Uso del procesamiento de lenguaje natural para derivar un modelo conceptual de Data Warehouses a partir del lenguaje del dominio

Carlos Antonio Carden[1] [0009-0003-8197-2513], Sandro Bimonte[2] [0000-0003-1727-6954], Stefano Rizzi[3] [0000-0002-4617-217X], Leandro Antonelli[1,4] [0000-0003-1338-0337]

[1] LIFIA, Fac. de Informatica, Universidad Nacional de La Plata, La Plata, Argentina
[2] TSCF - INRAE, University Clermont, Aubiere, France
[3] DISI, University of Bologna, Bologna, Italy
[4] CAETI – Fac. de Tecnología Informática - UAI, CABA, Argentina

carlos.carden@lifia.info.unlp.edu.ar, sandro.bimonte@inrae.fr, stefano.rizzi@unibo.it, lanto@lifia.info.unlp.edu.ar

Resumen: Un Data Warehouse (almacén de datos) es un sistema de almacenamiento diseñado para recopilar, organizar y analizar grandes volúmenes de datos provenientes de diversas fuentes. Se utiliza en el análisis de negocios y la toma de decisiones, ya que permite consultar información histórica de manera eficiente. Un Data Warehouse ayuda a las empresas a gestionar, analizar y visualizar datos de manera eficiente, mejorando la toma de decisiones basada en información confiable y consolidada. Uno de los modelos para diseñar un Data Warehouse es el modelo multidimensional, que permite estructurar los datos de manera eficiente para análisis y reportes. Este modelo facilita la navegación y exploración de la información desde diferentes perspectivas, optimizando el rendimiento en consultas analíticas. Aunque constituye un punto de partida valioso, no siempre resulta sencillo obtener una especificación de requerimientos ordenada y consistente, que describa de manera integral toda la funcionalidad del modelo. Este artículo propone una herramienta que permite la derivación del modelo multidimensional utilizando diversas técnicas de procesamiento de lenguaje natural a partir del glosario del dominio LEL.

Palabras claves: Procesamiento de Lenguaje Natural, glosario LEL**,** Data warehouse.

1 Introducción

En la actualidad, las organizaciones enfrentan el desafío de procesar y almacenar cantidades masivas de datos. Para esto es necesario integrar toda la información que manejan en un único modelo. Uno de los modelos más usados en la industria son los modelos multidimensionales que son utilizados en diversos sectores: (i) como negocios, para análisis de ventas, marketing, finanzas, etc., (ii) industria, para optimizar procesos de producción, control de calidad, etc., y (iii) gobierno, para análisis de datos demográficos, económicos, etc.

La construcción de un modelo no es una tarea fácil ya que se debe formar un equipo de trabajo bien coordinado y preparado para la resolución de los problemas de

Proceedings of the 28th Workshop on Requirements Engineering (WER2025), August 20-22, 2025, Rio de Janeiro-RJ, Brazil.
https://doi.org/10.29327/1588952.28-25

manera colaborativa y ordenada, integrando el conocimiento de muchos actores, por lo cual, hay que involucrar a mucha gente siguiendo un plan de trabajo que les asegure que el modelo multidimensional pueda ser el mejor posible para el tiempo determinado de resolución del problema. Una posible estrategia es construir el modelo multidimensional a partir del lenguaje [1]. El lenguaje es una síntesis del conocimiento, por lo tanto, se podría extraer todo ese conocimiento para construirlo.

Este artículo describe una herramienta que implementa el método presentado en [1]. La herramienta toma como input el glosario LEL y propone un modelo que se pueda editar. Al ser el LEL específico del dominio, las técnicas de procesamiento de lenguaje natural serán más precisas en su análisis para la producción final del modelo multidimensional

El resto del paper está organizado de la siguiente manera. La sección 2 proporciona una descripción detallada sobre los modelos multidimensionales, explicando su estructura y sus características. Además, se abordan los diccionarios LELs, su función dentro del sistema y cómo contribuyen a la organización y gestión eficiente de los datos. La sección 3 ofrece una visión completa de la herramienta, describiendo su propósito, funcionalidades y beneficios. También se profundiza en las distintas tecnologías que la conforman, destacando su arquitectura, componentes clave y cómo estas tecnologías interactúan para brindar una solución integral y eficiente. La sección 4 presenta un resumen de los principales logros alcanzados a lo largo del desarrollo y aplicación de la herramienta. Se destacan los avances más relevantes, los beneficios obtenidos y el impacto generado. Finalmente, se incluye un cierre general que sintetiza las conclusiones clave y plantea posibles direcciones para el futuro.

2 Background

2.1 Modelo multidimensional

El modelo multidimensional [8] representa una forma de organizar y visualizar datos que trasciende las limitaciones de las estructuras bidimensionales tradicionales. Fundamentalmente, se define por su capacidad para estructurar la información en más de dos dimensiones, lo que permite un análisis complejo desde múltiples perspectivas simultáneamente.

El modelo multidimensional se basa en los conceptos de hecho, dimensión, jerarquía y medida. Un hecho es un fenómeno relevante que los tomadores de decisiones desean monitorear y analizar (por ejemplo, ventas de automóviles). Las dimensiones actúan como coordenadas para identificar ocurrencias únicas de un hecho (por ejemplo, la fecha de una dimensión A puede describirse con un detalle progresivamente más grueso mediante una jerarquía de niveles categóricos (por ejemplo, la dimensión de tienda incluye niveles de ciudad, región y país); los niveles en una jerarquía forman un árbol enraizado en la dimensión. Los niveles pueden tener propiedades, es decir, atributos (típicamente numéricos) que describen un nivel pero no deben usarse para la agregación (por ejemplo, la capacidad del motor de un modelo). Cada hecho se describe cuantitativamente mediante algunas medidas numéricas (por ejemplo, la cantidad de automóviles vendidos y los ingresos correspondientes). Los posibles valo-

res de los niveles se llaman miembros. La Figura 1 describe el ejemplo mencionado.

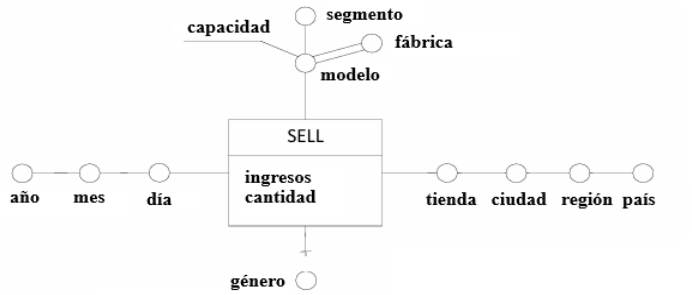


Fig. 1. Ejemplo de modelo multidimensional

2.2 Glosario LEL

El glosario LEL es un glosario que tiene como finalidad registrar la definición de términos (símbolos) que pertenecen a un dominio [2]. Cada símbolo tiene un "nombre" que lo identifica, una "noción" y un "impacto" que lo describen [3] La noción es el significado, la descripción del símbolo. Y el impacto, es la relación entre él y otros símbolos [4]. Cada símbolo del LEL puede pertenecer a una de las siguientes categorías: sujeto, objeto, verbo o estado. Los sujetos son elementos activos dentro del dominio. Los objetos son elementos pasivos en el dominio y son recursos o elementos que los sujetos utilizan. Los verbos son acciones realizadas por sujetos utilizando objetos. Y los estados son situaciones en qué sujetos, objetos o verbos pueden estar involucrados [5]. Cabe señalar que el método implementado por esta herramienta solo utiliza las nociones, sin embargo la tabla 1 indica cómo describir tanto nociones como impactos.

Tabla 1. Categorización y guía para describir los Ejemplos de símbolos del Glosario LEL

Categoría	Características	Noción	Respuestas conductuales
Sujeto	Elementos activos (personas u organización) que realizar acciones	Características o condición que el sujeto satisface	Acciones que realiza el sujeto
Objeto	Elementos pasivos (recursos, herramientas, datos) sobre los que se realizan acciones	Características o atributos que tiene el objeto	Acciones que se realizan en el objeto
Verbo	Acciones que los sujetos realizan en los objetos	Objetivo que persigue el verbo	Pasos necesarios para completar la acción
Estado	Situaciones en las que se pueden ubicar sujetos, objetos o incluso verbos	Situación representada	Acciones que deben realizarse para cambiar a otro estado

2.3 Reglas de derivación para la construcción de un modelo multidimensional

El método [1] consiste en una secuencia de dos pasos, donde el LEL es la entrada y

uno o más esquemas multidimensionales son la salida. Es importante destacar que intervienen dos roles distintos: el usuario final y el diseñador. Los usuarios finales conocen bien el dominio de la aplicación y accederán a los datos basándose en los esquemas multidimensionales; pueden ser analistas, tomadores de decisiones, científicos de datos, etc. Los diseñadores pueden tener un conocimiento nulo o limitado del dominio de la aplicación, pero son expertos en modelado multidimensional.

Los pasos del enfoque son los siguientes. En primer lugar, se aplican ciertas reglas. Este paso se basa en la aplicación de un conjunto de reglas de derivación a un LEL que describe el lenguaje del dominio para obtener elementos para construir uno o más borradores de esquemas multidimensionales. El segundo paso es la revisión. Este paso iterativo se basa en la colaboración entre diseñadores y usuarios finales, de modo que los primeros puedan ajustar los esquemas multidimensionales según las necesidades de los segundos. La tabla 2, muestra las 7 reglas que forman parte del paso 1 de este método y han sido implementadas por la aplicación que describe este paper.

Tabla 2. Reglas de derivación

Regla	Definición
1. Los verbos dan origen a hechos	Sea v un verbo de la LEL, entonces v debe definirse como un hecho.
2. Objetos numéricos y sujetos de verbos dan origen a medidas	Sea v un verbo definido como un hecho según la Regla 1 y n sea su noción. Sea M el conjunto de objetos y sujetos en n que denotan atributos numéricos, entonces todos los elementos en M deben definirse como medidas del hecho correspondientes a v.
3. Objetos categóricos y temas de verbos dar origen a los dímenes siones	Sea v un verbo definido como un hecho según la Regla 1 y n sea su noción. Sea D el conjunto de objetos y sujetos en n que denotan atributos categóricos, entonces los elementos en D deben definirse como dimensiones del hecho correspondientes a v.
4. Objetos categóricos y sujetos de objetos o sujetos dan origen a niveles	Sea o un objeto o sujeto definido como una dimensión (según la Regla 3) o nivel (según la Regla 4) y n sea su noción. Sea L el conjunto de objetos y sujetos en n que no se han definido como dimensiones, denotan atributos categóricos y están relacionados con o por semántica de agregación; entonces los elementos en L deben definirse como niveles de hijos de o.
5. Objetos numéricos y sujetos de objetos o los sujetos dan origen a propiedades	Sea un objeto o sujeto definido como una dimensión o nivel según a las Reglas 3 o 4 y n sea su noción. Sea L el conjunto de objetos y sujetos en n que denotan atributos numéricos, luego los elementos en L debe definirse como propiedades de o.
6. Objetos y sujetos plurales dan origen a múltiples arcos	Sea o un objeto o sujeto definido como una dimensión o nivel, n sea su noción, y o′ un objeto o sujeto en n definido como un nivel infantil de o. Si o′ es plural, entonces el arco de o a o′ es múltiple.
7. Las expresiones de posibilidad en objetos y sujetos determinan arcos opcionales	Sea o un objeto o sujeto definido como una dimensión o nivel, n sea su noción, y o′ un objeto o sujeto en n definido como un nivel infantil de o. Si el verbo usado en n para relacionar o con o' sugiere que algunas instancias de o pueden no estar asociadas a cada instancia de o', entonces el arco de o a o' es opcional.

3 Herramienta

3.1 Arquitectura y diseño

Esta herramienta se concentra en la aplicación de las reglas y proveer una interfase para permitir la edición del modelo resultante. La implementación fue realizada bajo el framework DJANGO de python desde el lado del backend y para las interacciones del usuario se utilizó javascript.

Para cumplir con el objetivo de las reglas se debe llevar a cabo sobre cada notion del Lel un procesamiento que se llevó a cabo con herramientas de procesamiento de lenguaje natural. En python una de las librerías para procesamiento natural más populares es SPACY [6] debido a su facilidad de uso al proporcionar objetos contenedores que representan elementos de textos en lenguaje natural. Estos objetos, a su vez, tienen atributos que representan características lingüísticas, como las partes del discurso. Cuenta con visualizadores integrados que se pueden invocar programáticamente para generar un gráfico de la estructura sintáctica de una oración o de las entidades nombradas en un documento.

Cuando un usuario selecciona los lels para generar el modelo multidimensional y selecciona la opción de generar el modelo, la petición se enviará al servidor, el cual se encarga de procesar los datos y enviar una respuesta al cliente, es decir, el navegador web. Cuando el cliente recibe la respuesta, el gráfico se crea utilizando el lenguaje de programación JavaScript junto a la librería go.js [7]. GoJS es una biblioteca de JavaScript que permite crear fácilmente diagramas interactivos en navegadores web. GoJS admite plantillas gráficas y enlace de datos de propiedades de objetos gráficos a datos del modelo. Solo necesitas guardar y restaurar el modelo, que consiste en objetos JavaScript simples con las propiedades que tu aplicación requiera. Muchas herramientas y comandos predefinidos implementan los comportamientos estándar que la mayoría de los diagramas necesitan. La personalización de la apariencia y el comportamiento se basa principalmente en la configuración de propiedades.

3.2 Funcionamieto

Derivación del modelo multidimensional. El modelo multidimensional está compuesto por diversos elementos fundamentales, entre ellos: hechos, verbos, medidas, niveles, propiedades, arcos múltiples y arcos opcionales. Cada uno de estos elementos desempeña un papel clave en la estructuración y representación de la información dentro del modelo.

Para su correcto procesamiento, se han definido siete reglas que varían en complejidad y en el nivel de análisis del lenguaje requerido. Estas reglas han sido diseñadas para derivar automáticamente todos los elementos del modelo multidimensional (MMD) siempre que su aplicación sea viable dentro de un marco algorítmico. De este modo, el algoritmo garantiza que las reglas se apliquen de manera consistente y precisa, permitiendo una generación estructurada y lógica de los componentes del modelo.Para cumplir con el objetivo de las reglas se debe llevar a cabo sobre cada notion del Lel un procesamiento que se llevará a cabo con herramientas de procesamiento de lenguaje natural.

El procedimiento utilizado para aplicar las reglas toma el glosario LEL como entrada y proporciona un borrador de esquema multidimensional como salida El algoritmo comienza aplicando la Regla 1 para encontrar los hechos Luego, para cada hecho, aplica las Reglas 2 y 3 para encontrar sus medidas y dimensiones. El LEL se navega iterativamente aplicando las Reglas 4 y 5 destinadas a construir jerarquías; las Reglas 6 y 7 se activan para reconocer arcos múltiples y opcionales. Para permitir una aplicación uniforme de la Regla 4, las dimensiones se tratan como niveles. Además,

dado que las propiedades no pueden tener hijos (es decir, siempre son hojas en esquemas multidimensionales), las Reglas 4 y 5 no se les aplican.

Visualización del esquema multidimensional. La visualización y revisión del modelo de datawarehouse es una fase crucial para asegurar que todos los stakeholders comprendan y puedan interactuar con el modelo de datos

La generación de este esquema dependerá de las categorizaciones realizadas desde el backend, las cuales definirán la estructura y disposición de los elementos. La herramienta encargada de representar visualmente dicho esquema es GoJS , que se encargará del dibujo y manipulación de los nodos.

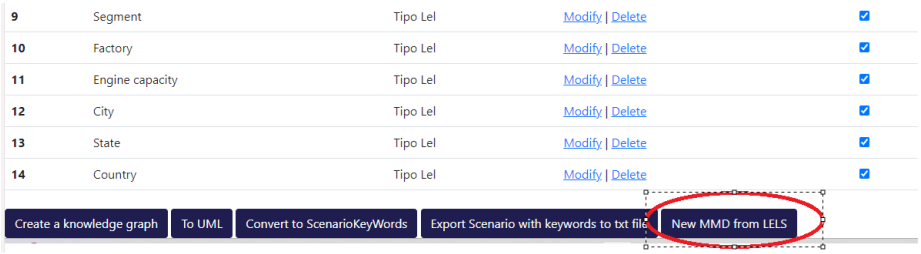
Todo el procesamiento relacionado con la posición de cada nodo y sus características es gestionado desde el backend. Por lo tanto, en esta fase, la tarea principal consiste en interpretar cada dato recibido para generar su objeto correspondiente, el cual será manipulado por GoJS para su representación gráfica.

Una vez que el esquema ha sido dibujado y configurado, la herramienta de modelado proporciona diversas funcionalidades, tales como la visualización interactiva, la posibilidad de mover cada uno de sus componentes, la edición de sus propiedades, así como la opción de agregar o eliminar información según sea necesario.

4 Ejemplo

Una vez cargados todos los símbolos del glosario LEL, el usuario podrá seleccionar mediante casillas de verificación (checkbox) aquellos que desee procesar para generar su respectivo modelo multidimensional. Al hacer clic en el botón "New MMD from LELS" (Figura 2), el sistema procesará automáticamente los elementos seleccionados y mostrará en pantalla el modelo multidimensional resultante.

Tabla 3. Símbolo de glosario LEL para derivar el modelo multidimensional de derivación

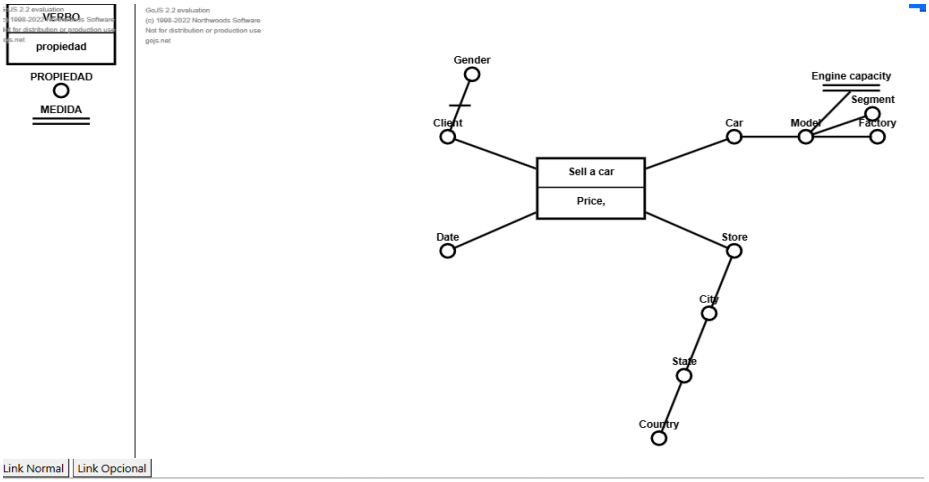


El programa generará el modelo (Figura 3) utilizando Spacy como la biblioteca de procesamiento de lenguaje natural (NLP) para aplicar las distintas reglas establecidas. Mediante el algoritmo predefinido, se creará el modelo multidimensional correspondiente. Adicionalmente, el sistema calculará automáticamente las coordenadas para representar gráficamente cada elemento en el canvas de GoJS.

En esta etapa, el usuario podrá modificar cada uno de los nodos y aristas del modelo generado. Las opciones de edición incluyen la eliminación de elementos existentes o la creación de nuevos componentes con los diferentes tipos de conexiones entre

ellos, ya sea un arco opcional, un arco normal o un arco múltiple.

Tabla 4. Modelo resultante



5 Conclusiones

Como todo proyecto es crucial y decisivo iniciar con una base construida en conceptos sólidos y probados en la industria y en el mundo real. Este enfoque se trata de una filosofía para el encaramiento de los problemas que se presenten y su manejo de manera correcta y ordenada.

Definir el lenguaje de dominio antes de especificar los requerimientos es una manera de hacer frente a este problema. El LEL es un documento que permite definir el lenguaje de un dominio particular. El uso de este documento es de gran apoyo para todas las etapas de la construcción de software, de gran valor por sí mismo para ordenar el conocimiento en común y los distintos términos que puedan aparecer

La herramienta construida permite a los analistas generar un modelo a partir de los distintas definiciones que se dieron en base al glosario técnico manejado en el proyecto que se quiera encarar, permitiéndoles utilizar el modelo generado de manera interactiva para subsiguientes mejoras del mismo

Con la potencia que nos brindan los LELs y utilizando herramientas de procesamiento del lenguaje natural, se pudo generar un modelo multidimensional completo, facilitando su construcción y no metiéndonos en la tediosa tarea de realizarlo a mano y sin la seguridad de tener un modelo completo por la variable del error humano

Pero esto no garantiza que se tenga un modelo sin errores o que no pueda ser mejorado más, por lo tanto, habiendo llegado a un modelo automatizado, se podrán hacer todas las mejoras posibles de manera interactiva por medio de una interfaz simple e intuitiva

Dentro de las principales fortalezas que tiene esta herramienta, es que procesa definiciones comunes para los miembros del proyecto, las procesa de manera coherente

y genera distintos componentes, generando valor en el proceso con un modelo que servirá de base para futuras discusiones entre el equipo que está trabajando en el sistema . Además de poder agregarle valor de una manera amena para el usuario sin que éste sea un experto en el uso de la herramienta.

Por último, se destaca el hecho de que es una ampliación de un proyecto en funcionamiento y construida para poder ser integrado con las herramientas que fueron ya implementadas. Además, se tendrá por trabajo futuro las pruebas evaluativas y validación empírica por parte de los usuarios finales

Referencias

1. Antonelli, L., Bimonte, S., Rizzi, S.: Multidimensional modeling driven from a domain language. Automated Software Eng. 30, 1, https://doi.org/10.1007/s10515-022-00375-5, (2023).
2. Sebastián, A., Hadad, G. D. S., Robledo, E.: Inspección centrada en Omisiones y Ambigüedades de un Modelo Léxico, CIbSE 2017 - XX Ibero-American Conf. Softw. Eng., pp. 71-84 (2017).
3. Leite, J. C. S. do. P., Franco, A. P. M.: A strategy for conceptual model acquisition, en Proceedings of the IEEE International Symposium on Requirements Engineering, , doi: 10.1109/ISRE.1993.324851., pp. 243-246 (1993).
4. Antonelli, L., Rossi, G., Oliveros, A.: A Collaborative Approach to Describe the Domain Language through the Language Extended Lexicon», J. Object Technol., vol. 15, nro 3, p. 3:1, doi: 10.5381/jot.2016.15.3.a3 (2016).
5. Gil, G. D., Figueroa, D. A., Oliveros, A., Producción del LEL en un Dominio Técnico. Informe de un caso, WER00 III Workshop en Ingenieria deRequerimientos, pp. 53-69 (2000).
6. Vasiliev, Y.: Natural Language Processing with Python and SpaCy: A Practical Introduction, No Starch Press, ISBN 978-1718500525 (2020).
7. Go.js, https://gojs.net/. Accedido el 5 de marzo de 2025
   Gallinucci, E., Golfarelli, M., Rizzi, S.: Interactive multidimensional modeling of linked data for exploratory OLAP, Information Systems 77, pp 86–104 (2018).