



Clasificador de perfiles de tratamientos de Diabetes Tipo 2 e Hipertensión Arterial utilizando métodos de recuperación vectorial de información: Caso IPS UNIPAMPLONA

Classifier of profiles of Type 2 Diabetes and Hypertension treatments using vector information retrieval methods: UNIPAMPLONA IPS Case

José Chacón, MgSc, PhD^{1*}, Miguel Vera, MgSc, PhD¹, Pedro Ariza, MgSc², Daniela Beltran, Ing², Yoleidy Huérfano, MgSc⁶, Joselyn Rojas, MD, MgSc^{3,4}, Johel Rodriguez, MgSc¹, Modesto Graterol-Rivas, MgSc, PhD⁷, Víctor Arias, Ing¹, Atilio Del Mar, MD⁸, Daniela Riaño-Wilches, BSc⁵, Julio Contreras-Velásquez, MgSc, PhD(c)¹, Sandra Wilches-Duran, MgSc, PhD(c)¹, Marco Cerda, MgSc¹, Valmore Bermúdez, MD, MPH, MgSc, PhD⁴

¹Grupo de Investigación Altos Estudios de Frontera (ALEF), Universidad Simón Bolívar, Cúcuta, Colombia. E-mail de correspondencia: j.chacon@unisimonbolivar.edu.co

²Grupo de investigación en Ciencias Computacionales CICOM, Universidad de Pamplona, Colombia.

³Pulmonary and Critical Care Medicine Department. Brigham and Women's Hospital. Harvard Medical School. Boston, MA. USA 02115.

⁴Centro de Investigaciones Endocrino-Metabólicas "Dr. Félix Gómez" Facultad de Medicina. Universidad del Zulia, Venezuela.

⁵Facultad de Medicina, Universidad de los Andes, Bogotá, Colombia

⁶Grupo de Investigación en Procesamiento Computacional de Datos (GIPCD-ULA) Universidad de Los Andes-Táchira, Venezuela.

⁷Centro de Estudios de la Empresa. Universidad del Zulia, Venezuela.

⁸Instituto de Bioingeniería y Diagnóstico Sociedad Anónima (IBIDSA), San Cristóbal, Venezuela.

Resumen

E

El objetivo de esta investigación es proponer un modelo que determine el perfil patológico de un paciente en función del tratamiento farmacéutico utilizado y asignado por IPS Unipamplona, determinando el consumo de medicamentos con más frecuencia, utilizando método vectorial de recuperación de información. La metodología utilizada es la recuperación de información vectorial (RIV) y está organizada en dos fases; la primera toma la necesidad de información, realiza la solicitud y esta se convierte en una declaración de búsqueda. En la segunda se tiene la información almacenada representada e indexada de alguna manera. El resultado fue un error del 8 % donde no coincidía con la enfermedad diagnosticada, se recopiló y parametrizó la información para que posteriormente pueda ser utilizada con otro modelo el diseño lógico. El modelo propuesto puede ser aplicado en otros sectores.

Palabras clave: Recuperación de información, modelo vectorial, información biomédica

Abstract

T

The objective of this research is to propose a model to determine the pathological profile of a patient depending on the pharmaceutical treatment used and assigned by IPS Unipamplona determining drug use more often, using vector information retrieval method. The methodology used is the recovery of vector information (RIV) and is organized in two phases; the first takes the need for information, making the request and this becomes a search statement. In the second stored information is represented and indexed in some way. The result was an error of 8% where it did not coincide with the diagnosed disease, the information was compiled and parameterized so that later the logical design could be used with another model. The proposed model can be applied in other sectors.

Keywords: Information Retrieval, vector model, biomedical information

Existe una necesidad mundial por construir sistemas informáticos que sirvan como soporte en la toma de decisiones en las áreas de la salud, debido a la cantidad de datos importantes que estas áreas generan y la recuperación de información en función estos datos para usos específicos, es de vital importancia en el sector salud¹. Las líneas de investigación en el desarrollo de aplicaciones biomédicas tienen como fin: el tratamiento digital de imágenes, interfaces avanzadas y gestión de la información biomédica. Esta línea de investigación tiene un gran interés dentro de los investigadores debido a lo complejo que pueden llegar a ser los datos que estas las representan. Los desarrollos de aplicaciones en gestión de la información biomédica orientadas al sector de la salud cada vez son más llamativos para la solución de problemas en la sociedad lo que refleja el interés y la importancia de crear nuevos sistemas que solucionen problemas en esta línea¹.

La recuperación de información (IR) consiste en encontrar material, por lo general documentos de una naturaleza no estructurada, generalmente texto, que satisfaga una necesidad de información dentro de grandes colecciones de datos almacenada en las computadoras^{2,3}. La IR tiene varios métodos para realizar sus búsquedas, los más utilizados son: Modelo booleano, Modelo probabilístico, Modelo Vectorial, Minería de datos y Big Data³.

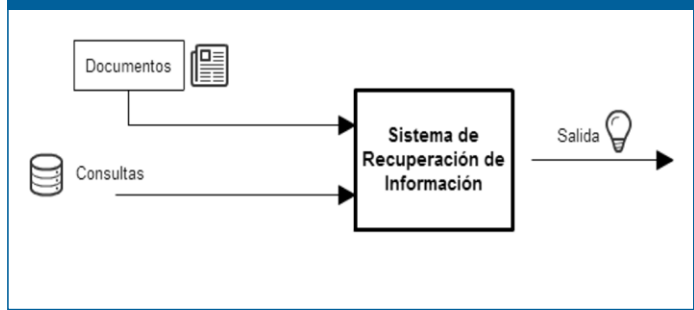
Para este proyecto se utilizó el modelo vectorial siendo el más usado por parte de la IR que permitirá saber que pacientes consumen un determinado medicamento teniendo en cuenta la frecuencia. Está enfocado a crear un software para modelar el perfil del paciente, utilizando recuperación de información. El dominio en el cual se aplica esta investigación será en el área de la salud; la relación entre paciente y medicamentos recetados y como caso de estudio la IPS UNIPAMPLONA.

Recuperación de información

La recuperación de información (IR) puede tener una definición muy amplia. Hechos tan cotidianos como buscar una persona en alguna red social, buscar sobre un personaje reconocido en algún buscador o examinar dentro del ordenador para encontrar un archivo, todo utiliza algún método o algoritmo para recuperar la información requerida⁴.

La recuperación de información se puede explicar con el siguiente problema típico (Figura 1), por un lado tenemos una persona con una necesidad de información, esta necesidad de alguna manera tiene que ser articulada con una solicitud que la describe, la solicitud se convierte a continuación en una declaración de búsqueda; por otro lado, tenemos la información almacenada en las bases de datos y se considera como un recurso potencialmente valioso, esta información debe estar representada e indexada de alguna manera. El reto es proporcionar una buena combinación entre estos dos con el fin de asegurar que la información presentada es de interés para la persona que realiza la consulta original.

Figura 1. Sistema de recuperación de información.



Modelos para la recuperación de la información

El diseño de un sistema de recuperación de la información se realiza bajo un modelo, donde queda definido como se obtienen las representaciones de los documentos y de la consulta, la estrategia para evaluar la relevancia de un documento respecto a una consulta y los métodos para establecer la importancia (orden) de los documentos de salida⁴. Como se describe en la Tabla 1, existen varias propuestas de clasificación de modelos, una de las síntesis más completas la realiza.

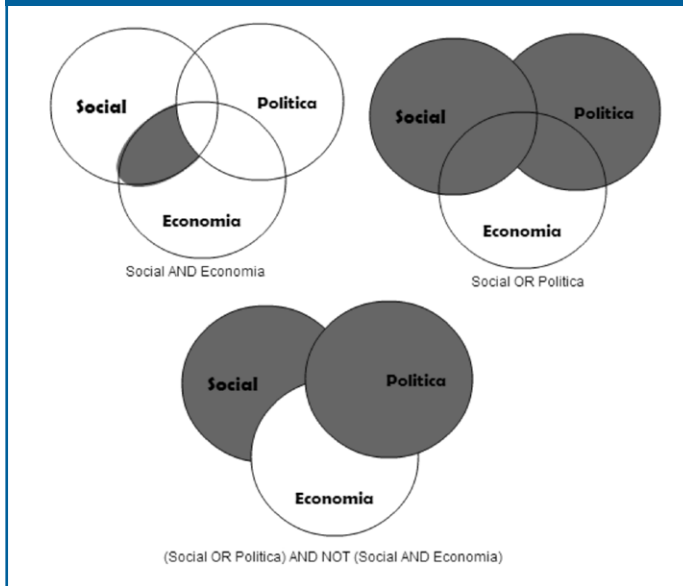
Tabla 1. Clasificación de los Modelos de Recuperación de Información

| Modelo | Descripción |
|---|---|
| Modelos clásicos | Incluye los tres más comúnmente citados: booleano, espacio vectorial y probabilístico. |
| Modelos alternativos | Están basados en la Lógica Fuzzy |
| Modelos lógicos | Basados en la Lógica Formal. La recuperación de información es un proceso inferencial. |
| Modelos basados en la Interactividad | Incluyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la relevancia de los documentos recuperados [SAL, 1989] |
| Modelos basados en la Inteligencia Artificial | Bases de conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural. |

El Modelo booleano constituye el primer modelo teórico, el más antiguo, empleado para establecer el subconjunto de documentos relevantes, en relación a una consulta específica, de entre todos los que configuran la colección, ya se trate del fondo de una biblioteca o de todas las páginas disponibles en la web⁵. Al mismo tiempo es, sin duda, uno de los más sencillos tanto desde un punto de vista teórico como práctico, al basarse en la teoría de conjuntos y en el álgebra de Boole –por una parte- y al ser fácil de diseñar e implementar en la práctica, por otra parte.

Se trata de un modelo basado en la teoría de conjuntos y en la lógica booleana. Las consultas, en este modelo, se especifican por medio de listas de palabras unidas por medio de conectivas lógicas: AND, OR y NOT (Ver Figura 2); de tal modo que se recuperan aquellos documentos que contengan (o no) las palabras indicadas en la consulta, pero de la forma indicada por las conectivas.

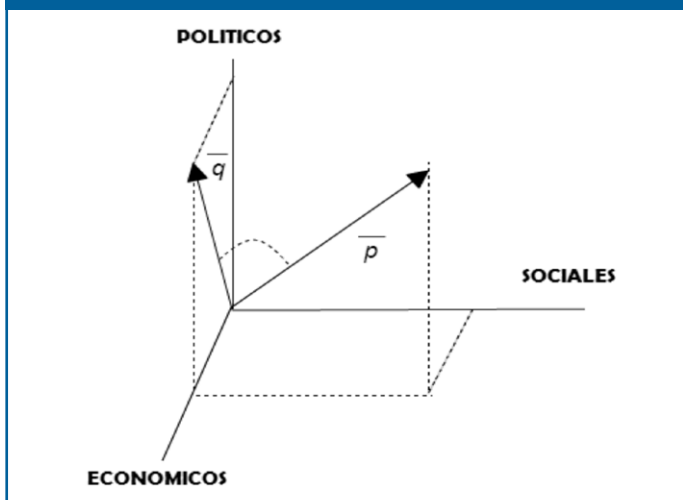
Figura 2. Combinaciones booleanas en conjuntos de diagramas de Venn



El modelo Probabilístico, introducido por Maron y Kuhns, a través de la probabilidad de relevancia, sin embargo fue Stephen Robertson quien convirtió la idea inicialmente en el principio de clasificación de probabilidad introducido en la década de los setenta por Robertson y Sparck Jones, también conocido como modelo de recuperación de independencia binaria (BIR)⁵.

El modelo vectorial es aplicado en operaciones de RI, así como también en operaciones de categorización automática, filtrado de información⁶. En el modelo vectorial se intenta recoger la relación de cada documento D_i de una colección de N documentos, con el conjunto de las m características de la colección. Formalmente un documento puede considerarse como un vector que expresa la relación del documento con cada una de esas características. Tomando el ejemplo del modelo booleano, la Figura 3 muestra un ejemplo de una consulta en el espacio en donde se utilizaron los tres términos: Sociales, económicos y políticos.

Figura 3. Representación de una consulta y documento en el modelo de espacio vectorial



Formalmente un documento puede considerarse como un vector que expresa la relación del documento con cada una de esas características (Ecuación 1).

$$D_i \rightarrow \vec{d}_i = (c_{i1}, c_{i2}, \dots, c_{im}) \quad \text{Ecuación 1}$$

Es decir, ese vector identifica en qué grado el documento D_i satisface cada una de las m características. En ese vector, c_{ik} es un valor numérico que expresa en qué grado el documento D_i posee la característica k . El concepto 'característica' suele concretarse en la ocurrencia de determinadas palabras o términos en el documento, aunque nada impide tomar en consideración otros aspectos. No obstante, una palabra puede aparecer más de una vez en el mismo documento, y además, unas palabras pueden considerarse con más peso, esto es, más significativas que otras, de forma que el valor numérico de cada uno de los componentes del vector obedece normalmente a cálculos más sofisticados que la simple asignación binaria. De otro lado, también es importante normalizar los vectores para no privilegiar documentos largos frente a otros documentos menos extensos. Esto se hace con la Ecuación 2

$$\vec{d}_i = \frac{1}{\sqrt{\sum_{j=1}^m w_{ij}^2}} (w_{i1}, w_{i2}, \dots, w_{im}) \quad \text{Ecuación 2}$$

Si la frecuencia de un término en toda la colección de documentos es extremadamente alta, se opta por eliminarlo del conjunto de términos de la colección (pertenecer al conjunto de palabras vacías). Podría decirse que la capacidad de recuperación de un término es inversamente proporcional a su frecuencia en la colección de documentos. Esto es lo que se conoce como IDF (Inverse Document Frequency). Así, para calcular el peso de cada elemento del vector que representa al documento se tiene en cuenta la frecuencia inversa del término en la colección, combinándola de alguna forma con la frecuencia del término dentro de cada documento. Normalmente se utiliza para ello el producto simple⁴.

Enfermedades

Para poder crear un modelado del paciente necesitamos que tenga dentro del tratamiento una enfermedad asociada en la cual su medicación sea frecuentemente con el fin de generar un perfil más acertado. Dichas enfermedades son crónicas o no tienen cura y el paciente tiene que acostumbrarse a su nuevo estilo de vida, por consiguiente hace que este tenga que recurrir continuamente a puestos de salud para recibir su medicación lo cual también puede representar que el tratamiento no es estable. Algunas enfermedades que cumplen con estas características son: gripa, diabetes tipo 2, hipertensión arterial, sida, cáncer, asma y enfermedad de Creutzfeldt-Jakob.

Todas las anteriores enfermedades tienen un plazo largo para su cura definitiva o simplemente no la tienen. Lo cual hace que las veces en que el paciente recurra al puesto de salud sean mayores y por ende el modelo puede ser más preciso. Las enfermedades que se eligieron fueron la Diabetes tipo 2 y la Hipertensión arterial, puesto que ambas requieren de monitoreo constante y su tratamiento no es tan complejo (ejemplo un sida o cáncer). A su vez son tratadas en la clínica IPS UNIPAMPLONA que suministró los datos.

Para la selección de los medicamentos se tuvieron en cuenta dos variables: 1. Los medicamentos que con mayor frecuencia se encontraban dentro de los tratamientos; 2. La opinión de un especialista.

Dentro de este proceso de elección de medicamentos que se realizó con en base a los datos suministrados por la IPS UNIPAMPLONA, para la enfermedad de Diabetes tipo 2, se listaron los cinco medicamentos correspondientes a la primera variable a tener en cuenta. Cuando se pidió la opinión a los especialistas, se descartaron, porque estas se debían a un problema pasajero no asociado a la enfermedad. Al finalizar, los especialistas indicaron 2 nuevos medicamentos que son relacionados a esta enfermedad pero que en el caso de estudio IPS UNIPAMPLONA no se habían suministrado hasta el momento. Por lo tanto, se contará con siete medicamentos (cinco que se eligieron en base a la primera variable más dos recomendados por el especialista) para generar el perfil del paciente

La diabetes tipo 2, también llamada diabetes mellitus 2, diabetes del adulto, diabetes no insulino-dependiente o sólo diabetes, es un trastorno frecuente que afecta el modo en que el cuerpo procesa y utiliza los carbohidratos, las grasas y las proteínas. Cada uno de estos nutrientes es una fuente de glucosa (azúcar), que es el combustible más básico para el cuerpo⁷. El signo más claro de diabetes es un nivel muy alto de azúcar en sangre. Los medicamentos son: dipirona, insulina zinc, insulina glargina, losartán, torvastatina. Los medicamentos sugeridos por el especialista: metformina, glibenclamida.

La hipertensión arterial es el aumento de la presión arterial de forma crónica. Es una enfermedad que no da síntomas durante mucho tiempo y, si no se trata, puede desencadenar complicaciones severas como un infarto de miocardio, una hemorragia o trombosis cerebral, lo que se puede evitar si se controla adecuadamente⁸. Las primeras consecuencias de la hipertensión las sufren las arterias, que se endurecen a medida que soportan la presión arterial alta de forma continua, se hacen más gruesas y puede verse dificultado al paso de sangre a su través. Esto se conoce con el nombre de arterosclerosis. Las medicinas para esta enfermedad son: betametil digoxina, atorvastatina, clopidogrel, espironolactona, furosemina, amlodipino.

Metodología

El perfil del paciente es construido gracias al modelo vectorial utilizado en la recuperación de información. Dicho modelo nos permite conocer cuáles son los pacientes que más consumen un medicamento. La Figura 4 muestra el proceso para la construcción del perfil de paciente se realizó, en cinco etapas relevantes⁹.

La base de datos IPS UNIPAMPLONA, es el asiento para la construcción del proceso de extracción de la información y modelamiento del perfil del paciente, en base a ésta, se diseñó una nueva base de datos la cual interactúa con la aplicación respectiva para obtener los datos relevantes. Esta base comprende los medicamentos y formulas dirigidas al paciente teniendo en cuenta su diagnóstico, consta de seis tablas: paciente, fórmula médica, fórmula y medicamento, medicamento, diagnóstico asociado y diagnóstico (Ver Figura 5).

Figura 4. Construcción del perfil del paciente

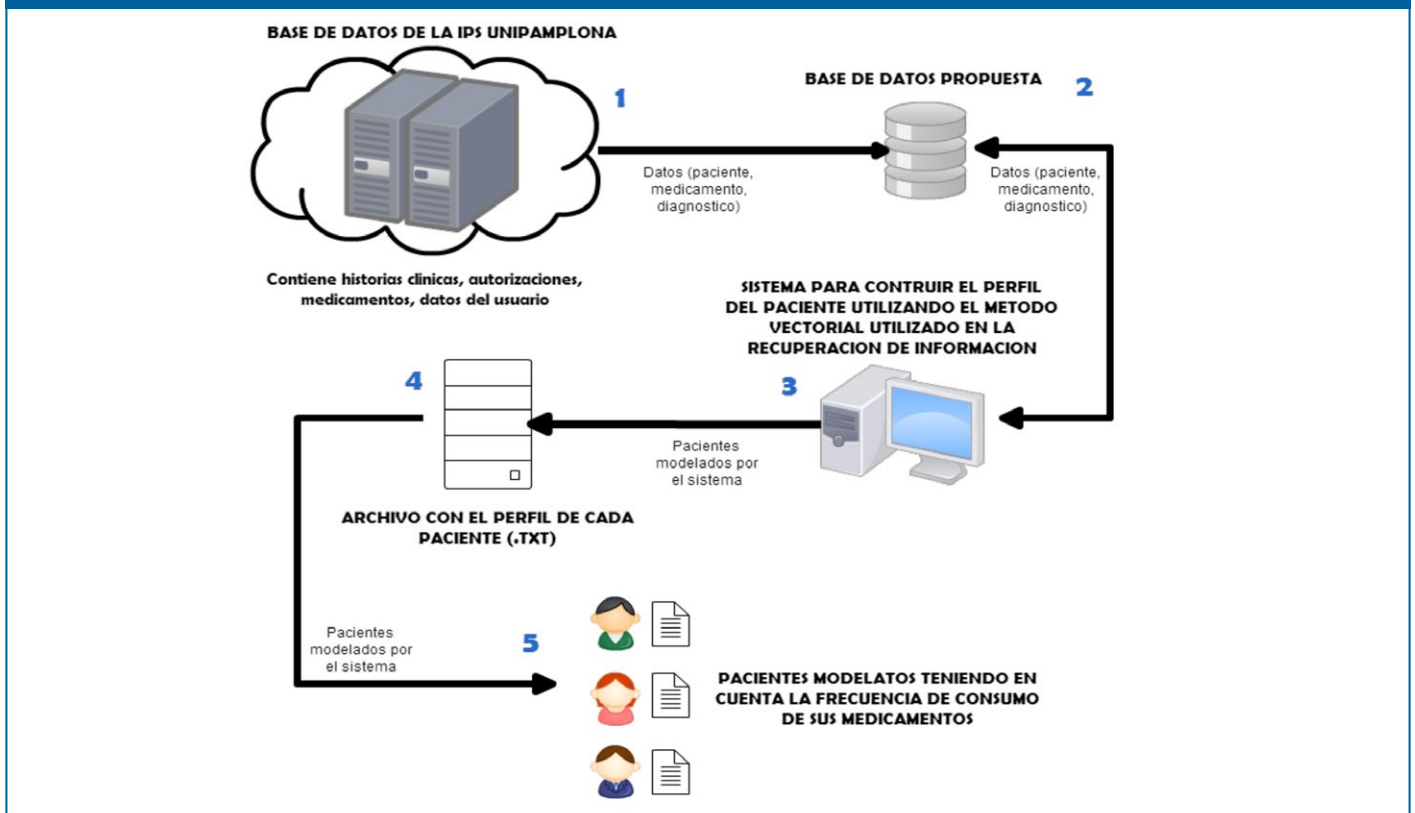
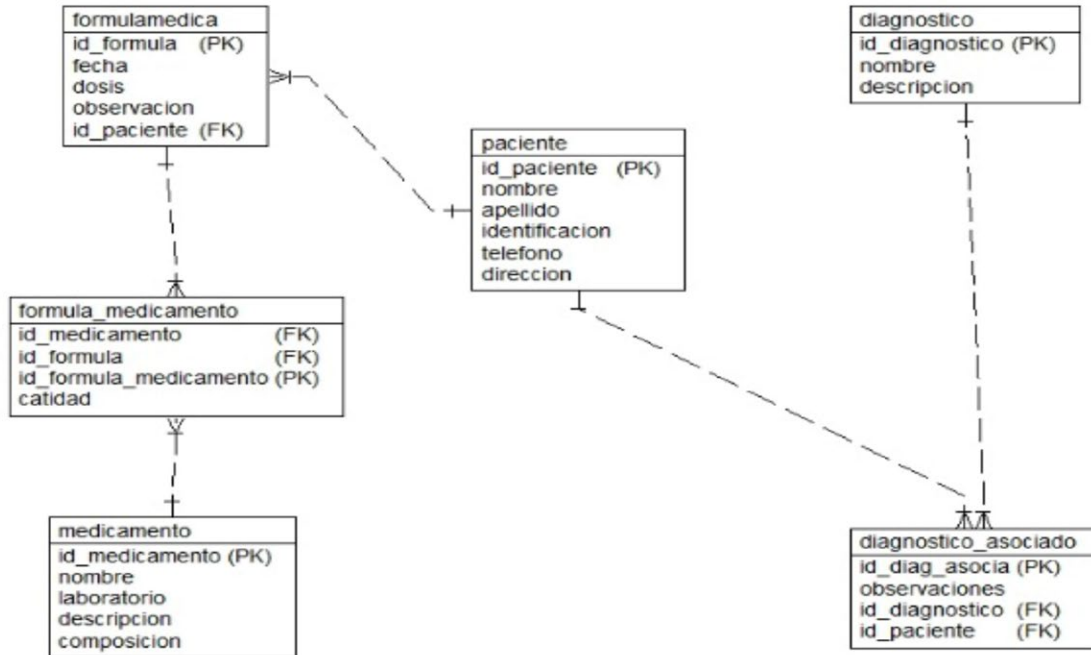


Figura 5. Tablas de la base de datos utilizadas

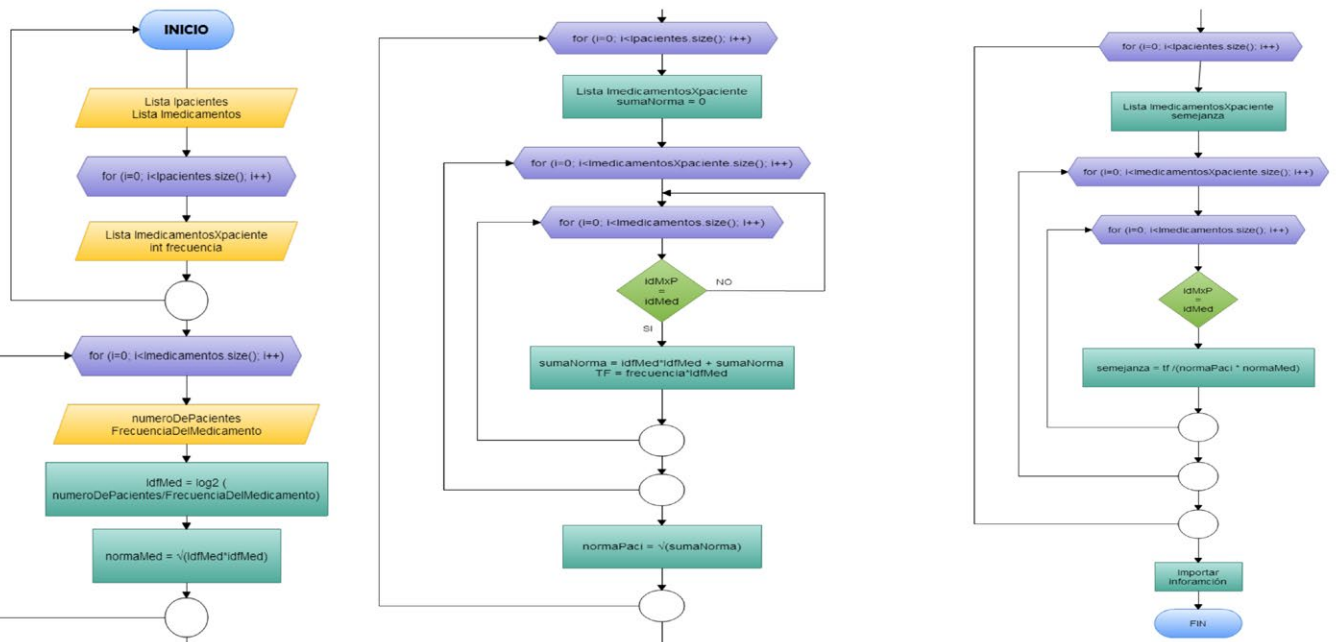


El sistema para modelar el perfil del paciente, se basa en el modelo vectorial utilizado en la recuperación de información, teniendo en cuenta la frecuencia en que los pacientes consumen un medicamento específico asociados a un diagnóstico. Los resultados observados una vez aplicado el modelo se reflejan así:

- a) Que medicamentos consume el paciente. Estos medicamentos se encuentran alojados en la base de datos propuesta correspondientes a cada enfermedad.
- b) Con qué frecuencia consume un medicamento el paciente. Esta frecuencia es el resultado de la semejanza o similitud utilizando el modelo vectorial.

El diagrama de flujo de la aplicación se muestra en la Figura 6.

Figura 6. Diagrama de flujo de la aplicación



Método vectorial aplicado

Se crea una lista de pacientes y medicamentos que se encuentran alojados en la base de datos propuesta. Una vez creadas las listas, se realiza una consulta por cada paciente

para conocer los medicamentos asociados a su diagnóstico y con qué frecuencia fueron suministrados. Lo anterior se denominará matriz de frecuencia^{10,11}. En la Figura 7 y 8, muestra una matriz donde se relaciona, los medicamentos ubicados en la columna y los pacientes en las filas.

Figura 7. Matriz de frecuencia (pacientes vs medicamentos)

| | Medicamentos | | | |
|------------|---------------|---------------|-----|---------------|
| | Medicamento 1 | Medicamento 2 | ... | Medicamento n |
| Paciente 1 | | | | |
| Paciente 2 | | | | |
| ... | | | | |
| Paciente m | | | | |

Cantidad formulada

Figura 8. IDF calculado para cada medicamento

| | Medicamento 1 | Medicamento 2 | ... | Medicamento n | NORMA |
|------------|---------------|---------------|-----|---------------|-------|
| Paciente 1 | | | | | |
| Paciente 2 | | | | | |
| ... | | | | | |
| Paciente m | | | | | |

TF x IDF

Luego se obtiene la tabla de IDF que se consigue mediante la Ecuación 3:

$$IDF = \log_2 \frac{N}{n_i} \quad \text{Ecuación 3}$$

Dónde:
 N = Número de pacientes
 ni = Frecuencia del medicamento entre los pacientes
 Como se muestra en la tabla 2

| Tabla 2. Resultado del perfil de paciente 1 | | |
|---|-----|-----------|
| MEDICAMENTO | COD | SEMEJANZA |
| Clindamicina (Fosfato) | 2 | 0.010178 |
| Dipirona | 3 | 0.018094 |
| Insulina Zinc | 5 | 0.180942 |
| Losartan | 6 | 0.011309 |
| Omeprazol | 17 | 0.002262 |
| Amiitipilina | 19 | 0.001131 |
| Carbamazepina | 23 | 0.005654 |

En la matriz anterior de la Figura 8 se agregó una nueva celda llamada norma, la cual corresponde a la longitud de los vectores y permite normalizar. Usando la siguiente ecuación se obtiene su resultado (Ecuación 4).

$$Norma (paciente_m) = \sqrt{(medicamento_1^2 + medicamento_2^2 + \dots + medicamento_n^2)}$$

Ecuación 4

El medicamento a consultar también debe pasar por un procedimiento similar. Para hallar la norma de cada medicamento se tiene en cuenta la Ecuación 5.

$$Norma (medicamento_m) = \sqrt{IDF^2} \rightarrow Norma (medicamento_m) = IDF$$

Ecuación 5

El modo más simple de calcular la similitud entre un medicamento y un paciente, utilizando el modelo vectorial¹¹ que consiste en realizar el producto escalar de los vectores que los representan esto se muestra en la ecuación 6.

$$sim(C, D_i) = \frac{\sum_{j=1}^t w_{qj} * w_{dij}}{\sqrt{\sum_{j=1}^t (w_{qj})^2 * \sum_{j=1}^t (w_{dij})^2}}$$

Ecuación 6

Dónde:

$D_i = w_{di1}, w_{di2}, w_{di3}, \dots, w_{dit}$ vector que representa el paciente

$C = w_{q1}, w_{q2}, w_{q3}, \dots, w_{qt}$ vector que representa el medicamento, este vector solo contiene el peso de un solo medicamento, ya que se realiza una consulta entre medicamento y paciente.

t: Número de términos en la colección

w_{dij} : Peso del término j en el paciente i

Para comprobar la efectividad del perfil del se tomaron datos de la base de datos pertenecientes a la IPS UNIPAMPLONA. El proceso de crear un diagnóstico a partir de los medicamentos consumidos integra varias tecnologías: La base de datos se realizó en PostgreSQL. La aplicación que realiza el modelado del paciente se hizo en lenguaje JAVA, utilizando como framework NetBeans. La base de datos contiene 10 pacientes (6 pacientes diagnosticados con Diabetes tipo II y 4 con hipertensión) y 25 medicamentos.

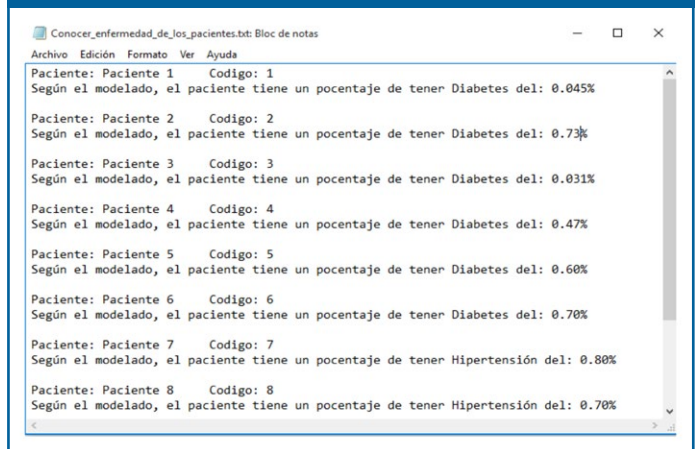
La similitud resultante es un valor entre el rango 0 y 1. En la Figura 9 se muestra la interfaz hombre máquina que inicia la aplicación.

Figura 9. Interfaz de la aplicación



Finalmente, el sistema genera un archivo plano con la información solicitada utilizando el método vectorial como se muestra en la Figura 10.

Figura 10. Archivo lineal generado por la aplicación



Resultados

Perfil del paciente

En la Tabla 2 se muestran los resultados del paciente 1, quien fue diagnosticado con Diabetes tipo II, quien toma los siguientes medicamentos: Clindamicina, Dipirona, Insulina Zinc, Losartan, Omeprazol, Amitiptilina y Carbamazepina. En la columna semejanza podemos observar el resultado de aplicar el modelo vectorial para este paciente

Conclusiones

En este proyecto se aplicó el modelo vectorial como forma para optimizar el manejo de perfil de un paciente teniendo en cuenta los medicamentos y la frecuencia con que son suministrados obteniendo un error del 8% donde no coincidía con su enfermedad diagnosticada. Se recopiló y parametrizó la información para que posteriormente pueda ser utilizada con otro modelo el diseño lógico, se alimentó la base de datos de fuentes reales provenientes de la IPS UNIPAMPLONA la cual contiene historias clínicas, autorizaciones, medicamentos y datos del paciente. El sistema creado en este proyecto puede ser aplicado a otros ámbitos, como comercio electrónico, áreas de educación y otras ciencias (informática, salud, industria).

Referencias

1. Naing L, Winn T, Rusli BN. Practical issues in calculating the sample size for prevalence studies. Archives of orofacial Sciences. 2006; 1(1): 9-14.
2. Raghavan VV, Wong SM. A critical analysis of vector space model for information retrieval. Journal of the American Society for information Science 1986;37(5): 279.
3. Xu J, Croft WB. Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems (TOIS). 2000;18(1):79-112.
4. Dominich, S. A unified mathematical definition of classical information retrieval. Journal of the American Society for Information Science. 2000;51(7):614-624.

5. Martínez-Méndez F. Recuperación de información: modelos, sistemas y evaluación. Murcia: kiosko; 2004.
6. Lindon JC, Nicholson JK. Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. *Annu. Rev. Anal. Chem.* 2008;1:45-69.
7. Bermudez V, Salazar J, Rojas J, Añez R, González R, Torres W, et al. Prevalencia de diabetes mellitus tipo 2 y factores asociados en la ciudad de Maracaibo, Venezuela. *Revista Latinoamericana de Hipertensión.* 2014;9(2):14-25.
8. Rojas M, Rosales Y, Guerrero N, Morillo J, Añez RJ, Bermúdez V, Rojas J. Comportamiento epidemiológico de la hipertensión arterial en individuos adultos del municipio San Cristóbal del estado Táchira-Venezuela. *Revista Latinoamericana de Hipertensión.* 2015;10(2): 25-35.
9. Martínez-Comeche J. Los modelos clásicos de Recuperación de información y su vigencia. Madrid: Universidad Complutense de Madrid; 2006.
10. Göker A, Davies J. Information retrieval: searching in the 21st century. India: Wiley; 2009.
11. Baeza-Yates, R., Ribeiro-Neto, B. Modern information retrieval. New York: ACM press; 1999.

Manuel Velasco (Venezuela) **Editor en Jefe** - Felipe Alberto Espino Comercialización y Producción

Reg Registrada en los siguientes índices y bases de datos:

SCOPUS, EMBASE, Compendex, GEOBASE, EMBiology, Elsevier BIOBASE, FLUIDEX, World Textiles,

OPEN JOURNAL SYSTEMS (OJS), REDALYC (Red de Revistas Científicas de América Latina, el Caribe, España y Portugal),

LATINDEX (Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal)

LIVECS (Literatura Venezolana para la Ciencias de la Salud), LILACS (Literatura Latinoamericana y del Caribe en Ciencias de la Salud)

PERIÓDICA (Índices de Revistas Latinoamericanas en Ciencias), REVENCYT (Índice y Biblioteca Electrónica de Revistas Venezolanas de Ciencias y Tecnología)

SCIELO (Scientific Electronic Library Online), SABER UCV, DRJI (Directory of Research Journal Indexing)

CLaCaLIA (Conocimiento Latinoamericano y Caribeño de Libre Acceso), EBSCO Publishing, PROQUEST.



Esta Revista se publica bajo el auspicio del
Consejo de Desarrollo Científico y Humanístico
Universidad Central de Venezuela.

