Instituto Tecnológico y de Estudios Superiores de Monterrey

Avance 0:
Propuesta de proyecto y firma de convenios

José Adan Vega Pérez - A01796093

Silvia Xochitl Ibañez Vara - A01795200

Diego Andrés Bernal Díaz - A01795975

Proyecto integrador

Fecha: 25 de Enero del 2025

Índice

Antecedentes.

El análisis de grafos ha sido ampliamente utilizado como una herramienta fundamental para modelar y representar relaciones complejas entre entidades en distintos dominios de información.

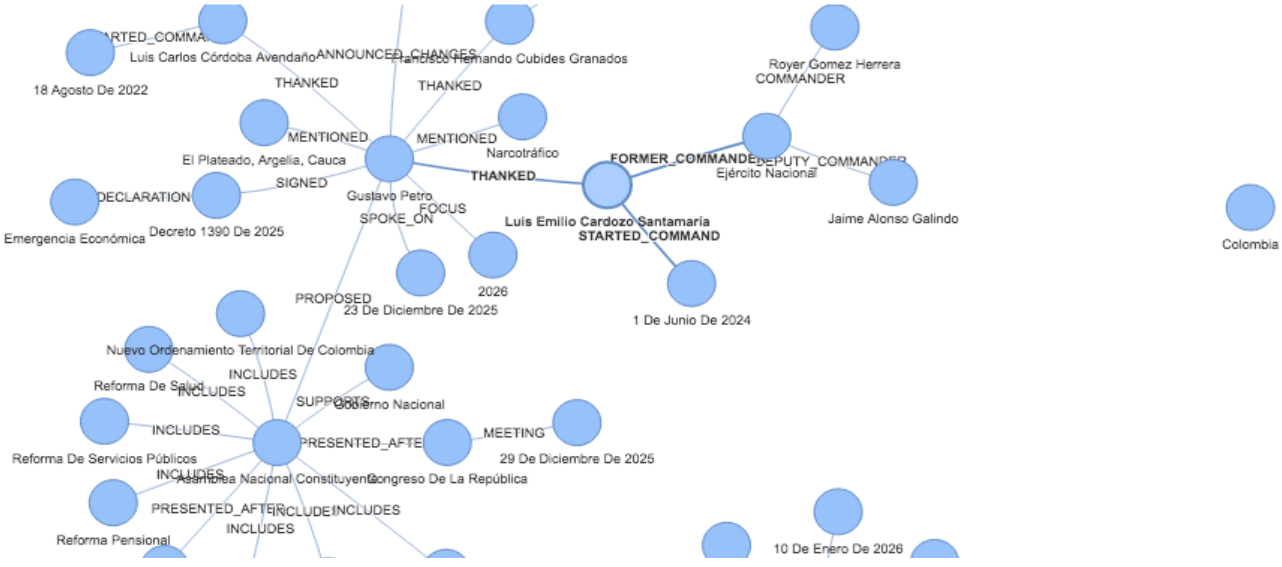
Tradicionalmente, la información se representa en tripletas simbólicas de tipo (sujeto, predicado, objeto); sin embargo, esta naturaleza simbólica presenta limitaciones críticas: los enfoques basados en reglas y símbolos son rígidos ante el ruido, difíciles de manipular matemáticamente de forma directa y carecen de escalabilidad al depender frecuentemente de la curación manual. (Wang et al., 2017)

En particular, los grafos de conocimiento han ganado relevancia al estructurar información en forma de entidades y relaciones, permitiendo tareas como la inferencia y la predicción de enlaces. Tradicionalmente, los métodos de Knowledge Graph Embedding (KGE) han buscado aprender representaciones vectoriales de estos grafos bajo el supuesto de que el conjunto de entidades y relaciones es fijo durante el entrenamiento.

Sin embargo, este supuesto resulta poco realista en escenarios del mundo real, donde la información es dinámica y continuamente emergen nuevos elementos que no estuvieron presentes durante el entrenamiento de los modelos. Ante esta limitación, investigaciones recientes han introducido el concepto de extrapolación de conocimiento, cuyo objetivo es generalizar modelos de grafos para identificar relaciones no explícitas que involucran entidades o relaciones previamente no observadas.

Figura 1:

Ejemplo de grafo de conocimiento (Knowledge graph)



Nota: Esta figura ejemplifica el problema de la información dinámica en los grafos de conocimiento, aunque todas las entidades del grafo hablan del país Colombia, este aparece como una entidad desconectada dado que el texto que lo mencionaba textualmente fue añadido posteriormente. Fuente: Elaboración propia.

Estas aproximaciones incorporan información estructural, textual u ontológica, así como técnicas inductivas basadas en subgrafos, reglas o aprendizaje meta, con el fin de inferir relaciones implícitas más allá del conjunto de entrenamiento. De esta manera, la extrapolación de conocimiento amplía las capacidades del análisis de grafos al permitir la detección de patrones y relaciones en contextos abiertos y evolutivos, incluyendo aquellas relaciones que involucran elementos no presentes durante la fase de entrenamiento. (Chen et al., 2023)

Siguiendo la terminología propuesta en la literatura reciente sobre extrapolación de conocimiento en grafos, es posible distinguir dos escenarios principales: la extrapolación de entidades y la extrapolación de relaciones. La extrapolación de entidades se presenta cuando aparecen entidades no observadas durante el entrenamiento y puede abordarse mediante enfoques que codifican dichas entidades a partir de información estructural del grafo, como sus conexiones o subgrafos de soporte, o bien a partir de información externa, como descripciones textuales u otras fuentes semánticas.

Por su parte, la extrapolación de relaciones se refiere a la aparición de relaciones previamente no observadas y puede resolverse mediante la codificación directa de las relaciones a partir de ejemplos estructurales, o bien a través de la inferencia basada en información adicional, como texto u ontologías. En ambos casos, la información utilizada para la extrapolación puede clasificarse de manera general en información estructural, derivada de la topología del grafo, y en otro tipo de información complementaria, lo que permite a los modelos generalizar y detectar relaciones no explícitas en escenarios abiertos y dinámicos.

El presente proyecto se desarrolla en el contexto de investigación académica del Departamento de computo del Tecnológico de Monterrey sede Santa Fé. La "organización" en este caso actúa como un laboratorio de I+D (Investigación y Desarrollo), cuyo interés principal es la generación de conocimiento científico y la mejora de algoritmos de Inteligencia Artificial para su posterior publicación o transferencia tecnológica. Los procesos de negocio impactados son la investigación en IA y la optimización de sistemas de gestión de conocimiento.

Metodología a emplear

El proyecto se regirá bajo la metodología estandarizada **CRISP-ML(Q)** (CRoss-Industry Standard Process model for the development of Machine Learning applications with Quality assurance), la cual asegura la calidad y robustez del ciclo de vida del modelo. Las fases se abordarán de la siguiente manera:

1. **Entendimiento del Negocio y Datos:** Fase actual donde se define el alcance, la problemática de los grafos dinámicos y la selección de datasets (Freebase/WordNet).
2. **Preparación de los Datos:** Limpieza de tripletas, generación de divisiones inductivas (entrenamiento con entidades vistas, validación/prueba con entidades no vistas) y construcción de grafos de soporte.
3. **Modelado:** Implementación de arquitecturas de redes neuronales de grafos (GNN) y técnicas de extrapolación.
4. **Evaluación:** Medición del desempeño utilizando métricas de clasificación (Accuracy, F1 y AUC) específicamente en el subconjunto de entidades nuevas.
5. **Despliegue (Simulado/Prototipo):** Documentación de la inferencia sobre nuevos nodos sin reentrenamiento.
6. **Monitoreo:** Planteamiento de normas para el análisis de la degradación del modelo ante cambios drásticos en la composición de los datos.

Entendimiento del negocio.

Formulación del problema:

Los modelos tradicionales de Knowledge Graph Embedding **no están diseñados para operar en grafos abiertos y dinámicos**, donde entidades y relaciones nuevas aparecen después del entrenamiento, lo que limita su aplicabilidad en escenarios reales. (Chen et al., 2023), (Hamaguchi et al., 2014), (Nickel et al., 2016)

El problema que se aborda en este trabajo consiste en la identificación de relaciones no explícitas de información dentro de grafos de conocimiento, particularmente en escenarios donde el conjunto de entidades y relaciones no es estático. En estos contextos, los modelos

tradicionales de análisis de grafos y machine learning presentan limitaciones al asumir que todos los elementos del grafo son conocidos durante la fase de entrenamiento, lo cual dificulta la inferencia de relaciones que involucran entidades o relaciones previamente no observadas.

Contexto:

En aplicaciones reales, los sistemas de información basados en grafos, como grafos de conocimiento, bases de datos semánticas o redes complejas, se encuentran en constante evolución. Nuevas entidades y relaciones emergen de manera continua, ya sea por la incorporación de nueva información, la integración de fuentes heterogéneas o la actualización del dominio de conocimiento. Resolver este problema es relevante porque permite extender las capacidades de inferencia de los sistemas basados en grafos, facilitando la detección de patrones y relaciones implícitas sin requerir un reentrenamiento completo del modelo ante cada cambio estructural.

En primera instancia se identifican tres razones que hacen relevante la solución de este problema:

- **Evitar el reentrenamiento costoso:** Los modelos tradicionales son transductivos, lo que significa que asumen un conjunto fijo de elementos y relaciones; para reconocer algo nuevo, requerirían volver a entrenar todo el sistema, lo cual es computacionalmente prohibitivo para modelos de producción. (Hamaguchi et al., 2017)
- **Falla en la generalización:** Los modelos clásicos no pueden mapear elementos que no vieron durante el entrenamiento a una posición adecuada en el espacio vectorial, lo que los deja "ciegos" ante cualquier información emergente. (Daza et al., 2020)
- **Incapacidad de razonamiento en tiempo real:** En escenarios reales (como e-commerce o análisis de noticias), la IA necesita predecir hechos sobre entidades nuevas de inmediato. Sin capacidades **inductivas**, el sistema no puede realizar tareas como la predicción de enlaces en grafos que cambian rápidamente. (Teru et al., 2020), (Wang et al., 2017)

Para resolver esto, ha surgido la **extrapolación de conocimiento**, que utiliza información adicional (como descripciones textuales o estructuras de subgrafos) para generar representaciones de elementos invisibles sin necesidad de reentrenar.

Objetivos:

El objetivo principal de este trabajo es analizar y aplicar técnicas de análisis de grafos y extrapolación de conocimiento para inferir relaciones no explícitas de información. De manera específica, se busca evaluar cómo distintos enfoques pueden generalizar a escenarios abiertos, permitiendo la identificación de relaciones que involucran entidades o relaciones no presentes durante el entrenamiento del modelo.

Objetivos del Negocio (Business Objectives):

- **Reducción de costos computacionales:** Eliminar la necesidad de reentrenar un grafo de conocimiento completo cada vez que ingresa una nueva entidad al sistema.
- **Capacidad de respuesta en tiempo real:** Habilitar la inferencia inmediata sobre información emergente (noticias, nuevos productos) sin latencia de aprendizaje.
- **Escalabilidad:** Permitir que la base de conocimiento crezca indefinidamente sin degradar la operatividad del sistema.

Objetivos del Proyecto/Minería de Datos (Project Objectives):

- **Diseñar e implementar** un modelo de aprendizaje profundo capaz de crear relaciones para nodos no vistos basándose en su vecindario o atributos.
- **Alcanzar métricas de desempeño** en escenarios inductivos comparables al estado del arte (baselines) en datasets como **FB15k-237** (Teru et al., 2020)**, WordNet11 y CoDEx** (Completing Definitions of Entities) (Safavi & Koutra, 2020).
- **Validar la generalización** demostrando que el modelo mantiene su precisión en la predicción de enlaces cuando el % de nodos nuevos aumenta.

Preguntas clave:

Las principales preguntas que guían este trabajo son:

- ¿Cómo pueden los modelos de análisis de grafos inferir relaciones no explícitas en grafos de conocimiento dinámicos?
- ¿Qué tipos de extrapolación de conocimiento resultan más adecuados para manejar entidades o relaciones no observadas?

- ¿Qué impacto tiene el uso de información estructural y de información adicional en la capacidad de generalización de los modelos?
- ¿Hasta qué punto estos enfoques permiten identificar patrones relevantes en contextos abiertos y evolutivos?

Involucrados:

Tabla 1

Involucrados en el proyecto ***Análisis de grafos y extrapolación de conocimiento para identificar relaciones no explícitas de información***

Nombre	Tipo de vinculación al proyecto
Dr. Gerardo Jesús Camacho González	Profesor Investigador
Dr. Eusebio Vargas Estrada	Profesor Investigador
Dr. Raúl Valente Ramírez	Profesor Asesor
José Adán Vega Pérez	Estudiante
Silvia Xóchitl Ibáñez Vara	Estudiante
Diego Andrés Bernal Díaz	Estudiante

Entendimiento de los datos.

Descripción de los datos:

Los datos utilizados en este trabajo se basan en grafos de conocimiento representados mediante tripletas del tipo (entidad origen, relación, entidad destino). Como referencia, se consideran datasets ampliamente utilizados en la literatura para tareas de completado de bases de conocimiento y extrapolación de entidades, tales como **WordNet11** y **Freebase13**, los cuales son subconjuntos de los grafos de conocimiento WordNet y Freebase, respectivamente.

También planteamos utilizar el dataset **CoDEx** (Completing Definitions of Entities) dadas sus propiedades de tamaños variables, y ejemplos falsos difíciles, ideales para entrenar modelos de clasificación robustos.

Estos datasets contienen un conjunto finito de entidades y relaciones, así como particiones de entrenamiento, validación y prueba, y han sido empleados para evaluar tanto escenarios estándar de predicción de enlaces como escenarios donde aparecen entidades no observadas durante el entrenamiento. Su estructura y nivel de complejidad los convierten en un punto de partida adecuado para analizar la inferencia de relaciones no explícitas en grafos de conocimiento dinámicos. (Hamaguchi et al., 2017)

Técnica de ML:

El problema abordado se enmarca principalmente dentro del aprendizaje supervisado, específicamente en tareas de clasificación de tripletas y predicción de enlaces en grafos de conocimiento. No obstante, el enfoque incorpora elementos de aprendizaje inductivo, ya que se busca generalizar el modelo a entidades o relaciones que no estuvieron presentes durante la fase de entrenamiento. Para ello, se consideran técnicas basadas en representaciones vectoriales y redes neuronales aplicadas a grafos, lo que sitúa el problema dentro del ámbito del deep learning sobre estructuras relacionales.

Identificación de las variables:

Entradas:

- Tripletas del grafo de conocimiento que representan relaciones conocidas.
- Información estructural derivada de la conectividad del grafo, como vecindarios o subgrafos asociados a las entidades.
- En escenarios de extrapolación, conjuntos auxiliares de tripletas que conectan entidades nuevas con entidades previamente observadas.

- Salida:
  - Clasificación de tripletas como válidas o no válidas dentro del grafo.
  - Inferencia de relaciones no explícitas, incluyendo aquellas que involucran entidades no presentes durante el entrenamiento del modelo.

Bibliografía.

- Chen, M., Zhang, W., Geng, Y., Xu, Z., Pan, J. Z., & Chen, H. (2023). *Generalizing to unseen elements: A survey on knowledge extrapolation for knowledge graphs*. arXiv. https://arxiv.org/abs/2302.01859
- Daza, D., Cochez, M., & Groth, P. (2020). *Inductive entity representations from text via link prediction*. arXiv. https://arxiv.org/abs/2010.03496
- Hamaguchi, T., Oiwa, H., Shimbo, M., & Matsumoto, Y. (2017). *Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach*. arXiv. https://arxiv.org/abs/1706.05674
- Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE, 104*(1), 11–33. https://doi.org/10.1109/JPROC.2015.2483592
- Safavi, T., & Koutra, D. (2020). *CoDEx: A comprehensive knowledge graph completion benchmark*. arXiv. https://arxiv.org/abs/2009.07810
- Teru, K. K., Denis, E., & Hamilton, W. L. (2020). Inductive relation prediction by subgraph reasoning. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 9448–9457). PMLR. https://arxiv.org/abs/1911.06962
- Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering, 29*(12), 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

Anexos:

Anexo 1 – Documento convenio firmado para la realización de este proyecto:

COMPROMISOS DE COLABORACIÓN PARA EL PROYECTO INTEGRADOR

Escuela de Ingeniería y Ciencias

Maestría en Inteligencia Artificial Aplicada

MNA

GENERALES

El *Proyecto Integrador*, corresponde al proyecto final que se realiza del programa de la Maestría en Inteligencia Artificial Aplicada. Éste se desarrollará y evaluará en la materia con clave TC5035, del mismo nombre.

El objetivo principal del proyecto es resolver una problemática real, vigente, sea del sector industrial o de instituciones gubernamentales, ONG u algún otro tipo de institución. Lo anterior, previa autorización del comité.

Para efectos del presente documento el programa de Maestría en Inteligencia Artificial Aplicada (en adelante MNA) y Tecnológico de Monterrey Campus Santa Fe (en adelante el cliente) reconocen los siguientes compromisos:

COMPROMISOS

Para realizar el proyecto integrador la MNA se compromete a lo siguiente:

1. Asignar un asesor de proyecto con experiencia y vinculación en la problemática presentada por el cliente.
2. Ofrecer las instalaciones y equipo existente requerido que contribuya a resolver la problemática identificada.
3. Proporcionar retroalimentación constante y sistemática durante el registro de la propuesta y desarrollo del proyecto en la materia TC5035.
4. Conformar el comité de evaluación una vez que se hayan alcanzado los objetivos presentados en la propuesta del Proyecto Integrador.

El cliente se compromete a:

1. Asignar a un patrocinador con reconocidas fortalezas en el área de dominio a atender por el proyecto.
2. Retroalimentar a(los) integrante(s) del proyecto con relación a los resultados parciales obtenidos.
3. Asegurar los recursos de información requeridos para la realización del proyecto.
4. Reconocer la participación del estudiante o los estudiantes, así como al Tecnológico de Monterrey, en las publicaciones o reportes que se generen del proyecto.

Ambas instituciones se comprometen a:

1. Guardar la confidencialidad con relación a los resultados obtenidos que sean sensibles/estratégicos para cualquiera de las instituciones.

2. Interacción bidireccional en reuniones de trabajo para optimizar el proceso de implementación del Proyecto Integrador.

La colaboración se realizará, (sin limitarse a ello) para la implementación del Proyecto de Construcción de Grafos de Conocimiento a partir del 12 de enero y concluyendo tentativamente el 27 de marzo del 2026.

Alumno(s) de Proyecto Integrador

Silvia Xochitl Ibañez Vara, A01795200

Nombre, matrícula y firma

José Adán Vega Pérez A01796093

Nombre, matrícula y firma

Diego Andres Bernal Diaz
A01795975

Nombre, matrícula y firma

Por la MNA

Dr. Luis Eduardo Falcón. Director Nacional de la Maestría en Inteligencia Artificial Aplicada

Nombre, puesto y Firma

Patrocinador asignado por el cliente

Dr. Gerardo Jesús Camacho González

Nombre, puesto y Firma