

## **Servidor de Terminología Médica para el Hospital de Clínicas de Paraguay utilizando Apache Lucene**

Aranda Acuña, Evelyn María<sup>1</sup>

Villalba, Cynthia<sup>2</sup>

Vázquez Noguera, José Luis<sup>3</sup>

<sup>1</sup> Universidad Nacional de Asunción/Facultad Politécnica, San Lorenzo, Paraguay, evearandag@gmail.com

<sup>2</sup> Universidad Nacional de Asunción/Facultad Politécnica, San Lorenzo, Paraguay, cynthiavillalbac@gmail.com

<sup>3</sup> Universidad Nacional de Asunción/Facultad Politécnica, San Lorenzo, Paraguay, joseluaster@gmail.com

**Resumen:** En el Hospital de Clínicas de Paraguay, el proceso de búsqueda de terminologías para la codificación médica en estándares de salud se realiza manualmente. Se propone, como objetivo principal, optimizar el proceso actual de búsqueda a través de la implementación de un servidor de terminología médica utilizando servicios web y una librería de motor de búsqueda de texto. Con este proyecto se pretende: reducir el tiempo de búsqueda de terminologías médicas codificadas; realizar búsquedas de terminologías codificadas a través de términos amigables y comparar el tiempo de respuesta del servidor de terminología contra el tiempo de respuesta de otra herramienta existente, denominada Metamorphosys. Se propone una arquitectura cliente - servidor de tres capas: capa de presentación, negocios y capa de datos. Se eligió utilizar este patrón por la independencia entre las capas y la clara definición de cada una de ellas. El servidor de terminología se encuentra representado en la capa de negocios. Está compuesta por un conjunto de servicios web de tipo REST y una librería de motor de búsqueda de texto, denominada Apache Lucene. Fueron realizados dos experimentos acordes a los objetivos propuestos. El nuevo servidor de terminología responde hasta 19 veces más rápido y resultó ser bastante competitivo contra Metamorphosys. Si bien ambas herramientas presentan un tiempo de respuesta promedio similar, el servidor de terminología es hasta 5 veces más rápido que Metamorphosys en sus valores atípicos. El servidor de terminología implementado reduce el tiempo de búsqueda siendo más rápido que el proceso actual de búsqueda.

**Palabras clave:** servidor de terminología, Metamorphosys codificación médica.

## I. INTRODUCCIÓN

Un servidor de terminología médica es un sistema de software que mapea el texto ingresado, a una lista de terminologías médicas completa, detallada, formal y codificada en estándares. El objetivo principal del servidor de terminología es la representación de datos médicos como datos estructurados, a través de la codificación en estándares médicos, para que puedan ser utilizados en una base de datos para el gestionamiento de la información. En la actualidad existen distintos tipos de terminologías a ser utilizadas dentro de un servidor de terminología, su organización, estructura y granularidad depende de su propósito. Por ejemplo, las terminologías de clasificaciones tienen fines estadísticos, mientras que los vocabularios controlados o terminologías de referencia buscan normalizar el registro clínico. El uso de las terminologías está clasificado en tres principales fases: términos de entrada, terminologías de referencias, y clasificaciones administrativas o estadísticas. Para lograr su propósito, el servidor de terminología se vale de estas tres terminologías que se organizan como capas independientes entre sí como se puede observar en la Fig. 1.

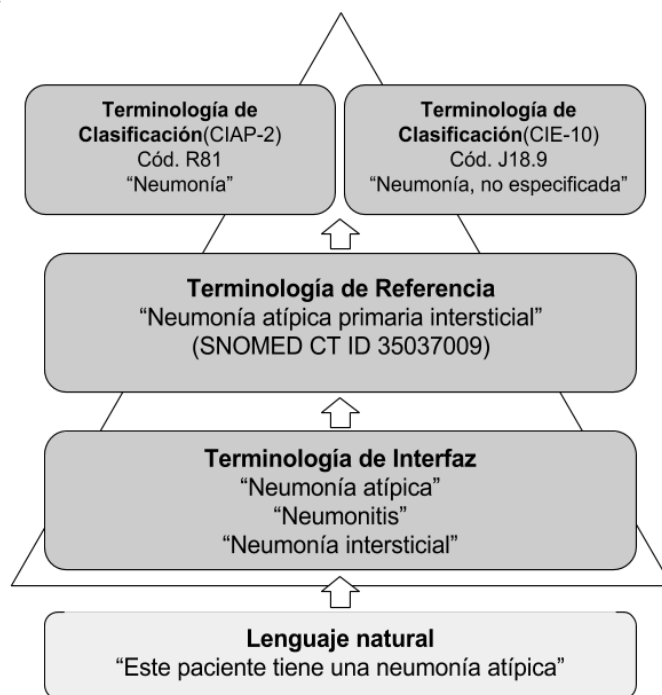


Fig. 1. División de los vocabularios de un servidor de terminología [1].

Las terminologías se encuentran organizadas de esa manera porque las capas inferiores sirven como punto de entrada a las capas superiores. Se parte de términos sencillos (lenguaje normal o natural), hacia términos más estructurados, estandarizados, como los de la terminología de referencia y de clasificación.

Debajo de la pirámide y como punto de partida, se encuentra el lenguaje natural. El lenguaje natural comprende el lenguaje expresado comúnmente, en el ejemplo: "Neumonía atípica", y está muy relacionado a la capa más baja de la pirámide (terminología de interfaz) debido a que ambos utilizan términos del lenguaje común.

La terminología de interfaz representa el dominio y la jerga local, es el lenguaje utilizado por los médicos en el registro, tiene la ventaja de utilizar términos médicos amigables y familiares. Esta terminología comúnmente se puebla con términos utilizados en el lenguaje natural y se enlaza a una terminología médica completa y detallada que se encuentra en la capa superior (terminología de referencia).

En la capa de terminología de interfaz de la Fig. 1, se observa como los médicos utilizan términos comunes y amigables como ``Neumonitis'', ``Neumonía atípica'' y ``Neumonía intersticial'' para referirse a un mismo concepto estandarizado ``Neumonía atípica primaria intersticial'' de la capa superior. Sin la terminología de interfaz que enlaza estos distintos términos comúnmente utilizados, a un mismo concepto formal de referencia, estos términos terminan “sueltos” provocando ambigüedad entre los conceptos. La terminología de interfaz permite contar con una rica sinonimia que permite a los médicos representar datos clínicos utilizando las palabras o frases que prefieran pero haciendo referencia a un mismo concepto de salud.

Las terminologías de referencia son terminologías designadas para proveer representaciones exactas de un dominio de conocimiento dado, típicamente optimizadas para apoyar el almacenamiento y la recuperación de datos clínicos [4]. A menudo poseen mapeos a la terminología de clasificación a falta de la especificidad en la terminología de referencia.

Las terminologías de salida (o de clasificación) permiten, como su nombre lo indica, clasificar los datos como por ejemplo, diagnósticos, enfermedades y otros, para luego poder realizar un análisis sobre los mismos.

Los diagnósticos médicos se codifican diariamente en el Hospital de Clínicas del Paraguay a través de terminologías médicas codificadas en estándares de salud. El Ministerio de Salud Pública y Bienestar Social dictaminó que los diagnósticos médicos fueran codificados utilizando el estándar de Clasificación Internacional de Enfermedades Versión 10 (CIE-10). Este estándar permite obtener información estadística sobre enfermedades o problemas, que pueden ser utilizados para la elaboración de reportes y toma de decisiones en el área de la salud de un país o región. Los médicos utilizan manuales de codificación o el internet de sus teléfonos celulares para la búsqueda de terminologías codificadas en dicho estándar. Este proceso toma mucho tiempo y por ello, como objetivo principal de este trabajo, se pretende diseñar e implementar un servidor de terminología médica ágil que a partir de un texto ingresado en lenguaje común, proporcione un listado de terminologías médicas codificadas en estándares internacionales de salud.

Los objetivos específicos del proyecto son:

1. Reducir el tiempo de búsqueda de terminologías médicas codificadas, respecto al proceso actual.
2. Comparar el tiempo de respuesta del servidor de terminología implementado contra el tiempo de respuesta de otra herramienta existente.

## II. MÉTODO

Se propone una arquitectura cliente - servidor de tres capas (también conocida como arquitectura multinivel), organizada de la siguiente manera: capa de presentación, de negocios y capa de datos. Se eligió utilizar este patrón por la independencia entre las capas y la clara definición de cada una de ellas en cuanto al objetivo que persigue. De esta manera, es posible implementar cada capa de forma total-

mente independiente a las otras. En la Fig. 2. se observa la distribución de los componentes principales del sistema en estas tres capas.

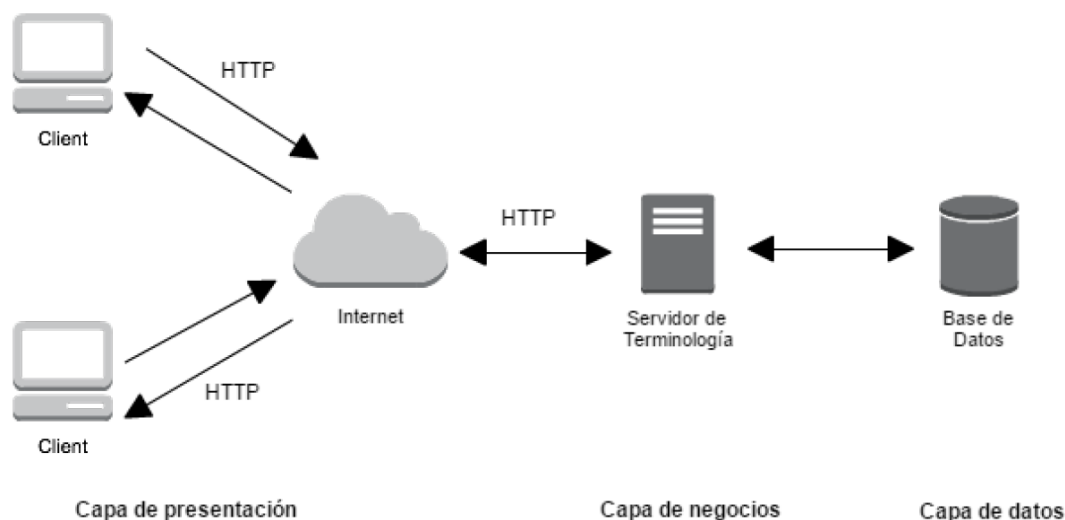


Fig. 2. Arquitectura cliente - servidor de tres niveles.

La capa de presentación es la que se expone del lado del cliente, está compuesta por interfaces que proveen acceso a la capa de negocios.

La capa de negocios está formada por el servidor de terminología (conjunto de servicios web y otras herramientas) y la capa de datos está conformada por las bases de datos. En la Fig. 3. se detalla los componentes de cada capa.

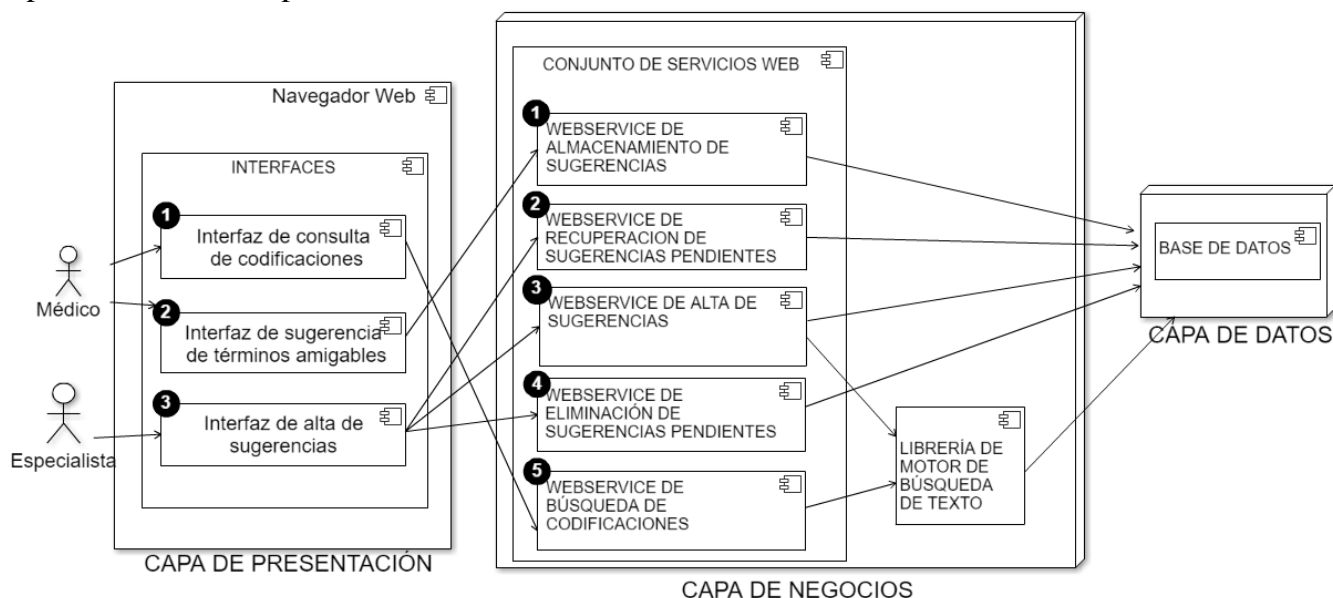


Fig 3. Diagrama de componentes de la arquitectura propuesta.

El servidor de terminología se encuentra representado en la capa de negocios. Está compuesta por un conjunto de servicios web de tipo REST y una librería de motor de búsqueda de texto, denominada Apache Lucene. Lucene está enfocada en el almacenamiento y recuperación de información de manera ágil [2].

Como fuente de información para los datos se utilizó el Metatesauro. El Metatesauro es la integración de

tesauros y ontologías biomédicas desarrolladas independientemente a lo largo de años. Integra cerca de

2.000.000 de nombres, unos 900.000 conceptos de más de 60 familias de vocabularios biomédicos [3].

Forma parte de un sistema unificado de lenguaje médico (UMLS, por sus siglas en Inglés), creado para la

aplicación de sistemas informáticos en la medicina.

### III. EXPERIMENTOS Y RESULTADOS

Fueron realizados dos experimentos acordes a los objetivos específicos mencionados anteriormente.

#### *A. Experimento 1: Medición del tiempo de respuesta del servidor de terminología*

El objetivo de este experimento es medir el tiempo promedio de respuesta del servidor de terminología, ante la búsqueda de diagnósticos, y comparar resultado con el tiempo promedio que toma el proceso actual de búsqueda utilizando el internet de los teléfonos celulares. Para llevar a cabo este experimento, se reunió a 13 médicos residentes (entre R1, R2 y R3) del área de Pediatría del Hospital de Clínicas quienes procedieron a realizar la búsqueda de 5 diagnósticos cada uno, a fin de obtener las terminologías codificadas en estándares. En total, en el servidor de terminología implementado, se buscaron 64 diagnósticos.

En la Tabla 1. se observan los resultados de los tiempos de respuesta.

Tabla 1. Velocidad del servidor de terminología sobre el proceso actual de búsqueda de terminologías codificadas

Tiempo promedio de búsquedas a través del internet de los celulares (en segundos)	Tiempo de respuesta promedio del servidor de terminología (en segundos)	Rapidez del servidor implementado sobre la búsqueda a través del celular
18,37	0,97	$18,37 / 0,97 = 19$ veces más rápido

En la Tabla 1. se puede observar que el servidor de terminología resultó ser hasta 18 veces más rápido que el tiempo que la búsqueda de terminologías codificadas a través del internet de los celulares.

*B. Experimento 2: Comparación del tiempo de respuesta del servidor implementado contra el buscador Metamorphosys*

Se realizó una comparación lo más justa posible contra un buscador de terminologías denominado Metamorphosys. Se tomaron todos los textos ingresados por los médicos en el experimento 1 y se realizaron las mismas búsquedas en Metamorphosys. Se registraron estos tiempos y se compararon contra los tiempos arrojados por el servidor de terminología implementado. Metamorphosys dispone de 4 opciones de búsqueda utilizando 4 algoritmos diferentes. Un algoritmo por opción de búsqueda. Ellos son:

- Coincidencia en la frecuencia más baja (Algoritmo A).
- Descartar coincidencias que solo contienen palabras con mayor frecuencia (Algoritmo B).
- Algoritmo básico de coincidencia (Algoritmo C).
- Coincidencia en al menos dos palabras (Algoritmo D).

Se comparó el tiempo de respuesta del servidor de terminología contra el tiempo de respuesta utilizando cada una de opciones de búsqueda de Metamorphosys mencionadas anteriormente. Los resultados se observan en el diagrama de cajas de la Fig. 3.

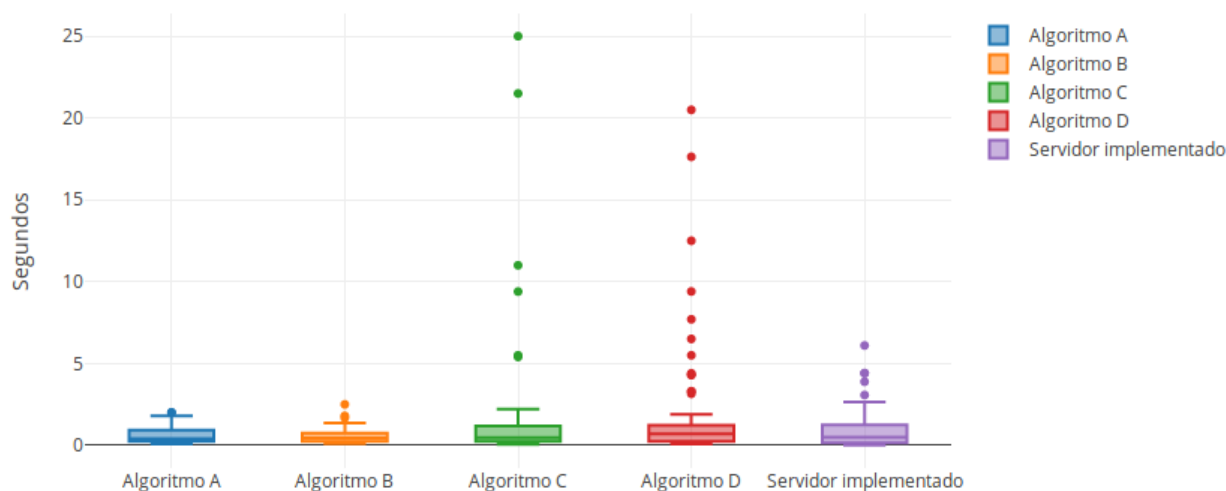


Fig. 3. Diagramas de caja de Metamorphosys y del servidor implementado.

La distribución es asimétrica, todas las cajas presentan sesgo positivo (datos alineados en el extremo inferior en la representación vertical), lo que implica un tiempo de respuesta bajo.

En la Tabla 2. Se aprecia que el servidor de terminología implementado se muestra bastante competitivo contra Metamorphosys al tener prácticamente el mismo tiempo de respuesta en cuanto a la media-

na. En el 75% de los casos (tercer cuartil), tanto Metamorphosys como el servidor de terminología implementado responden en aproximadamente 1 segundo.

Tabla 2. Tiempos de respuesta en la mediana y el tercer cuartil del servidor de terminología implementado y Metamorphosys.

Tiempos de repuesta	Metamorphosys				Servidor implementado	
	Algoritmo A	Algoritmo B	Algoritmo C	Algoritmo D		
	Mediana	0,4	0,4	0,4	0,7	0,5
	Tercer cuartil (75% de los casos)	1	0,7	1	1	1

En la Tabla 3. se observa que el servidor de terminología implementado presenta un valor atípico máximo de 6 segundos. Metamorphosys, en los algoritmos C y D, presenta valores atípicos máximos de 25 y 21 segundos, respectivamente. En estos casos, el servidor de terminología implementado es hasta 4 veces más veloz.

Tabla 3. Valores atípicos de Metamorphosys y el servidor implementado.

Valor atípico máximo (en segundos)	Metamorphosys				Servidor implementado
	Algoritmo A	Algoritmo B	Algoritmo C	Algoritmo D	
	2	2,5	25	21	6

#### IV. CONCLUSIONES

Con la implementación de este trabajo se concluye que el servidor de terminología implementado reduce el tiempo de búsqueda del proceso actual siendo hasta 19 veces más rápido que el proceso actual de búsqueda. Finalmente, ante la comparación del servidor implementado contra el buscador Metamorphosys, el servidor implementado se muestra competitivo contra dicho buscador ya que tienen tiempos de respuesta similares. Cabe mencionar además que Metamorphosys presenta valores atípicos de hasta 25 segundos en algunos casos. El servidor de terminología implementado en este trabajo, sin embargo, presenta un valor atípico máximo de 6 segundos.

#### REFERENCIAS

- (1) Rector, A.L., W.D. Solomon, W.A. Nowlan, T.W. Rush, P.E. Zanstra and W.M. Claassen, A Terminology Server for medical language and medical information systems. *Methods Inf Med*, 1995. 34(1-2): p. 147-57.
- (2) Qian, L., & Wang, L. (2010, June). An evaluation of Lucene for keywords search in large-scale short text storage. In *Computer Design and Applications (ICCD), 2010 International Conference on* (Vol. 2, pp. V2-206). IEEE.

- (3) Schuyler, P. L., Hole, W. T., Tuttle, M. S., & Sherertz, D. D. (1993). The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2), 217.
- (4) Rosenbloom, S. T., Brown, S. H., Froehling, D., Bauer, B. A., Wahner-Roedler, D. L., Gregg, W. M., & Elkin, P. L. (2009). Using SNOMED CT to represent two interface terminologies. *Journal of the American Medical Informatics Association*, 16(1), 81-88.