ciencias aplicadas

MDPI

Artículo

Una metodología para el descubrimiento de conocimiento en lenguaje etiquetado y Gráficos heterogéneos

Víctor H. Ortega Guzmán, Luis Gutiérrez Preciado, Francisco Cervantes * y Mildreth Alcaraz-Mejía

Departamento de Electrónica, Sistemas e Informática, ITESO—Universidad Jesuita de Guadalajara, Tlaquepaque 45604, México; vortega@iteso.mx (VHO-G.); lgutierrez@iteso.mx (LG-P.); levereth@iteso.mx (MA-M.)
* Correspondencia: fcervantes@iteso.mx; Tel.: +52-33-3669-3517

Resumen: La minería de grafos se ha convertido en un campo de investigación importante con aplicaciones que abarcan múltiples ámbitos, como el marketing, el análisis de la corrupción, los negocios y la política. La exploración del conocimiento dentro de los grafos ha recibido considerable atención debido al crecimiento exponencial de los datos modelados en grafos y su potencial en aplicaciones donde las relaciones entre los datos son un componente crucial, pudiendo ser incluso más importantes que los propios datos. Sin embargo, el creciente uso de grafos para el almacenamiento y modelado de datos presenta desafíos únicos que han impulsado avances en algoritmos de minería de grafos, modelado y almacenamiento de datos, lenguajes de consulta para bases de datos de grafos y técnicas de visualización de datos. Si bien existen diversas metodologías para el análisis de datos, estas se centran principalmente en datos estructurados y podrían no ser óptimas pa En consecuencia, este trabajo presenta una nueva metodología específicamente diseñada para el descubrimiento de conocimiento en gráficos etiquetados y heterogéneos (KDG), y presenta tres estudios de caso que demuestran su aplicación exitosa para abordar diversos desafíos en diferentes dominios de aplicación.

Palabras clave: minería de grafos; descubrimiento de conocimiento; grafos etiquetados; grafos heterogéneos

Citación: Ortega-Guzmán, VH; Gutiérrez-Preciado, L.; Cervantes, F.; Alcaraz-Mejía, M. Una metodología para el descubrimiento de conocimiento en grafos etiquetados y heterogéneos. Appl. Sci. 2024, 14, 838. https://doi.org/10.3390/app14020838

Editor académico: Donghai Guan

Recibido: 11 de diciembre de 2023
Revisado: 14 de enero de 2024
Aceptado: 16 de enero de 2024
Publicado: 18 de enero de 2024

1. Introducción

El campo de la minería de grafos ha experimentado un auge en popularidad en los últimos años, impulsado principalmente por el crecimiento exponencial de datos que pueden representarse eficazmente como grafos y su amplia gama de aplicaciones. Esto ha impulsado el auge de bases de datos de grafos como Neo4J, AllegroGraph y OrientDB [1], que ofrecen plataformas robustas para modelar y almacenar datos como nodos y relaciones interconectadas. Aprovechando estas bases de datos, investigadores y profesionales pueden aplicar algoritmos avanzados de minería de grafos para realizar diversas tareas, como el análisis de comunidades, la identificación de centralidad, la búsqueda de rutas y la exploración de patrones estructurales. En diversos dominios de aplicación, como los sistemas de información empresarial [2], los sistemas de detección de delitos financieros [3], los sistemas de información de transporte [4] y los sistemas de recomendación [5], las bases de datos de grafos han ganado una amplia aceptaci Sin embargo, depender únicamente de bases de datos gráficas no es suficiente. Se requiere una metodología integral para definir las tareas y las etapas necesarias para realizar eficazmente el análisis de la información y extraer información valiosa.

Este artículo revisa los trabajos existentes sobre minería de grafos y presenta los marcos y metodologías más utilizados para el descubrimiento de conocimiento, como KDD [6], CRISP-DM [7] y SEMMA [8]. Describimos las tareas necesarias para obtener valor de grafos grandes, ausentes en las metodologías y marcos utilizados en la literatura. Finalmente, con base en KDD, CRISP-DM y las tareas necesarias para la minería de grafos, proponemos una nueva metodología para el descubrimiento de conocimiento en grafos.

Las contribuciones del trabajo son las siguientes:

- Una metodología nueva y específica denominada KDG (descubrimiento de conocimiento en gráficos) para guiar a los usuarios a encontrar información a partir de datos representados como gráficos.

• Tres casos de uso aplicando la metodología propuesta.

El resto del artículo se estructura de la siguiente manera. La Sección 2 ofrece una visión general de los marcos y metodologías empleados en la minería de datos, incluyendo los procesos, herramientas, técnicas de visualización y algoritmos de análisis de grafos que han surgido en las últimas dos décadas. La Sección 3 presenta los conceptos relacionados que se emplean a lo largo del artículo. La Sección 4 presenta nuestra metodología para el descubrimiento de conocimiento en grafos, describiendo sus componentes y tareas esenciales. Posteriormente, en la Sección 5, mostramos la aplicación de esta metodología a través de tres casos de uso diferentes. Finalmente, en la Sección 6, analizamos nuestros hallazgos y las líneas de investigación futuras.

2. Trabajos relacionados

La minería de grafos se refiere al conjunto de herramientas y algoritmos utilizados para modelar grafos que se ajustan a patrones del mundo real, analizar sus propiedades y predecir cómo las estructuras y propiedades de un grafo determinado podrían afectar ciertas aplicaciones [9]. Revisamos trabajos en el campo de la minería de grafos de las últimas dos décadas, desde marcos y metodologías para el desarrollo de proyectos hasta herramientas de modelado y algoritmos para el análisis de grafos. La siguiente sección ofrece una visión general de las metodologías y los marcos de minería de datos comúnmente empleados para el descubrimiento de conocimiento en grafos.

2.1. Marcos y metodologías

Fayyad et al. [6] definieron el descubrimiento de conocimiento en bases de datos (KDD) como un proceso que identifica patrones útiles, valiosos y comprensibles en los datos. Abordaron cómo escalar los algoritmos para que funcionen correctamente en conjuntos de datos masivos y cómo visualizarlos e interpretarlos. Además, incluyeron cómo modelar y respaldar las interacciones hombre-máquina. En el marco del KDD, la minería de datos es un paso singular dentro del proceso, donde los datos preprocesados correctamente se transforman en patrones que generan conocimiento valioso. Sin embargo, es importante destacar que varias metodologías desarrolladas después del KDD, como CRISP-DM [7] y SEMMA [8], utilizan la minería de datos como sinónimo de KDD.

CRISP-DM es ampliamente reconocida como la metodología para la mayoría de las propuestas de modelos de procesos de minería de datos a medida que la ciencia de datos ha evolucionado. Esta metodología replantea los pasos de la propuesta original de KDD: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Por otro lado, la metodología SEMMA fue propuesta por SAS [8] y significa Muestrear, Explorar, Modificar, Modelar y Evaluar. Esta metodología se centra principalmente en los aspectos técnicos del desarrollo de productos de minería de datos y está estrechamente vinculada al minero de datos de SAS. Una diferencia significativa entre SEMMA y las metodologías anteriores es el requisito previo de conocer y comprender los requisitos del negocio y las bases de datos.

CRISP-DM, SEMMA y otras metodologías relacionadas, como ASUM-DM [10], CASP- DM [11], FMDS [12] y TDSP [13], se basan en una perspectiva orientada a objetivos, que aborda procesos, tareas y roles. Estas metodologías se desarrollaron principalmente para procesos de minería de datos con objetivos de negocio relativamente claros. Además, trabajan principalmente con datos estructurados, que pueden provenir de archivos planos u obtenerse mediante consultas a bases de datos relacionales. Sin embargo, actualmente, en el contexto del big data y la ciencia de datos, la gran diversidad de fuentes de información, la cantidad y la heterogeneidad de los datos, así como la complejidad de los problemas, representan un desafío para las metodologías clásicas.

En los últimos años, han surgido nuevas propuestas para cerrar la brecha entre las metodologías clásicas y la complejidad de los problemas actuales. Martínez-Plumed et al. [14] proponen un modelo de trayectoria de datos (DST) y categorizan los proyectos de ciencia de datos en dirigidos a objetivos, exploratorios y de gestión de datos. Su modelo extiende CRISP-DM, incluyendo seis actividades exploratorias centradas en el objetivo, la fuente de datos, el valor de los datos, los resultados, la narrativa y el producto. Aunque DST es una propuesta que puede reducir la brecha entre las metodologías de minería de datos y los proyectos de ciencia de datos, la propuesta se limita a proponer actividades exploratorias para datos tabulares, que describe en sus casos de uso. Studer et al. [15] propusieron un modelo de proceso para desarrollar aplicaciones de aprendizaje automático.

En cuanto a los marcos, han surgido varias propuestas centradas en diferentes fases del descubrimiento de conocimiento en grafos, por ejemplo, comprensión del negocio [16], modelado de datos [17], minería de grafos [18-20], visualización de grafos [21,22] y evaluación [23,24].

Algunas propuestas se centran en cómo modelar grafos realistas que coincidan con los patrones encontrados en grafos del mundo real. Shrivastava et al. [25] propusieron un marco de minería de grafos que captura entidades y relaciones entre ellas de diferentes fuentes de datos. Su marco ofrece un enfoque integral con cinco módulos que abarcan el preprocesamiento de grafos , la base de datos de grafos, la extracción de subestructuras densas, el descubrimiento frecuente de subestructuras y la visualización de grafos. Sin embargo, carece de una adecuada consideración de los objetivos de negocio y del proceso de evaluación dentro del alcance.

Nasiri et al. [26] presentaron un marco con dos componentes de modelado: modelado de contexto y modelado de alternativas de diseño analítico. El primero justifica la necesidad del análisis predictivo en el contexto organizacional. Este componente extiende el modelo de inteligencia de negocios (BIM) [16], y el segundo identifica los requisitos para adaptar el marco presentado [27], pero no aborda adecuadamente la valiosa información que se puede extraer de los datos de grafos. Por lo tanto, la integración de bases de datos de grafos en el marco podría mejorar su eficacia para descubrir patrones y relaciones ocultos en los datos.

El modelado de datos como un grafo y su análisis encontraron una buena área de aplicación en las redes sociales. Autores como Kumar [17] y Schroeder [28] han desarrollado marcos que se centran en redes sociales como Twitter. Kumar et al. [17] propusieron un marco para analizar redes sociales. Utilizaron el procesamiento del lenguaje natural para aislar las características de los nodos y los metadatos para los bordes. Las fases principales del marco son la adquisición de datos, el preprocesamiento, la creación de grafos multiatribuidos, la transformación del grafo multiatribuido en un grafo de similitud y la agrupación en clústeres. Sin embargo, el alcance se limita a las redes sociales. Schroeder et al. [28] presentaron un marco para recopilar estructuras gráficas de redes de seguidores, publicaciones y perfiles en la red social Twitter. El objetivo es detectar bots sociales mediante el análisis de datos estructurados en grafos.

2.2. Herramientas y algoritmos para el análisis de gráficos

Además de las metodologías y los marcos propuestos, los trabajos relevantes se centran en algoritmos de minería de gráficos con diferentes objetivos, como resumir y visualizar gráficos, realizar operaciones de agregación en los datos, buscar subestructuras en el gráfico, construcción automática de gráficos y desarrollar propuestas para optimizar las operaciones realizadas en los gráficos.

Algunos trabajos en la literatura se centran en el análisis de subgrafos y la transformación de la topología del grafo. Qiao [29] propuso un algoritmo de minería de subgrafos frecuentes en paralelo en un único grafo grande utilizando Spark. La propuesta emplea una estrategia de búsqueda heurística, balanceo de carga, poda de investigación y poda descendente en el soporte. Qiao también propuso un marco de dos fases. En la primera fase, la extensión de subgrafos paralelos utiliza una estrategia que genera todos los subgrafos en paralelo. En la segunda fase, utiliza un método de evaluación de soporte similar para encontrar isomorfismos de subgrafos. Zhang [30] propuso un marco para la clasificación ponderada basada en metagrafos de redes de información heterogéneas. El núcleo es un algoritmo que clasifica iterativamente los objetos en redes de información heterogéneas para capturar la información oculta en la semántica y la estructura del grafo. Lee et al. [31] propusieron un método para extraer subgrafos frecuentes mientras se mantiene la información semántica y se considera la escalabilidad en grafos a gran escala. La información semántica generada incluye conteos de frecuencia para tareas como la predicción de calificación o la recomendación.

Pienta et al. propusieron Vigor [18], una herramienta interactiva para la exploración de grafos que incluye la construcción de sentido tanto ascendente como descendente para los analistas, facilitando la revisión de subgrafos mediante un proceso de resumen. Esta propuesta proporciona valiosas contribuciones al campo de la exploración de grafos. Sin embargo, existe la oportunidad de explorar más a fondo los requisitos del negocio. Además, el trabajo de Dunne y Shneiderman [32] se centró en mejorar la visualización de grafos mediante la simplificación de motivos, que reemplaza los comunes

patrones de nodos y enlaces con glifos compactos y significativos, mejorando aún más la exploración de gráficos.

Yin y Hong investigaron el problema de agregación en diferentes tipos de nodos y relaciones proponiendo una función basada en la entropía del grafo para medir la similitud de los nodos. También demostraron que el problema de agregación basado en las funciones es NP-hard y propusieron un algoritmo heurístico para realizar la agregación, incluyendo aspectos informativos y estructurales. A pesar del alcance limitado de la propuesta del algoritmo, su trabajo es significativo al abordar la operación fundamental de realizar agregaciones en datos almacenados en nodos del grafo.

La búsqueda, el procesamiento y la visualización de subgrafos pueden ser tareas que requieren muchos recursos computacionales. Por esta razón, autores como Bok et al. [19] han desarrollado propuestas para mejorar la accesibilidad de los subgrafos. Proponen una estrategia de caché de dos niveles que almacena en caché los subgrafos a los que es probable acceder según su patrón de uso. Su propuesta evita el almacenamiento en caché de subgrafos de bajo uso y la sustitución frecuente de subgrafos en memoria. Una de las principales aplicaciones de su propuesta es permitir el procesamiento y el análisis de consultas de grafos grandes en entornos computacionales con memoria reducida.

En la última década, han surgido diversas propuestas que se centran en las diferentes fases del descubrimiento de conocimiento en bases de datos de grafos. Estas propuestas abarcan áreas como la comprensión del negocio [16], el modelado de datos [17], la minería de grafos [18-20], la visualización de grafos [21,22] y la evaluación [23,24].

Las metodologías, marcos, herramientas y algoritmos mencionados en las Secciones 2.1 y 2.2 han demostrado su utilidad, pero aún se necesita una metodología integral que incorpore los avances y algoritmos existentes. Dicha metodología debería proporcionar una guía detallada y específica para el análisis de datos relacionados mediante grafos, abordando las diversas tareas implicadas en el descubrimiento de conocimiento. Al incorporar estos elementos, los usuarios contarán con instrucciones claras para maximizar el potencial de los grafos en sus proyectos de análisis de datos.

Adoptar una metodología integral que abarque tareas críticas es fundamental para aprovechar el conocimiento integrado en las relaciones de datos. Estas tareas implican modificar un grafo mediante la aplicación de reglas u operaciones a sus nodos y aristas, e incorporar nuevos atributos estructurales que antes no existían en los nodos ni en las relaciones. Además, la representación visual de grafos y redes desempeña un papel fundamental en este proceso. La creación de representaciones visuales de grafos permite a los usuarios identificar patrones, anomalías y tendencias en conjuntos de datos complejos, lo que permite una comprensión más profunda de la información

La metodología que se presenta proporciona una guía detallada y específica para el análisis de datos relacionados mediante grafos. Abarca una amplia gama de operaciones y tareas , ofreciendo una completa gama de opciones. Además, presenta un orden coherente de etapas y pasos, lo que sirve como una valiosa referencia para realizar una exploración y un análisis exhaustivos de forma estructurada. La metodología enfatiza el modelado y el análisis efectivos de la información mediante grafos, a la vez que garantiza la flexibilidad en la elección de herramientas y algoritmos. Al seguir esta guía, los usuarios recibirán instrucciones claras para maximizar el potencial de los grafos en sus proyectos de análisis de datos.

3. Conceptos

relacionados En esta sección, introducimos las nociones principales de gráficos necesarias para comprender la Trabajo propuesto. Para más información, véase [33,34].

3.1. Grafos.

Un grafo es una estructura matemática compuesta por elementos llamados vértices o nodos, conectados por enlaces llamados aristas. Un vértice representa una entidad que puede ser cualquier objeto, como personas, productos o ciudades. Una arista representa la relación entre dos entidades. Algunas aplicaciones comunes de los gráficos incluyen redes sociales, donde los vértices son personas y los bordes representan amistad u otras relaciones entre personas, y transporte.

redes de coordenadas, donde los vértices son lugares (por ejemplo, paradas de autobús y aeropuertos) y los bordes son los caminos entre esos lugares.

Según la naturaleza del problema, podemos modelar los datos utilizando diferentes tipos de grafos. Si las relaciones entre vértices son bidireccionales, el grafo se conoce como grafo no dirigido; por lo tanto, la arista se representa como una línea recta. De lo contrario, el grafo se conoce como grafo dirigido y la arista se representa como una flecha. En este trabajo, nos centramos en grafos donde los nodos y las aristas pueden tener uno o más atributos asociados, denominados grafos etiquetados. Además, podemos utilizar nodos o aristas de diferentes entidades y atributos, conocidos como grafos heterogéneos, además de grafos homogéneos, donde los nodos y las aristas son del mismo tipo y atributos.

3.2 Estructura de los

gráficos Para analizar los gráficos, es útil considerar la información estructural de los mismos.

gráficos, por lo que es importante introducir algunas nociones básicas como las que se indican a continuación.

La topología de gráficos es la disposición y la forma en que se conectan los nodos.
Mediante el análisis topológico, pudimos encontrar algunas características importantes para el análisis de grafos, como el grado del nodo, que se refiere al número de relaciones de un nodo. El grado de entrada se refiere a las relaciones que entran en un nodo, y el grado de salida se refiere a las relaciones que salen de él. En grafos no dirigidos, el grado de entrada, el grado de salida y el grado...
tienen el mismo valor

La centralidad de los nodos se utiliza para identificar los nodos esenciales o centrales de una red. Una comunidad se utiliza para describir un clúster o un conjunto de nodos altamente conectados dentro de una red. La similitud es una forma de medir o comparar la similitud entre dos nodos diferentes en una red, considerando sus relaciones, información estructural o atributos. La ruta es una secuencia de nodos conectados a lo largo de la red.

Otro concepto importante para comprender el trabajo propuesto en este trabajo es el subgrafo, una porción del grafo original que consiste en un subconjunto de nodos y aristas seleccionados, lo cual reduce el tamaño del modelo y, por lo tanto, la complejidad del análisis. Un subgrafo puede ser el resultado de aplicar un filtro. Dos operaciones útiles para analizar la estructura de un grafo son el filtrado y el resumen. Un filtro puede ser una herramienta útil para seleccionar o mostrar un subconjunto de nodos o relaciones que cumplen ciertas condiciones o criterios. El uso de filtros puede reducir el tamaño y la complejidad del grafo y facilitar el análisis y la visualización de información importante. Los filtros se pueden aplicar en función de atributos, grado, comunidad, peso o distancia, entre otros. Los atributos estructurales de un grafo son características que describen la topología. Algunos ejemplos de atributos estructurales son el grado, la comunidad y la ruta, entre otros. El resumen de grafos es un proceso que permite una representación más compacta del grafo original para facilitar el análisis y la comprensión de grafos complejos.

Se pueden utilizar varios algoritmos para resumir gráficos, agrupando nodos o relaciones según atributos específicos o tipos de entidades para presentar una versión más concisa del gráfico completo.

3.3. Visualización de grafos

El modelado de grafos crea una representación abstracta y estructurada de elementos representados por nodos y relaciones representadas por aristas para analizar sistemas complejos en ciencias de la computación, ciencias sociales, ingeniería y más. El modelo de grafo puede ser mostrado usando diferentes diseños. Un diseño se refiere a la forma y posición en la que los nodos y relaciones de un grafo son representados en un plano de visualización. Varios algoritmos de visualización apuntan a minimizar el cruce o superposición de relaciones y nodos. En el contexto del análisis y visualización, operaciones de desglose y acumulación son dos técnicas usadas para explorar datos jerárquicamente y obtener diferentes niveles de detalle o resumen. El desglose implica desglosar datos de un nivel más alto o más general a un nivel más bajo o más detallado en una jerarquía de datos. La acumulación implica resumir datos de un nivel más bajo o más detallado a un nivel más alto o más general.

Algunos problemas, por su naturaleza, requieren considerar la evolución del grafo a lo largo del tiempo. En este caso, podemos analizar el grafo mediante una línea de tiempo, que es una representación gráfica y secuencial de eventos a lo largo del tiempo. Muestra cómo se posicionan los nodos y las relaciones en momentos específicos y cómo evolucionan.

4. Propuesta metodológica.

Analizar grandes cantidades de información estrechamente relacionada es complejo, y en ocasiones es necesario visualizarla de forma resumida. Algunos ejemplos son las recomendaciones de productos, la detección de fraudes, el análisis de relaciones como amistades, la colaboración, la coautoría, las líneas de autoridad, la detección de influenciadores, la logística en la cadena de suministro de productos, la lista de materiales, el enrutamiento en una red informática o la búsqueda de carreteras.

Proponemos una metodología denominada KDG para el descubrimiento de conocimiento en grafos como referencia para extraer, transformar, cargar, procesar, modelar, visualizar y analizar información altamente conectada que pueda ser representada por grafos etiquetados y heterogéneos; aplicando algoritmos de minería para descubrir nuevas características estructurales útiles para mejorar el análisis que soporta la toma de decisiones.

Nuestra metodología proporciona una visión global de las etapas que pueden requerirse en el proceso de descubrimiento de conocimiento modelado por grafos, que considera todas las tareas posibles para explorar la información, identificar patrones, realizar recomendaciones mediante la aplicación de consultas, utilizar herramientas de visualización y resolver preguntas, entre otras.

La Figura 1 ilustra la metodología de nuestro enfoque, que consta de seis etapas: (1) Comprensión del Proceso Analítico, (2) Construcción de Gráficos, (3) Minería de Gráficos, (4) Transformación de Gráficos, (5) Visualización de Gráficos y (6) Evaluación. Cada etapa contiene varias tareas. El usuario puede optar por completar todas las tareas de cada etapa o, como mínimo, centrarse en las resaltadas explícitamente en negrita en la figura, ya que representan las tareas esenciales necesarias para la metodología.

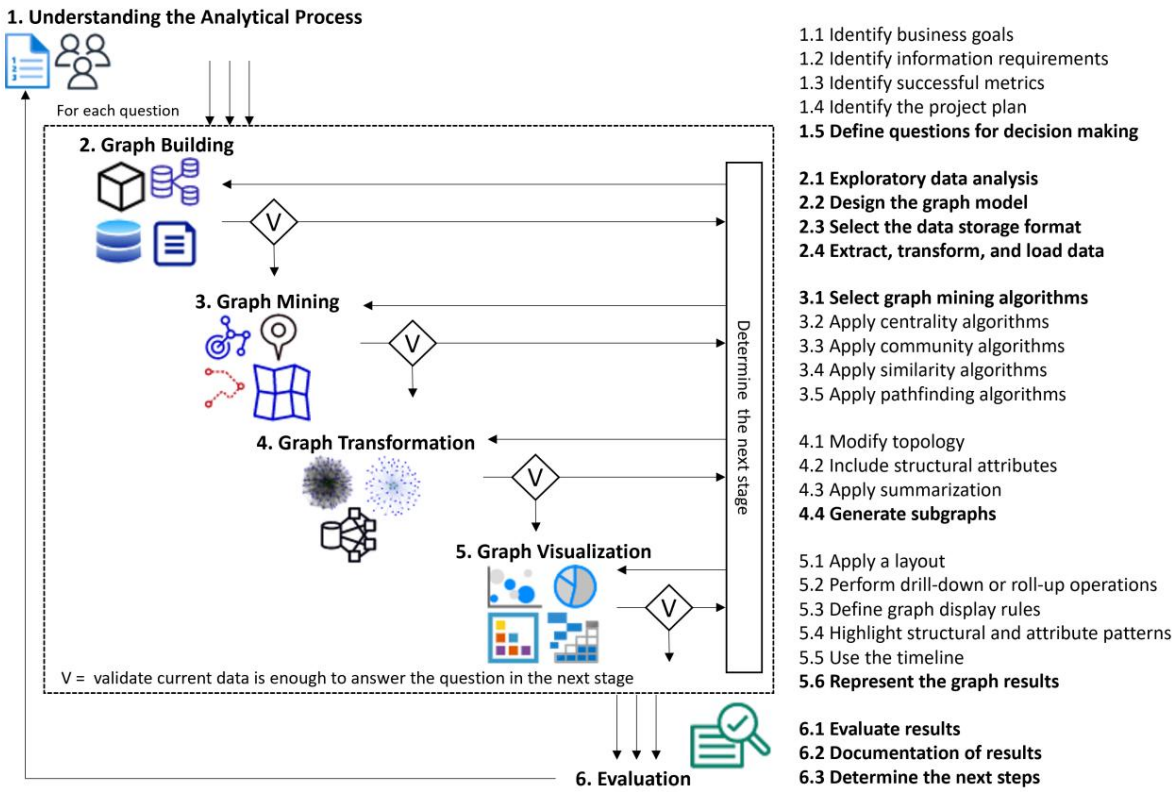


Figura 1. La metodología para el descubrimiento de conocimiento en gráficos etiquetados y heterogéneos (KDG).

4.1. Etapa 1: Comprensión del proceso analítico. El

objetivo de esta etapa es comprender exhaustivamente todos los pasos del análisis de datos antes de emprender cualquier acción, análisis o pasar a etapas posteriores de la metodología. Además, es fundamental establecer un hilo conductor coherente que guíe al usuario de forma coherente durante toda la aplicación de la metodología.

Esta etapa consta de cinco tareas: (1.1) Identificar los objetivos del negocio, que consiste en alinear los objetivos del análisis con la misión y visión de la empresa. Estos objetivos suelen seguir el marco de objetivos SMART, lo que significa que son específicos, medibles, alcanzables, relevantes y limitados en el tiempo; se puede encontrar más información en [35]. (1.2) Identificar los requisitos de información que sean concretos y factibles, describiendo las funcionalidades necesarias. Algunos ejemplos son: usar una tecnología, algoritmo o herramienta específica, confidencialidad de los datos y usar fuentes de datos específicas. Si el lector desea profundizar en la información requerida para los requisitos, puede consultar [36]. (1.3) Identificar métricas exitosas para obtener datos objetivos que ayuden a los usuarios a descubrir áreas de mejora y garantizar que el progreso del proyecto esté de acuerdo con los objetivos establecidos. Pueden ser medidas cuantitativas o cualitativas utilizadas para evaluar y medir varios aspectos de un proyecto, su progreso y su rendimiento. Algunos ejemplos de métricas son el costo, el tiempo y la calidad. Si el lector desea profundizar en la información requerida para las métricas, puede consultar [37,38]. (1.4) Identificar el plan del proyecto es un documento guía para todas las partes interesadas, proporcionando una dirección clara. Define el alcance del proyecto, la duración, las actividades requeridas y los entregables acordados. Los componentes esenciales de un plan de proyecto incluyen la visión general del negocio, el alcance del proyecto, los objetivos a alcanzar, los miembros del equipo involucrados, los roles y responsabilidades asignados, los entregables acordados, el cronograma de actividades, el presupuesto asociado y la aprobación por escrito del patrocinador del proyecto. Si el lector desea profundizar en la información requerida para el plan del proyecto, puede consultar [39]. (1.5) Definir preguntas para la toma de decisiones. Su objetivo principal es facilitar la recopilación de información relevante, evaluar las opciones disponibles y considerar los factores que influyen en la toma de decisiones bien informada y efectiva. Estas preguntas son aplicables en contextos personales y profesionales, abordando dilemas específicos y contribuyendo a resolver problemas altamente significativos. Se caracterizan por su naturaleza reflexiva y su enfoque preciso, centrándose en objetivos, alternativas, consecuencias y valores personales, lo que a menudo requiere una profunda reflexión. Si el lector desea profundizar en la forma adecuada de definir preguntas para la toma de decisiones, puede consultar [40,41]. Las preguntas particularmente útiles para el análisis deben considerar información altamente conectada y ser lo suficientem Algunos ejemplos de preguntas son: ¿Cómo se relacionan los datos entre sí? ¿Existen rutas entre elementos específicos? ¿Cuáles son los grupos de elementos más relacionados o similares? ¿Cuáles son los elementos con más conexiones en la red? ¿Cuáles son las posibles nuevas conexiones entre elementos según patrones de datos ocultos? ¿Qué tan similares son los elementos en cuanto a sus conexiones? Es importante destacar que estas preguntas son el resultado de esta etapa, y las etapas 2 a 5 se realizan para cada pregunta.

4.2. Etapa 2: Construcción

del grafo. El objetivo de esta etapa es crear un modelo gráfico que considere las preguntas formuladas en la etapa anterior y validar que los datos sean suficientes para responderlas. Además, es fundamental implementar este modelo gráfico en un soporte de almacenamiento de datos adecuado.

Esta etapa consta de cuatro tareas: (2.1) Análisis exploratorio de datos, que consiste en analizar la información disponible mediante la aplicación de una o más de las siguientes opciones: estadística, visualización, agregación de datos e identificación de datos faltantes o corruptos. Esta exploración y comprensión de los datos se realiza en las primeras etapas del proceso, precediendo a un análisis más profundo. En el contexto de la construcción de grafos, el objetivo de esta actividad es identificar las entidades, propiedades y sus relaciones útiles para responder a cada pregunta definida en la etapa anterior. Para orientación en el análisis exploratorio de datos (AED),

el lector puede referirse a [42,43], y para una revisión, un estudio comparativo de cuatro métodos empleados en EDA, pueden referirse a [44]. (2.2) El diseño del modelo de grafos define qué datos serán representados como nodos, establece las relaciones entre ellos, y define los atributos asociados con cada elemento. Es posible diseñar grafos homogéneos o heterogéneos, así como grafos etiquetados o no etiquetados y dirigidos o indirectos. Para orientación en el proceso del modelo, el lector puede referirse a [34,45,46]. (2.3) Selección de formatos de almacenamiento de datos, que implica elegir la tecnología para almacenarlos en bases de datos de grafos, tales como relacionales, columnas, clave-valor, documentos, grafos, o cualquier otro tipo de formato de archivo. Para aprovechar al máximo el potencial analítico del análisis de grafos, se recomienda optar por una base de datos de grafos pero no es obligatorio. El lector puede referirse a [47] para orientación en este proceso de selección. Algunos ejemplos son NEO4J, TigerGraph, archivos JSON, MongoDB y Apache Cassandra. Finalmente, (2.4)
Extraer, transformar y cargar datos implica la adquisición de datos de diversas fuentes, como bases de datos, hojas de cálculo, archivos planos o servicios web. Los datos se recopilan de diversos orígenes y se almacenan en un área de almacenamiento para su posterior procesamiento. Durante la fase de transformación, los datos se depuran, se formatean y se reestructuran para adecuarse a los requisitos del grafo. Finalmente, los datos refinados y transformados se cargan en el medio de almacenamiento. El lector puede consultar [48–50] para obtener orientación sobre esta tarea. Entre las tecnologías que admiten las tareas de extracción, transformación y carga se incluyen Kettle Pentaho, AWS Glue, Knime y Apache Spark.

El resultado de esta etapa es la implementación exitosa del gráfico en el formato de almacenamiento de datos seleccionado.

4.3. Etapa 3: Minería de

grafos. La minería de grafos es una técnica potente que permite analizar diversas propiedades de grafos, predecir sus estructuras y relaciones, y modelar patrones presentes en grafos reales [9]. Esta etapa se centra en la aplicación de algoritmos de centralidad, similitud y comunidad, así como de búsqueda de rutas, basándose en las preguntas específicas a responder. En ciertos casos, puede ser necesario aplicar múltiples algoritmos para comprender y analizar completamente el grafo .

Esta etapa consta de cinco tareas: (3.1) Seleccionar algoritmos de minería de grafos, que pueden ayudar a los usuarios a analizar y extraer información de estructuras de grafos, como redes sociales, redes de corrupción y sistemas de transporte. Dependiendo del contexto específico, los usuarios pueden aplicar algoritmos de centralidad para identificar los nodos más influyentes, emplear algoritmos de comunidad para comprender mejor las posibles agrupaciones de nodos, utilizar algoritmos de similitud para evaluar las similitudes de los nodos en función de sus relaciones o utilizar algoritmos de búsqueda de rutas para determinar las rutas más cortas o ventajosas entre dos o más nodos. Por ejemplo, podríamos utilizar un algoritmo de comunidad para encontrar productos similares que se puedan recomendar como paquete, emplear un algoritmo de búsqueda de rutas para determinar la ruta más corta entre dos puntos en diferentes ciudades, implementar un algoritmo de centralidad para localizar a un amigo de Facebook con el mayor número de "me gusta" en sus publicaciones, o utilizar un algoritmo de similitud para descubrir proveedores alternativos para productos que ya están en stock en una tienda. El lector puede consultar [51] para obtener orientación en esta tarea. (3.2) La aplicación de algoritmos de centralidad desempeña un papel esencial en el análisis de redes, ya que permite la evaluación y clasificación de nodos según su influencia e importancia. Estos algoritmos son válidos para la identificación de nodos influyentes, la detección de líderes de opinión y el análisis de la estructura de la red. En esencia, proporcionan una medida de la proximidad o distancia de un nodo a otros nodos del grafo, lo cual es esencial para comprender la dinámica y el
Algunos ejemplos son la centralidad de grado, la centralidad de intermediación, la centralidad de cercanía y el PageRank. El lector puede consultar [52] para obtener orientación en esta tarea. (3.3) Aplicar algoritmos de comunidad para delinear clústeres o comunidades de nodos dentro de un grafo. Estos clústeres comprenden nodos que presentan conexiones más fuertes entre sí que con nodos externos a sus respectivas agrupaciones. Los algoritmos de comunidad buscan patrones en la conectividad de los nodos para identificar estos grupos y pueden ayudar a comprender cómo se organizan los nodos en un grafo. Dichos algoritmos encuentran aplicaciones pertinentes para identificar comunidades dentro de las redes soci

redes o detectar cohortes de usuarios que muestran intereses compartidos en plataformas digitales. Algunos ejemplos son Louvain, propagación de etiquetas, componentes débilmente conectados, componentes fuertemente conectados, conteo de triángulos y coloración K-1. El lector puede consultar [53] para obtener orientación en esta tarea. (3.4) Los algoritmos de similitud de aplicación cuantifican y contrastan la similitud estructural entre nodos, grafos o subgrafos distintos. Estos algoritmos permiten evaluar hasta qué punto dos grafos se parecen entre sí en cuanto a su estructura, patrones e interconexiones entre nodos. Algunos ejemplos son la similitud de nodos, el índice de Jaccard y los K vecinos más cercanos. El lector puede consultar [54] para obtener orientación en esta tarea. (3.5) Los algoritmos de búsqueda de rutas de aplicación determinan la ruta más corta u óptima entre dos puntos específicos en un grafo o mapa. Estos algoritmos encuentran aplicación en diversos dominios, incluidos los sistemas de navegación GPS, donde permiten una planificación de rutas eficiente, los videojuegos para el movimiento de personajes y las redes de transporte para optimizar las rutas de entrega de paquetes. Estos algoritmos emplean técnicas de búsqueda y heurísticas para identificar la ruta más favorable, garantizando así una solución eficiente. Algunos ejemplos son Dijkstra, la ruta más corta, la búsqueda en amplitud, la búsqueda en profundidad y el recorrido aleatorio. El lector puede consultar [55] para obtener orientación en esta tarea.

Los resultados de la etapa de Minería de Grafos son útiles para un análisis posterior. Uno de los resultados clave es la generación de información para crear nuevos atributos estructurales que brindan información adicional sobre las propiedades del grafo. Estos atributos pueden utilizarse para el resumen del grafo, lo que ayuda a reducir su complejidad mediante la creación de abstracciones de alto nivel y facilita una comprensión más completa de los patrones y estructuras subyacentes. Además, los nuevos atributos derivados pueden utilizarse para abordar algunas de las preguntas iniciales que impulsaron el análisis, proporcionando información y respuestas valiosas .

4.4. Etapa 4: Transformación del gráfico

El objetivo de esta etapa es incorporar nuevos atributos, nodos o aristas al grafo original, seleccionando subgrafos o ejecutando consultas para abordar las preguntas formuladas en la fase inicial.

Esta etapa consta de cuatro tareas guiadas por las preguntas definidas en la etapa 1: (4.1) Modificar la topología, que puede mejorar la representación del grafo, mejorando el soporte para su proceso analítico. Esta tarea se vuelve necesaria cuando los nodos o relaciones inicialmente presentes en el grafo son insuficientes para abordar las preguntas de análisis. Un ejemplo es agregar nuevas relaciones para todos los nodos de producto comprados juntos. El lector puede referirse a [56] para orientación en esta tarea. (4.2) Incluir atributos estructurales, que proporciona información adicional en nodos y atributos de relación para una comprensión más completa de la estructura del grafo para mejorar el análisis. Los ejemplos abarcan la integración de atributos tales como centralidad de grado, centralidad de PageRank, comunidad de Louvain y similitud de Jaccard en atributos de nodo, así como incluir la ruta de Dijkstra como atributos de relación. El lector puede referirse a [52] para orientación en esta tarea. (4.3) Aplicar resumen, que es útil para reducir la complejidad del grafo mediante la creación de abstracciones de alto nivel, permitiendo a los usuarios obtener información e identificar patrones más rápidamente al agrupar nodos, relaciones o ambos. Por ejemplo, el grafo se puede resumir utilizando atributos como la centralidad de grado, la comunidad de Lovaina u otras medidas relevantes. El lector puede consultar [57] para obtener orientación en esta tarea. (4.4) Generar subgrafos, que contiene elementos filtrados y modificados del grafo original, atributos estructurales y resultados del resumen. Cada subgrafo sirve como una representación enfocada del grafo original, adaptada para cumplir con los requisitos analíticos específicos de los usuarios. Los ejemplos incluyen la estructura general del grafo, un grafo homogéneo y nodos específicos de interés. El lector puede consultar [58] para obtener orientación en esta tarea.

El resultado de esta etapa consiste en uno o más subgráficos o gráficos con nodos o aristas con nuevos atributos.

4.5. Etapa 5: Visualización gráfica

El núcleo de la etapa de visualización radica en elegir representaciones apropiadas del gráfico para respaldar la respuesta a las preguntas definidas en la etapa 1. Los usuarios pueden elegir varias formas de visualización, como tablas, redes o gráficos.

Esta etapa tiene una fuerte dependencia de las etapas anteriores. Se prevé que pueda haber múltiples iteraciones entre la etapa de visualización y las etapas 1, 2 y 3.
Permite incorporar los resultados de tareas previas, como modificaciones topológicas, nuevos atributos estructurales u otros resultados de algoritmos de minería de grafos. Cada alteración o transformación del grafo conlleva una representación visual que guía el proceso de análisis en curso.

Esta etapa consta de seis tareas: (5.1)

Aplicar diseño se refiere a configurar la forma, disposición y presentación de los nodos, y sus relaciones dentro de un grafo. El objetivo es mejorar la legibilidad y comprensión de la estructura de la red. Algunos ejemplos son el diseño Force Atlas, el diseño circular, el diseño jerárquico y el diseño de árbol. El lector puede referirse a [59] para orientación en esta tarea. (5.2) Realizar operaciones de desglose/recopilación son dos enfoques para navegar y examinar estructuras de datos jerárquicas o en capas. Son útiles para la visualización y el análisis de grafos, permitiendo la exploración en profundidad de los detalles o la síntesis de información dentro de los niveles superiores de la jerarquía. Además, ofrecen la capacidad de extraer información pertinente a través de diferentes niveles de granularidad o abstracción. El desglose implica un examen exhaustivo de la estructura jerárquica de datos, comenzando desde una perspectiva de nivel superior y profundizando progresivamente en niveles más finos de complejidad. En el campo del análisis de grafos, esto implica una exploración cada vez más específica de nodos y conexiones, comenzando desde un nodo inicial. Por ejemplo, se puede iniciar un análisis detallado a nivel de categoría, descendiendo progresivamente a productos individuales. Por el contrario, la acumulación se refiere a la condensación de datos desde los niveles inferiores a los superiores dentro de la jerarquía. En un grafo, podría ser necesaria la combinación o abstracción de nodos y relaciones para presentar una perspectiva más amplia o de nivel superior de la estructura de datos. Por ejemplo, es posible iniciar un análisis de acumulación, comenzando con productos individuales, agregándolos en categorías y, posteriormente, agrupando varias categorías según criterios específicos. El lector puede consultar [60] para obtener orientación en esta tarea.

(5.3) Definir reglas de visualización de gráficos comprende la descripción, el diseño y la implementación de las reglas de representación de gráficos mediante scripts y herramientas de visualización. Las decisiones tomadas para implementar estas reglas pueden influir profundamente en la claridad y eficacia del análisis de gráficos, destacando efectivamente patrones, relaciones y facetas dentro del conjunto de datos. Estas reglas pueden incorporar una serie de configuraciones y personalizaciones, cada una de las cuales puede aplicarse a distintas facetas de la visualización del gráfico. Algunos ejemplos de reglas de visualización son: (a) establecimiento apropiado de esquemas de color y dimensiones para nodos y sus relaciones, lo cual puede ser útil en un gráfico heterogéneo, donde los nodos pueden codificarse por color por tipo, como productos en azul, categorías en amarillo y clientes en verde, mejorando la distinción visual; (b) definición de restricciones para determinar la inclusión o exclusión de nodos y relaciones específicos relevantes para un análisis particular, lo cual puede ser útil para mostrar exclusivamente productos con ventas mayores a 200 unidades en un solo día, limitado a la categoría de mariscos; (c) criterios de agrupamiento para construir comunidades entre nodos o relaciones que comparten atributos distintivos, identificando características estructurales resultantes dentro del gráfico; por ejemplo, uno podría necesitar agrupar a todos los clientes que compraron en la categoría de cereales los viernes por la mañana, alentándolos a identificar patrones de compra; (d) mostrar etiquetas de nodos y relaciones para mostrar información complementaria del gráfico; por ejemplo, utilizando esta regla, la visualización de nodos de productos podría enriquecerse al mostrar sus atributos, como la fech
El lector puede consultar [61] para obtener orientación en esta tarea.

(5.4) Destacar patrones estructurales y de atributos implica la capacidad de reconocer patrones recurrentes dentro de la estructura de un grafo y los atributos vinculados a sus nodos y relaciones. Estos patrones son esenciales para comprender los nodos y las relaciones.

inherentes a un grafo, que ofrecen información sobre cómo se interrelacionan sus elementos. Los patrones estructurales se refieren a la organización de nodos y relaciones dentro del grafo. Abarcan la detección de subgrafos, ciclos, jerarquías, clústeres o cualquier otra configuración estructural en el grafo. Por ejemplo, dentro de una red de aeropuertos, se podría identificar el aeropuerto o conjunto de aeropuertos que sirven como centro pivotal, conectando a muchos otros y facilitando el tránsito entre múltiples destinos. Los patrones de atributos se centran en los atributos o características afiliadas a los nodos y relaciones del grafo. Estos atributos pueden ser instrumentales en la agrupación de nodos que comparten características o propósitos comunes. Por ejemplo, podría ser ventajoso identificar atributos que agrupan aeropuertos comúnmente utilizados para mejorar la congestión del tráfico. El lector puede referirse a [62] para obtener orientación en esta tarea. (5.5) Usar una línea de tiempo se refiere a una representación visual de una secuencia de datos que ilustra la progresión o alteraciones dentro de un grafo durante un período de tiempo específico. Esta representación rastrea y comprende la trayectoria de desarrollo de nodos, relaciones y atributos asociados a un grafo. La línea de tiempo resulta especialmente útil cuando los datos del grafo incorporan dimensiones temporales. Estas dimensiones son fundamentales para monitorear y analizar las fluctuaciones de los elementos del grafo en las relaciones o características a lo largo de distintos Además, es indispensable para obtener información dinámica y comprender mejor la dinámica temporal inherente a los datos durante el análisis de gráficos. Por ejemplo, una línea de tiempo puede revelar cómo se han transformado las conexiones entre productos y clientes, identificando productos que han experimentado cambios en su atractivo a lo largo de períodos específicos del año. Otro escenario ilustrativo implica exponer la propagación de una enfermedad dentro de un gráfico a lo largo de días, semanas o meses. El lector puede consultar [63] para obtener orientación en esta tarea. (5.6) Representar los resultados del gráfico implica la presentación y visualización de los resultados tras el análisis de un gráfico. Esta tarea implica transformar la información recopilada del gráfico en un formato comprensible y expresivo para los usuarios. Comunicar eficazmente los hallazgos y resultados del análisis del gráfico garantiza su comprensibilidad y valor para el usuario. La representación de estos resultados puede manifestarse en diversos formatos, dependiendo de la naturaleza de la información y los objetivos del análisis. Esta representación puede abarcar una vista de red que resalte la estructura del gráfico mostrando sus nodos y relaciones, y ayudas visuales como gráficos de barras y mapas de calor. Además, las representaciones pueden ser tablas o listas que resaltan los detalles de los nodos y los valores de los atributos. Significativamente, estas representaciones pueden ser estáticas o dinámicas y cumplir diversos propósitos. Por ejemplo, un gráfico de barras puede aclarar los cinco productos más vendidos, mientras que un mapa de calor podría proporcionar información sobre los aeropuertos más utilizados. El lector puede consultar [64] para obte en esta tarea.

El resultado de esta etapa consiste en representaciones del gráfico adecuadas para apoyar la respuesta de las preguntas definidas en la etapa uno.

4.6. Etapa 6: Evaluación La

etapa tiene como objetivo evaluar los resultados obtenidos, desarrollar la documentación y determinar estratégicamente las acciones posteriores.

Esta etapa consta de tres tareas: (6.1) Evaluar los Resultados: requiere un examen exhaustivo y un análisis crítico de los resultados obtenidos a lo largo del proceso de análisis para determinar su validez, relevancia y utilidad para cada pregunta planteada en la etapa uno. Esta tarea tiene el doble propósito de salvaguardar la credibilidad y la alineación de los hallazgos con los objetivos predefinidos. Para ello, la evaluación de los resultados prioriza la necesidad de su fiabilidad, pertinencia y valor. Además, esta tarea subraya la importancia de contextualizar y comunicar estos hallazgos de manera eficiente a las partes interesadas. La tarea de evaluación es un proceso continuo que se inicia de las etapas 2 a 5, e implica un proceso de validación para garantizar la idoneidad de los datos actuales para abordar la indagación de la etapa posterior, como se ilustra en la Figura 1. Además, en la etapa 1, si se requiere el contexto del caso, se pueden establecer métricas para complementar las preguntas de decisión. En consecuencia, al llegar a la etapa de evaluación, las métricas estarán preparadas para ofrecer una resolución cuantitativa junto con la evaluación cualitativa. Para garantizar la validez de los resultados,

El usuario necesita un análisis minucioso para confirmar la correcta aplicación de los algoritmos y el preprocesamiento adecuado de los datos de entrada. Para garantizar la relevancia de los resultados, el usuario debe determinar si la información derivada de estos aborda eficazmente las preguntas planteadas en la primera etapa. Para garantizar la utilidad de los resultados, el usuario debe realizar un análisis exhaustivo para evaluar en qué medida esta información contribuye al logro de los resultados de la primera etapa, como los objetivos de negocio y las preguntas para la toma de decisiones.
(6.2) La Documentación de Resultados implica la recopilación, el registro y la presentación sistemática y estructurada de todos los hallazgos, conclusiones, perspectivas y datos pertinentes derivados de un análisis, experimento o investigación. Esta documentación exhaustiva cumple el doble propósito de comunicar eficazmente los resultados a las partes interesadas relevantes, a la vez que salvaguarda la integridad del trabajo al permitir la trazabilidad y la reproducibilidad. Permite a otros examinar y evaluar el trabajo, estableciendo una base sólida para una toma de decisiones bien informada. Los elementos esenciales que deben documentarse incluyen una descripción general del proyecto, su ejecución, las fuentes de datos, los resultados, las limitaciones, las consideraciones, las conclusiones, las referencias y los anexos. (6.3) Determinar los Próximos Pasos implica el análisis de los resultados actuales y la identificación de las necesidades no satisfechas. Si no es posible responder a ninguna de las preguntas planteadas en la etapa 1, el análisis puede dar lugar a nuevas preguntas y objetivos o al inicio de un nuevo proyec
Sin embargo, si todas las preguntas planteadas en la etapa 1 se han respondido satisfactoriamente, entonces se puede dar por concluido el proyecto.

Los resultados de la etapa de evaluación incluyen las respuestas a las preguntas orientadoras, documentación del proyecto y la formulación de nuevas preguntas para su posterior análisis.

Tras revisar las seis etapas de la metodología para el descubrimiento de conocimiento en grafos etiquetados y heterogéneos (Figura 1), hemos establecido una guía completa para abordar el análisis de grafos. La siguiente sección describe cómo aplicar esta metodología a tres casos de estudio, mostrando su aplicación práctica.

5. Estudios de caso

En esta sección se explica la aplicación de la metodología a través de tres estudios de caso. Cada caso pretende proporcionar un ejemplo ilustrativo sin pretender una cobertura exhaustiva. El primer caso se centra en la información almacenada en una base de datos relacional, que rastrea las ventas de productos a diversos clientes. El segundo caso analiza un conjunto de datos que abarca aeropuertos globales y sus conexiones nacionales e internacionales. Por último, el tercer caso de estudio analiza un conjunto de datos georreferenciados de un cuadrante específico en Guadalajara, Jalisco, México, para determinar la ruta óptima para visitar varios museos.

A través de estos casos prácticos, demostramos la versatilidad y aplicabilidad de la metodología propuesta en diferentes ámbitos. Al aprovechar las técnicas de análisis de grafos, podemos extraer información valiosa, descubrir patrones ocultos y tomar decisiones basadas en datos, impulsando así la innovación y el progreso en diversos campos.

Cada etapa de la metodología contiene múltiples tareas, y cada tarea puede tener varias opciones para aplicar; la selección de estas opciones depende del conocimiento de los analistas y de los requerimientos técnicos del escenario.

En todos los casos de estudio, la base de datos gráfica utilizada es NEO4J. Si el usuario desea utilizar una opción diferente, puede seguir el proceso descrito [47] para elegir la que mejor se adapte a su caso.

5.1. Recomendación de productos

Supongamos que el dueño de un supermercado solicita a los analistas que identifiquen algunos productos para recomendar para aumentar las ventas con base en los productos más vendidos. Por lo tanto, este caso se centra en proporcionar recomendaciones de compra de productos. Inicialmente, la información se almacena en una base de datos relacional (RDB). Seguiremos la metodología KDG para llevar a cabo el análisis. Específicamente, utilizamos las tareas 1.5, 2.1, 2.2, 2.3, 2.4, 3.1, 3.3, 4.1, 4.2, 4.3, 4.4, 5.1, 5.2, 5.3, 5.6, 6.1, 6.2 y 6.3. A lo largo del caso práctico, proporcionaremos el razonamiento detrás de cada uno de estos pasos seleccionados.

Según la metodología, deberíamos verificar todas las tareas de la etapa 1. Sin embargo, en este caso, no hay objetivos comerciales definidos, como "aumentar las ganancias en un 5% en el primer trimestre del año en curso", pero tampoco hay requisitos de información, como seleccionar una tecnología de almacenamiento específica o seleccionar marcas de productos específicas, proveedores de productos o cadenas de distribución, etc. Además, no hay métricas, por ejemplo, lograr el objetivo en un período de menos de tres meses u otras métricas relacionadas con el costo o la calidad de los productos. Además, falta un plan de proyecto, que es un documento que incluye el alcance, las partes interesadas, el tiempo y el costo. Aunque faltan las tareas 1.1 a 1.4, la metodología indica que debemos aplicar la tarea 1.5, en la que necesitamos definir al menos una pregunta para la toma de decisiones que guíe el análisis para cumplir con la solicitud inicial. Para este caso, definimos las siguientes preguntas de toma de decisiones: (1) ¿Cuáles son las 5 categorías principales de productos que se compran junt (2) ¿Cuáles productos se encuentran entre los 5 más comúnmente comprados juntos? (3) Dado un producto específico, ¿cuáles son las recomendaciones de compra proporcionadas? (4) Si un producto no está disponible, ¿cuál es el sustituto sugerido para una recomendación?

Según la metodología, pasamos a la etapa 2, donde todas las tareas son obligatorias. La tarea inicial (2.1) implica realizar un análisis exploratorio de la información. La base de datos relacional (RDB) comprende 13 tablas: productos, categorías, clientes, proveedores, pedidos, detalles de pedidos, empleados, transportistas, empleados-territorios, territorios, región, datos demográficos de los clientes y cliente-cliente-demostración. La tarea (1.5) plantea preguntas sobre las recomendaciones de productos, y analizar la información en formato gráfico resulta más eficiente que examinarla como tablas dentro de una base de datos relacional. Pasando a la tarea (2.2), formulamos el modelo gráfico. Como estrategia de transformación simplificada, convertimos cada tabla en un nodo y cada relación en una arista. La Figura 2 muestra el modelo comple En la tarea (2.3), se selecciona el formato de almacenamiento. Dado el grafo mostrado en la Figura 2, se aprovechan las capacidades de una base de datos basada en grafos que almacena información de forma nativa, ya que un grafo se considera más efectivo. En este caso, optamos por NEO4J, aunque existen varias alternativas, como Tigergraph, ArangoDB o Dgraph. Para la tarea (2.4), exportamos cada tabla de la base de datos relacional a un archivo CSV y cada relación a otro archivo similar. Posteriormente, se crea una nueva base de datos en NEO4J, se importan los archivos CSV y se genera un script en lenguaje Cypher para transformar cada tabla en un nodo. Se crean índices por nodo, se establecen las relaciones entre los nodos y se incluyen los atributos de cada nodo y relación. Finalmente, la información se carga en la base de datos basada en grafos.
La Figura 2 describe la estructura general del gráfico. Al completar todas las tareas de la etapa 2, verificamos la suficiencia de la información y pasamos a la etapa 3.

Dada nuestra búsqueda de recomendaciones de productos, se podrían aplicar algoritmos comunitarios al grafo para recomendar productos de la misma comunidad (3.1). Sin embargo, debido a las preguntas de decisión específicas que guían el caso, los algoritmos de búsqueda de rutas (tarea 3.5) son innecesarios, ya que no buscamos la ruta más corta y económica que recorra todo el grafo. De igual manera, los algoritmos de centralidad (tarea 3.2) no son aplicables, ya que no buscamos los nodos más influyentes. Si bien se podría emplear un algoritmo de similitud (tarea 3.4) para identificar patrones entre productos en las facturas, optamos por no seguir esta ruta.
Nos centramos únicamente en aplicar algoritmos comunitarios (tarea 3.3) para abordar las preguntas. Al finalizar la etapa 3, verificamos la idoneidad de la información y pasamos a la etapa 4.

En la etapa 4, para abordar las preguntas, generamos un subgrafo que incluye los nodos de pedidos, productos y categorías, y las relaciones que los conectan (tarea 4.4). Para facilitar la recomendación de productos sustitutos (cuarta pregunta), introducimos dos nuevas relaciones : “comprar productos” y “comprar categorías”. También añadimos un nuevo atributo, “Louvain”, al nodo de productos, resultante de la aplicación de algoritmos comunitarios de minería de grafos (tareas 3.1 y 3.3). Este nuevo atributo nos permite recomendar productos sustitutos dentro del mismo grupo (tareas 4.1 y 4.2). A continuación, aplicamos el resumen basado en el nuevo atributo “Louvain ”, generando un subgrafo (tarea 4.3). La Figura 3 ilustra el subgrafo, que inicialmente contiene 77 productos. El grafo se resume en cuatro grupos: el primer clúster incluye 9 productos, el segundo clúster tiene 24 productos, el tercer clúster tiene 20 produ

El último grupo consta de 24 productos. Tras completar la etapa 4, verificamos la idoneidad de la información y pasamos a la etapa 5.

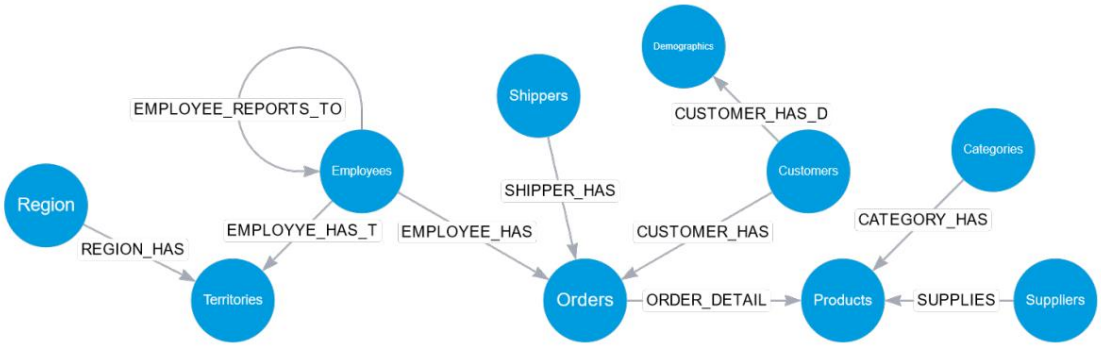


Figura 2. El modelo general del gráfico Northwind es heterogéneo, dirigido y con atributos. Consta de diez tipos de nodos y diez tipos de relaciones. Estos nodos representan diferentes entidades dentro del grafo, mientras que las relaciones denotan sus conexiones e interacciones. Este modelo integral proporciona una representación estructurada del conjunto de datos Northwind, lo que permite el análisis práctico y la exploración de sus elementos interconectados.

Figura 3. La vista muestra el subgrafo de pedidos, productos y categorías, revelando las nuevas relaciones establecidas entre "comprar productos juntos" y "comprar categorías juntas". Esta representación condensada ofrece una visión general completa de las conexiones dentro del conjunto de datos, destacando las asociaciones entre los productos comprados y la coexistencia de categorías en los pedidos de los cl

En la etapa 5, para responder las preguntas 1 y 2, ejecutamos dos consultas en el lenguaje Cypher sobre el subgrafo. Debido a que queremos representar la distribución espacial de nodos y relaciones, utilizamos un diseño basado en fuerza. Sin embargo, hay otros algoritmos, por ejemplo, jerárquico . La Figura 4 muestra el subgrafo de pedidos, productos y categorías. En esta vista, aplicamos un diseño basado en fuerza (tarea 5.1). Además, necesitamos otra visualización para saber cuántos productos pertenecen a cada categoría y cuántos pedidos tuvo cada producto. Eso nos permitiría seleccionar uno o más productos, tener los datos actualizados dinámicamente y hacer lo mismo con categorías o pedidos. Al examinar la Figura 5, podemos observar que al seleccionar las 5 categorías más vendidas juntas, se han vendido 59 productos y 786 pedidos. Al realizar un desglose, centrándonos en la categoría número 4, descubrimos 10 productos vendidos y 304 pedidos asociados (tarea 5.2). Además, queremos aplicar una regla para mostrar solo el subgráfico que contiene los 5 productos más vendidos. Si tomamos la Figura 4 como base y aplicamos la regla, podemos ver el resultado en la Figura 6 (tarea 5.3), que ilustra el gráfico resultante con 7 productos (debido a que algunas combinaciones se repiten), 240 pedidos y 3 categorías. Al abordar las preguntas, no encontramos necesidad de enfatizar patrones y atributos estructurales (5.4) ni de establecer una línea de tiempo que proporcione información sobre cambios hacia adelante o hacia atrás (5.5); por lo tanto, estos componentes no están incluidos. Finalmente, optamos por utilizar un tablero para presentar los resultados, como se muestra en la Figura 7. Este tablero incorpora dos tablas: la primera muestra los cinco pares de categorías de productos más vendidos y la segunda muestra los cinco pares de productos individuales más vendidos. Además, presenta un elemento dinámico, un cuadro combinado, que permite a los usuarios seleccionar un producto específico. Tras la selección, el tablero se actualiza dinámicamente para presentar tanto los productos recomendados como las opciones alternativas asociadas con el producto elegido. Al finalizar la etapa 5, verificamos la idoneidad de la información y pasar

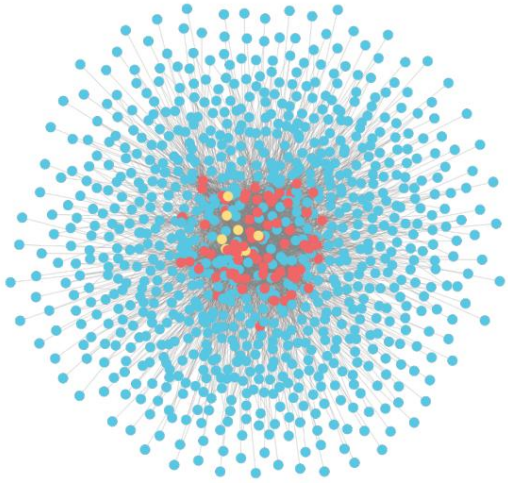
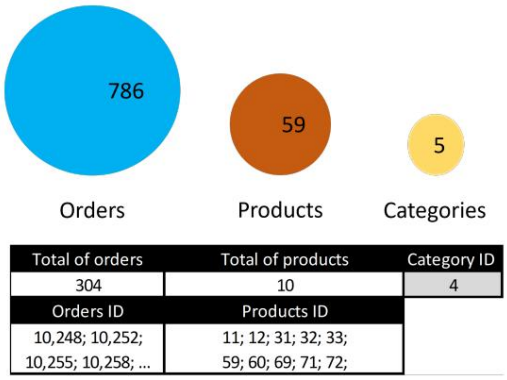


Figura 4. El subgráfico consta de 830 pedidos (representados en azul), 77 productos (representados en rojo) y 8 categorías (representadas en amarillo), cada uno con nuevas relaciones que indican productos comprados juntos y categorías compradas juntas.



Total of orders	Total of products	Category ID
304	10	4
Orders ID	**Products ID**	
10,248; 10,252; 10,255; 10,258; ...	11; 12; 31; 32; 33; 59; 60; 69; 71; 72;	

Figura 5. Se han identificado y seleccionado las 5 categorías más vendidas dentro de los círculos. Si seleccionamos exclusivamente la categoría con un identificador igual a cuatro, se muestran los detalles completos, incluyendo el número total de productos, el identificador de cada producto, el total de pedidos y el identificador de cada pedido.

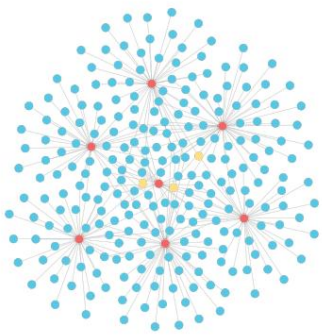


Figura 6. Este subgráfico consta de los 7 artículos más vendidos, representados por nodos rojos, junto con los 242 pedidos en los que aparecen, indicados por nodos azules. Además, el subgráfico destaca las tres categorías a las que pertenecen estos artículos, distinguidas por nodos amarillos. Esta representación visual explica claramente las relaciones entre los artículos más vendidos, sus frecuencias de pedido y sus categorías correspondientes. Al examinar este subgráfico, se puede obtener información valiosa sobre la popularidad y la clasificación de estos artículos dentro del conjunto de datos.

Top 5 Categories	
Category_1	**Category_2**
Dairy Products	Confections
Dairy Products	Beverages
Seafood	Beverages
Condiments	Beverages
Confections	Beverages

Top 5 Products	
Product_1	**Product_2**
Sirop d'rable	Sir Rodney's Scones
Gorgonzola Telino	Pavlova
Camembert Pierrot	Pavlova
Flotemysost	Camembert Pierrot
Mozzarella di Giovanni	Gorgonzola Telino

Product	Product Substituted	Product Recommended
Chang	Chai	Anissed Syrup
	Chartreuse verte	Chef Anton's Cajun Seasoning
	cate de Blaye	Gradma's Boysenberry Spread

Figura 7. El panel ofrece una visión general de los resultados de las preguntas. Incluye la siguiente información: 5 categorías principales; 5 productos principales; Productos sustitutos; Productos recomendados. Al analizar esta información, los usuarios pueden identificar rápidamente las categorías y productos más populares, explorar alternativas y recibir recomendaciones personalizadas según sus preferencias. Este panel es una herramienta valiosa para la toma de decisiones, ya que permite a los usuarios elegir con conocimiento de causa y optimizar su proceso de selección de productos.

En la Etapa 6, la finalización de todas las tareas es obligatoria. Las preguntas descritas en la Etapa 1 son la base para evaluar los resultados. Las preguntas de este caso fueron las siguientes: (1) ¿Cuáles son las 5 categorías principales de productos que se compran frecuentemente juntos? (2) ¿Cuáles productos se encuentran entre los 5 más comúnmente comprados juntos? (3) Dado un producto específico, ¿cuáles son las recomendaciones de compra proporcionadas? (4) Si un producto no está disponible, ¿cuál es el sustituto sugerido para una recomendación? Con base en el análisis , las 5 categorías principales que se compran frecuentemente juntas son Productos lácteos y dulces, Productos lácteos y bebidas, Mariscos y bebidas, Condimentos y bebidas, y Dulces y bebidas. Los cinco productos principales que se compran a menudo juntos son Sirop d'rable y Sir Rodney's Scones, Gorgonzola Telino y Pavlova, Camembert Pierrot y Pavlova, Flotemysost y Camembert Pierrot, y Mozzarella di Giovanni y Gorgonzola Telino.
Si seleccionamos Chang como producto específico, los productos recomendados para comprar juntos son el jarabe de anís, el condimento cajún del chef Anton, la crema de moras Boysen de la abuela, las peras secas orgánicas del tío Bob o la salsa de arándanos Northwoods. Si consideramos la mozzarella di Giovanni, los productos sustitutos podrían ser el camembert pierrot, el flotemysost, el geitost, el gorgonzola telino, el gudbrandsdalsost, el mascarpone fabioli, el queso de cabrales, el queso manchego La Pastora o la raclette Courdavault. El panel de control (Figura 7) permite seleccionar cualquier otro producto y obtener recomendaciones o productos sustitutos (tarea 6.1).

Si se actualiza la información, las categorías y productos recomendados pueden variar según los pedidos de los clientes. Dado que el análisis se realizó con datos extraídos de una base de datos relacional dentro de un período específico, es fundamental documentar los resultados asociados a una fecha determinada. Al documentar esta información, podemos realizar un análisis basado en la cronología del comportamiento de los productos y categorías de productos (tarea 6.2). Los siguientes pasos serían crear nuevas preguntas basadas en los resultados (tarea 6.3).

Gracias a nuestra metodología, hemos identificado las categorías y productos clave que se compran juntos con frecuencia. Además, hemos desarrollado un panel que integra los resultados y permite a los usuarios seleccionar productos específicos para recomendaciones de compra o sugerencias de productos sustitutos.

Este enfoque mejora significativamente la comprensión del comportamiento del consumidor y la experiencia de compra. Las empresas pueden obtener información valiosa para optimizar su oferta de productos, realizar campañas de marketing dirigidas y mejorar la satisfacción del cliente mediante el análisis de datos basados en los pedidos y las interconexiones entre productos y categorías.

Es importante destacar que este caso práctico se basa en la extracción de datos de una base de datos relacional durante un período específico. Por lo tanto, es crucial documentar los resultados asociados a una fecha determinada y realizar un análisis más detallado basado en la evolución de los pedidos y las categorías de productos. Actualizar la información también es esencial, ya que los productos y categorías recomendados pueden variar según los pedidos de los clientes.

En resumen, la aplicación de la metodología propuesta ha demostrado ser eficaz en el ámbito de la recomendación de compra de productos. Esta metodología se adapta a diferentes contextos y bases de datos, proporcionando a las empresas una herramienta eficaz para comprender y optimizar el proceso de compra de sus clientes.

5.2.

Aeropuertos. Los aeropuertos son un componente crítico de la conectividad global, y es crucial identificar los aeropuertos más importantes dentro de su red de conexiones. Este estudio de caso busca identificar estos aeropuertos clave y recomendar inversiones para su mejora o expansión. Inicialmente, la información se almacena en dos archivos separados por comas. Seguimos la metodología KDG para su análisis. Específicamente, utilizamos las tareas 1.5, 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 4.1, 4.2, 4.4, 5.6, 6.1, 6.2 y 6.3.

Según la metodología, deberíamos comprobar todas las tareas de la etapa 1. Sin embargo, al igual que en el caso 1, no hay objetivos de negocio definidos, requisitos de información, métricas o planes de proyecto. Aunque faltan las tareas 1.1 a 1.4, la metodología indica que debemos aplicar al menos la tarea 1.5, en la que debemos definir al menos una pregunta para la toma de decisiones que guíe el análisis para cumplir con la solicitud inicial. En este escenario, formulamos la siguiente pregunta de toma de decisiones: ¿Cuáles son los aeropuertos más importantes en nuestra red de conexiones? Para refinar nuestra respuesta, establecemos tres subpreguntas: (1) ¿Cuáles cinco aeropuertos tienen el mayor número de conexiones inmediatas con otros aeropuertos? (2) ¿Cuáles cinco aeropuertos podrían desempeñar un papel fundamental en el mantenimiento de conexiones posteriores debido a sus amplios vínculos de red? (3) ¿Cuáles son los cinco aeropuertos principales idealmente adecuados para las escalas? (tarea 1.5).

Siguiendo la metodología, avanzamos a la etapa 2, donde es obligatorio completar todas las tareas. La tarea principal (2.1) consiste en realizar un análisis exploratorio de la información.
El archivo CSV inicial contiene detalles del aeropuerto, como nombre, país, latitud, longitud y un identificador único, mientras que el segundo archivo proporciona información sobre las conexiones del aeropuerto. Cabe destacar que todas las conexiones son bidireccionales, representando rutas tanto de salida como de llegada. Los archivos incluyen 225 países con al menos un aeropuerto registrado. Para modelar el grafo, representaremos cada aeropuerto como un nodo conectado consigo mismo mediante la relación "Conectado a", como se ilustra en la Figura 8 (tarea 2.2). En la tarea (2.3), optamos por NEO4J como solución de almacenamiento de datos, ya que es una base de datos basada en grafos. Posteriormente, creamos un script Cypher para extraer información de los archivos CSV, transformando los detalles del aeropuerto en nodos con atributos y las conexiones en aristas. Finalmente, rellenamos una nueva instancia de la base de datos con los datos procesados (tarea 2.4). Tras completar todas las tareas de la etapa 2, verificamos la suficiencia de la información y pasamos a la etapa 3.

En la Etapa 3, decidimos utilizar algoritmos de centralidad para abordar las subpreguntas relacionadas con los aeropuertos con conexiones más inmediatas, aquellos cruciales para la continuidad y los aeropuertos ideales para escalas (tarea 3.1). Los algoritmos de centralidad resultan valiosos para identificar nodos críticos dentro de la red, lo que los hace idóneos para nuestros objetivos en esta etapa. Los aplicamos específicamente a los nodos aeroportuarios (tarea 3.2), excluyendo la aplicación de algoritmos de comunidad, similitud o búsqueda de rutas (tareas 3.2 a 3.4). Tras completar la Etapa 3, evaluamos la suficiencia de la información antes de avanzar a la Etapa 4.

En la Etapa 4, no es necesario generar un subgrafo para responder las preguntas, ya que solo contamos con nodos de tipo aeropuertos y aristas de tipo conectadas, como se muestra en la Figura 8 (tarea 4.4). Modificamos la topología e incluimos los atributos de grado, centralidad de intermediación y centralidad de proximidad, que se añadirán a los nodos de aeropuerto. El atributo de grado ayudará a responder la subpregunta 1, mientras que la centralidad de intermediación proporcionará información para la subpregunta 2. El atributo de centralidad de proximidad abordará la subpregunta 3 (tareas 4.1 y 4.2). Finalmente, al aplicar el resumen tanto a los nodos como a las aristas, el resultado se ilustra en la Figura 8. Al completar la Etapa 4, verificamos la idoneidad de la información y procedemos a la Etapa 5.

En la Etapa 5, nuestro enfoque se centra en la implementación de un panel de control (tarea 5.6), extrayendo información de las tareas realizadas en la Etapa 4. Este panel de control está diseñado para incorporar tres

Tablas. La tabla inicial presentará los nombres de los aeropuertos y sus respectivos países, priorizando los 5 mejores según su centralidad de grado. De igual forma, la segunda tabla presentará la misma información, organizada según la centralidad de intermediación, mientras que la tercera se estructurará según la centralidad de proximidad (Figura 9). Además, se podrá acceder a un panel secundario. A la izquierda de este panel se encuentra un gráfico de barras que muestra los 5 países con mayor número de aeropuertos registrados. Al seleccionar uno de los países a la derecha, la pantalla se actualiza dinámicamente, mostrando los 5 mejores aeropuertos según la centralidad de grado, la centralidad de intermediación y la centralidad de proximidad (véase la Figura 10).
Tras concluir la Etapa 5, una verificación exhaustiva de la idoneidad de la información precede a nuestro avance a la Etapa 6.

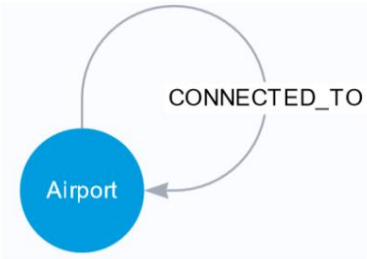


Figura 8. El modelo general del gráfico del aeropuerto es homogéneo, dirigido e incluye atributos. Los nodos en el gráfico representan aeropuertos, mientras que las relaciones entre nodos son del tipo conectado-con, lo que indica la conectividad entre aeropuertos.

Dergree	
Airport	**Country**
Frankfurt	Germany
Paris	France
Amsterdam	Netherlands
Istanbul	Turkey
Atlanta	United States

BetweennessCentrality	
Airport	**Country**
Paris	France
Los Angeles	United States
Dubai	United Arab Emirates
Anchorage	United States
Frankfurt	Germany

ClosenessCentrality	
Airport	**Country**
Melbourne	Australia
Sao Felix do Araguaia	Brazil
Vancouver	Canada
Victoria	Canada
Ittoqqortoormiit	Greenland

Figura 9. El panel presenta la siguiente información: Los 5 aeropuertos con conexiones más inmediatas y sus países correspondientes. Los 5 aeropuertos más críticos para mantener la continuidad, incluyendo sus respectivos países. Los 5 aeropuertos recomendados para escalas y sus países asociados.

En la Etapa 6, es obligatorio completar todas las tareas. Nuestra pregunta inicial fue: ¿Cuáles son los aeropuertos más importantes de nuestra red de conexiones? Para acotar la respuesta, definimos tres subpreguntas. (1) ¿Cuáles son los cinco aeropuertos con las conexiones más inmediatas ? Los elegidos son Fráncfort, París, Ámsterdam, Estambul y Atlanta.
(2) ¿Qué cinco aeropuertos podrían ser vitales para dar continuidad a las siguientes conexiones debido a sus conexiones con los demás? París, Los Ángeles, Dubái, Anchorage y Fráncfort obtienen la mejor puntuación. (3) ¿Cuáles son los cinco aeropuertos ideales para hacer escalas? Melbourne, São Félix do Araguaia, Victoria, Vancouver e Ittoqqortoormiit obtienen la mejor puntuación. Si compartimos los datos, los aeropuertos que aparecen en más de una categoría son Fráncfort y París (tarea 6.1)

Según los resultados de la Figura 9, los 17 aeropuertos más importantes de la red de conexiones son Fráncfort, París, Ámsterdam, Estambul, Atlanta, Los Ángeles, Dubái, Anchorage, Melbourne, São Félix do Araguaia, Victoria, Vancouver, Ittoqqortoormiit, Neerlerit Inaat, Windhoek, Numea y Unalaska. Según el análisis de centralidad, estos son los que deben considerarse al invertir o expandirse.

Finalmente, podemos replicar este análisis en cualquier país. Si tomamos como ejemplo Estados Unidos , que cuenta con el mayor número de aeropuertos registrados, podemos observar en la Figura 10 que los aeropuertos de Atlanta y Chicago destacan por tener el mayor número de conexiones inmediatas, pero también aparecen entre los aeropuertos con mayor flujo. Cabe destacar que Seattle aparece en dos categorías (tarea 6.1). Si la información se actualiza, los resultados pueden variar. Considerando que el análisis se realizó con datos

Al extraer datos de un archivo plano dentro de un período específico, es fundamental documentar los resultados correspondientes a una fecha específica. Esta documentación nos permite realizar un análisis alineado con la cronología de los aeropuertos (tarea 6.2). Además, en proyectos futuros se podría considerar la incorporación de datos sobre el tráfico de pasajeros en cada aeropuerto para enriquecer el análisis y explorar conjuntos de datos alternativos dentro de la metodología (tarea 6.3).

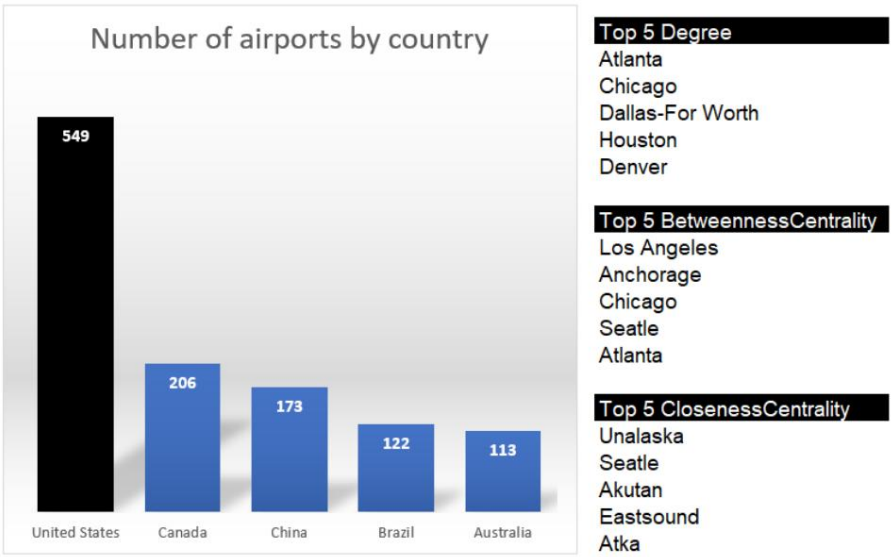


Figura 10. El panel proporciona la siguiente información: Los 5 países con más aeropuertos. Las 5 mejores conexiones inmediatas, exclusivamente dentro de Estados Unidos. Los 5 aeropuertos más críticos para garantizar la continuidad en EE. UU. Los 5 aeropuertos más recomendados para escalas en EE. UU.

En conclusión, el análisis del conjunto de datos aeroportuarios mediante la metodología propuesta proporciona información valiosa sobre la estructura de la red e identifica con éxito los aeropuertos más importantes. Al invertir en la mejora y expansión de estos aeropuertos, podemos mejorar la conectividad global, optimizar las rutas de viaje e impulsar el crecimiento económico.

5.3. Recomendación de ruta

En este caso, nuestro objetivo es establecer una serie de rutas que conecten todas las atracciones turísticas preinscritas, con especial énfasis en minimizar el costo total en lugar de priorizar la orientación de las rutas. Inicialmente, los datos para el análisis provienen de OSM [65], específicamente de un cuadrante (Figura 11) dentro de Guadalajara, Jalisco, México. Seguimos la metodología KDG. Específicamente, utilizamos las tareas 1.5, 2.1, 2.2, 2.3, 2.4, 3.1, 3.5, 4.1, 4.2, 4.4, 5.6, 6.1, 6.2 y 6.3.

Según la metodología, debemos verificar todas las tareas de la etapa 1. Sin embargo, al igual que en el caso 1, no se han definido objetivos de negocio, requisitos de información, métricas ni planes de proyecto. Si bien faltan las tareas 1.1 a 1.4, la metodología indica que debemos aplicar al menos la tarea 1.5, en la que debemos definir al menos una pregunta para la toma de decisiones que guíe el análisis para cumplir con la solicitud inicial. En este escenario, formulamos la siguiente pregunta de decisión: ¿Al menos una ruta nos permitirá visitar todos los atractivos turísticos? (tarea 1.5).

Siguiendo la metodología, avanzamos a la etapa 2, donde es obligatorio completar todas las tareas. La primera tarea (2.1) implica realizar un análisis exploratorio de la información. Los datos para el análisis provienen de OSM [65], específicamente de un cuadrante (Figura 11) en Guadalajara, Jalisco, México. Nuestro análisis se centra en la información georreferenciada de un cuadrante en Guadalajara, Jalisco, México. Específicamente, buscamos identificar las rutas que conectan los museos dentro de este cuadrante. Utilizamos OpenStreetMaps (OSM) para obtener los datos necesarios, lo que proporciona valiosa información geoespacial (tarea 2.1). Al avanzar a la tarea 2.2, nuestro modelo gráfico consta de dos tipos de nodos: usuario y punto. Dado que OSM es una plataforma impulsada por la comunidad, rastrea a los usuarios que editan cada punto o conjunto de puntos en el

Mapa. Además, los puntos del mapa se identifican mediante un identificador único y coordenadas de latitud y longitud. Algunos puntos también pueden tener atributos adicionales, como la marca de negocios como Oxxo o 7-Eleven, números de teléfono de contacto, nombres de calles, descripciones detalladas o indicadores de atracción turística. Aunque los puntos no tienen una conexión inherente, estableceremos relaciones e incorporaremos información de distancia. El modelo gráfico se ilustra en la Figura 12. Continúe con la tarea 2.3. Elegimos NEO4J como solución de almacenamiento de datos por ser una base de datos basada en grafos. Posteriormente, extrajimos las coordenadas de latitud y longitud de los sitios clasificados por OSM como atracciones turísticas.

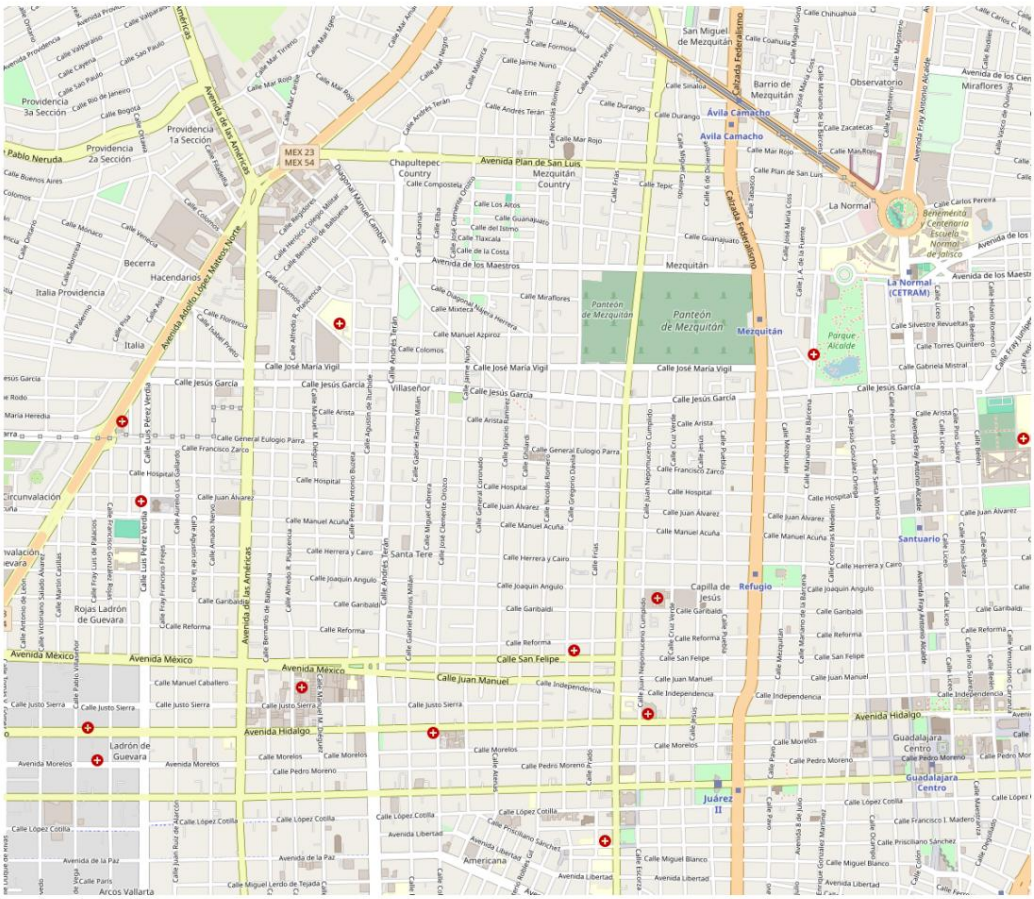


Figura 11. La imagen proporcionada representa el cuadrante asociado a Guadalajara, Jalisco, México, abarcando la información georreferenciada de OpenStreetMaps.

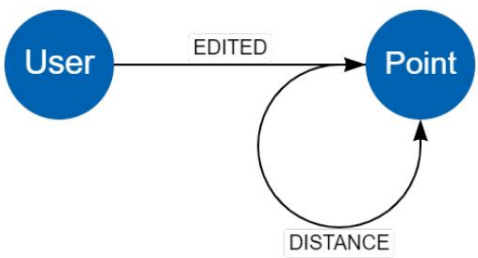


Figura 12. El modelo general del grafo de atracciones consiste en un grafo homogéneo y dirigido con atributos. Incluye dos tipos de nodos y dos tipos de relaciones.

Además, transformamos la información de OSM en una base de datos gráfica. Primero, exportamos los datos del cuadrante seleccionado en Guadalajara como un archivo XML. Segundo, desarrollamos un script utilizando el lenguaje Cypher para convertir cada punto en un nodo, extraer sus atributos e insertar la información en NEO4J. Cada nodo incluye detalles esenciales como un identificador único, latitud y longitud. Tercero, calculamos la geodésica.

Distancia entre cada par de nodos y establecer una relación denominada "distancia" con el atributo correspondiente "geodesicDist", que contiene el valor calculado. Estos datos se incorporan a la base de datos del grafo (tarea 2.4). Tras completar todas las tareas de la etapa 2, verificamos la suficiencia de la información y pasamos a la etapa 3.

En la Etapa 3, se decidió emplear algoritmos de búsqueda de rutas para determinar la ruta que permite visitar todos los museos (tarea 3.1). Estos algoritmos son especialmente ventajosos cuando el objetivo es identificar rutas óptimas, determinar los caminos más cortos o analizar patrones de flujo dentro de la red. Por lo tanto, en este contexto no se aplican algoritmos de centralidad, comunidad ni similitud. En su lugar, se implementa el algoritmo de Árbol de Expansión Mínimo (MST) para identificar la ruta con el menor peso total, lo que facilita la visita a todos los museos (tarea 3.5). Tras completar la Etapa 3, se realiza una evaluación exhaustiva de la adecuación de la información antes de pasar a la Etapa 4.

En la Etapa 4, generamos un subgrafo para identificar los puntos con el atributo turismo, cuyo valor corresponde a un museo. El resultado genera 11 nodos y 121 relaciones (Figura 13) (tarea 4.4). Para determinar la ruta más corta que conecta todos los museos, modificamos la topología del grafo introduciendo las relaciones con los pesos más bajos entre los museos (tarea 4.1). Además, incorporamos una relación denominada MST, resultante de aplicar el algoritmo de árbol de expansión mínimo (MTR), para identificar la ruta con el menor peso total que permite visitar todos los museos. Estas modificaciones topológicas y la inclusión de un nuevo atributo estructural dieron lugar a la creación del subgrafo que se muestra en la Figura 13 (tarea 4.2). Tras completar la Etapa 4, verificamos la idoneidad de la información y pasamos a la Etapa 5.

Figura 13. El subgrafo está compuesto por puntos registrados en la base de datos de grafos, que representan específicamente museos y sus correspondientes distancias georreferenciadas.

En la Etapa 5, enfatizamos la implementación de un tablero de control (Figura 14) para mostrar los once museos y el Árbol de Expansión Mínimo, permitiendo la visita a cada uno de ellos (tarea 5.6). Una vez completada la Etapa 5, una verificación exhaustiva de la idoneidad de la información precede nuestra transición a la Etapa 6.

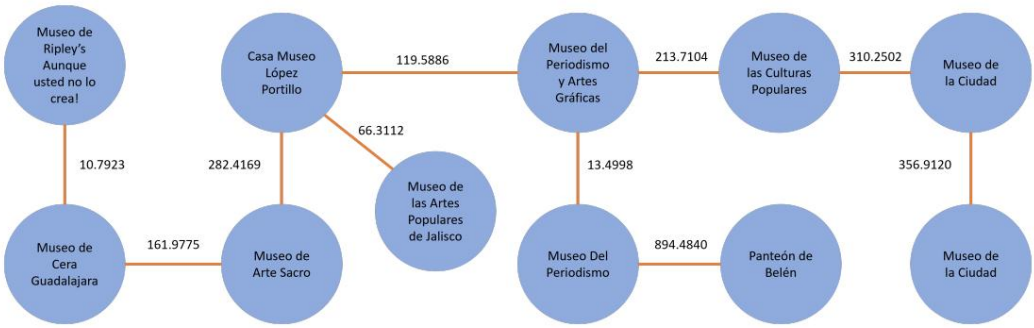


Figura 14. Este subgráfico incorpora los once museos y el árbol de expansión mínimo que permite visitarlos todos.

En la Etapa 6, es obligatorio completar todas las tareas. La cuestión inicial era buscar una red de conexiones entre atractivos turísticos preinscritos para minimizar el coste total sin tener en cuenta el sentido de las conexiones. Logramos este objetivo obteniendo el árbol de expansión mínimo que nos permita visitar los siguientes once museos: Museo de Ripley's Aunque usted no lo crea!, Panteón de Belén, Museo Del Periodismo, Museo del Periodismo y Artes Gráficas, Museo de la Ciudad, Museo de las Culturas Populares, Museo de Cera Guadalajara, Museo de Arte Sacro, Casa Museo López Portillo, Ex Convento Del Carmen y Museo de las Artes Populares de Jalisco. Así, cumplimos el requisito y respondimos la pregunta (tarea 6.1).

Dada la naturaleza dinámica de OSM, es crucial documentar con precisión los hallazgos con una fecha específica, considerando la continua incorporación de nuevas atracciones turísticas. Dado que el análisis se basó en un extracto de OSM de un período específico, es esencial asegurar que los resultados estén asociados a la fecha correspondiente. Mediante el monitoreo y análisis continuos del comportamiento de la información, podemos adaptar y optimizar el sistema de recomendaciones de rutas para futuras actualizaciones (tarea 6.2). En los siguientes pasos, podríamos ampliar el cuadrante del mapa o seleccionar otra ciudad (tarea 6.3).

En conclusión, la implementación de la metodología propuesta ha arrojado resultados significativos en el establecimiento de una red de conexiones entre las atracciones turísticas preinscritas, lo que ha permitido minimizar los costos de las visitas a los once museos. El análisis del extracto de OSM ha proporcionado información valiosa sobre los patrones de conectividad de estas atracciones.
El uso de esta metodología en el ámbito turístico muestra su potencial para mejorar la experiencia del visitante y optimizar las rutas de viaje en diversas regiones.

6. Discusión

El artículo tiene como objetivo abordar aspectos críticos del descubrimiento de conocimiento en bases de datos de grafos de manera integral. En primer lugar, identifica actividades de transformación y exploración que aprovechan el conocimiento existente de las relaciones de datos. A pesar de su importancia en la teoría de grafos, las metodologías actuales han abordado inadecuadamente estas actividades. Además, el artículo reconoce las limitaciones de metodologías bien conocidas como CRISP-DM cuando se aplican al análisis de grafos, ya que se centra en datos tabulares y las directrices proporcionadas por estas metodologías son muy generales; además, no tiene en cuenta los desafíos que se abordan cuando se requiere modelar datos como un grafo. Por lo tanto, existe una brecha en esas metodologías entre el tipo de información, las tareas, la visualización utilizada y lo que se requiere para el análisis basado en grafos.

Para abordar estas limitaciones, el artículo presenta KDG, una novedosa metodología diseñada específicamente para el descubrimiento de conocimiento en bases de datos de grafos. Esta metodología ofrece un enfoque sistemático para extraer información valiosa de datos estructurados en grafos, lo que permite a investigadores y profesionales extraer patrones y relaciones significativas de forma eficaz.

Para ilustrar la metodología, presentamos tres casos prácticos que demuestran su aplicación para abordar diversos problemas empresariales. Estos ejemplos describen cómo la metodología puede abordar eficazmente diversos desafíos, destacando su versatilidad y practicidad.

La metodología KDG se puede comparar con otras en términos de las tareas que describimos en Sección 4. Hemos seleccionado las tres metodologías más representativas de la literatura para el descubrimiento de conocimiento. Presentamos en la Tabla 1 un mapeo de cada tarea con los pasos de CRISP-DM [7], DST [14] y KDD [6]; cada fila muestra la relación entre uno La tarea de KDG y un paso o pasos en otras metodologías. Sin embargo, en cada etapa de nuestro metodología, no existe una correspondencia uno a uno entre nuestras tareas y los pasos descrito en otras metodologías. Esto se debe a que las metodologías no se desarrollaron trabajar específicamente con información representada como un gráfico. Además, los pasos de Estas metodologías asociadas con la tarea 3.1 (Seleccionar algoritmos de minería de grafos) están relacionadas para algoritmos de minería de datos en general, no útil para la minería de gráficos.

Tabla 1. Asignación de las tareas del KDG a los pasos CRISP-DM, DST y KDD.

Etapas de KDG	Tareas de KDG	DM CRUJIENTE [7]	Horario de verano [14]	KDD [6]
1. Comprender la Proceso analítico	1.1 Identificar los objetivos del negocio	Exploración de objetivos de comprensión empresarial		Desarrollar una comprensión de el dominio de aplicación y el conocimientos previos relevantes, y Identificar el objetivo del KDD proceso desde el punto de vista del cliente.
	1.2 Identificar la información requisitos	Comprensión empresarial	Comprensión empresarial	-
	1.3 Identificar métricas exitosas 1.4	Comprensión empresarial	Comprensión empresarial	-
	Identificar el plan del proyecto 1.5	Comprensión empresarial	Comprensión empresarial	-
	Definir preguntas para Toma de decisiones	Exploración del valor de los datos de comprensión empresarial		-
2. Construcción de gráficos	2.1 Análisis exploratorio de datos 2.2	Comprensión de datos	Exploración de fuentes de datos	Creación de un conjunto de datos de destino
	Diseño del modelo gráfico 2.3	-	-	-
	Selección del formato de almacenamiento de datos Preparación de		Preparación de datos	Creación de un conjunto de datos de destino
	datos 2.4 Extraer, transformar y cargar datos	Preparación de datos	Preparación de datos	Limpieza y preprocesamiento de datos. Reducción y proyección de datos
3. Minería de gráficos	3.1 Seleccionar gráfico algoritmos de minería	Modelado *	Modelado *	Coincidencia de los objetivos del mosaico KDD proceso para minería de datos particulares Método. Elección de los datos algoritmos de minería. Minería de datos
	3.2 Aplicar algoritmos de centralidad 3.3	-	-	-
	Aplicar algoritmos de comunidad 3.4 Aplicar		-	-
	algoritmos de similitud 3.5 Aplicar	-	-	-
	búsqueda de rutas algoritmos	-	-	-
4. Gráfico Transformación	4.1 Modificar la topología	-	-	-
	4.2 Incluir atributos estructurales	-	-	-
	4.3 Aplicar resumen 4.4 Generar	-	-	-
	subgrafos	-	-	-
5. Gráfico Visualización	5.1 Aplicar un diseño	-	-	-
	5.2 Realizar un desglose o una acumulación operaciones	-	-	-
	5.3 Definir reglas de visualización de	-	-	-
	gráficos 5.4 Resaltar reglas estructurales y patrones de atributos	-	-	-
	5.5 Utilice la línea de tiempo	-	-	-
	5.6 Representar los resultados gráficos	-	Exploración narrativa	Interpretación de patrones extraídos
6. Evaluación	6.1 Evaluar resultados	Evaluación	Exploración de resultados	Consolidando lo descubierto conocimiento
	6.2 Documentación de resultados	Evaluación	Exploración de productos	Consolidando lo descubierto conocimiento
	6.3 Determinar los próximos pasos	Evaluación	Evaluación	-

* Indica que los pasos son parcialmente similares a la tarea del KDG, ya que otras metodologías no son específicas de los gráficos.

En resumen, el artículo hace contribuciones significativas al descubrimiento de conocimientos en bases de datos gráficas mediante la identificación de actividades de transformación y exploración, abordando la limitaciones de las metodologías existentes, introduciendo una metodología novedosa y proporcionando Ejemplos concretos de su aplicación en situaciones empresariales reales. La metodología...

es ideal para analizar información altamente interconectada. Sin embargo, si el análisis implica otros tipos de información, esta metodología puede no ser la opción más adecuada. El usuario debe comprender las capacidades inherentes al trabajo con gráficos. Mientras Se proporcionará al usuario una guía para realizar el análisis, conocimientos fundamentales. El uso de técnicas, como los algoritmos de minería de grafos, es esencial. El usuario no debe ser un experto pero familiarizado con el tema. Como parte de nuestro trabajo futuro, planeamos integrar Operaciones de aprendizaje automático (MLOps) y prácticas de DevOps en nuestra metodología para Mejorar su eficacia y eficiencia. Al incorporar los principios de MLOps, buscamos... para optimizar el desarrollo, la implementación y la gestión de modelos de aprendizaje automático, Garantizar operaciones fluidas y escalables. Además, pretendemos validar la robustez de nuestra metodología aplicándola a una gama más amplia de estudios de casos, lo que nos permite recopilar obtener más información y validarla en diferentes escenarios de negocio.

Contribuciones de los autores: Conceptualización, VHO-G., LG-P., FC y MA-M.; Metodología, VHO-G., LG-P. y FC; Software, VHO-G.; Validación, VHO-G., LG-P., FC y MA-M.; Análisis formal, LG-P.; Investigación, LG-P. y FC; Curación de datos, VHO-G.; Redacción—borrador original, VHO-G.; Redacción, revisión y edición, VHO-G., LG-P., FC y MA-M.; Visualización, VHO-G. y MA-M.; Supervisión, LG-P. y FC Todos los autores han leído y están de acuerdo con la versión publicada. del manuscrito.

Financiación: Esta investigación no recibió financiación externa.

Declaración de la Junta de Revisión Institucional: No aplicable.

Declaración de consentimiento informado: No aplica.

Declaración de disponibilidad de datos: Los datos presentados en este estudio están disponibles a pedido. autor (vortega@iteso.mx).

Conflictos de intereses: Los autores declaran no tener ningún conflicto de intereses.

Abreviaturas

En este manuscrito se utilizan las siguientes abreviaturas:

SQL	Lenguaje de consulta estructurado
NOSQL	No sólo lenguaje de consulta estructurado
CRISP-DM: proceso estándar intersectorial para minería de datos	
BIM	Modelo de inteligencia empresarial
Operaciones de MLO	Operaciones de aprendizaje automático
OSM	Mapas de calles abiertas
CSV	Valores separados por comas
BGF	Base de datos gráfica
Base de datos de recursos	Base de datos relacional
ELEGANTE	Específico, medible, alcanzable, relevante y con plazos determinados
KDD	Descubrimiento de conocimiento en bases de datos
SEMMA Muestrear, explorar, modificar, modelar y evaluar	
Horario de verano	Herramienta de apoyo a la toma de decisiones
EDA	Análisis exploratorio de datos
PDA	Análisis predictivo de datos
ASUM	Método unificado de soluciones analíticas
CASP-DM Proceso estándar contextual para minería de datos	
FMDS	Metodología fundamental para la ciencia de datos
TDSP	Proceso de ciencia de datos en equipo
Horario de verano	Trayectoria de Ciencia de Datos
KDG	Descubrimiento de conocimiento en gráficos

Referencias

1. Fernandes, D.; Bernardino, J. Comparación de bases de datos de gráficos: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J y OrientDB. En Actas de la 7ª Conferencia Internacional sobre Ciencia de Datos, Tecnología y Aplicaciones (DATA 2018), Volterra, Italia, 13-16 de septiembre de 2018; págs. 373-380.
2. Lysenko, A.; Roznov˘a¸t, IA; Saqi, M.; Mazein, A.; Rawlings, CJ; Auffray, C. Representación y consulta de redes de enfermedades mediante bases de datos de grafos. BioData Min. 2016, 9, 1–19. [CrossRef]
3. Do˘gan, B. La importancia de las bases de datos gráficas en la detección de delitos financieros organizados. En El impacto de la inteligencia artificial en la gobernanza, la economía y las finanzas; Springer: Berlín/Heidelberg, Alemania, 2022; Volumen 2, págs. 147-155. 4. Czerepicki, A. Aplicación de bases de datos de grafos para fines de transporte. Bull. Pol. Acad. Sci. Tech. Sci. 2016, 64, 457–466. [CrossRef]
5. Sayeb, Y.; Jebri, M.; Ghezala, HB. Un sistema de recomendación basado en grafos para la gestión de la crisis de la COVID-19. Procedia Comput. Sci. 2022, 196, 348–355. [Referencia cruzada] [PubMed]
6. Fayyad, UM; Piatetsky-Shapiro, G.; Smyth, P. Descubrimiento de conocimiento y minería de datos: hacia un marco unificador. En Actas de la Segunda Conferencia Internacional sobre Descubrimiento de Conocimiento y Minería de Datos (KDD-96), Portland, Oregón, 2-4 de agosto de 1996; Volumen 96, págs. 82-88.
7. Wirth, R.; Hipp, J. CRISP-DM: Hacia un modelo de proceso estándar para la minería de datos. En Actas de la 4.ª Conferencia Internacional sobre Aplicaciones Prácticas del Descubrimiento de Conocimiento y la Minería de Datos, Mánchester, Reino Unido, 11-13 de abril de 2000; Volumen 1, págs. 29-39.
8. Sarma, KS Modelado predictivo con SAS Enterprise Miner: soluciones prácticas para aplicaciones empresariales; SAS Institute: Cary, NC, Estados Unidos, 2017.
9. Chakrabarti, D. Minería de grafos. En Enciclopedia de Aprendizaje Automático; Sammut, C., Webb, GI, Eds.; Springer: Boston, MA, EE. UU. 2010; págs. 469–471. [CrossRef]
10. Método Unificado de Soluciones Analíticas de IBM (ASUM). Disponible en línea: http://gforge.icesi.edu.co/ASUM-DM_External/index. htm#cognos.external.asum-DM_Teaser/procesosdeentrega/ASUM-DM_8A5C87D5.html_desc.html?proc=_0eKIHlt6EeW_y7 k3h2HTng&path=_0eKIHlt6EeW_y7k3h2HTng (consultado el 14 de septiembre de 2023).
11. Martínez-Plumed, F.; Ochando, LC; Ferri, C.; Flach, Pensilvania; Hernández-Orallo, J.; Kull, M.; Lachiche, N.; Ramírez Quintana, MJ CASP-DM: Proceso estándar contextual para minería de datos. CoRR 2017, arXiv:1709.09003.
12. Metodología Fundamental para la Ciencia de Datos. Disponible en línea: https://www.ibm.com/downloads/cas/WKK9DX51 (accedido el 14 de septiembre de 2023).
13. Proceso de Ciencia de Datos en Equipo. Disponible en línea: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-proceso (consultado el 14 de septiembre de 2023).
14. Martínez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; Hernández-Orallo, J.; Kull, M.; Lachiche, N.; Ramirez-Quintana, MJ; Flach, P. CRISP-DM veinte años después: De los procesos de minería de datos a las trayectorias de la ciencia de datos. IEEE Trans. Knowl. Data Eng. 2019, 33, 3048–3061. [Referencia cruzada]
15. Studer, S.; Bui, TB; Drescher, C.; Hanuschkin, A.; Winkler, L.; Peters, S.; Müller, KR. Hacia CRISP-ML (Q): Un modelo de proceso de aprendizaje automático con metodología de control de calidad. Mach. Learn. Knowl. Extr. 2021, 3, 392–413. [CrossRef]
16. Horkoff, J.; Barone, D.; Jiang, L.; Yu, E.; Amyot, D.; Borgida, A.; Mylopoulos, J. Modelado estratégico de negocios: representación y razonamiento. Softw. Syst. Model. 2014, 13, 1015–1041. [CrossRef]
17. Kumar, D.; Showrov, MIH. Un marco de minería de datos para la generación y el análisis de grafos sociales. En Actas de la 2.ª Conferencia Internacional sobre Innovación en Ingeniería y Tecnología (ICIET), Harbin, China, 20-22 de enero de 2019; págs. 1-6.
18. Pienta, R.; Hohman, F.; Endert, A.; Tamersoy, A.; Roundy, K.; Gates, C.; Navathe, S.; Chau, D.H. VIGOR: exploración visual interactiva de resultados de consultas de grafos. IEEE Trans. Vis. Comput. Graph. 2017, 24, 215–225. [CrossRef] [PubMed]
19. Bok, K.; Yoo, S.; Choi, D.; Lim, J.; Yoo, J. Almacenamiento en caché en memoria para mejorar la accesibilidad de subgrafos. Appl. Sci. 2020, 10, 5507. [Referencia cruzada]
20. Chen, C.; Yan, X.; Zhu, F.; Han, J.; Philip, SY Graph OLAP: un marco multidimensional para el análisis de datos de grafos. Knowl. Inf. Syst. 2009, 21, 41–63. [Referencia cruzada]
21. Mcgee, F.; Ghoniem, M.; Melançon, G.; Otjacques, B.; Pinaud, B. El estado del arte en la visualización de redes multicapa. En Computer Graphics Forum; Wiley Online Library: Hoboken, NJ, EE. UU., 2019; Volumen 38, págs. 125–149.
22. Nararatwong, R.; Kertkeidkachorn, N.; Ichise, R. Visualización de gráficos de conocimiento: desafíos, marco e implementación. En Actas de la 3.ª Conferencia Internacional IEEE sobre Inteligencia Artificial e Ingeniería del Conocimiento (AIKE), Laguna Hills, CA, EE. UU., 9-11 de diciembre de 2020; págs. 174-178.
23. Shchur, O.; Mumme, M.; Bojchevski, A.; Günnemann, S. Dificultades en la evaluación de redes neuronales gráficas. arXiv 2018, arXiv:1811.05868.
24. Alshahrani, M.; Thafar, MA; Essack, M. Aplicación y evaluación de incrustaciones de grafos de conocimiento en datos biomédicos. PeerJ Comput. Sci. 2021, 7, e341. [PubMed]
25. Shrivastava, S.; Pal, SN. Marco de minería de grafos para la búsqueda y visualización de subestructuras mediante bases de datos de grafos. En Actas de la Conferencia Internacional sobre Avances en Análisis y Minería de Redes Sociales, Atenas, Grecia, 20-22 de julio de 2009; págs. 379-380.
26. Nasiri, A.; Nalchigar, S.; Yu, E.; Ahmed, W.; Wrembel, R.; Zimanyi, E. De los indicadores a la analítica predictiva: Un marco de modelado conceptual. En Actas de la Conferencia de Trabajo de la IFIP sobre la Práctica del Modelado Empresarial, Lovaina, Bélgica, 22-24 de noviembre de 2017; págs. 171-186.

27. Yu, E. Modelado de relaciones estratégicas para la reingeniería de procesos. Soc. Model. Requir. Eng. 2011, 11, 66–87.
28. Schroeder, DT; Pogorelov, K.; Langguth, J. Fact: un marco para el análisis y la captura de grafos de Twitter. En Actas de la Sexta Conferencia Internacional sobre Análisis, Gestión y Seguridad de Redes Sociales (SNAMS) de 2019, Granada, España, 22-25 de octubre de 2019; págs. 134-141.
29. Qiao, F.; Zhang, X.; Li, P.; Ding, Z.; Jia, S.; Wang, H. Un enfoque paralelo para la minería frecuente de subgrafos en un único grafo grande Usando Spark. Appl. Sci. 2018, 8, 230. [CrossRef]
30. Zhang, J.; Li, T.; Jiang, Z.; Hu, X.; Jazayeri, A. Un método de metagrafo ponderado de Noval para la clasificación en redes de información heterogéneas. Appl. Sci. 2020, 10, 1603. [CrossRef]
31. Lee, K.; Jung, H.; Hong, JS; Kim, W. Aprendizaje de conocimiento mediante minería de subgrafos frecuentes a partir de datos de grafos de ontología. Apl. Ciencia. 2021, 11, 932. [Referencia cruzada]
32. Dunne, C.; Shneiderman, B. Simplificación de motivos: mejora de la legibilidad de la visualización de redes con glifos de abanico, conector y grupo . En Actas de la Conferencia SIGCHI sobre Factores Humanos en Sistemas Informáticos, París, Francia, 27 de abril–2 de mayo de 2013; págs. 3247–3256.
33. West, DB Introducción a la teoría de grafos; Prentice Hall: Upper Saddle River, NJ, EE. UU., 2001; Volumen 2.
34. Robinson, I.; Webber, J.; Eifrem, E. Bases de datos gráficas: Nuevas oportunidades para datos conectados; O'Reilly Media, Inc.: Sebastopol, CA, Estados Unidos, 2015.
35. Les MacLeod, Doctor en Educación. Haciendo que los objetivos SMART sean más inteligentes. Physician Exec. 2012, 38, 68.
36. ISO/IEC/IEEE 29148:2018(E); Norma Internacional ISO/IEC/IEEE - Ingeniería de Sistemas y Software - Procesos del Ciclo de Vida - Ingeniería de Requisitos. IEEE: Piscataway, NJ, EE. UU., 2018; págs. 1–104. [CrossRef]
37. Lovett, J. Secretos de las métricas de las redes sociales; John Wiley & Sons: Hoboken, NJ, EE. UU., 2011.
38. Pendleton, M.; Garcia-Lebron, R.; Cho, J. H.; Xu, S. Una encuesta sobre métricas de seguridad de sistemas. ACM Comput. Surv. (CSUR) 2016, 49, 1–35. [Referencia cruzada]
39. Reich, BH; Wee, SY: Búsqueda de conocimiento en la Guía del PMBOK®. Proj. Manag. J. 2006, 37, 11–26. [CrossRef]
40. Hammond, JS; Keeney, RL; Raiffa, H. Decisiones inteligentes: Una guía práctica para tomar mejores decisiones; Harvard Business Review Prensa: Brighton, MA, EE.UU., 2015.
41. Bowell, T.; Kemp, G. Pensamiento crítico: una guía concisa; Routledge: Londres, Reino Unido, 2014.
42. Kojima, R.; Legaspi, R.; Wada, S. Predicción del destino de un viaje mediante un enfoque de análisis exploratorio de datos entre ciudades en datos de flujo de personas. En las Actas de la Conferencia Internacional del IEEE sobre Big Data (Big Data), Osaka, Japón, 17-20 de diciembre de 2022; págs. 6547-6552. [CrossRef]
43. Fuentes, A. Conviértase en un analista de datos de Python: realice análisis de datos exploratorios y obtenga conocimientos sobre computación científica usando Python; Packt Publishing Ltd.: Birmingham, Reino Unido, 2018.
44. Uzhga-Rebrov, O.; Grabusts, P. Evaluación comparativa de cuatro métodos para el análisis exploratorio de datos. En Actas de la 62.ª Conferencia Científica Internacional sobre Tecnologías de la Información y Ciencias de la Gestión de la Universidad Técnica de Riga (ITMS), Riga, Letonia, 14-15 de octubre de 2021; págs. 1-5. [CrossRef]
45. Mostajabi, F.; Safaei, AA; Sahafi, A. Revisión sistemática de modelos de datos para el problema de los macrodatos. IEEE Access 2021, 9, 128889–128904. [Referencia cruzada]
46. Lal, M. Modelado de datos gráficos Neo4j; Packt Publishing Ltd.: Birmingham, Reino Unido, 2015.
47. Ortega, V.; Ruiz, L.; Gutiérrez, L.; Cervantes, F. Un proceso de selección de bases de datos de grafos basado en requisitos de negocio. En Actas de la Conferencia Internacional sobre Mejora de Procesos de Software, León, México, 23-25 de octubre de 2019; págs. 80-90.
48. Bansal, SK; Kagemann, S. Integración de Big Data: Un marco semántico de extracción-transformación-carga. Computer 2015, 48, 42–50. [Referencia cruzada]
49. Maria Carina, R. Aprendizaje de Pentaho Data Integration 8 CE (tercera edición): Comience a utilizar la herramienta de integración de datos de Pentaho con esta guía práctica y de fácil lectura; Packt Publishing: Birmingham, Reino Unido, 2017.
50. Dr. Tirthajyoti, S.; Shubhadeep, R. Manejo de datos con Python: Creación de datos procesables a partir de fuentes sin procesar; Packt Publishing: Birmingham, Reino Unido, 2019.
51. Koutra, D.; Faloutsos, C. Minería de grafos individual y colectiva: principios, algoritmos y aplicaciones; Springer Nature: Berlín/Heidelberg, Alemania, 2022.
52. Needham, M.; Hodler, AE Algoritmos gráficos: ejemplos prácticos en Apache Spark y Neo4j; O'Reilly Media: Sebastopol, CA, EE. UU., 2019.
53. Chintalapudi, SR; Prasad, MHMK. Un estudio sobre algoritmos de detección de comunidades en redes reales a gran escala. En las Actas de la 2.ª Conferencia Internacional sobre Computación para el Desarrollo Sostenible Global (INDIACom) de 2015, Nueva Delhi, India, 11-13 de marzo de 2015; págs. 1323-1327.
54. Buttler, D. Un breve estudio de los algoritmos de similitud de estructuras de documentos; Informe técnico; Laboratorio Nacional Lawrence Livermore. (LLNL): Livermore, CA, EE. UU., 2004.
55. Lawande, SR; Jasmine, G.; Anbarasi, J.; Izhar, LI Una revisión sistemática y análisis de la búsqueda de rutas basada en inteligencia Algoritmos en el campo de los videojuegos. Appl. Sci. 2022, 12, 5499. [CrossRef]
56. Barabási, AL Red de ciencia. Filos. Trans. R. Soc. Una Matemática. Física. Ing. Ciencia. 2013, 371, 20120375. [Referencia cruzada] [PubMed]
57. Liu, Y.; Safavi, T.; Dighe, A.; Koutra, D. Métodos y aplicaciones de resumen de grafos: Un estudio. ACM Comput. Surv. (CSUR) 2018, 51, 1–34. [Referencia cruzada]
58. Erciyes, K. Redes complejas: una perspectiva algorítmica; CRC Press: Boca Raton, FL, EE. UU., 2014.

59. Cherven, K. Dominando la visualización de redes Gephi; Packt Publishing Ltd.: Birmingham, Reino Unido, 2015.
60. Ward, MO; Grinstein, G.; Keim, D. Visualización interactiva de datos: fundamentos, técnicas y aplicaciones; CRC Press: Boca Raton, Florida, Estados Unidos, 2010.
61. S, S.; Dileep, S.; Manoj, R.; M, A.; Harikumar, S. Comparación de la eficacia de las técnicas de visualización de datos para descubrir relaciones entre enfermedades en un conjunto de datos de red complejo. En las Actas de la 7.ª Conferencia Internacional sobre Tendencias en Electrónica e Informática (ICOEI) de 2023, Tirunelveli, India, 11-13 de abril de 2023; págs. 1486-1492. [CrossRef]
62. Wajahat, A.; Nazir, A.; Akhtar, F.; Qureshi, S.; ullah, F.; Razaque, F.; Shakeel, A. Visualización y análisis interactivo de redes sociales Gephi. En las Actas de la 3.ª Conferencia Internacional sobre Tecnologías de Computación, Matemáticas e Ingeniería (iCoMET), Sukkur, Pakistán, 29-30 de enero de 2020; págs. 1-9. [CrossRef]
63. Chaudhary, A.; Jain, N.; Kumar, A. Herramientas para el análisis y la minería de datos de redes sociales. En las actas de la 11.ª Conferencia Internacional sobre Modelado de Sistemas y Avances en las Tendencias de Investigación (SMART), Moradabad, India, 16-17 de diciembre de 2022; págs. 1063-1067. [CrossRef]
64. Islam, M.; Jin, S. Una visión general de la visualización de datos. En Actas de la Conferencia Internacional sobre Ciencias de la Información y Tecnologías de la Comunicación (ICISCT), Karachi, Pakistán, 9 y 10 de marzo de 2019; págs. 1–7. [CrossRef]
65. OpenStreetMap. 2023. Disponible en línea: https://www.openstreetmap.org (consultado el 14 de septiembre de 2023).