Artículo original

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

Medical Terminology Server for the Hospital de Clínicas de Paraguay using Apache Lucene

Evelyn María Aranda Acuña [1*]
Cynthia Villalba [1]
José Luis Vázquez Noguera [1]

[1] Universidad Nacional de Asunción, Facultad Politécnica. San Lorenzo, Paraguay.

[*]Autor para la correspondencia: evearandag@gmail.com

RESUMEN

Introducción: En el Hospital de Clínicas de Paraguay, el proceso actual de búsqueda de terminologías para la codificación médica en estándares de salud toma mucho tiempo ya que se realiza manualmente. Se propone, 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.

Método: Se propone una arquitectura cliente - servidor de tres capas (también conocida como arquitectura multi-nivel), 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. 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.

Experimentos y Resultados: Fueron realizados dos experimentos acordes a los objetivos específicos mencionados anteriormente. El servidor de terminología implementado responde hasta 19 veces más rápido que el proceso actual de búsqueda 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.

Conclusiones: El servidor de terminología implementado reduce el tiempo de búsqueda del

proceso actual siendo 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.
Palabras clave: servidor de terminología, Metamorphosys codificación médica.

ABSTRACT

Introduction: In the Hospital Clínicas of Paraguay, the current process of searching for terminologies for medical coding in health standards takes a long time since it is done manually. It is proposed to optimize the current search process through the implementation of a medical terminology server using web services and a text search engine library.

Method: Three layer client-server architecture is proposed (also known as multilevel architecture), organized as follows: presentation layer, business layer and data layer. The use of this pattern was due to its contribution to the independence between the layers and the clear definition of them in terms of the objective pursued. The terminology server is represented in the business layer. It is composed of a set of REST web services and a text search engine library, called Apache Lucene.

Experiments and Results: Two experiments were carried out according to the objective mentioned above. The implemented terminology server responds up to 19 times faster than the current search process and proved to be quite competitive against Metamorphosys. While both tools have a similar average response time, the terminology server is up to 5 times faster than Metamorphosys in their outliers.

Conclusions: The terminology server implemented reduces the search time of the current process being faster than the current search process. Finally, before the comparison of the server implemented against the Metamorphosys search engine, the implemented server is competitive since they have similar response times.
Keywords: terminology server, Metamorphosys medical coding.

Introducción

Un servidor de terminología médica es un sistema de software que mapea un 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 la gestión de la información. En la actualidad existen diferentes tipos de terminologías para 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.

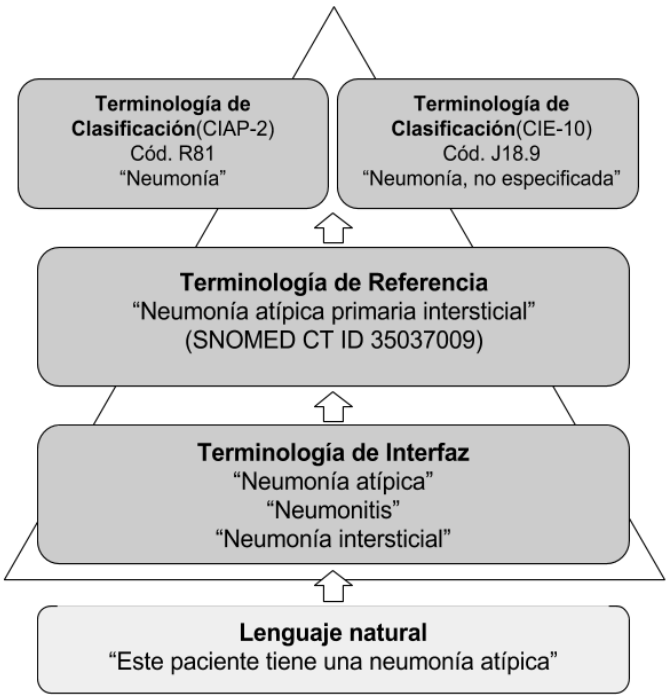

Fig. 1- División de los vocabularios de un servidor de terminología. [(1)]

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 Figura 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. [(2)] 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 son codificados diariamente en el Hospital de Clínicas del Paraguay a través de terminologías médicas codificadas según estándares de salud. El Ministerio de Salud Pública y Bienestar Social dictaminó que los diagnósticos médicos debían ser codificados utilizando el estándar de la Versión 10 de la Clasificación internacional de enfermedades (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 el internet a través de sus teléfonos celulares o manuales de codificación para la búsqueda de terminologías codificadas en dicho estándar. Este proceso toma mucho tiempo y por ello, como objetivo principal nos propusimos diseñar e implementar un servidor de terminología médica ágil que permita a partir de un texto en lenguaje común, proporcionar un listado de terminologías médicas codificadas 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.

Método

Se propone una arquitectura cliente - servidor de tres capas, conocida como multinivel, organizada de la siguiente manera: capa de presentación, capa 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 totalmente independiente a las otras. En la Figura 2 se observa la distribución de los componentes principales del sistema en estas tres capas.

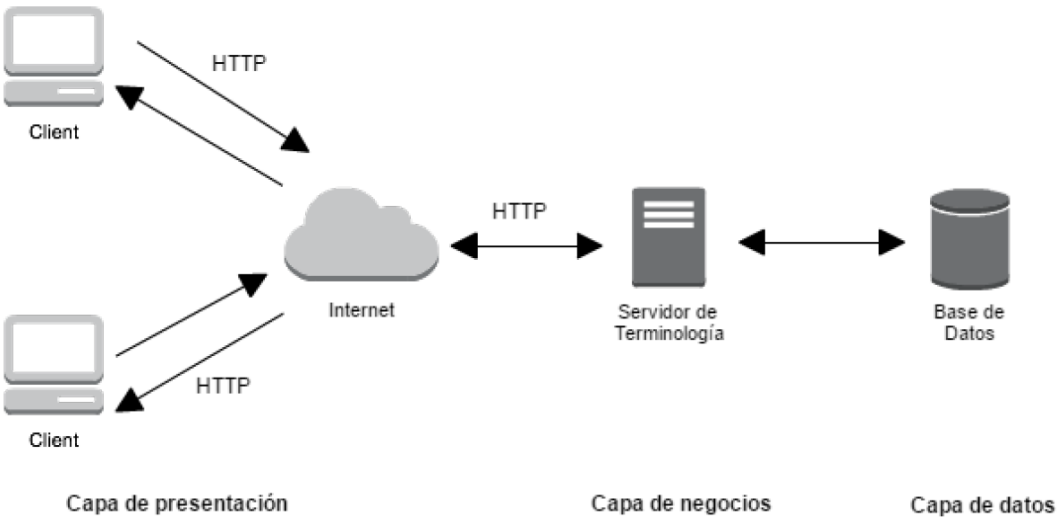

Fig. 2- Arquitectura cliente - servidor de tres niveles.

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

La capa de negocios, por su parte, 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 Figura 3 se detallan los componentes de cada capa.

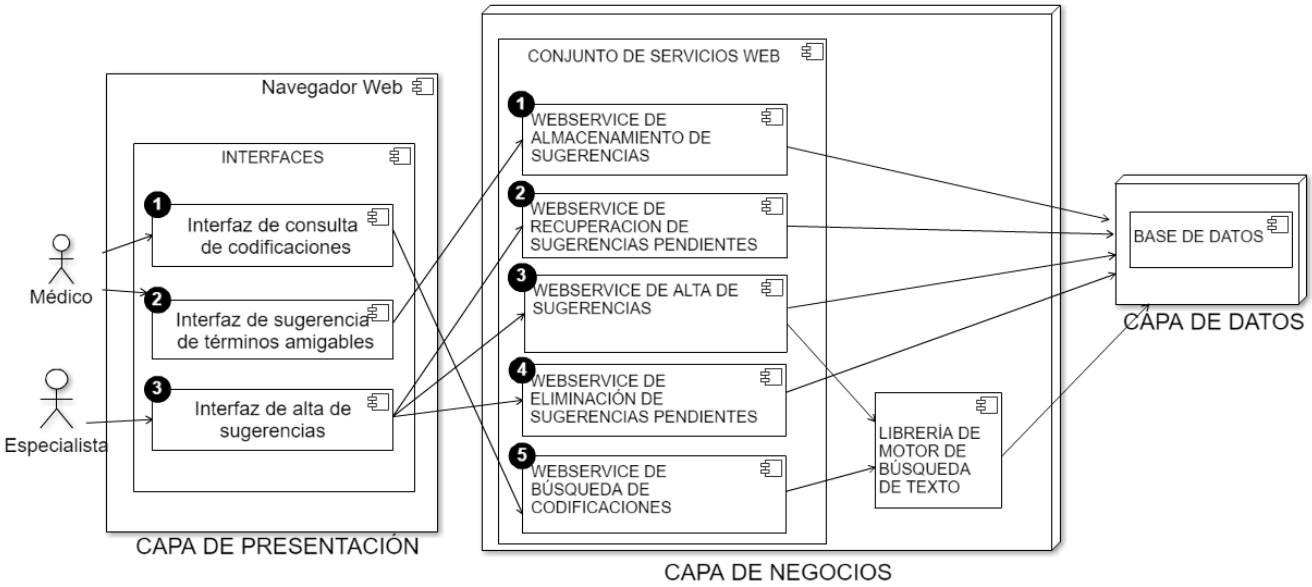

Fig 3- Diagrama de componentes de la arquitectura propuesta.

El servidor de terminologías 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, Apache Lucene, que está enfocada en el almacenamiento y recuperación de información de manera ágil [(3)].

Como fuente de información para los datos se utilizó el Metatesauro. Metatesauro que integra Tesauros y ontologías biomédicas que están desarrolladas independientemente a lo largo de años.

Integra cerca de 2.000.000 de nombres, unos 900.000 conceptos de algo más de 60 familias de vocabularios biomédicos [(4)]. 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.

Resultados

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

Experimento 1: Medición del tiempo de respuesta del servidor de terminologías.

El objetivo de este primer experimento fue 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 terminologías 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 implemetado sobre la búsqueda a través del celular**
18,37	0,97	18,37 / 0,97 = 19 veces más rápido

Como se ve, el tiempo utilizado por el servidor de terminologías resultó ser hasta 18 veces más rápido que el tiempo que tomó la búsqueda a través del internet de los celulares.

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 terminologías

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 implementado contra el tiempo de respuesta usando cada una de opciones mencionadas de Metamorphosys. Los resultados se observan en el diagrama de cajas de la Figura 3.

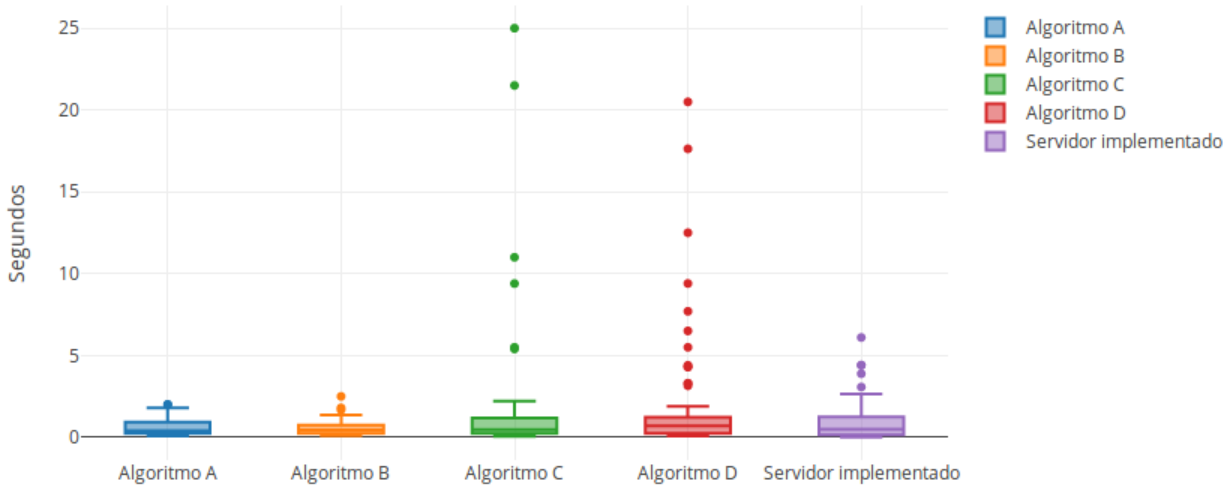

Fig. 3. Diagramas de caja de Metamorphosys y del servidor implementado.

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 puede apreciar que el servidor de terminología implementado se muestra bastante competitivo contra Metamorphosys al tener prácticamente el mismo tiempo de respuesta en cuanto a la mediana. En el 75% de los casos, que constituye el tercer cuartil, tanto Metamorphosys como el servidor implementado responden en aproximadamente 1 segundo.

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

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

Como puede observarse el servidor de terminologías 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 terminologías que se implementó es hasta 4 veces más veloz. (Tabla 3)

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

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

Conclusiones

Con la implementación de este trabajo se concluye que el servidor de terminologías 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, este se muestra competitivo ya que presentan tiempos similares de respuesta. 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 AL, Solomon WD, Nowlan WA, Rush TW, Zanstra PE, Claassen WM. A Terminology Server for medical language and medical information systems. Methods Inf Med. 1995:34(1-2):147-57.

2. Rosenbloom S T, Brown SH, Froehling D, Bauer BA, Wahner-Roedler DL, Gregg WM, Elkin P. Using SNOMED CT to represent two interface terminologies. JAMA. 2009;16(1):81-8.

3. Qian L, Wang L. An evaluation of Lucene for keywords search in large-scale short text storage. Vol. 2, International Conference on Computer Design and Applications 2010: IEEE; 2010: 206 p.

4. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bulletin of the Medical Library Association. 1993;81(2):217.