Grafos de Conocimiento y Sistemas de Recomendación para Prevención de Sobrepeso

Miguel Massiris[1] and Claudio Delrieux[2]

[1] CONICET - UNS miguel.massiris@cs.uns.edu.ar
[2] CONICET - UNS cad@cs.uns.edu.ar

Resumen La medicina 4P (predictiva, de precisión, personalizada y participativa) requiere la comprensión exhaustiva diversos factores genéticos, congénitos, sociales, medioambientales y biomédicos, los cuales están interrelacionados de manera compleja. Además, la información relevante en contextos médicos está dispersa en diferentes investigaciones y foros específicos, lo que complica su aprovechamiento para prever y abordar condiciones de salud con intervenciones tempranas. En este trabajo se propone investigar métodos Procesamiento del Lenguaje Natural (NLP) y Grafos de Conocimiento (KG) para inferir, representar y visualizar las complejas relaciones entre el Origen de la Salud y Enfermedad en el Desarrollo (DOHAD) y las trayectorias adversas para la salud, con el objetivo de un ulterior desarrollo de una herramienta con uso clínico en neonatología. En particular, este trabajo se enfoca en el sobrepeso/obesidad infantil, dada su relevancia como uno de los principales riesgos de salud a nivel mundial. Se realizó una búsqueda bibliográfica exhaustiva con la cual se alimentó un modelo de tópicos (Topic Modeling), el cual fue validado por un grupo de especialistas de diversas disciplinas. Con dichos tópicos, se construyó un KG probabilístico, el cual ofrece una comprensión profunda de la etiología de esta condición, así como las posibles intervenciones durante la historia gestacional o los primeros años de vida, con vista a mitigar sus aspectos adversos.

Palabras clave: Grafos de Conocimiento, Procesamiento de Lenguaje Natural, Origen de la Salud y Enfermedad en el Desarrollo (DOHAD).

1. Introducción

En el ámbito médico, la detección temprana de enfermedades y la toma de decisiones informadas son pilares fundamentales para garantizar la salud materno-fetal durante el embarazo. La atención prenatal no solo se centra en el bienestar de la madre, sino también en el desarrollo óptimo del feto y la prevención de posibles complicaciones a largo plazo. En este contexto, surge la importancia de comprender cómo las experiencias de la vida temprana pueden influir en la salud tanto de la madre gestante como del futuro hijo. Es en esta intersección entre el cuidado médico y el impacto de los factores ambientales donde emerge el campo de los Orígenes del Desarrollo de la Salud y la Enfermedad (DOHaD). Esta exploración requiere la integración de vastos conjuntos de datos, como los

conservados por PubMed Central (PMC), un repositorio repleto de conocimientos científicos médicos. Como complemento al enfoque del DOHaD, se recurre al uso de técnicas avanzadas de Procesamiento del Lenguaje Natural (NLP) y Grafos de Conocimiento (KG). Estas herramientas no solo amplían el espectro del análisis, sino que también abren nuevas vías para extraer información práctica de los datos textuales.

2. Materiales y métodos

2.1. Área de estudio y adquisición de datos

El conjunto de datos, compuesto por 2203 artículos, se obtuvo a partir de una búsqueda exhaustiva en el repositorio PubMed Central (PMC) [1]. Esta búsqueda se centró en artículos relacionados con las siguientes palabras clave: SOBREPESO, OBESIDAD y EMBARAZO, con el objetivo de analizar la interacción entre estos factores durante el período gestacional y su impacto en la salud materno-fetal.

2.2. Preprocesamiento y Grafo de Conocimiento

De cada artículo científico se extrajo la siguiente información: título completo, resumen, revista cientifica, identificación única de PubMed (pmid), identificación única de PubMed Central (pmcid), el identificador de objeto digital (doi) y texto completo. Después de recolectar toda esta información, se procedió a preprocesar el texto de cada artículo científico. Para lograrlo se realizó la tokenización del texto, convirtiendo cada palabra en valores numéricos. Luego, se eliminaron todas las palabras vacías (stopwords), que no aportan un significado relevante al análisis. Además, se eliminaron los signos de puntuación, manteniendo únicamente las palabras alfanuméricas.

Una vez completado el preprocesamiento de los datos, se implementan las siguientes técnicas de Procesamiento de NLP para cada artículo:

1. Reconocimiento de entidades nombradas (NER): Es una sub-tarea de extracción de información (análisis de texto) que tiene como objetivo encontrar y categorizar entidades específicas en el texto [2].
2. Word2Vec: Es una técnica popular de NLP que representa palabras como vectores en un espacio de alta dimensión [3].
3. TF-IDF: Es una técnica estadística que refleja la importancia de una palabra en un documento en relación con toda la colección de documentos. Tiene en cuenta tanto la frecuencia de la palabra dentro de un documento específico (TF-Frecuencia de términos) como su frecuencia global en todos los documentos (IDF-Frecuencia inversa de documentos) [4].

Estas técnicas en conjunto permiten abarcar una multitud de escenarios que forman parte deL tratamiento del paciente. Al utilizar el modelo Bio-Epidemiology-NER [5], se pueden identificar hasta 84 entidades biomédicas, cuya importancia

Nombre de la relación	Nodos Participantes
Mentions	Paper, Entity Node
Authorship	Paper, Author
Published in	Paper, Scientific Journal
Co-occurrence	Entity Node, Entity Node
Similar to	Entity Node, Entity Node

Cuadro 1: Relaciones dentro del grafo y sus nodos participantes

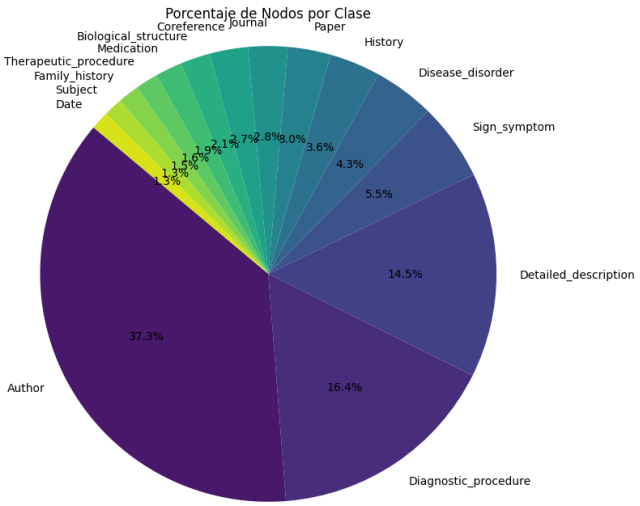

Figura 1: Las 15 clases más relevantes en el KG.

relativa se pondera mediante TF-IDF en cada artículo. Además, la aplicación de Word2Vec facilita la comparación y análisis de similitud entre estas entidades, ofreciendo nuevas oportunidades dentro del grafo. Con el corpus ya procesado se almacena toda esta información en la base de datos orientada a grafos, Neo4J. A partir de esta información, se generan diversos tipos de nodos, como se muestra en la Fig.1 . Dichos nodos se relacionan como lo indica el Cuadro 1.

3. Resultados y Discusión

Como resultado de todas las técnicas aplicadas de NLP se logra un grafo de 77870 nodos de 42 clases distintas y 266464 enlaces de 7 tipos distintos.
Para investigar la obesidad y el sobrepeso, realizamos un análisis de modularidad en el grafo de Neo4j con Gephi, identificando módulos o comunidades de nodos con funciones similares. Este análisis reveló elementos clave en el KG y mecanismos subyacentes de la obesidad y el sobrepeso, utilizando parámetros

como la randomización, pesos en los bordes y una resolución de 1.5, resultando en una estructura modular fuerte y 22 comunidades específicas como se muestra en la Fig.2.

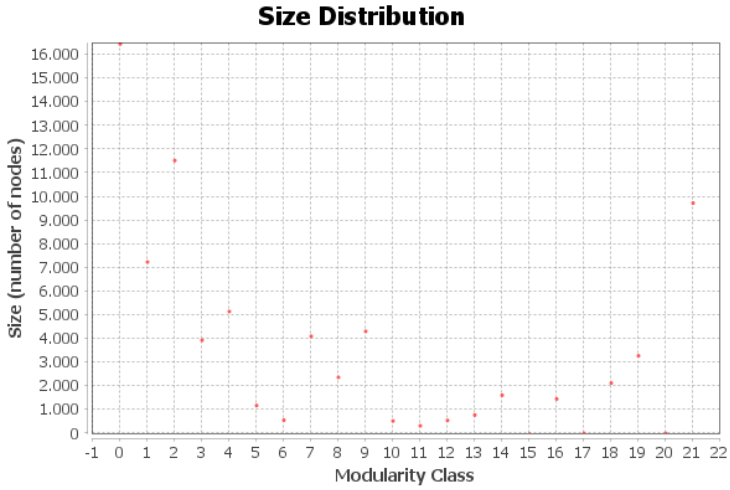

Figura 2: Cantidad de nodos para cada comunidad.

Clase	Palabra 1	Palabra 2	Palabra 3
0	'obesity'	'women'	'overweight'
1	'smoking'	'parents'	'delivery'
2	'diabetes'	'chronic'	'mother'
3	'offspring'	'diet'	'female'
4	'calcium'	'carbohydrates'	'plasma'
5	'twins'	'nicotinamide'	'sirt1'
6	'clinical'	'collagen'	'sathyanarayana sheela'
7	'study'	'rural'	'2013'
8	'dna'	'cpg'	'methylation'
9	'glucose'	'healthy'	'systolic'
10	'tsh'	'thyroid'	'after'
11	'sampling'	'surgery'	'open'
12	'direct'	'project'	'violence'
13	'activity'	'fitness'	'metropolitan'
14	'multiple'	'analyses'	'serum'
15	'transcripts'		
16	'exercise'	'home'	'continuous'
17	'armadillo-repeat'	'plakophilin-2'	'desmosomes'
18	'physiologic'	'kuang alan'	'lowe lynn p.'
19	'pcbs'	'phthalates'	'edcs'
20	'glucosa'	'médicos'	'glucemia'
21	'year'	'year'	'year'

Cuadro 2: Las 3 palabras más relevantes por clase.

Para finalizar se muestran las palabras más relevantes para cada clase en el Cuadro 2.

4. Conclusiones

El KG generado para la obesidad infantil puede aportar información sobre otras enfermedades crónicas y comorbilidades. Esta capacidad de los KG no solo facilita una comprensión más profunda de la salud de un individuo, sino que también promueve una atención médica más efectiva y personalizada. Es por eso que los KG pueden ser herramientas valiosas para la toma de decisiones del personal médico al personalizar el tratamiento de cada paciente, teniendo en cuenta sus características únicas y las interacciones complejas entre diferentes condiciones médicas.

Referencias

1. PubMed Central Homepage. (s. f.). Recuperado de https://www.ncbi.nlm.nih.gov/pmc/
2. Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named Entity Recognition and Relation Detection for Biomedical Information Extraction. In Frontiers in Cell and Developmental Biology (Vol. 8). Frontiers Media SA. https://doi.org/10.3389/fcell.2020.00673
3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.
4. Fan, H., & Qin, Y. (2018). Research on Text Classification Based on Improved TF-IDF Algorithm. In Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018). 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018). Atlantis Press. https://doi.org/10.2991/ncce-18.2018.79
5. Raza, S., Reji, D. J., Shajan, F., & Bashir, S. R. (2022). Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1(12), e0000152. https://doi.org/10.1371/journal.pdig.0000152