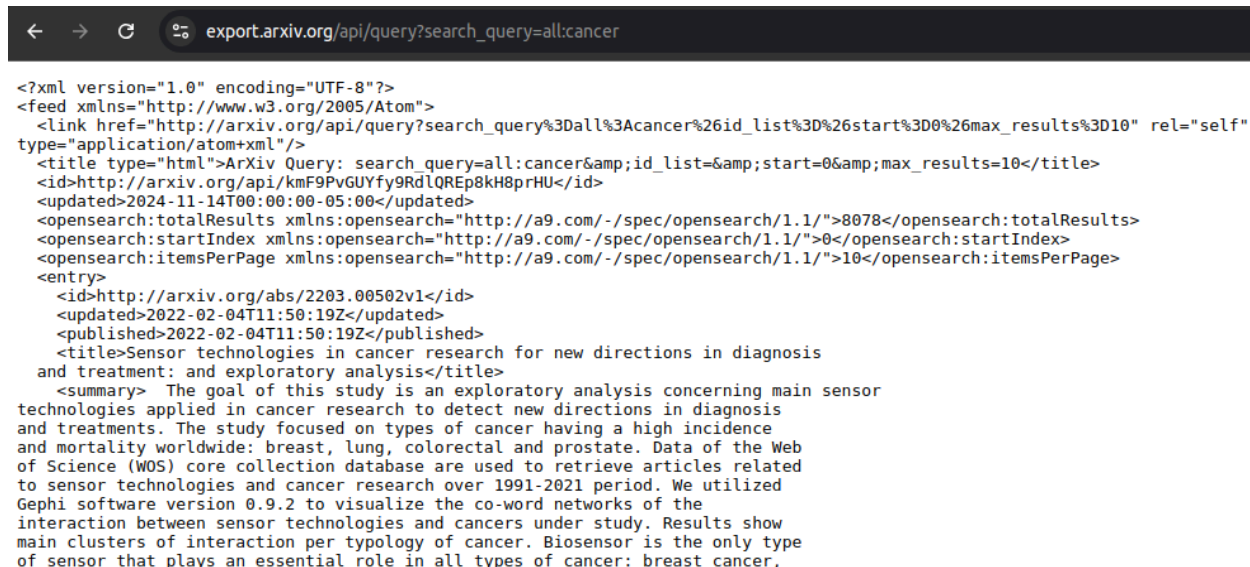


Figura 7.1. Acceso al endpoint del API de arXiv.

También era de nuestro interés disponer de la información de SciELO dado que tiene publicaciones en español. Sin embargo el acceso a los trabajos por medio del protocolo OAI-PMH implicaba listar identificadores de las diferentes instituciones y hacer consultas a varios niveles que no ofrecían la practicidad de las demás implementaciones. En este caso encontramos que la implementación de CORE ya hace este proceso y como muestra [40] han logrado indexar diversas instancias de SciELO así como gran cantidad publicaciones en español provenientes de revistas e instituciones universitarias incluso de latinoamérica. De acuerdo con lo anterior, se decidió que PubMed, ArXiv y Core fueran nuestra fuente de consulta de literatura científica.

Librerías para interactuar con las APIs

Para mayor compatibilidad con los diferentes recursos de código abierto disponibles en línea, se ha seleccionado Python [41] como lenguaje de programación. Python ha emergido como un lenguaje de programación muy atractivo en el campo de la bioinformática y la

quimioinformática debido a su rápida curva de aprendizaje y su fácil integración en diversos flujos de análisis de datos [42]. Por esta razón se realizó una búsqueda de implementaciones y librerías relacionadas a las API en GitHub [43] y en PyPI [44]. De esta manera se logró encontrar herramientas probadas para el acceso a las API de interés. Luego de diversas pruebas y teniendo en cuenta restricciones de acceso, facilidad de uso y disponibilidad de librerías compatibles con Python, se definieron las tecnologías a usar de acuerdo a la tabla 7.2.

Agregador	Url	Librería / API
Arxiv	https://arxiv.org/	https://pypi.org/project/arxiv/
Core	https://core.ac.uk/	https://api.core.ac.uk/docs/v3
Pubmed	https://pubmed.ncbi.nlm.nih.gov/	https://pypi.org/project/metapub/

Tabla 7.2. Relación de agregadores de literatura científica y librerías de acceso seleccionados para usar en el proyecto.

PyPI - arXiv

Es una librería de python [45] para acceso a la API de arXiv. A continuación se brinda un script en Python para una consulta sencilla de artículos usando la librería de arXiv:

```
import arxiv
client = arxiv.Client()
    search = arxiv.Search(
        query = "cancer",
        max_results = 10,
        sort_by = arxiv.SortCriterion.SubmittedDate
    )
results = client.results(search)
for r in client.results(search):
    print(r.title)
```

Implementación Core

Basados en un webinar publicado por Core en GitHub [46], una forma recomendada para acceder a la API de Core por medio de un script de Python sería la siguiente:

```
import json
import requests

api_key = "INCLUDE_YOUR_API_KEY"

def query_api(url_fragment, query, limit=100):
    headers={"Authorization":"Bearer "+api_key}
    query = {"q":query, "limit":limit}
    response = requests.post(f"{api_endpoint}{url_fragment}", data = json.dumps(query),
headers=headers)
    if response.status_code ==200:
        return response.json(), response.elapsed.total_seconds()
    else:
        print(f"Error code {response.status_code}, {response.content}")

query = "cancer"
results, elapsed = query_api("search/works", query, limit=1)
pretty_json(results)
```

PyPI - Metapub

Metapub [47] es una biblioteca de Python que proporciona objetos de Python obtenidos a través de eutils que representan artículos y conceptos de PubMed encontrados en las bases de datos del NCBI. Actualmente, Metapub proporciona capas de abstracción sobre Medgen, Pubmed, ClinVar y CrossRef, y pretende abarcar tantos tipos de búsquedas y resúmenes de bases de datos como se puedan proporcionar a través de Eutils/Entrez. Metapub también puede ayudar a crear bibliotecas de artículos académicos mediante el uso de la clase FindIt.

El siguiente es un script de Python probado para la consulta del id de artículos de cáncer y la búsqueda de url del pdf relacionado.

```
from metapub import PubMedFetcher
fetch = PubMedFetcher()
pmids = fetch.pmids_for_query( "cancer[Title/Abstract]", retmax=1000 )
print(pmids)
```

El siguiente script de Python muestra una url de descarga de artículo en pdf a partir del DOI usando la librería Metapub:

```
from metapub import FindIt
src = FindIt('18381613')
print(src.url)
```

Ecuaciones de búsqueda

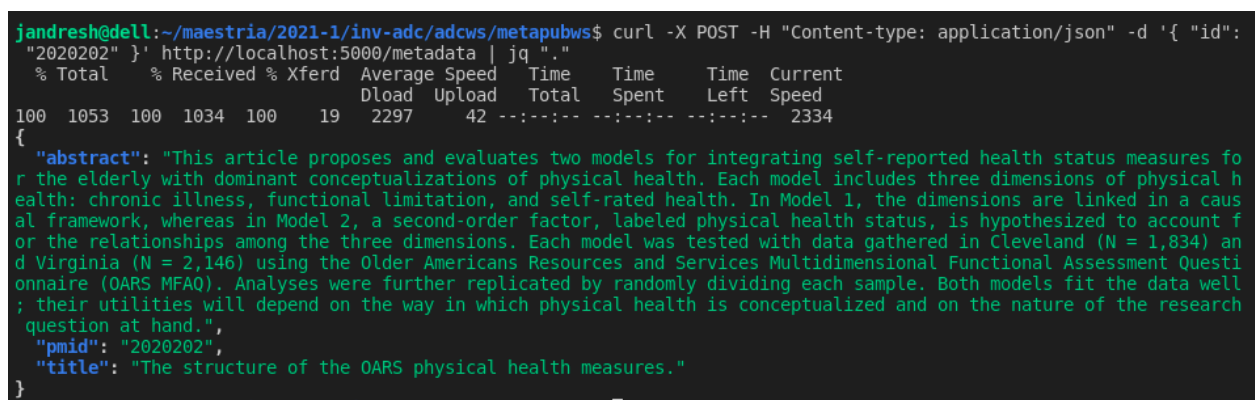
Cada agregador tiene sus propios formatos para definir la cadena de búsqueda. Se identificó para el presente trabajo la recuperación de documentos que tengan todas las palabras de interés en el abstract. La tabla 7.3 muestra el envío de peticiones a las API para consulta de documentos dependiendo del agregador de literatura científica consultado.

Agregador	Ecuación de búsqueda
ArXiv	abs:breast AND abs:carcinoma
Core	abstract(breast AND carcinoma)
Pubmed	breast[Title/Abstract] AND carcinoma[Title/Abstract]

Tabla 7.3. Ecuaciones de búsqueda en el formato de los agregadores seleccionados en la investigación.

Sistema prototipo para la descarga de artículos científicos

A partir de las APIs y librerías relacionadas identificadas, se desarrolló un sistema prototipo que automatiza la obtención del enlace al texto completo del artículo científico y sus metadatos buscando por palabras clave del tema de interés. Este sistema fue programado en Python utilizando bibliotecas como requests para hacer solicitudes HTTP y json para manejar las respuestas de las APIs. Se decidió usar un servidor Flask [48], y la implementación para arXiv [49] que utiliza un contenedor Docker [15] para arranque del microservicio. Se logró pasar las funcionalidades de las tecnologías seleccionadas a una implementación basada en servicios web para los agregadores de literatura científica seleccionados Pubmed, arXiv y Core. La figura 7.2 muestra el acceso al microservicio desde una terminal linux usando el comando curl.



```
jandresh@deli:~/maestria/2021-1/inv-adc/adaws/metapubws$ curl -X POST -H "Content-type: application/json" -d '{"id":
"2020202"}' http://localhost:5000/metadata | jq "."
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
   Dload  Upload   Total   Dload  Upload   Total   Spent    Left    Speed
100 1053    100 1034    100    19   2297    42  --:--:--  --:--:--  --:--:--  2334
{
  "abstract": "This article proposes and evaluates two models for integrating self-reported health status measures fo
r the elderly with dominant conceptualizations of physical health. Each model includes three dimensions of physical h
ealth: chronic illness, functional limitation, and self-rated health. In Model 1, the dimensions are linked in a caus
al framework, whereas in Model 2, a second-order factor, labeled physical health status, is hypothesized to account f
or the relationships among the three dimensions. Each model was tested with data gathered in Cleveland (N = 1,834) an
d Virginia (N = 2,146) using the Older Americans Resources and Services Multidimensional Functional Assessment Questi
onnaire (OARS MFAQ). Analyses were further replicated by randomly dividing each sample. Both models fit the data well
; their utilities will depend on the way in which physical health is conceptualized and on the nature of the research
question at hand.",
  "pmid": "2020202",
  "title": "The structure of the OARS physical health measures."
}
```

Figura 7.2. Consulta de servicio web para acceso a PubMed usando el comando curl.

En el caso presentado el microservicio solo devuelve el abstract, el título del artículo y el Id. El prototipo puede ser accedido a través de github [50] donde se documenta el arranque y uso de los servicios configurados.

Tecnologías para implementación de integración y despliegue continuo

Para este proceso se consideró probar proveedores de servicios en la nube escogiendo la capa gratuita para propósitos investigativos. Se realizaron pruebas en AWS [51] y Google Cloud

[16]. Ambos proveedores ofrecen capa gratuita, permiten arrancar cargas de trabajo y probar los diversos componentes de su servicio. Se seleccionó Google Cloud dado que tiene un sistema de créditos en dólares con una duración de 3 meses habilitando el acceso a la mayoría de los recursos de su servicio. En tal prueba fue posible el despliegue de máquinas virtuales y una posterior prueba de creación de un cluster Kubernetes [52] usado en el despliegue y manejo de ambientes productivos.

Como repositorio de código se escogió GitHub [43] debido a la disponibilidad de implementaciones de software de libre acceso, y en particular, porque alberga aquellas librerías relacionadas a los agregadores Pubmed, Arxiv y Core. Posteriormente, en cuanto a la integración y el despliegue continuo, se realizaron pruebas con Jenkins [53] y GitLab CI/CD [54] para soportar la implementación del prototipo en diferentes entornos. Aunque GitLab ofrece una prueba gratuita por 60 días, y tiene la facilidad de permitir la integración con el repositorio por medio de su interfaz web y unos archivos de configuración, se escogió Jenkins dado que fue posible instalar el servidor en el cluster Kubernetes suministrado por Google Cloud. Este servidor no tenía restricciones de prueba y la configuración para trabajar en conjunto con Google Cloud estaba muy bien documentada [55], [56]. De esta manera se encontró que, usando la herramienta Jenkins configurada en Google Cloud, se puede automatizar el despliegue de diversos entornos bien sea para pruebas o para pasar las implementaciones a un entorno productivo.

Sistema prototipo basado en servicios web I-ADC-UNIVALLE

Después de haber seleccionado las diversas tecnologías de se procedió a configurar y documentar el proceso de implementación del sistema prototipo que soportará los microservicios y características propuestas. Se ha creado un documento [57] que detalla el proceso para

implementar el sistema bajo la metodología de integración y despliegue continuos. El repositorio usado para alojar la base de código y enlazar con el servidor de integración continua se puede consultar en [31]. Este sistema tiene como características el despliegue de máquinas virtuales, el despliegue de un cluster Kubernetes y la instalación del servidor Jenkins de integración continua en el clúster. De manera alternativa, en el repositorio de código se tienen las instrucciones para hacer el arranque del sistema en un ambiente local usando contenedores.

Implementación integración continua

Una vez se ha completado el proceso y se ha logrado el arranque del sistema, es el momento de continuar el desarrollo de nuevas características. Para ello, como muestra la figura 7.3 se empieza definiendo una característica a implementar. Esto se puede realizar apoyado en sistemas gestores de proyectos como Jira [58]. Un desarrollador es asignado a implementar la característica y usa un IDE de desarrollo como Visual Studio Code [59] para hacer la escritura de código. Mediante una implementación local basada en contenedores, el desarrollador puede iterar en su nueva característica probando que la implementación sea apropiada. En este momento, le corresponde al desarrollador hacer la publicación de sus cambios en el repositorio [31], para ello se utiliza Git [60] que permite el acceso, modificación y control de versiones de repositorios.

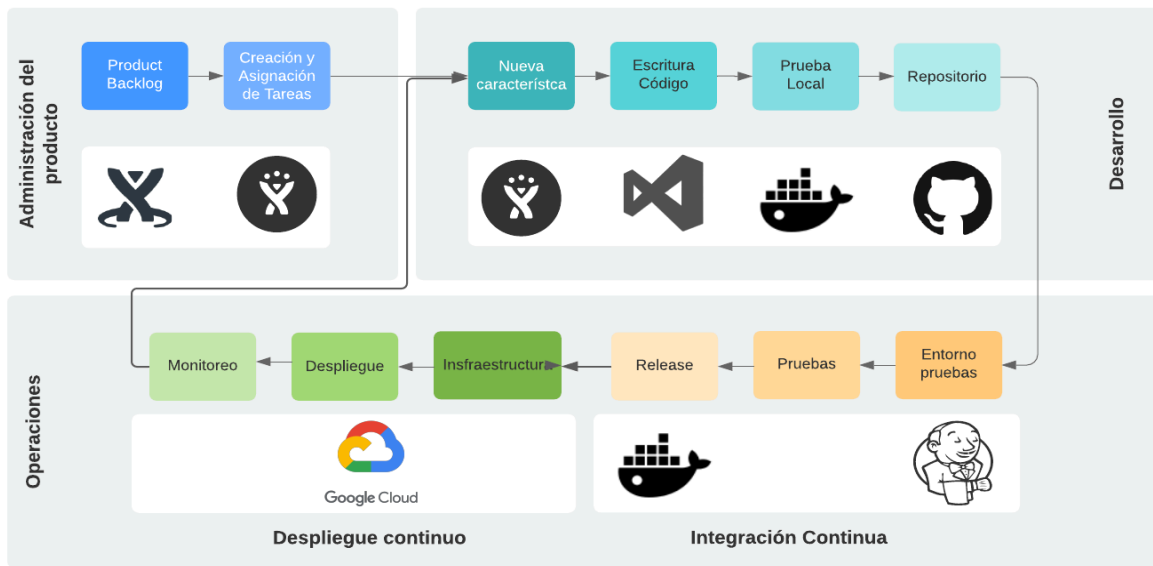


Figura 7.3. Pipeline de integración y despliegue continuo del sistema prototipo I-ADC-UNIVALLE.

Cuando se actualizan los nuevos cambios en el repositorio del proyecto, el servidor de integración continua Jenkins [53] realiza un seguimiento de estos cambios y ejecuta un pipeline de integración y despliegue continuo configurado en el archivo “Jenkinsfile” que se encuentra en la raíz del repositorio. En primera medida, el servidor de integración valida el código nuevo con pruebas de integración que consisten en compilar y arrancar los contenedores Docker de los microservicios usando el comando “docker compose” en una máquina virtual en la nube de Google Cloud. A continuación, para cada contenedor se crea una imagen y se sube al repositorio de contenedores Docker Hub [61] donde permanece disponible públicamente. Este proceso es la base para la creación de los diversos servicios y contenedores mostrados en la tabla 7.4 y los insumos requeridos para el posterior proceso de despliegue continuo. Finalmente, el servidor de integración continua despliega los contenedores llamado la imagen desde Docker Hub en un entorno de prueba de aceptación previo al entorno de producción y se verifica su correcto

arranque ejecutando unos test funcionales que se encargan de llamar los endpoints en cada contenedor para verificar su correcto funcionamiento. El script Bash empleado para pruebas funcionales está disponible en el archivo “functional.sh” en la raíz del repositorio.

Nombre contenedor	Función
metapub	Soporta la conexión con PubMed y descarga de metadatos
arxiv	Soporta la conexión con el agregador arXiv y descarga de metadatos
core	Soporta la conexión con el agregador Core y descarga de metadatos
preprocessing	Implementa técnicas de IA, expresiones regulares y herramientas que permiten la descarga de los PDF, extracción de texto, e identificación de patrones correo electrónico, idioma principal del documento, ubicación
db	se encarga de implementar la base de datos que usará el sistema para guardar metadatos de los documentos e información del sistema
orchestrator	Corresponde a un contenedor que permite obtener información de los diversos microservicios para integrar las acciones de ellos por ejemplo para ejecutar un pipeline de captura de metadatos o de análisis de redes
gui	Este contenedor soporta la interfaz de usuario que facilita acceso a la información que puede ser procesada por el sistema

Tabla 7.4. Contenedores Docker implementados en el sistema para el soporte de los microservicios.

Implementación despliegue continuo

Se investigó e implementó un cluster de kubernetes con el fin de desplegar ambientes de prueba y de producción. Fruto de ello se identifica y construye una estrategia de despliegue canary como la mostrada en la figura 7.4.

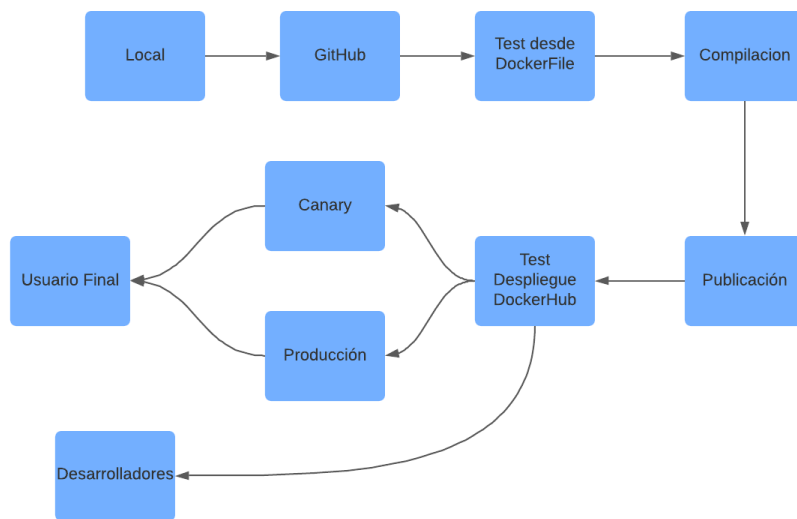


Figura 7.4. Estrategia canary para el despliegue del entorno de producción.

Cuando ha pasado satisfactoriamente el proceso de integración continua el servidor Jenkins ejecuta la estrategia canary para liberar paulatinamente las nuevas características del aplicativo al usuario final. Esta estrategia consiste entonces en que la versión más actualizada del aplicativo va a estar en menor cantidad de pods que la versión que está actualmente en producción. Utilizar kubernetes permite generar un entorno de producción confiable y escalable, al detectar alguna falla en algún pod relacionado con una versión nueva del contenedor, es capaz de deshacer el cambio y regresar a la versión anterior. Cuando se completa este proceso, se tiene disponibilidad de un aplicativo web desplegado en la nube de Google Cloud y con acceso global.

Arquitectura

De esta manera, la arquitectura del sistema sería la mostrada en la figura 7.5. Se parte de agregar cadenas de búsqueda del tema de interés. Las cadenas de búsqueda pueden ser agregadas directamente en el contenedor db o a través del contenedor gui por medio de la interfaz de usuario web.

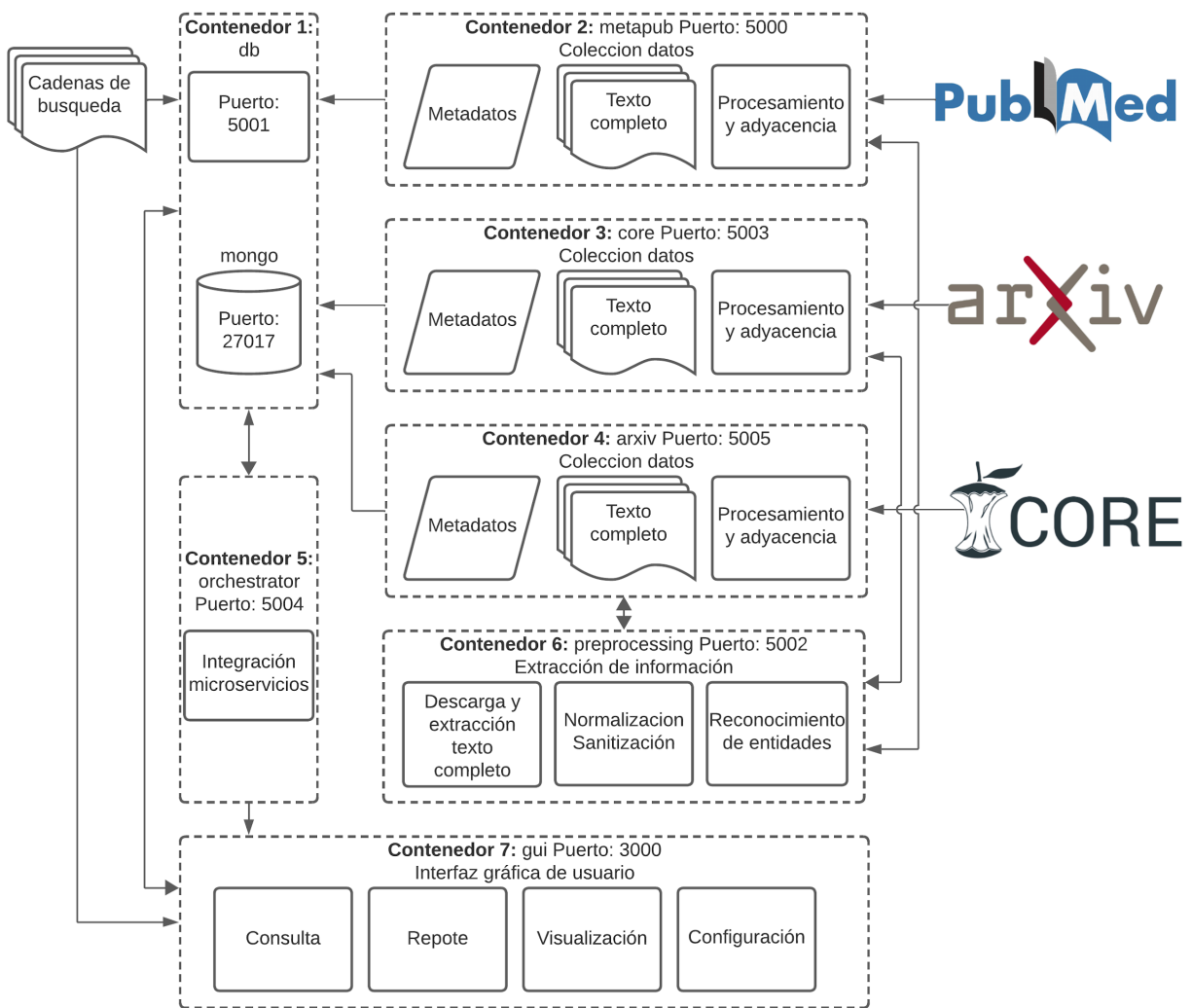


Figura 7.5. Arquitectura de software del sistema prototipo I-ADC-UNIVALLE.

Cuando se ejecuta un pipeline de metadatos configurado en el contenedor orchestrator, se consulta en el contenedor db las cadenas de búsqueda y por cada cadena de búsqueda se realiza la consulta de los diferentes agregadores a través de los contenedores metapub, arxiv y core. Estos contenedores a su vez utilizan el contenedor preprocessing para obtener metadatos faltantes. Cuando se ha recuperado la información de metadatos se calcula de manera adicional las relaciones de cada documento comparando con la información que ha sido previamente agregada en la base de datos para tener información disponible para análisis de redes. A

continuación se envía información obtenida a la base de datos. Un segundo pipeline de adyacencia configurado en el contenedor orchestrator sirve para llamar información de la base de datos, construir la matriz de adyacencia y ejecutar el análisis de redes entregando gráficos y datos de interés dependiendo de la configuración y parámetros empleados. Por medio de la interfaz web, se puede hacer uso de los servicios en el contenedor orchestrator para configurar los diversos pipelines y para la administración de usuarios y proyectos a manejar en el sistema. Cuando se ejecuta el pipeline de adyacencia desde la interfaz de usuario, la respuesta obtenida se usa para mostrar al usuario las gráficas y datos obtenidos en el análisis de redes.

Extracción de información de artículos científicos

Cada contenedor metapub, arxiv, core y sus microservicios nos permiten acceder a la información de las bases de datos y agregadores. Allí encontramos diferentes implementaciones de datos cuando son consultados en el agregador. Los propios contenedores se encargan de modificar estos metadatos para hacerlos compatibles con la información del sistema prototipo, sin embargo alguna información como correos electrónicos, ubicaciones, palabras clave e idioma no están disponibles. Se decidió entonces usar el documento PDF para capturar el fulltexto de los artículos consultados y de allí intentar la obtención de la información. En el contenedor preprocessing usando librerías y entrenados de IA como regex, gensim, geograpy3, langdetect, nltk, numpy, pdftotext se logra este este propósito por lo cual se puede obtener una serie de metadatos de interés para el posterior análisis de redes.

Metadatos a extraer

En la tabla 7.5 se listan los metadatos a emplear en el sistema, el tipo de datos, la descripción y la fuente de donde se extrajo el metadato. Toda esta información es obtenida luego

de la consulta y el procesamiento de la información cuando se han usado los contenedores metapub, arxiv, core y preprocessing.

Nombre	Tipo dato	Función	Fuente
dbid	str	Identificador del documento en el agregador	Agregador
doi	str	Identificador digital del documento	Agregador
title	str	Título del artículo o documento consultado	Agregador
abstract	str	Resumen del documento consultado	Agregador
authors	list[str]	Lista de autores	Agregador
emails	list[str]	Lista de emails detectados en el texto completo	preprocessing
keywords	list[str]	Lista de palabras clave detectadas en el documento	preprocessing
organizations	list[str]	Lista de organizaciones obtenida del dominio de correo	preprocessing
countries	list[str]	Países detectados en el documento	preprocessing
url	str	Dirección de acceso al texto completo en PDF	Agregador
year	int	Año de la publicación	Agregador
language	str	Idioma principal del documento	preprocessing

Tabla 7.5. Listado de metadatos obtenidos y agregados al sistema.

Implementación de la base de datos

Aunque se inició usando una base de datos relacional en las primeras iteraciones del proyecto, luego se escogió la base de datos mongoDB [62]. Fue interesante que los problemas de compatibilidad de datos se solucionaron, en particular los provenientes de la codificación de los textos recuperados. Adicionalmente, mostró mayor facilidad de consulta de información. MongoDB es una base de datos no relacional que permite crear información a través de una estructura que comprende la tripleta base de datos, colección y documentos. La figura 7.6

presenta un esquema de la base de datos implementada. Se observa que para los nombres de las colecciones se utiliza el carácter numeral para a través del nombre de la colección indexar la información.

Database: global Collection: users _id first_name last_name org org_id email password created updated is_admin is_active is_verified	Database: org_id Collection: projects _id name description maxDocs status Database: org_id Collection: patterns#project_name _id pattern	Database: org_id Collection: metadata#project_name#pattern_id _id title abstract authors countries dbid doi emails keywords language organizations pat_id source url year	Database: org_id Collection: author_info#project_name#pattern_id _id author authors countries doc_id doi keywords language organizations year	Database: org_id Collection: organization_info#project_name#pattern_id _id organization authors countries doc_id doi keywords language organizations year	Database: org_id Collection: country_info#project_name#pattern_id _id country authors countries doc_id doi keywords language organizations year
Database: org_id Collection: keyword_info#project_name#pattern_id _id keyword authors countries doc_id doi keywords language organizations year	Database: org_id Collection: language_info#project_name#pattern_id _id language authors countries doc_id doi keywords organizations year	Database: org_id Collection: year_info#project_name#pattern_id _id year authors countries doc_id doi keywords organizations	Database: org_id Collection: authors#project_name#pattern_id _id author related Database: org_id Collection: organizations#project_name#pattern_id _id organization related	Database: org_id Collection: countries#project_name#pattern_id _id country related Database: org_id Collection: keywords#project_name#pattern_id _id keyword related	Database: org_id Collection: wordcloud#project_name#pattern_id _id keyword

Figura 7.6. Esquema de la base de datos implementada en el proyecto.

Como esta implementación se basa en proyectos y se desea que sirva para múltiples usuarios y organizaciones y con la finalidad de que el aplicativo pueda ser reusado, se creó una base de datos global en la cual se aloja la colección de usuarios. El usuario se registra por su correo electrónico y el sistema utiliza el dominio de email como org_id. Múltiples usuarios de un

mismo dominio pueden unirse pero quedarían marcados bajo la misma organización. Para cada organización se usa el `org_id` para crear una base de datos independiente. Posteriormente, se tiene la creación de proyectos y se definió la colección `projects` para esta finalidad. El agregado de patrones que son las ecuaciones de búsqueda se definió en la colección `patterns#project_name`, donde `project_name` es el nombre del proyecto.

También, como se observa en la figura 7.6, cuando se arranca el pipeline de metadatos se crea la colección `metadata#project_name#pattern_id` cuyo nombre es dinámico, `project_name` corresponde al nombre del proyecto y `pattern_id` corresponde al identificador de una ecuación de búsqueda. Entonces, para cada ecuación de búsqueda se guarda un documento por cada artículo PDF procesado. Paralelamente, se crea una colección llamada `metadata#project_name#global` en la cual también se llenan los metadatos obtenidos de los artículos pero en esta colección se incluyen los artículos de todas las cadenas de búsqueda. Similar a los metadatos de los artículos, se tienen otras colecciones en las que se va llenando metadatos de entidades como autores, organizaciones, países, palabras clave, idioma y año cuyo nombre de colección se definió como `entity_info#project_name#pattern_id` donde `entity` puede tomar los valores de `author`, `organization`, `country`, `keyword`, `language`, `year`. Estos metadatos de entidades también son llenados en colecciones independientes por cada cadena de búsqueda y en una colección global llamada `entity_info#project_name#global`.

Otra parte de las colecciones generada por el pipeline de metadatos tiene el formato `entities#project_name#pattern_id` donde `entities` puede tomar valores como `authors`, `organizations`, `countries` y `keywords`. La finalidad de estas colecciones es llenar en la base de datos una información que usará un pipeline posterior para construir matrices de adyacencia que serán la base para construir las redes. Como en el caso anterior, también se llena la información

por cada cadena de búsqueda y de manera global. Una colección adicional llamada `wordcloud#project_name#pattern_id` permite agregar todas las palabras clave que se van encontrando, sirve para un proceso posterior que genera un gráfico de nube de palabras. De esta manera, se estructura una información que permite soportar múltiples usuarios, organizaciones, proyectos y cadenas de búsqueda, en la figura 7.6 también quedan especificados los campos de información manejados en cada documento de las colecciones construidas en el proyecto.

Revisión de literatura sobre análisis de información científica en el dominio del cáncer

Se realizó una revisión sistemática de literatura con el fin de identificar estudios previos que aplicaran análisis de redes complejas en el contexto de corpus de literatura científica sobre el cáncer. Esta revisión permitió definir las bases metodológicas para implementar el análisis de redes en los datos extraídos y se logró encontrar el avance tecnológico al respecto y los trabajos relacionados con la idea investigativa. Como parte de la metodología se desarrolló una estrategia de búsqueda que permitió categorizar los términos encontrando una ecuación para obtener resultados relacionados con la investigación propuesta. Las categorías fueron las siguientes:

- Enriquecimiento semántico
- Ciencia de redes
- Redes de coautoría

Como resultado de agregar y reducir términos con criterios de inclusión y exclusión, se encontró un corpus de 48 documentos que al revisar su contenido resultaron relacionados 25 como muestra la tabla 7.6. Después de procesar la información de la búsqueda de literatura, se creó el estado del arte que se encuentra en las primeras secciones del presente documento.

Tema	Configuración	Cadena de búsqueda	Total	Relacionado
Semantic Enrichment	Scopus filtrado año: 2018-2023 Acceso abierto tipo artículo y review,	OA (all) AND ABS (semantic OR ontology OR knowledge OR biomedical) AND ABS (enrichment OR annotation OR automat* OR mining OR discover*) AND ABS (scopes OR pubmed OR medline OR shield OR literature OR search) AND ABS (cancer) AND NOT ABS (drug OR gen* OR metabolism OR nucleic OR security OR imag*) AND (LIMIT-TO (SUBJAREA,"COMP"))	21	9
	WOS Filtrado año:2018-2023 Acceso abierto, refinado por categorías de web of science	AB=(Semantic OR Ontology OR Knowledge OR Biomedical) AND AB=(Enrichment OR Annotation OR Automat* OR Mining OR Discover*) AND AB=(Scopus OR Pubmed OR Medline OR Scielo OR Literature OR Search) AND AB=(Cancer) NOT AB=(Drug OR Gen* OR Metabolism OR Nucleic OR Security OR Imag*)	4	4
Network Science/ Cancer	Scopus filtrado año: 2018-2023 Acceso abierto	OA (all) AND ABS ("Network science" OR "Network Analysis") AND ABS (cancer) AND ABS (centralit*) AND (LIMIT-TO (SUBJAREA,"COMP"))	8	7
	WOS Filtrado año:2018-2023 Acceso abierto	AB=("Network science" OR "Network Analysis") AND AB=(Cancer) AND AB=(centralit*)	5	5
Co-auth* Networks/ Cancer	Scopus filtrado año: 2018-2023	OA (all) AND ABS ("Co-auth* Networks") AND (cancer) AND (LIMIT-TO (SUBJAREA,"COMP"))	5	5
	WOS Filtrado año:2018-2023	AB=("Co-auth* Networks") AND AB=(Cancer)	5	4
Duplicados				9
Total				25

Tabla 7.6. Documentos encontrados en la revisión sistemática de literatura.

Obtención de matrices de adyacencia

Las matrices de adyacencia son requeridas para la construcción de redes a partir de la información. En una primera iteración, luego de descargar los metadatos, se utilizó un script que

consultaba la información disponible en la base de datos e intentaba crear las interacciones entre entidades como autores. Sin embargo, por limitaciones de memoria de Python, las matrices empleadas duraban mucho tiempo y en algunos casos no se podían terminar de procesar. Se encontró que la forma apropiada para evitar este procesamiento era crear la matriz dentro de la misma base de datos y en tiempo de consulta a los agregadores de literatura ir llenando la información de todas las relaciones para la construcción de la red.

La figura 7.7 muestra el proceso seguido para la creación de una colección en la base de datos que será la base para la construcción de las matrices de adyacencia. Cuando alguno de los contenedores metapub, arxiv o core consultan los metadatos de un artículo científico y se envía una petición al contenedor preprocessing para extraer la información de los metadatos adicionales como se muestra en la figura 7.5, se dispone entonces del metadato emails que consiste en una lista de los correos electrónicos detectados en el texto completo del documento. El dato emails se guarda en los metadatos del artículo pero adicionalmente llena la colección authors#project_name#pattern_id, esta colección tiene los dos campos author y related. Para cada ítem de la lista emails se verifica si el email existe en el campo author de la colección, en caso de no existir se crea el nuevo registro. Luego se toman los demás correos en la lista y se llenan en el campo related, teniendo en cuenta que la lista no quede con ítems duplicados.

De esta manera, se obtiene una lista de correos electrónicos de autores anotada con las relaciones que tienen con otros correos electrónicos. El mismo proceso se realiza para el caso de entidades como organizaciones, países, y palabras clave. En un proceso posterior, a través del pipeline de adyacencia del contenedor orchestrator, se consultan estas colecciones y usando la función parse_adjlist de la librería NetworkX de Python se logra agregar al pipeline la información de la matriz de adyacencia para el posterior análisis de redes.

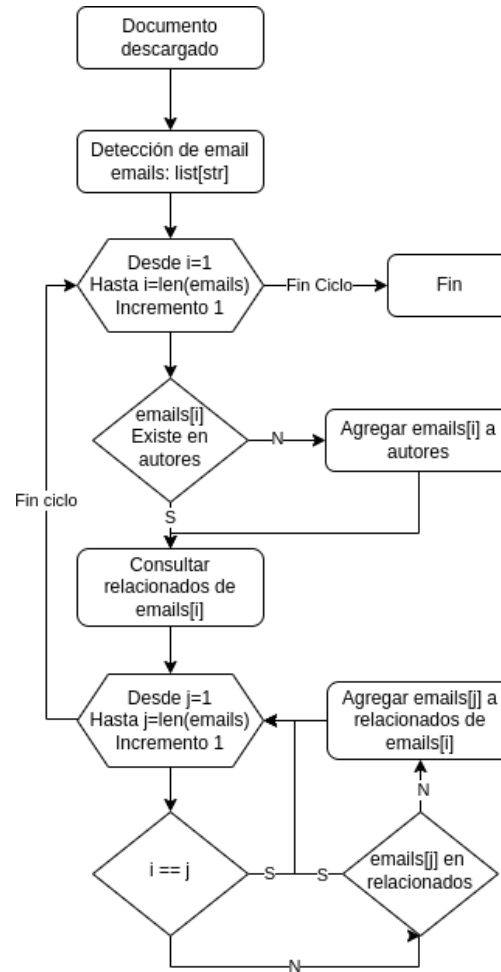


Figura 7.7. Construcción de colección de autores en la base de datos para soporte de las matrices de adyacencia.

Sistema prototipo para el análisis de redes complejas

Se implementó un sistema prototipo que permite a los usuarios generar análisis de redes complejas a partir de los datos descargados. Este sistema proporciona información como la mostrada en la figura 7.8 e incluye la generación, cálculo de métricas y visualización de las redes ya disponibles en la base de datos.

usuario la generación de visualizaciones interactivas para las redes, se seleccionó la librería 3d-force-graph [64]. Esta librería usa el objeto `node_link_data` para construir una visualización que permite navegar de manera tridimensional en las redes ofreciendo una mejor vista para cuando se tienen gran cantidad de nodos y es difícil representarlos en las gráficas obtenidas desde el microservicio implementado.

Estudio de UX/UI basado en encuestas a personal médico

En enero de 2022, se realizó un estudio [65] que consistió en entrevistas a personal e investigadores médicos, con el fin de implementar una plataforma acorde a las necesidades del sector salud. Estas encuestas permitieron identificar los aspectos más relevantes para los usuarios finales, entre ellas estuvo definir tres roles como personal médico, investigador y administrador del sistema. De acuerdo a ello se identificó las proto personas como muestra la figura 7.9.

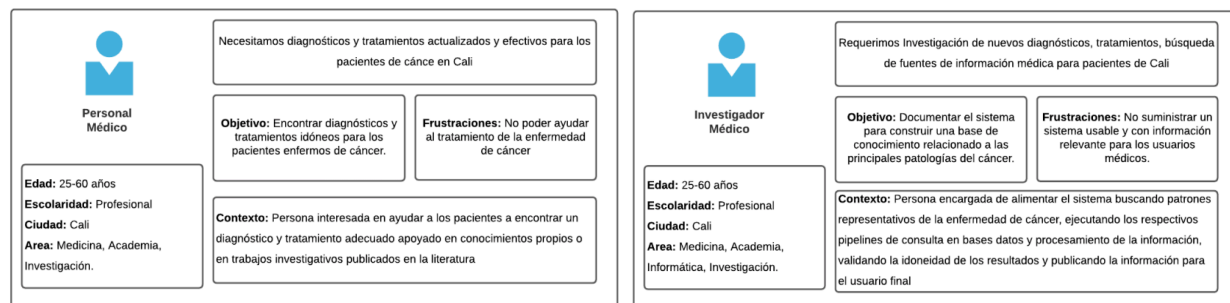


Figura 7.9. Proto Personas y roles iniciales del sistema propuesto.

Por otra parte, a los entrevistados les interesa tener una interfaz clara y sencilla, que facilite el acceso a la información. Por ejemplo disponer de opciones para filtrar la literatura por características del paciente, diagnóstico, alternativas terapéuticas. Que la información presentada tenga datos como autor, revisita, país, clasificación Q1-Q4, fecha de publicación, enlace directo al artículo. Esta actividad ayudó a conocer el público objetivo y definir un diseño [66] para la interfaz de usuario cuya arquitectura de información propuesta se presenta en la figura 7.10.

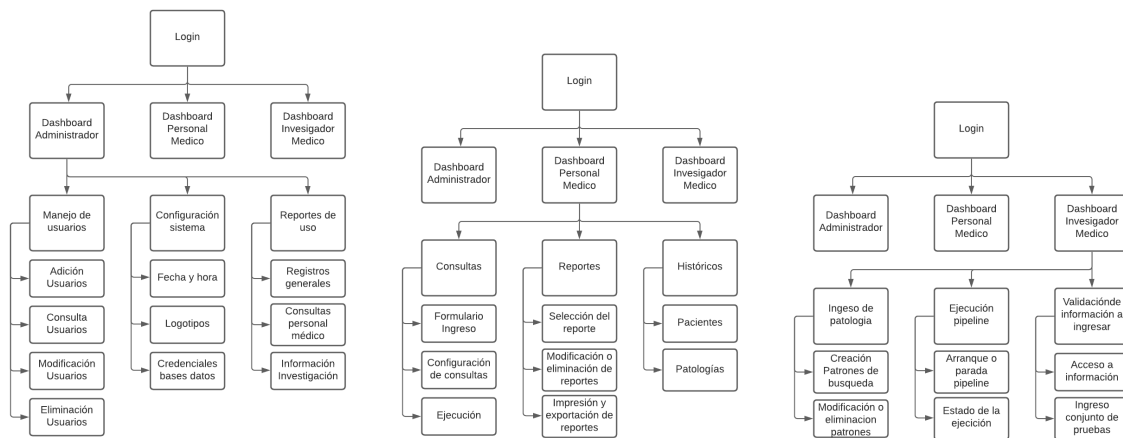


Figura 7.10. Arquitectura de información propuesta para la interfaz de usuario para el proyecto.

Desarrollo de un aplicativo web prototipo basado en microservicios I-ADC-UNIVALLE

En este punto el sistema cuenta con las herramientas necesarias para recuperar y analizar la información. Sería posible a través de los microservicios implementados y sin usar una interfaz gráfica de usuario, realizar un estudio del tema de interés. Sin embargo, tal proceso como el mostrado en la figura 7.11 implicaría integrar las diversas implementaciones presentes en los contenedores. Esto no es algo práctico y quizás sea necesario desarrollar un script adicional para lograrlo. Por lo tanto surge la necesidad de facilitar el proceso de uso del sistema y es la razón por la cual se implementó el contenedor gui. En este contenedor se ha instalado el aplicativo web I-ADC-UNIVALLE basado en Create React App [67], TypeScript [68] y Material UI [69]. Este aplicativo permite realizar el proceso desde la creación de nuevas cuentas de usuario, el ingreso de proyectos y cadenas de búsqueda hasta la obtención de los patrones de interés de la literatura consultada por medio de una interfaz gráfica de usuario integrando todos los contenedores configurados en el proyecto.



Figura 7.11. Proceso de obtención de patrones de literatura científica en el dominio del cáncer usando el sistema propuesto. Implica las siguientes etapas: Definir patrones de búsqueda relacionados con el cáncer, ingreso de los patrones de búsqueda y creación de usuarios organizaciones y proyectos para acceso a la plataforma, ejecución del pipeline de metadatos para capturar información del sistema, ejecución del pipeline de adyacencia para obtención de datos de análisis de redes, uso de los patrones obtenidos para recuperar los documentos e información relevante.

Características del aplicativo

El aplicativo fue implementado pensando en soportar múltiples usuarios y proyectos, a su vez permite ejecutar pipelines para múltiples cadenas de búsqueda y ofrecer información de valor a los usuarios del sistema. Por tal razón tiene las siguientes características:

Manejo de usuarios y proyectos

El sistema inicia con un usuario administrador que tiene unos datos de acceso definidos en el arranque del sistema. Por razones de seguridad el usuario administrador debe cambiar su información de acceso. A medida que los usuarios requieren acceder al sistema, pueden entrar a la interfaz que se observa en la figura 7.12 y llenar un formulario de registro. El usuario administrador al ingresar por configuración de usuarios puede activar las cuentas que se van registrando y definir si corresponden al rol usuarios médico o investigador médico. Las organizaciones de usuario son creadas de manera automatizada en el momento del registro

usando el dominio de correo del usuario registrado. Dependiendo del rol creado, los usuarios acceden a características de la aplicación de acuerdo a la arquitectura mostrada en la figura 7.10.

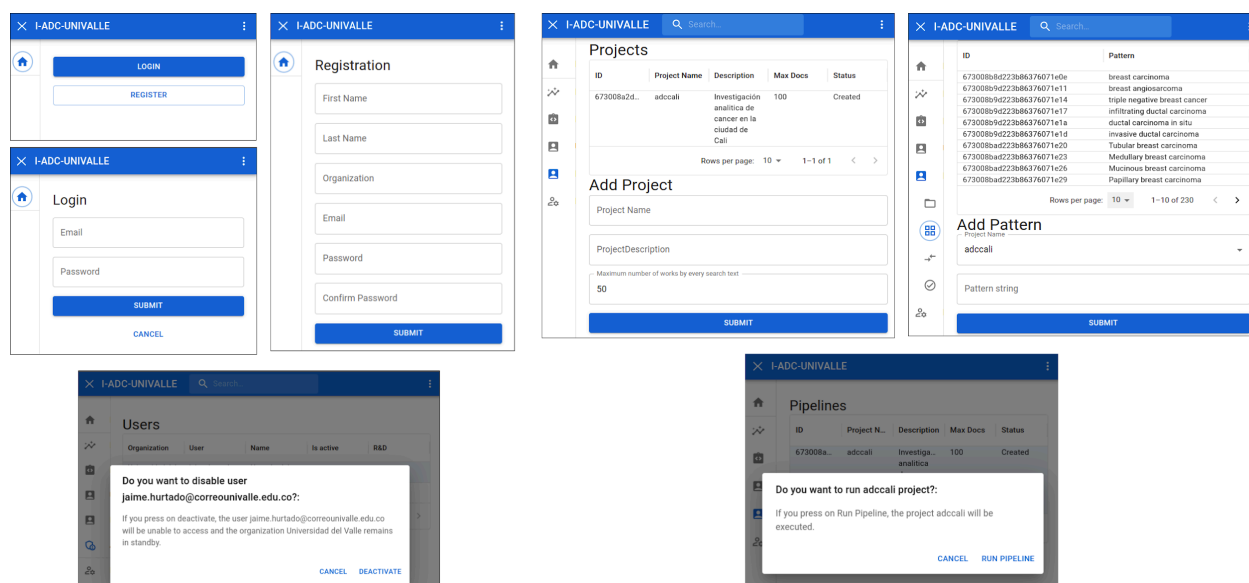


Figura 7.12. Manejo de usuarios y administración de proyectos patrones y pipelines.

En la derecha de la figura 7.12 también se muestran los formularios a llenar en el proceso de configuración del sistema. El usuario administrador o el investigador médico tienen acceso a esta característica y deben primero crear un proyecto introduciendo el nombre del proyecto, la descripción y el número de documentos máximo a consultar por cada cadena de búsqueda. Una vez se tiene confirmada la creación del proyecto se ingresa a la sección de patterns que presenta dos formularios para ingresar las cadenas de búsqueda, uno de los formularios ingresa un texto a la vez, la otra opción permite cargar un archivo de texto donde cada línea corresponde a una cadena de búsqueda diferente. Se puede listar en una tabla superior los patrones configurados. Una vez ya se han introducido todas las cadenas de búsqueda de interés, se procede a ejecutar el pipeline de metadatos. Ingresando por la sección de pipelines, se observa una tabla con todos los proyectos configurados y al hacer click en la sección status se abre una confirmación para iniciar la ejecución del proyecto.

Visualización de las redes construidas y sus métricas

Desde el momento en que se ejecuta el proyecto, se habilita la visualización de información de las redes construidas. Sin embargo desde la sección de pipeline se puede ir consultando el avance de la ejecución para poder visualizar la información completa cuando el pipeline haya finalizado. En la figura 7.13 se presenta la interfaz de usuario para la visualización de las redes construidas. En la parte superior de esta visualización se puede introducir la configuración de la red a consultar seleccionando el tipo de gráfico (autores, organizaciones, países y palabras clave), el nombre del proyecto y el nombre de la cadena de búsqueda. Adicional a todas las cadenas de búsqueda se tiene la opción global para consultar una red que une toda la literatura consultada en las diferentes cadenas de búsqueda.

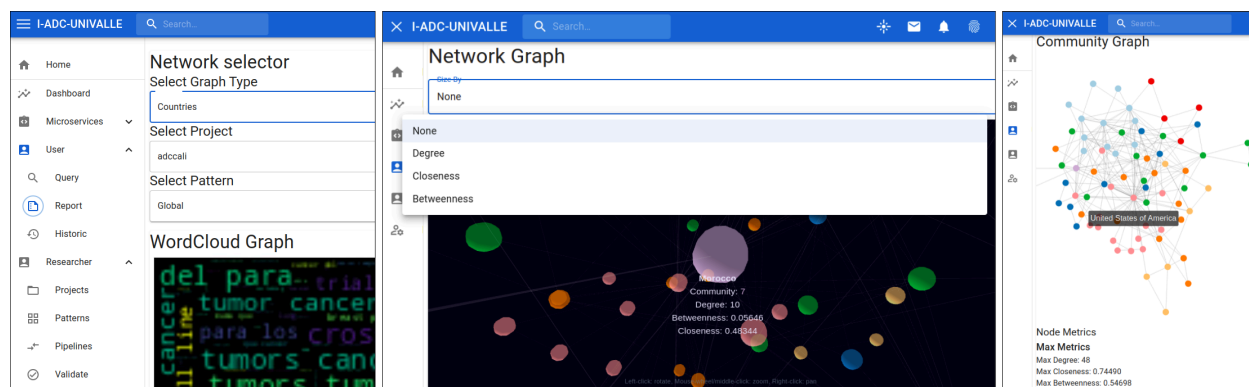


Figura 7.13. Consulta de información de analítica de redes.

Posterior al selector de red, aparece el gráfico de la nube de palabras cuyo tamaño de la palabra representa la frecuencia en que se ha detectado esa coincidencia en todos los documentos. Cabe aclarar que por cada documento solo se detectan las tres palabras clave más importantes. Haciendo scroll hacia abajo, aparecen el gráfico de red que consiste en una visualización tridimensional que permite navegación a través de la red y el gráfico de comunidades. Ambos gráficos son interactivos pero se ha implementado en el gráfico de red la

opción size by que permite cambiar el tamaño de los nodos dependiendo de la métrica de red empleada. Las opciones para esta característica son el grado del nodo, la centralidad por intermediación y la centralidad por cercanía.

Consulta de los nodos de interés sus metadatos e información relacionada

La interacción de estos gráficos también se ha configurado para que al colocar el cursor sobre un nodo despliegue información con las métricas de redes, y al oprimir sobre un nodo de interés se pueda acceder a los metadatos registrados para el nodo. En la figura 7.14 se observa un modal con la información correspondiente al nodo seleccionado y sus metadatos corresponden a enlaces con información relacionada al nodo. Al oprimir en estos enlaces se abre un segundo modal de información detallada en el que, si la instancia consultada corresponde a un identificador de un documento, se muestra el enlace de descarga o visualización del texto completo.

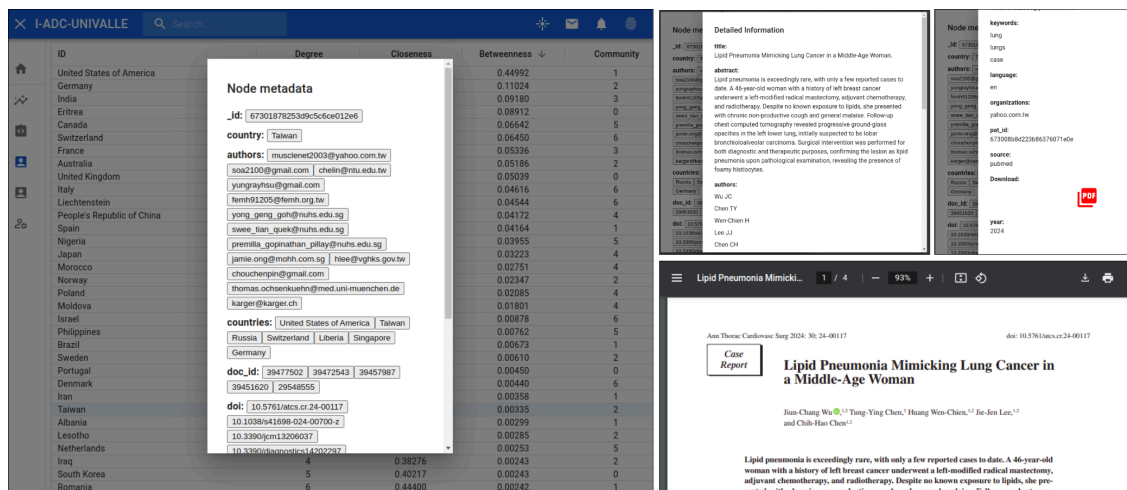


Figura 7.14. Identificación de entidades más importantes. La tabla permite ordenar la lista por cada métrica, al seleccionar el ítem de interés despliega información del nodo con enlaces para ampliar información importante de los metadatos relacionados. En caso de elegir un identificador de documento se tiene acceso al texto completo en formato PDF.

La figura 7.14 también muestra la tabla de métricas de red en la que es posible ordenar por cada métrica para encontrar los nodos más importantes. Al seleccionar un nodo en esta tabla, también se abre el modal de información de los metadatos y se enfoca el nodo en la gráfica de red. De esta manera se facilita la búsqueda de patrones en los datos y se tiene acceso inmediato a la información de cada nodo para facilitar los procesos de investigación.

Patrones en literatura científica de cáncer de mama usando técnicas de redes complejas

Se analizaron registros patológicos de varias instituciones médicas en Cali, Colombia. Se pudo identificar y extraer patrones como nombres de medicamentos, tipos de tumores, hallazgos clínicos y otros términos específicos relacionados con el cáncer de mama y la enfermedad mamaria. Con base en la información allí obtenida se creó una lista de palabras clave relacionadas al cáncer de mama disponible en [70]. Este listado contiene 111 cadenas de búsqueda en inglés y 121 cadenas de búsqueda en idioma español.

Se creó una organización en el sistema I-ADC-UNIVALLE usando la interfaz gráfica. Posteriormente, se creó un nuevo proyecto en esta organización y se agregó la lista de palabras clave al sistema prototipo. Se configuró como parámetro de búsqueda un máximo de 100 documentos por cada palabra clave. Con las diversas visualizaciones permitidas en la plataforma, se obtuvo la visualización de redes y se extrajo patrones obtenidos de la literatura consultada como lista de autores, instituciones, países y palabras clave relevantes mostradas en el anexo 1.

En el pipeline de metadatos fue posible procesar 11348 documentos de los cuales a 8726 eran en idioma inglés. La fuente de los artículos obtenidos fue 58% de Core, 36% de PubMed y 6% de arXiv. Se estima una duración de 15 segundos para procesar cada documento en el pipeline de metadatos, sin embargo el proceso se realiza de manera paralela en cada uno de los contenedores metapub, arxiv y core. El tiempo de procesamiento también depende del tamaño

del texto completo, en el proceso fue importante limitar la cantidad máxima de caracteres a tener en cuenta por cada documento y se configuró en 10000. Como resultado de este estudio de caso se resumen los patrones obtenidos de la literatura científica consultada y relacionada con la enfermedad del cáncer de mama en la tabla 7.7. Se recopilan los nodos más relevantes para cada caso teniendo en cuenta las métricas de grado, centralidad por cercanía y centralidad por intermediación.

Métrica	Autores	Organizaciones	Países	Palabras clave
Grado	mruchala@ump.edu.pl	meduniwien.ac.at	United States of America	cancer
Grado	m.delarentiis@istitutotumori.na.it	mdpi.com	People's Republic of China	cell
Grado	shleem@dau.ac.kr	uniroma1.it	Italy	breast
C.C.	lucia.delmastro@hsanmartino.it	imperial.ac.uk	Eritrea	tumor
C.C.	giuseppe.curigliano@ieo.it	wustl.edu	Germany	patient
C.C.	dgenerali@units.it	mdanderson.org	Canada	carcinoma
C.I.	rupert.bartsch@meduniwien.ac.at	oup.com	Moldova	image
C.I.	jacortes@vhio.net	sydney.edu.au	United Kingdom	lesion
C.I.	g.jerusalem@chuliege.be	petermac.org	Norway	gene

Tabla 7.7. Patrones de literatura científica obtenidos en el estudio de caso sobre la enfermedad del cáncer de mama. C.C. y C.I corresponden a centralidad por cercanía y centralidad por intermediación, respectivamente. Se escogieron los nodos más relevantes, 3 por cada métrica evitando resultados repetidos.

Autores

Se utilizó la interfaz gráfica para generar la red de autores para todas las cadenas de búsqueda encontrando 9385 autores. En este proceso se observa un limitante dado que se usan los email detectados en el documento. Algunas veces este email no corresponde necesariamente

al autor, puede estar relacionado con la revista u otra entidad. También se observa que para algunos documentos no fue posible encontrar el email. La decisión de tomar este parámetro para la construcción de la red obedeció a la ambigüedad que hay en el nombre de los autores que es definido de manera diferente por cada agregador. En la figura 7.15 se muestran imágenes de la red construida, mruchala@ump.edu.pl fue el autor con más conexiones en la red, 180 en total. En esta visualización de redes los nodos del mismo color pertenecen a la misma comunidad, se observa como el autor pertenece a la comunidad de color azul pero también tiene muchos trabajos en común con la comunidad de color amarillo.

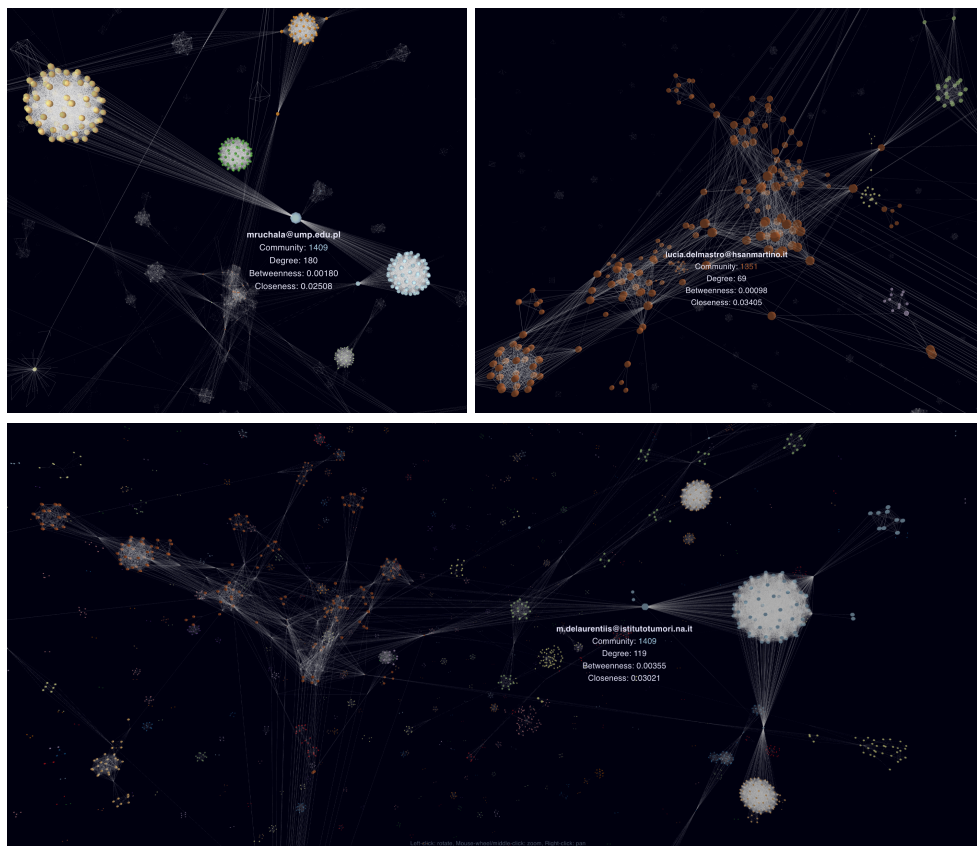


Figura 7.15. Visualización de la red de autores. Superior izquierda: grado, superior derecha: centralidad por cercanía, inferior: centralidad por intermediación.

El nodo de lucia.delmastro@hsanmartino.it se resalta como el más importante usando la métrica de centralidad por cercanía, esto indica que desde este nodo se encuentra a menos pasos de llegar a cualquier otro nodo de la red. Se observa que nodos cercanos a este también tienen un peso similar de intermediación por cercanía y la comunidad de color naranja puede ser entonces de importancia en el tema investigado. En la figura 7.15 también se puede observar el nodo m.delaurentiis@istitutotumori.na.it que es el más importante en la métrica de centralidad por intermediación. Indica que por este nodo hay un flujo importante de información dado que permite conectar diversas zonas de la red. En la visualización se observa que las conexiones cercanas a este nodo viajan a otros puntos de la red evidenciando el comportamiento esperado de la métrica.

Organizaciones

El grafo de organizaciones se muestra en la figura 7.16 y es construido con los dominios de los email detectados en cada documento. Los resultados del estudio de caso muestran que una parte de los investigadores usan correos no institucionales para referenciar su contacto, esto explica la presencia de nodos grandes en la visualización perteneciente a dominios como gmail, yahoo, hotmail, 163, etc. Todos ellos son proveedores de email con dominios no institucionales. Esto nubla ligeramente la detección de organizaciones relevantes dado que estos dominios están muy conectados.

Sin embargo, el cálculo de métricas de red también revela nodos de interés con valores más altos de estas métricas. Descartando aquellos nodos de correo no institucional, según las métricas, de los 2404 nodos obtenidos, el nodo meduniwien.ac.at fue el más importante tanto en grado como en centralidad por cercanía. Este nodo corresponde a la Universidad Médica de Viena indicando que posiblemente su trabajo es ampliamente realizado en coordinación con otras instituciones.

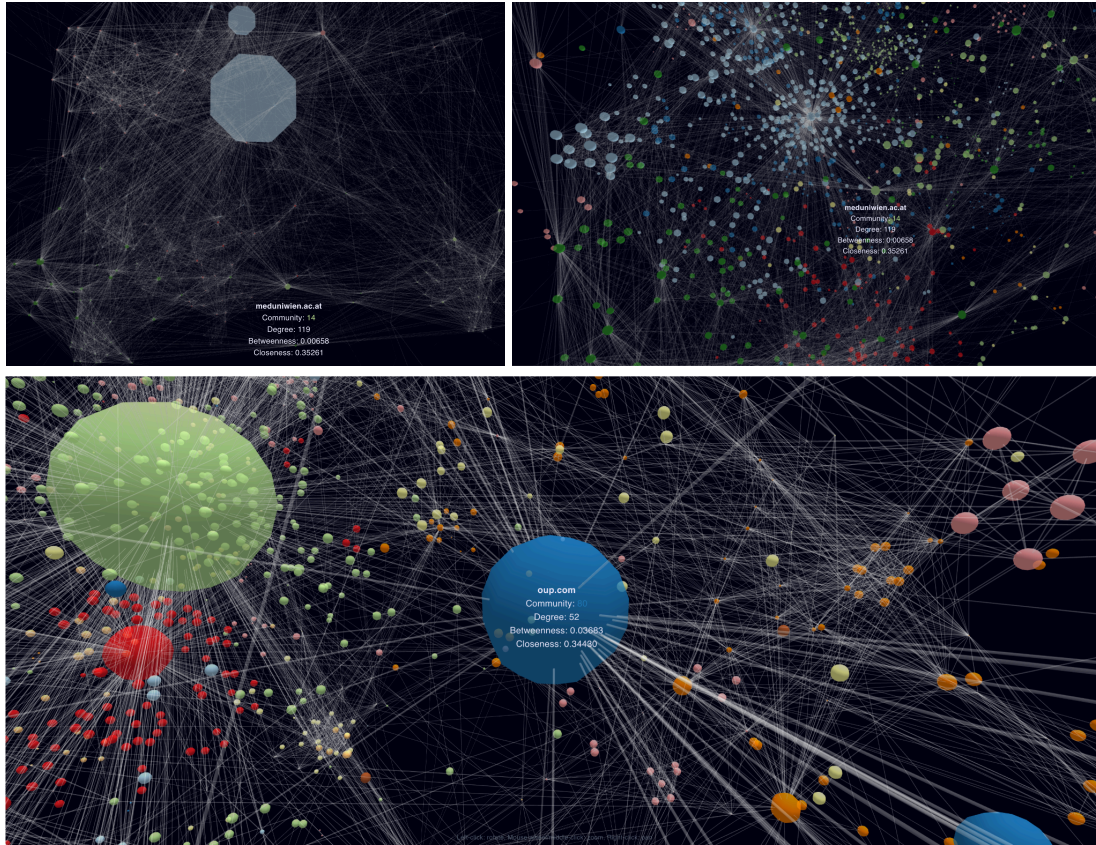


Figura 7.16. Visualización de la red de organizaciones. Superior izquierda: grado, superior derecha: centralidad por cercanía, inferior: centralidad por intermediación.

Por otra parte, se encontró el nodo `oup.com` perteneciente a Oxford University Press un editor universitario. Aunque se observa que este editor tiene servicios enfocados en la temática de cáncer, también nos indica que aparte de email no institucional podemos tener interferencia de los email puestos por editoriales en el documento. También la visualización nos muestra que esta red está más conectada comparada con la red de autores.

Países

Según la red obtenida para países, Estados Unidos, China, Italia y Alemania están entre los más relevantes en el tema de este estudio. Aunque Estados Unidos mostró ser el nodo más

importante en todas las métricas analizadas como se observa en la figura 7.17. También se ve que de los 200 nodos de países detectados, Estados Unidos está relacionado con 197.

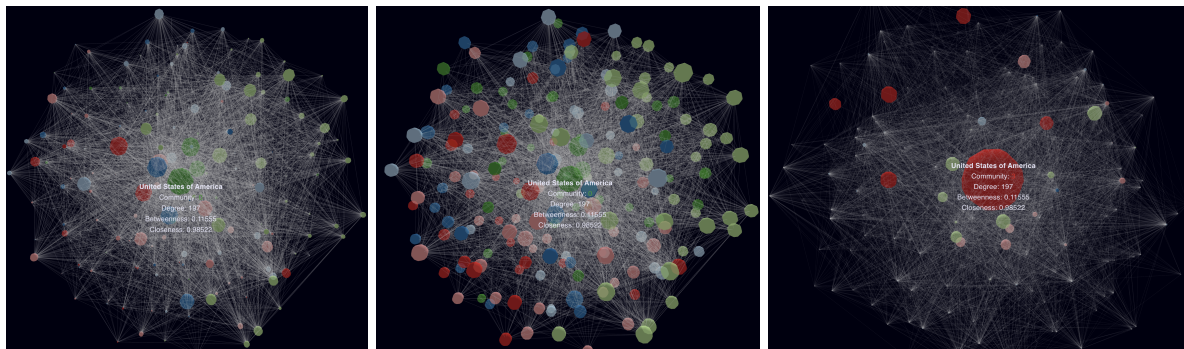


Figura 7.17. Visualización de la red de países. Izquierda: grado, mitad: centralidad por cercanía, derecha: centralidad por intermediación.

La gran cantidad de interacciones entre los nodos hacen que la métrica de centralidad por cercanía sea de valores semejantes para la mayoría de los nodos en la red de países. Por eso, en este caso la visualización del nodo de Estados Unidos no se mostró de tamaño más grande. De manera similar a Estados Unidos, China mostró estar en segundo lugar en todas las métricas seguido en tercer lugar por Italia. Esto muestra la importancia de estos países en estudios sobre el cáncer de mama.

Palabras clave

En cuanto a palabras clave, esta métrica se construyó en el sistema utilizando un método keywords de la librería gensim [71]. Este método devuelve una lista de palabras clave y un score que indica las palabras más rankeadas. Al observar la red obtenida, se encontró que este método keywords no es apropiado para idiomas por fuera del inglés dado que clasifica pronombres y conectores al parecer del español o portugueses, adicionalmente en los datos guardados faltó implementar la obtención de la raíz de las palabras para evitar clasificaciones ambiguas. Aun así, fue muy interesante ver que en la red obtenida representada por la figura 7.18 porque la

disposición mostrada de la red clasifica por ejemplo las palabras del español a la izquierda del gráfico y palabras de entidades o significados parecidos las ubica muy cercanas entre sí. Las clasificaciones de acuerdo a las métricas de redes mostraron que de los 4578 nodos de palabras clave, los más importantes fueron cancer, cell, breast y tumor clasificados de manera similar por todas las métricas empleadas.

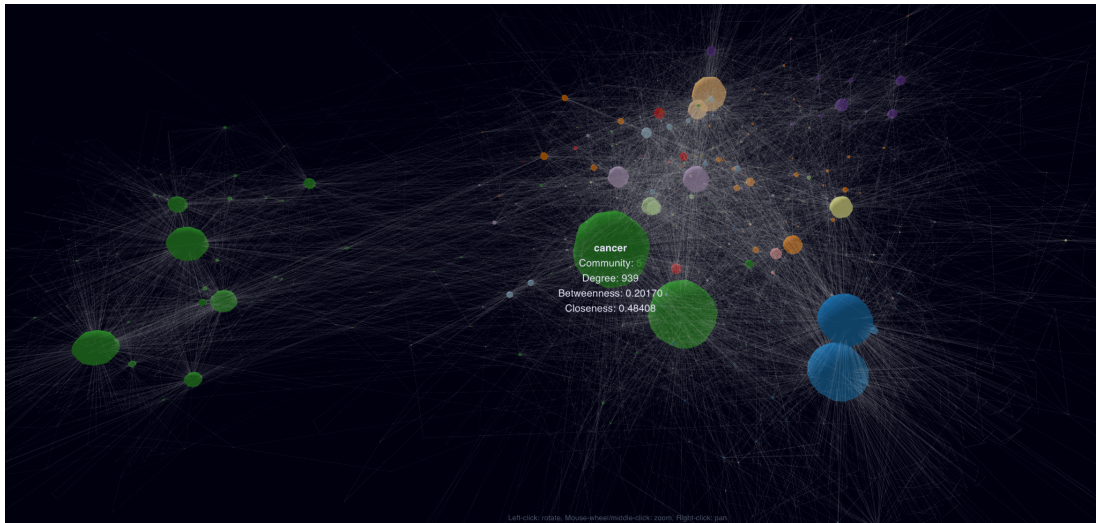


Figura 7.18. Visualización de la red de palabras clave, el tamaño de los nodos está definido por la métrica grado del nodo.

8. Conclusiones

Por medio de una metodología DSRM se abordó el problema de tener una gran cantidad de literatura científica y la dificultad que hay para correlacionar la información de diferentes estudios para encontrar trabajos prometedores. El desarrollo metodológico por medio de la construcción de diversos prototipos, permitió consolidar un pipeline tecnológico para solucionar esta problemática. Mediante el uso de un sistema prototipo, basado en tecnologías de integración y despliegue continuo fue posible el desarrollo de los diversos componentes del sistema.

El sistema principal se compone por servicios web con funcionalidades diversas. Por una parte están los contenedores metapub, arxiv y core que brindan acceso a bases de datos de literatura científica, extraen los metadatos y el url de acceso al texto completo de cada documento. Adicionalmente estos contenedores usan un contenedor preprocessing encargado de extraer información del texto completo de los documentos. Finalmente, con toda la información obtenida, los contenedores llenan la base de datos que se encuentra en el contenedor db construyendo una base de conocimiento apropiada para el posterior análisis de la información. Un contenedor llamado orchestrator se encarga de ejecutar pipelines implementados de metadatos y de adyacencia, este último es empleado en el análisis y visualización de redes.

El sistema completo se integra por medio de una interfaz gráfica de usuario implementada en el contenedor gui. Por medio de esta interfaz se facilita el manejo de usuarios, proyectos, configuración del sistema y ejecución de las pipelines configuradas. Adicionalmente, la interfaz permite acceder a la información recopilada y a los diferentes análisis implementados con la finalidad de encontrar patrones en literatura científica.

Durante el desarrollo del proyecto fue de importancia el estudio e implementación de tecnologías para integración y despliegue continuo. Con esto se logró que el sistema pueda funcionar tanto en ambientes locales como en un ambiente productivo implementando en un cluster Kubernetes en la nube de Google Cloud.

Un estudio de caso fue realizado para poner a prueba el sistema desarrollado, se utilizaron 232 cadenas de búsqueda que permitieron construir un corpus de 11348 documentos relacionados con el cáncer de mama. El sistema fue capaz de crear relaciones entre los metadatos obtenidos y por medio de técnicas de análisis de redes encontrar patrones representativos de la literatura consultada.

9. Trabajo futuro

Algunas limitaciones encontradas pueden ser abordadas por trabajos futuros que permitan complementar y optimizar el sistema como incluir otros agregadores de literatura, mejorar la experiencia de usuario y la calidad de la información a ser consultada.

Como se mostró en la tabla 7.2, la presente investigación seleccionó los agregadores de literatura ArXiv, Core y Pubmed dado que presentaban características de acceso gratuito al texto completo de los documentos así como a librerías de software que facilitaban el proceso de desarrollo. Sin embargo hay agregadores de relevancia que pueden ser de interés de algún usuario y podrían ser agregados al sistema creando un microservicio para cada nuevo agregador e integrarlo en el sistema. Relacionado a esto, se tiene que este prototipo de búsqueda podría ser usado en investigaciones de otras áreas de estudio diferentes a la medicina, por lo que se podría usar agregadores de literatura afines con esas áreas.

Por otra parte, el presente trabajo utilizó unas ecuaciones de búsqueda sencillas para encontrar literatura científica que se basan en usar las palabras de interés para buscar en el resumen o abstract del documento (Ver tabla 7.3). Como la búsqueda avanzada de los agregadores ofrece una diversidad de configuraciones, se podría mejorar las consultas incluyendo búsqueda por áreas, fecha de publicación, afiliación, manejo de criterios de inclusión y exclusión, y demás opciones que estén permitidas.

El prototipo desarrollado no tiene limitación para la cantidad de ecuaciones de búsqueda a procesar en cada pipeline. En la obtención de patrones para el cáncer de mama, se usaron 232 cadenas de búsqueda y procesar esta información no fue posible de manera inmediata. A pesar de paralelizar el procesamiento para tener 3 instancias del microservicio preprocesing, el tiempo de procesamiento mínimo fue de 1 día. Este hecho puede indicar que la arquitectura de la figura 7.5

podría ser modificada con otras estrategias de procesamiento paralelizado que reduzcan el tiempo de las pipelines y consideren limitaciones relacionadas a cantidad de peticiones y tiempo de procesamiento del texto completo.

Así como se puede crear un microservicio para aumentar los agregadores de literatura científica, también se pueden modificar los metadatos que se extraen del texto completo y con esta información crear nuevas redes que complementen las obtenidas para autores, organizaciones, países y palabras clave. Se sugiere también extender el sistema para usar otros conceptos de redes que orienten la obtención de otros patrones o tendencias en la información sobre la literatura analizada.

Finalmente, otros trabajos futuros podrían abordar el tema de la usabilidad e interacción con el usuario, complementando pruebas de usabilidad del prototipo con el desarrollo de nuevas características en el frontend del aplicativo. Podría requerir diversas iteraciones con realimentación del usuario final para consolidar la presentación de información cambiando o agregando gráficas y tablas, aumentando la información que se presenta, creando nuevos dashboard, agregando nuevas visualizaciones para alguna necesidad específica, explorando opciones de accesibilidad, entre otras acciones encaminadas a tener una buena experiencia de usuario.

10. Referencias

- [1] A. Al-Aamri, K. Taha, Y. Al-Hammadi, M. Maalouf, y D. Homouz, «Analyzing a co-occurrence gene-interaction network to identify disease-gene association», *BMC Bioinformatics*, vol. 20, n.º 1, p. 70, feb. 2019, doi: 10.1186/s12859-019-2634-7.
- [2] A. Mehmood, B.-W. On, I. Lee, H. W. Park, y G. S. Choi, «Corroborating social media echelon in cancer research», *Qual. Quant.*, vol. 52, n.º 2, pp. 801-813, mar. 2018, doi: 10.1007/s11135-017-0489-4.
- [3] V. M. Gupta y C. V. P. Krishna, «K-core analysis and modeling for network centralities», *Int. J. Eng. Technol.*, vol. 7, n.º 2.7, Art. n.º 2.7, mar. 2018, doi: 10.14419/ijet.v7i2.7.10285.

- [4] O. Solarte, O. Montenegro, M. Torrente, A. R. González, M. Provencio, y E. Menasalvas, «Negation and uncertainty detection in clinical texts written in Spanish: a deep learning-based approach», *PeerJ Comput. Sci.*, vol. 8, p. e913, mar. 2022, doi: 10.7717/peerj-cs.913.
- [5] M. Thilakaratne, K. Falkner, y T. Atapattu, «A Systematic Review on Literature-based Discovery: General Overview, Methodology, & Statistical Analysis», *ACM Comput. Surv.*, vol. 52, n.º 6, p. 129:1-129:34, Diciembre 2019, doi: 10.1145/3365756.
- [6] A. Kastrin y D. Hristovski, «Scientometric analysis and knowledge mapping of literature-based discovery (1986–2020)», *Scientometrics*, vol. 126, n.º 2, pp. 1415-1451, feb. 2021, doi: 10.1007/s11192-020-03811-z.
- [7] B. Raza, R. Ahmad, M. H. N. B. M. Nasir, y S. S. M. Fauzi, «Socio-Technical Congruence as an Emerging Concept in Software Development: A Scientometric Analysis and Critical Literature Review», *IEEE Access*, vol. 9, pp. 129051-129077, 2021, doi: 10.1109/ACCESS.2021.3113637.
- [8] M. G. Galety, C. A. Atroshi, B. Balabantaray, S. N. Mohanty, y an, «Social Network Analysis», *Soc. Netw. Anal.*, 2022.
- [9] A. S. D. Mata, «Complex Networks: a Mini-review», *Braz. J. Phys.*, vol. 50, n.º 5, pp. 658-672, oct. 2020, doi: 10.1007/s13538-020-00772-9.
- [10] W. Fan y M. D. Gordon, «The power of social media analytics», *Commun. ACM*, vol. 57, n.º 6, pp. 74-81, jun. 2014, doi: 10.1145/2602574.
- [11] J. Murack, «LibGuides: Resources and Tools for Computational Research: Home». Accedido: 24 de octubre de 2023. [En línea]. Disponible en: <https://libguides.mit.edu/c.php?g=1253643&p=9182313>
- [12] Amazon Web Services, «Integración continua», Amazon Web Services, Inc. Accedido: 17 de octubre de 2023. [En línea]. Disponible en: <https://aws.amazon.com/es/devops/continuous-integration/>
- [13] Gitlab, «What is CI/CD?» Accedido: 16 de noviembre de 2023. [En línea]. Disponible en: <https://about.gitlab.com/topics/ci-cd/>
- [14] Gitlab, «What is a continuous integration server?» Accedido: 17 de noviembre de 2023. [En línea]. Disponible en: <https://about.gitlab.com/topics/ci-cd/continuous-integration-server/>
- [15] J. Shah y D. Dubaria, «Building Modern Clouds: Using Docker, Kubernetes & Google Cloud Platform», en *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, ene. 2019, pp. 0184-0189. doi: 10.1109/CCWC.2019.8666479.
- [16] E. Bisong, «An Overview of Google Cloud Platform Services», en *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, E. Bisong, Ed., Berkeley, CA: Apress, 2019, pp. 7-10. doi: 10.1007/978-1-4842-4470-8_2.
- [17] J. I. Garcia-Gathright *et al.*, «Evaluating Casama: Contextualized semantic maps for summarization of lung cancer studies», *Comput. Biol. Med.*, vol. 92, pp. 55-63, ene. 2018, doi: 10.1016/j.combiomed.2017.10.034.
- [18] H. Kilicoglu, G. Rosemblat, M. Fiszman, y D. Shin, «Broad-coverage biomedical relation extraction with SemRep», *BMC Bioinformatics*, vol. 21, n.º 1, p. 188, may 2020, doi: 10.1186/s12859-020-3517-7.
- [19] H.-Y. Kim, K.-A. Kang, S.-J. Han, y J. Chun, «Web-Based Research Trends on Child and Adolescent Cancer Survivors Over the Last 5 Years: Text Network Analysis and Topic

- Modeling Study», *J. Med. Internet Res.*, vol. 24, n.º 2, p. e32309, feb. 2022, doi: 10.2196/32309.
- [20] Y. Ahmed y N. Miskov-Zivanov, «Guided assembly of cellular network models from knowledge in literature», en *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, nov. 2021, pp. 4458-4464. doi: 10.1109/EMBC46164.2021.9630181.
- [21] J. Fagan, K. S. Eddens, J. Dolly, N. L. Vanderford, H. Weiss, y J. S. Levens, «Assessing Research Collaboration through Co-authorship Network Analysis», *J. Res. Adm.*, vol. 49, n.º 1, pp. 76-99, 2018.
- [22] G. González-Alcaide, H. Pinargote, y J. M. Ramos, «From cut-points to key players in co-authorship networks: a case study in ventilator-associated pneumonia research», *Scientometrics*, vol. 123, n.º 2, pp. 707-733, may 2020, doi: 10.1007/s11192-020-03404-w.
- [23] R. Khademi, «Visualizing Garfield's Scientific Performance», *J. Scientometr. Res.*, vol. 7, n.º 3, pp. 210-214, ene. 2019, doi: 10.5530/jscires.7.3.33.
- [24] S. S. M. Khairi *et al.*, «Deep Learning on Histopathology Images for Breast Cancer Classification: A Bibliometric Analysis», *Healthcare*, vol. 10, n.º 1, Art. n.º 1, ene. 2022, doi: 10.3390/healthcare10010010.
- [25] M. S. Chaudhary, V. V. H. Pham, y T. D. Le, «NIBNA: a network-based node importance approach for identifying breast cancer drivers», *Bioinformatics*, vol. 37, n.º 17, pp. 2521-2528, sep. 2021, doi: 10.1093/bioinformatics/btab145.
- [26] J. Rafferty *et al.*, «Ranking sets of morbidities using hypergraph centrality», *J. Biomed. Inform.*, vol. 122, p. 103916, oct. 2021, doi: 10.1016/j.jbi.2021.103916.
- [27] C. T. Hoyt, A. L. Hoyt, y B. M. Gyori, «Prediction and curation of missing biomedical identifier mappings with Biomappings», *Bioinformatics*, vol. 39, n.º 4, p. btad130, abr. 2023, doi: 10.1093/bioinformatics/btad130.
- [28] F. Borchert *et al.*, «Knowledge bases and software support for variant interpretation in precision oncology», *Brief. Bioinform.*, vol. 22, n.º 6, p. bbab134, nov. 2021, doi: 10.1093/bib/bbab134.
- [29] T. Luechtefeld, T. Bozada, R. Goel, L. Wang, y C. J. Paller, «Applications for open access normalized synthesis in metastatic prostate cancer trials», *Front. Artif. Intell.*, vol. 5, 2022, Accedido: 19 de junio de 2023. [En línea]. Disponible en: <https://www.frontiersin.org/articles/10.3389/frai.2022.984836>
- [30] K. Peffers, T. Tuunanen, M. A. Rothenberger, y S. Chatterjee, «A Design Science Research Methodology for Information Systems Research».
- [31] J. Hurtado, *jandresh/inv-adc*. 2024. Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://github.com/jandresh/inv-adc>
- [32] «Elsevier Developer Portal». Accedido: 13 de noviembre de 2024. [En línea]. Disponible en: <https://dev.elsevier.com/>
- [33] Rosa-Clark, «Metadata Retrieval», Crossref. Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://www.crossref.org/documentation/retrieve-metadata/>
- [34] «Clarivate Developer Portal». Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://developer.clarivate.com/>
- [35] «APIs - Develop - NCBI». Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://www.ncbi.nlm.nih.gov/home/develop/api/>
- [36] «CORE API». Accedido: 14 de noviembre de 2024. [En línea]. Disponible en:

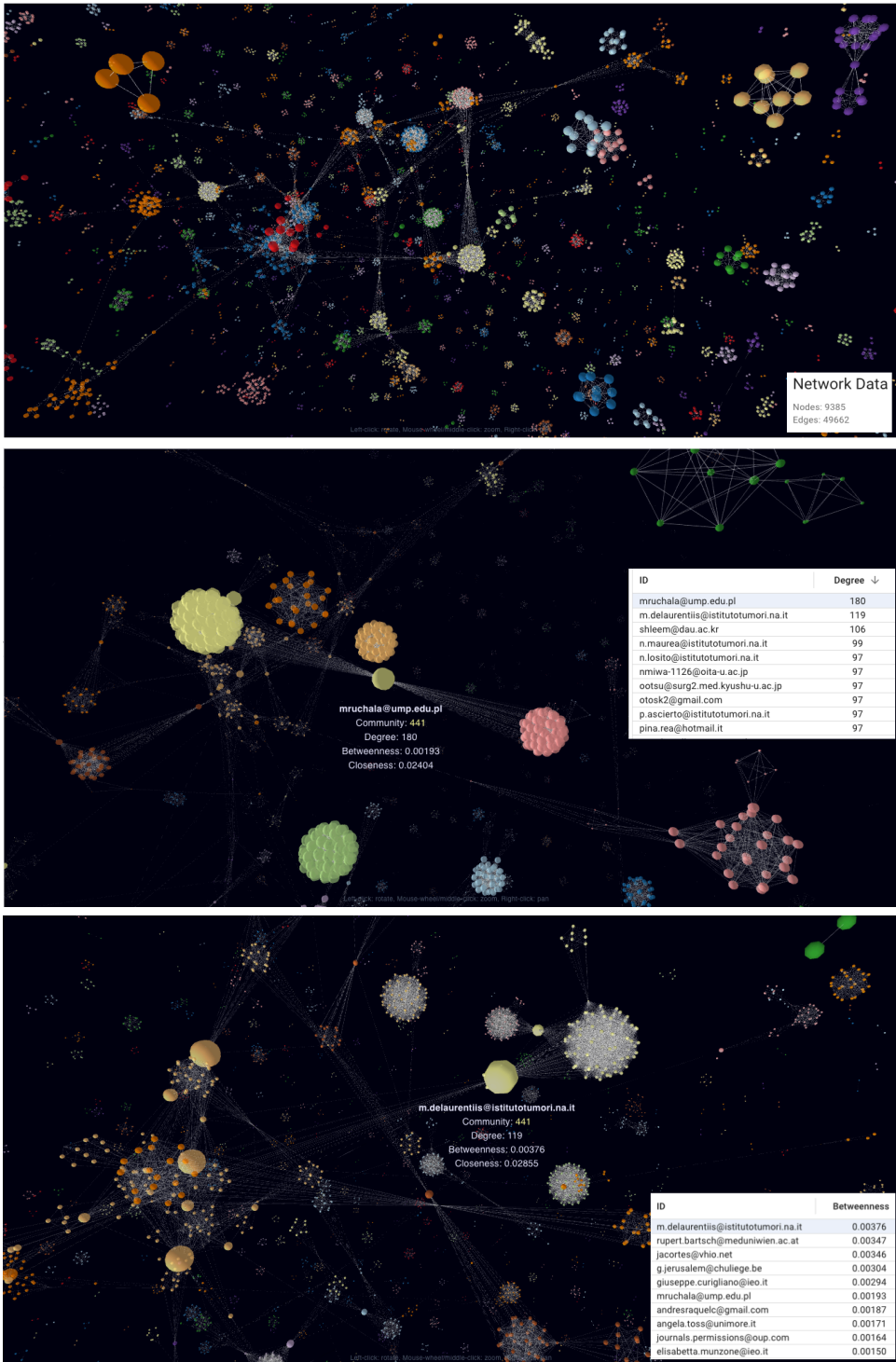
- <https://core.ac.uk/services/api>
- [37] A. Becerril Garcia, R. Lozano Espinosa, y J. M. Molina Espinosa, «Enfoque semántico para el descubrimiento de recursos sensible al contexto sobre contenidos académicos estructurados con OAI-PMH», *Comput. Sist.*, vol. 20, n.º 1, pp. 127-142, mar. 2016, doi: 10.13053/cys-20-1-2189.
 - [38] «arXiv API Basics - arXiv info». Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://info.arxiv.org/help/api/basics.html>
 - [39] S. U. MartinAlba, «Lista Alfabética de Bases de Datos de Libros y Revistas», División de Bibliotecas / Universidad del Valle / Cali, Colombia. Accedido: 7 de diciembre de 2024. [En línea]. Disponible en: <https://biblioteca.univalle.edu.co/servicios/busquedas/libros-y-revistas-en-formato-electronico-a-c>
 - [40] «Data providers search». Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://core.ac.uk/data-providers/>
 - [41] «Welcome to Python.org», Python.org. Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://www.python.org/>
 - [42] J. Joshi, «Chapter 9 - Python, a reliable programming language for chemoinformatics and bioinformatics», en *Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences*, N. Sharma, H. Ojha, P. K. Raghav, y R. k. Goyal, Eds., Academic Press, 2021, pp. 279-304. doi: 10.1016/B978-0-12-821748-1.00013-0.
 - [43] «Build software better, together», GitHub. Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://github.com>
 - [44] «PyPI · El Índice de paquetes de Python», PyPI. Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://pypi.org/>
 - [45] *arxiv: Python wrapper for the arXiv API*. Accedido: 14 de noviembre de 2024. [OS Independent]. Disponible en: <https://pypi.org/project/arxiv/>
 - [46] *oacore/apiv3-webinar*. CORE, 2024. Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://github.com/oacore/apiv3-webinar>
 - [47] *metapub: Pubmed / NCBI / eutils interaction library, handling the metadata of pubmed papers*. Accedido: 14 de noviembre de 2024. [En línea]. Disponible en: <https://pypi.org/project/metapub/>
 - [48] *Flask: A simple framework for building complex web applications*. [OS Independent]. Disponible en: <https://pypi.org/project/Flask/>
 - [49] «josanabr/RunCLI at arxiv», GitHub. Accedido: 15 de noviembre de 2024. [En línea]. Disponible en: <https://github.com/josanabr/RunCLI>
 - [50] «inv-adc/metapub/prototype at master · jandresh/inv-adc · GitHub». Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://github.com/jandresh/inv-adc/tree/master/metapub/prototype>
 - [51] «Capa gratuita de AWS | Cloud computing gratis |AWS», Amazon Web Services, Inc. Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://aws.amazon.com/es/free/>
 - [52] «Production-Grade Container Orchestration», Kubernetes. Accedido: 7 de diciembre de 2024. [En línea]. Disponible en: <https://kubernetes.io/>
 - [53] «Jenkins», Jenkins. Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://www.jenkins.io/>

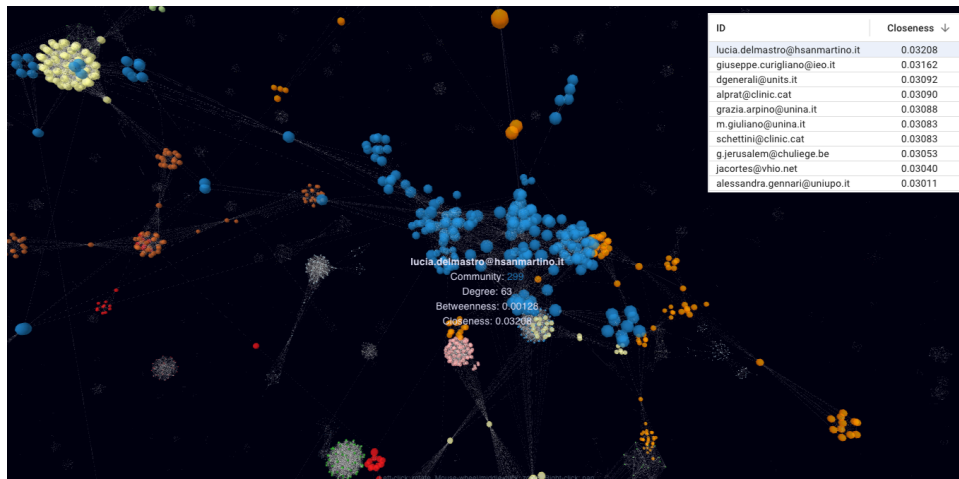
- [54] «Get started with GitLab CI/CD | GitLab». Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://docs.gitlab.com/ee/ci/>
- [55] Qwiklabs, «Cómo configurar Jenkins en Kubernetes Engine | Google Cloud Skills Boost», Qwiklabs. Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://www.cloudskillsboost.google/focuses/1776?parent=catalog>
- [56] *GoogleCloudPlatform/continuous-deployment-on-kubernetes*. Google Cloud Platform, 2024. Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://github.com/GoogleCloudPlatform/continuous-deployment-on-kubernetes>
- [57] «I-ADC-UnivalleArranque», Google Docs. Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://github.com/jandresh/inv-adc/blob/master/jenkins/I-ADC-UnivalleArranque.pdf>
- [58] «Jira | Issue & Project Tracking Software | Atlassian». Accedido: 7 de diciembre de 2024. [En línea]. Disponible en: https://www.atlassian.com/software/jira?campaign=18442480203&adgroup=140479881486&targetid=kwd-855725830&matchtype=e&network=g&device=c&device_model=&creative=687972959756&keyword=jira&placement=&target=&ds_eid=700000001558501&ds_e1=GOOGLE&gad_source=1&gclid=Cj0KCCQiAgdC6BhCgARIsAPWNWH13NI557A9dkQIc_huU4b7bly7Tf_TAMfH281vQHuhu7IpI39YSVacaAmgUEALw_wcB
- [59] «Visual Studio Code - Code Editing. Redefined». Accedido: 7 de diciembre de 2024. [En línea]. Disponible en: <https://code.visualstudio.com/>
- [60] «Git». Accedido: 7 de diciembre de 2024. [En línea]. Disponible en: <https://git-scm.com/>
- [61] «jandresh's Profile | Docker Hub». Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://hub.docker.com/u/jandresh>
- [62] «MongoDB: La plataforma de datos para desarrolladores», MongoDB. Accedido: 6 de diciembre de 2024. [En línea]. Disponible en: <https://www.mongodb.com/>
- [63] *networkx: Python package for creating and manipulating graphs and networks*. Accedido: 8 de diciembre de 2024. [OS Independent]. Disponible en: <https://networkx.org/>
- [64] «3D Force-Directed Graph», 3d-force-graph. Accedido: 8 de diciembre de 2024. [En línea]. Disponible en: <https://vasturiano.github.io/3d-force-graph/>
- [65] «inv-adc/gui/design/EntrevistaUXUI.pdf at master · jandresh/inv-adc · GitHub». Accedido: 8 de diciembre de 2024. [En línea]. Disponible en: <https://github.com/jandresh/inv-adc/blob/master/gui/design/EntrevistaUXUI.pdf>
- [66] «Figma», Figma. Accedido: 8 de diciembre de 2024. [En línea]. Disponible en: <https://www.figma.com/proto/QtXkc50t3vpvD5MtmeuYj5/adccali?node-id=627-126&scaling=min-zoom&page-id=27%3A2323&starting-point-node-id=202%3A53>
- [67] «Create React App». Accedido: 9 de diciembre de 2024. [En línea]. Disponible en: <https://create-react-app.dev/>
- [68] «JavaScript With Syntax For Types.» Accedido: 9 de diciembre de 2024. [En línea]. Disponible en: <https://www.typescriptlang.org/>
- [69] «Overview - Material UI». Accedido: 9 de diciembre de 2024. [En línea]. Disponible en: <https://mui.com/material-ui/getting-started/>
- [70] «inv-adc/db/app/patterns.txt at master · jandresh/inv-adc · GitHub». Accedido: 9 de diciembre de 2024. [En línea]. Disponible en: <https://github.com/jandresh/inv-adc/blob/master/db/app/patterns.txt>
- [71] *gensim: Python framework for fast Vector Space Modelling*. Accedido: 9 de diciembre de 2024.

2024. [OS Independent]. Disponible en: <https://radimrehurek.com/gensim/>

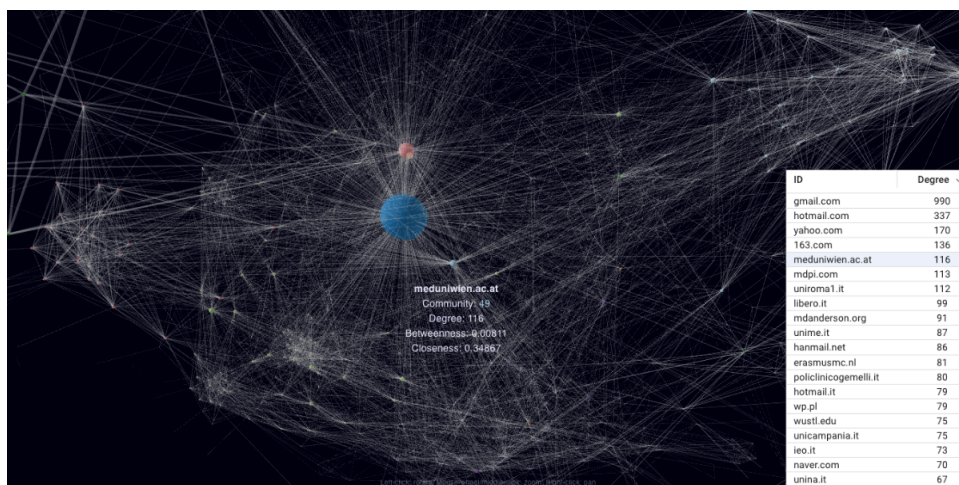
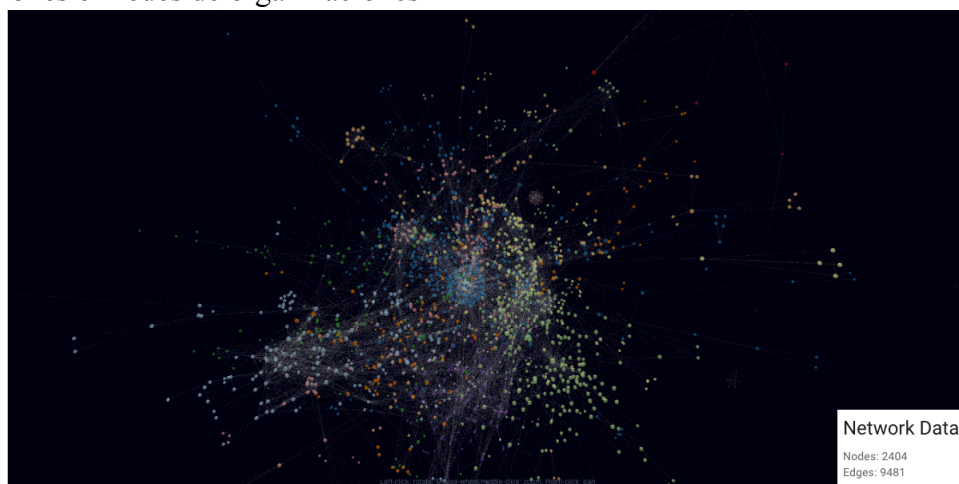
Anexo 1. Patrones de literatura científica relacionados al cáncer de mama encontrados por el prototipo.

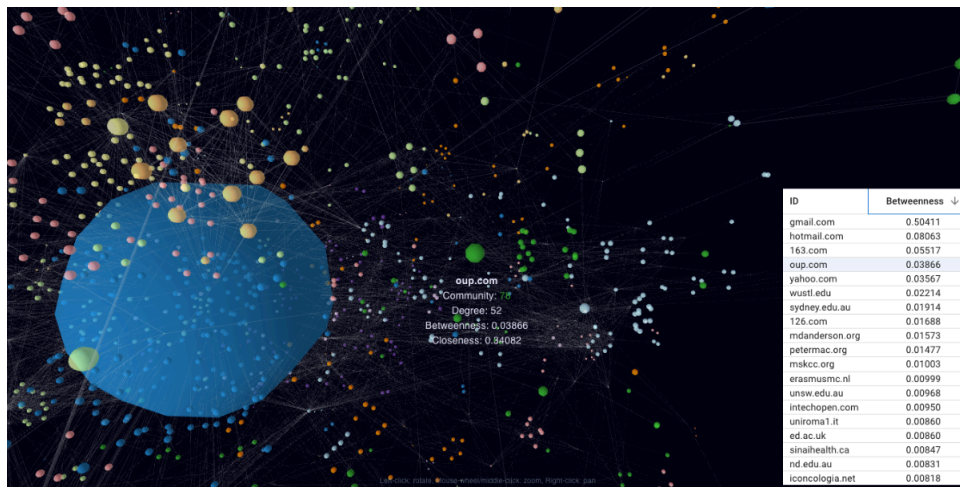
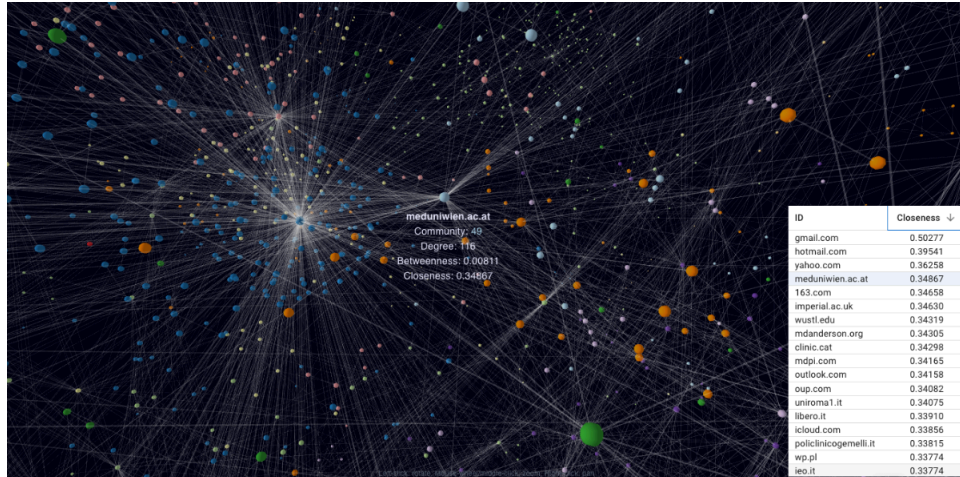
- Patrones en redes de coautoría



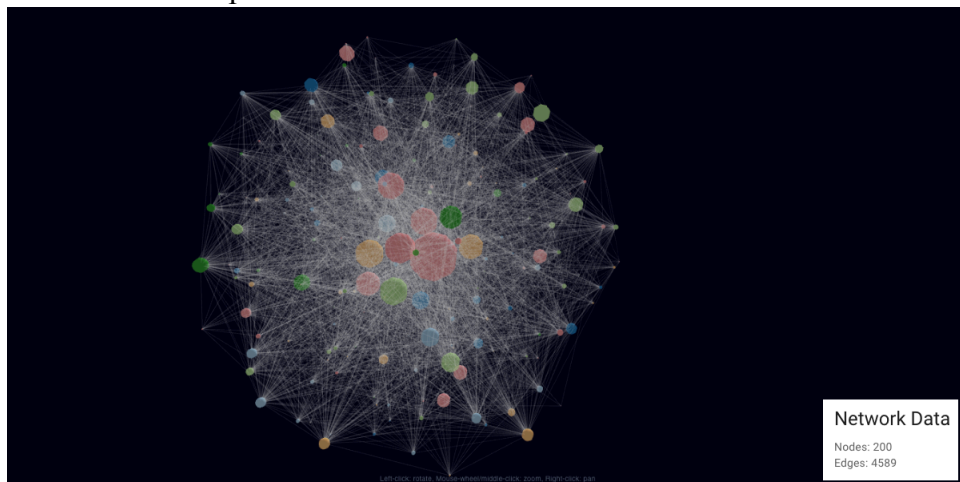


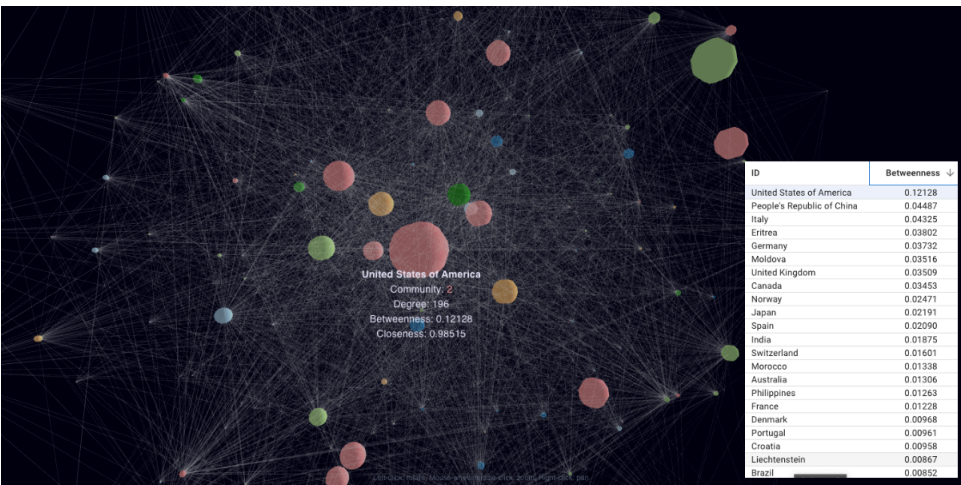
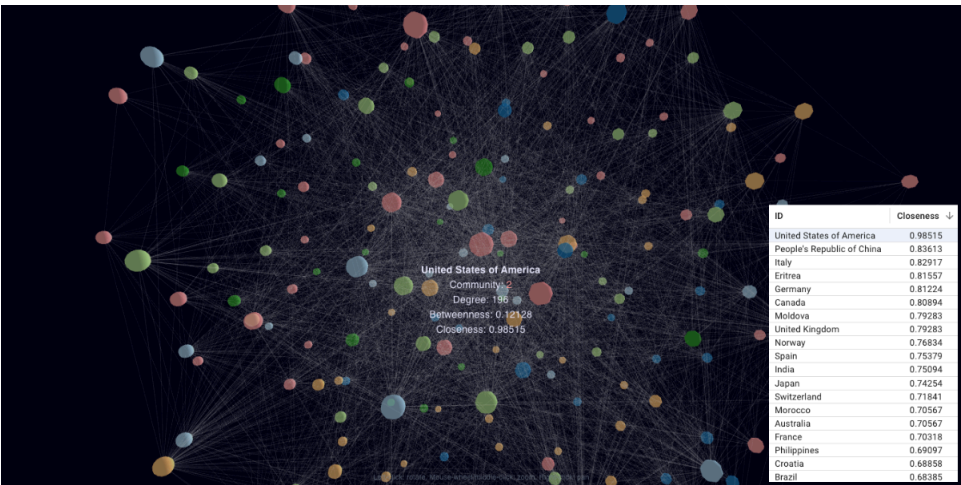
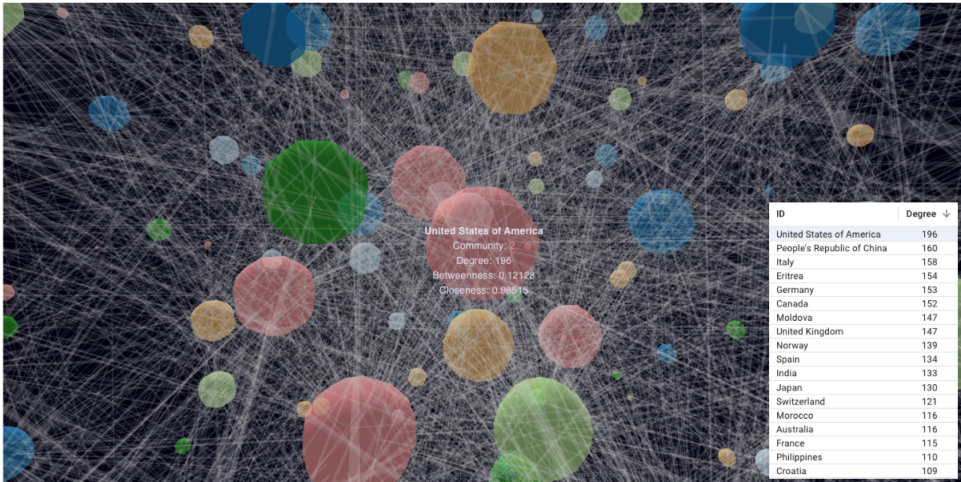
- Patrones en redes de organizaciones





- Patrones en redes de países





-

