# An Ontology-Oriented Architecture for Dealing with Heterogeneous Data Applied to Telemedicine Systems (Una arquitectura orientada a la ontología para el tratamiento de datos heterogéneos aplicados a los sistemas de telemedicina)

**Abstracto:**

Las tendencias actuales en medicina con respecto a los problemas de accesibilidad y la cantidad y calidad de la información y la calidad del servicio son muy diferentes en comparación con las décadas anteriores. El estado actual requiere nuevos métodos para abordar el desafío de lidiar con enormes cantidades de datos presentes y en crecimiento en la Web y otras fuentes de datos heterogéneas como sensores y redes sociales y datos no estructurados, normalmente conocidos como big data. Los enfoques tradicionales no son suficientes, al menos por sí solos, aunque se utilizaron con frecuencia en arquitecturas híbridas en el pasado. En este documento, proponemos una arquitectura para procesar grandes datos, incluidas fuentes heterogéneas de información. Hemos definido una arquitectura orientada a la ontología, donde se ha utilizado una ontología central como base de conocimiento y permite la integración de datos de diferentes fuentes heterogéneas. Hemos utilizado métodos de procesamiento de lenguaje natural e inteligencia artificial para procesar y extraer datos en el sector de la salud para descubrir el conocimiento oculto en diversas fuentes de datos. Nuestro enfoque se ha aplicado al campo de la medicina personalizada (estudio, diagnóstico y tratamiento de enfermedades personalizadas para cada paciente) y se ha utilizado en un sistema de telemedicina. Se presenta un estudio de caso centrado en la diabetes para demostrar la validez del modelo propuesto. y tratamiento de enfermedades personalizadas para cada paciente) y se ha utilizado en un sistema de telemedicina. Se presenta un estudio de caso centrado en la diabetes para demostrar la validez del modelo propuesto. y tratamiento de enfermedades personalizadas para cada paciente) y se ha utilizado en un sistema de telemedicina. Se presenta un estudio de caso centrado en la diabetes para demostrar la validez del modelo propuesto.

[***Tema:*** Entornos de inteligencia ambiental con redes inalámbricas de sensores desde el punto de vista de Big Data y ciudades inteligentes y sostenibles](https://ieeexplore.ieee.org/xpl/topics.jsp?isnumber=8274985&punumber=6287639&refinements=SpecialSection:Ambient%20Intelligence%20Environments%20with%20Wireless%20Sensor%20Networks%20from%20the%20Point%20of%20View%20of%20Big%20Data%20and%20Smart%20&%20Sustainable%20Cities)

**Publicado en:**[Acceso IEEE](https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6287639)( Volumen: 6 )

**Página (s):** 41118 - 41138

**Fecha de publicación:** 19 de julio de 2018

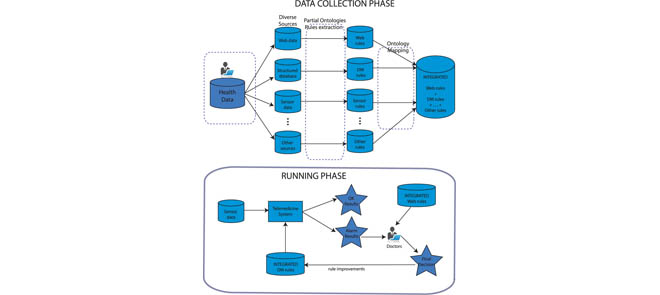
**ISSN electrónico:** 2169-3536

**Número de acceso de INSPEC:** 18035455

**DOI:**[10.1109 / ACCESS.2018.2857499](https://doi.org/10.1109/ACCESS.2018.2857499)

**Editorial:**IEEE

**Agencia fundadora:**



La arquitectura general aplicada a los sistemas de telemedicina. Recolección de datos y fases de ejecución.

Ocultar resumen completo

**SECCIÓN I.**

Introducción y motivación

La asistencia sanitaria ha estado generalmente muy conectada a la tecnología. Sin embargo, esta relación se ha fortalecido en las últimas dos décadas. Una de las principales razones de esto es la proliferación de todo tipo de dispositivos que se pueden instalar fácilmente en la mayoría de los centros de salud. Además, la telemedicina, que se mencionó por primera vez hace varias décadas, ahora es una realidad y se ha desarrollado mucho y esta evolución también se ha extendido a otros sectores de la salud. Además, en los últimos años, los dispositivos móviles pueden obtener una serie de datos biomédicos y se han desarrollado cada vez más aplicaciones relacionadas con la salud [1] - [2] [3]. Además de los teléfonos inteligentes, hay muchos dispositivos que usan sensores para recopilar información de diferentes partes del cuerpo humano. A este respecto, el documento “Infografía: ¿estás listo para la atención médica de los sensores?” Proporciona un enfoque novedoso para la distribución de sensores alrededor del cuerpo para reunir la información más precisa [4] . El mercado y la publicidad de los sensores portátiles se está desarrollando y creciendo rápidamente, como se estudió en el emergente Internet de las cosas, IoT, donde se puede recopilar una gran cantidad de datos en todas partes, todo el tiempo [5] - [6] [7] , que requiere almacenar la información heterogénea generada en soluciones de almacenamiento en la nube como las propuestas en [8] - [9] [10]. Los dispositivos están siendo diseñados para los siguientes propósitos: ayudar a las personas a manejar condiciones particularmente crónicas; recuperarse más rápido de las lesiones, un sector bien investigado en el mercado deportivo; analizar anomalías físicas y ambientales que pueden conducir a problemas de salud más graves; y para detectar hábitos poco saludables antes de que causen problemas, un aspecto se toma muy en serio en el análisis de las condiciones de trabajo según Pathfinder [11] , incluso con aquellos problemas relacionados con la eficiencia energética para las redes de sensores ubicuos [12] .

Las nuevas tendencias en la ciencia cognitiva siguen siendo un desafío, ya que requieren un alto grado de interdisciplinariedad. Con los cambios en los modelos de la sociedad y las nuevas tecnologías utilizadas en la telemedicina, así como en todos los sectores de la salud, los métodos tradicionales parecen insuficientes para esta nueva situación. Este nuevo estado requiere el uso de nuevas técnicas para abordar el problema. Aunque los métodos tradicionales siguen siendo valiosos, deben usarse juntos, proporcionando nuevos modelos de arquitectura.

El objetivo principal de este documento es ensamblar una arquitectura compuesta que integre las diversas fuentes de información actualmente disponibles en el sector de la salud. La arquitectura propuesta se centra en el procesamiento de las enormes cantidades de datos presentes y crecientes en la Web y otras fuentes de datos heterogéneas, como sensores y redes sociales y datos no estructurados, normalmente conocidos como Big Data. Hemos utilizado una arquitectura orientada a la ontología, donde la ontología central se ha definido como una base de conocimiento (KB), que permite la integración de datos de las diferentes fuentes. Hemos descrito las fuentes de datos en términos de la ontología central a través de conceptos y relaciones equivalentes (mapeo ontológico) entre la ontología central y las ontologías de dominio especializado de las diferentes fuentes.

Nuestra propuesta se ha aplicado al campo de la medicina personalizada, en el que el tratamiento de enfermedades se personaliza para cada paciente, y se ha utilizado en un sistema de telemedicina. Finalmente, presentamos un estudio de caso basado en el tratamiento de la diabetes para demostrar la validez del modelo propuesto.

Las contribuciones de nuestro artículo se pueden resumir de la siguiente manera:

* Una arquitectura que procesa cantidades masivas de datos de diversos tipos de fuentes (fuentes de datos heterogéneas). Además, permite la integración a nivel semántico de fuentes heterogéneas de información.
* Se ha definido una arquitectura orientada a la ontología donde la ontología central se ha utilizado como una KB que permite la integración de datos de las diferentes fuentes.
* Los métodos de procesamiento de lenguaje natural e inteligencia artificial se han utilizado para procesar y extraer datos en el sector de la salud para descubrir el conocimiento de diversas fuentes de datos.
* Aplicación del enfoque al campo de la medicina personalizada (estudio, diagnóstico y tratamiento de enfermedades personalizadas para cada paciente). Se resolvió un problema de salud real existente.
* El modelo propuesto fue validado a través de un estudio de caso centrado en la diabetes.
* En este estudio de caso se ha presentado un sistema de telemedicina que ayuda a los médicos en el proceso de toma de decisiones en el tratamiento de la diabetes. El sistema permite al médico mejorar las reglas para los procedimientos de tratamiento aprendidos utilizando técnicas de IA.
* El uso de diferentes fuentes de información (información almacenada en bases de datos, información web e información de sensores) permite mejoras (orientadas a la personalización del paciente) en las reglas que se aplicarán a cada paciente específico.

Las principales novedades presentadas son:

* Una arquitectura orientada a la ontología que utiliza una ontología central que permite la comunicación entre diferentes fuentes de datos, cada una con su propia ontología.
* Una mejora de los sistemas tradicionales de IA en el tratamiento de la diabetes. El tratamiento personalizado de cada paciente y la mejora de los sistemas tradicionales de IA han sido posibles al incluir diferentes fuentes de datos.
* La aplicación de la arquitectura propuesta en el sistema de telemedicina para mejorar su rendimiento.

La parte restante del documento se organiza de la siguiente manera: en la Sección II , se resume la literatura científica relacionada con la salud; en la Sección III , presentamos nuestro enfoque, es decir, la arquitectura propuesta; La Sección IV describe la aplicación de nuestra propuesta al estudio de caso del sector salud; finalmente, la Sección V saca las conclusiones relevantes y presenta trabajos futuros.

**SECCION II.**

Lo último

Como se presentó en la sección de motivación anterior, presentamos una arquitectura que integra diferentes fuentes heterogéneas disponibles actualmente en el sector de la salud. En este sentido, como se presenta en el artículo de De Buenaga Rodríguez *et al.*[13] , los últimos avances y descubrimientos en el campo biomédico, tanto en términos de tecnología como de investigación básica, han llevado a un progreso importante en el enfoque y las prácticas clínicas modernas de Medicina Basada en la Evidencia, EBM [14]y medicina personalizada. La medicina personalizada busca identificar terapias personalizadas para proporcionar un tratamiento individualizado seguro y efectivo de pacientes específicos. Para permitir un tratamiento personalizado, primero es necesario hacer un diagnóstico personalizado correcto. Actualmente, este parece ser el marco más adecuado debido a la gran cantidad de información disponible (estudios experimentales, ensayos clínicos, práctica clínica diaria, sensores biomédicos, grandes conjuntos de datos y texto de libre acceso, datos abiertos y vinculados, etc.). Sin embargo, la situación real es que no hay sistemas de información flexibles capaces de proporcionar conocimiento preciso, actualizado e interrelacionado basado en el acceso estratificado a múltiples tipos de fuentes de datos heterogéneas [15]. El manejo adecuado de esta fuente extraordinaria de conocimiento proporcionaría un avance en el diagnóstico correcto y el tratamiento personalizado. Por ejemplo, el proyecto de investigación actualmente en curso e informado en [13] prevé tres escenarios de uso: (i) asistencia a profesionales de la salud durante el proceso de toma de decisiones en entornos clínicos, (ii) acceso a información relevante sobre su estado de salud y crónico dependiente pacientes y (iii) apoyo para la capacitación basada en evidencia de nuevos estudiantes de medicina. Proponen la integración de sensores biomédicos como se detalla en [16] . Además, eso no es todo porque los últimos años se han desarrollado diferentes esfuerzos de investigación en la amplia gama del sector de la salud. Podemos mencionar algunos ejemplos de estos, como los trabajos desarrollados por Medina*et al.*[17] , [18] . Presentan la aplicación de un enfoque lingüístico difuso sobre dispositivos de monitoreo médico en flujos de datos en el marco de desarrollo de un controlador de medicamentos multidosis para la fiebre. Su enfoque define términos lingüísticos difusos en dispositivos de monitoreo médico individuales y aplica un motor de inferencia basado en reglas diseñado para analizar los flujos de datos. Por otro lado, la solución exhibida por Urzaiz *et al.*[19] detalla la combinación de entornos de hardware y software para desarrollar un dispensador automático de medicamentos para pacientes con enfermedad de Alzheimer.

Hemos dividido las diversas fuentes de información en diferentes etiquetas que son: (1) datos no estructurados (datos web) y las técnicas utilizadas para procesarlos (procesamiento del lenguaje natural); (2) información estructurada (bases de datos estructuradas) y los procesos llevados a cabo para extraer información de ellas (Minería de datos); y (3) datos recopilados de sensores y dispositivos móviles (datos de sensores). Vale la pena señalar que, aunque los datos del sensor podrían clasificarse como bases de datos estructuradas, hemos decidido analizarlos por separado (como un nuevo tipo de fuente de datos) debido a sus propias características específicas.

Las siguientes tres subsecciones discuten el estado del arte de los temas mencionados con respecto al sector de la salud: datos web, bases de datos estructuradas y minería de datos, y datos de sensores. Finalmente, la cuarta subsección muestra el estado del arte del problema de la integración de datos.

**A. Procesamiento de datos web y lenguaje natural**

Hoy en día, se utilizan diferentes sistemas y técnicas de procesamiento del lenguaje natural (PNL) en los sectores biomédico y de la salud. Específicamente, Savova *et al.*[20] presentó un análisis de texto clínico y un sistema de extracción de conocimiento para extraer información del texto libre de registros médicos electrónicos. En la tarea de reconocimiento de entidades nombradas, un diccionario que es un subconjunto del Unified Medical Language System, UMLS [21] se utiliza para incluir conceptos SNOMED CT 1 y RxNORM 2 guiados por amplias consultas con investigadores clínicos y profesionales. Del mismo modo, Pivovarov *et al.*[22]presentan un modelo gráfico probabilístico para el descubrimiento a gran escala de modelos computacionales de enfermedades o fenotipos en los que las observaciones se extraen directamente de datos de registros de pacientes heterogéneos (notas, pruebas de laboratorio, medicamentos y códigos de diagnóstico) y Soguero-Ruiz *et al.*[23] desarrolló un sistema de aprendizaje que utiliza la narrativa clínica en forma de texto libre para la predicción de una complicación postoperatoria común (fuga de anastomosis).

El trabajo de investigación desarrollado por Zeng *et al.*[24] presentó una herramienta de extracción de texto de información de salud que se utiliza para extraer hallazgos clave para un estudio de investigación sobre enfermedades de las vías respiratorias. Esta herramienta de PNL también asigna las cadenas de texto a los conceptos UMLS. UMLS [25] es un recurso que integra y distribuye terminología clave, estándares de clasificación y codificación, y recursos asociados para promover la creación de sistemas y servicios de información biomédica más efectivos e interoperables, incluidos los registros electrónicos de salud. UMLS contiene tres fuentes de conocimiento: el Metathesaurus, la Red Semántica y el Léxico Especialista. En particular, el UMLS Metathesaurus [26]se usa para reconocer entidades con nombre médico en el texto usando un proceso similar al presentado por Terol *et al.*[27] .

Wu *y col.*[28] mostró una técnica de PNL que caracteriza las instancias empíricas de las cadenas de términos de Metathesaurus UMLS en un corpus clínico grande e ilustra qué tipos de características de términos son generalizables en las fuentes de datos. El estudio desarrollado por Xu *et al.*[29] analizaron los términos UMLS Metathesaurus analizando sus ocurrencias en más de 18 millones de resúmenes de MEDLINE. Este estudio concluyó con un Metathesaurus UMLS aumentado que potencialmente puede usarse para mejorar la eficiencia y la precisión de las tareas de recuperación de información y PNL basadas en UMLS. Jiang *y col.*[30]desarrolló diferentes enfoques basados ​​en el aprendizaje automático para extraer entidades clínicas (incluidos problemas médicos, pruebas y tratamientos, así como su estado afirmado) de los resúmenes de alta hospitalaria escritos con lenguaje natural. Los estudios desarrollados por Carroll *et al.*[31] concluyó con la portabilidad de un algoritmo de fenotipo publicado para identificar pacientes con artritis reumatoide (AR) de Electronic Health Records (EHR) en tres instituciones con diferentes sistemas de EHR. Llegaron a la conclusión de que los sistemas genéricos UMLS NLP pueden ser suficientes para un buen rendimiento en al menos algunas tareas específicas de identificación de fenotipos. Otro trabajo interesante es el de [32], que extrae automáticamente el conocimiento explícito de las bases de datos en forma de reglas IF-THEN que contienen cláusulas conectadas AND. Este conocimiento se aplica para el diagnóstico en el dominio médico.

Por lo tanto, al analizar todos estos sistemas y técnicas de PNL, podemos concluir que la mayoría de ellos usan UMLS como fuente de conocimiento en los dominios biomédicos y de salud.

**B. Bases de datos estructuradas y minería de datos**

Las bases de datos estructuradas permiten que las técnicas de DM y Machine Learning (ML) se apliquen directamente para extraer el conocimiento relevante. En resumen, DM es el proceso de analizar datos desde varias perspectivas y resumirlos en información útil [33] . La información oculta que se pone a disposición a través de la minería de datos puede beneficiar a la persona involucrada al proporcionar un soporte eficiente para la toma de decisiones. Se afirma que proporcionar apoyo para la toma de decisiones en el sector de la salud puede ayudar a salvar vidas humanas [34] . La asistencia sanitaria basada en datos abre nuevas oportunidades en medicina personalizada, atención preventiva, manejo de enfermedades crónicas y en telemonitorización y manejo de pacientes con dispositivos implantados [35] .

Además, extraer el conocimiento de la información y los datos ha sido el objetivo principal en una gran cantidad de trabajo relacionado con la gestión de datos que se ocupa de varias áreas de aplicación diferentes, incluida la asistencia sanitaria [36] . Las técnicas de aprendizaje automático se están aplicando intensamente al sector sanitario para el análisis predictivo. Estas técnicas se clasifican en tres categorías: (1) regla de asociación, (2) clasificación y (3) agrupamiento [37] . Una encuesta sobre los enfoques de DM para la asistencia sanitaria se puede encontrar en [38] .

Las reglas de asociación se refieren al descubrimiento de relaciones entre elementos. Por ejemplo, pueden descubrir que un conjunto de indicaciones o síntomas ocurren frecuentemente junto con otro conjunto de síntomas. A priori es un método típico en esta categoría. Algunos ejemplos recientes de la aplicación de la regla de asociación DM a diferentes sectores de la salud se pueden encontrar en [33] , [39] y [40] .

La clasificación asigna elementos de datos en una de varias clases predefinidas. Por ejemplo, las reglas de clasificación sobre una enfermedad pueden extraerse de casos conocidos anteriores y luego usarse para diagnosticar nuevos pacientes de la enfermedad en función de sus síntomas. Árboles de decisión [41] , [42] , Redes neuronales artificiales [42] , [43] , Máquinas de vectores de soporte [ **?**], [44] , [45] , Bayesian Networks [46] y Naive Bayes [47] son ejemplos de enfoques de clasificación aplicados a la asistencia sanitaria.

La agrupación reconoce la clase para un conjunto de elementos sin clasificar según sus atributos. Por ejemplo, un conjunto de enfermedades puede agruparse en varios grupos según las similitudes en sus síntomas, y los síntomas comunes de las enfermedades en un grupo pueden usarse para describir o predecir ese grupo de enfermedades. El vecino más cercano a K (K-nn) es uno de los métodos más populares de agrupamiento. Se pueden encontrar ejemplos de aplicación en la asistencia sanitaria en [48] o [49] .

Otras técnicas de inteligencia artificial incluyen métodos evolutivos [50] . Por ejemplo, en [51] , se utiliza un algoritmo genético como parte del proceso para la identificación de cohortes de fenotipos de pacientes, extrayendo datos textuales de registros de salud.

Otro conjunto importante de nuevas técnicas proviene de la minería de datos web. La minería de datos web tiene como objetivo descubrir información útil y conocimiento de la estructura del hipervínculo web, el contenido de la página y los datos de uso [52] . Un ejemplo reciente de su aplicación para la predicción de la enfermedad se puede encontrar en [53] , donde se propone una remodelación del algoritmo HIT.

Los conjuntos de datos utilizados en muchos estudios informáticos de salud actuales se pueden clasificar como Big Data. Big Data se puede definir como que incluye las siguientes cualidades: Volumen, Velocidad, Variedad, Veracidad y Valor [54] . El volumen se refiere al gran tamaño de los conjuntos de datos, la velocidad se refiere a la gran velocidad con la que se incorporan nuevos datos, la variedad se refiere a los diferentes formatos y estructuras utilizados para la representación de datos, la veracidad se refiere a la correspondencia con la realidad y el valor se refiere a la calidad del conjunto de datos en correlación con el resultado previsto.

En el sector de la salud, hay un gran número de grandes conjuntos de datos procedentes, entre otros, de la amplia aceptación de los registros electrónicos de salud (HME) [34] , [55] ; sensores clínicos, algunos de ellos portátiles, utilizados en nuevos sistemas de vigilancia de la salud [56] - [57] [58] ; datos ambientales relevantes para la salud, como la contaminación, el humo del tabaco, el polen, ahora disponibles a través de dispositivos móviles y redes inalámbricas [59] ; redes sociales y servicios en línea [60] , [61] .

Un problema en relación con los conjuntos de datos mencionados es el diseño de arquitecturas y marcos que permiten el procesamiento de registros a un rendimiento razonable de manera escalable. Las soluciones genéricas de big data como el marco MapReduce, los sistemas de archivos distribuidos o las bases de datos NoSQL se han aplicado con éxito a algunos problemas informáticos de salud en diversas propuestas ( [62] - [63] [64] ). En [65] , se describen estas tecnologías y sus posibilidades en el análisis de datos clínicos.

Concluimos esta sección haciendo referencia a un problema adicional causado por la disponibilidad general de una gran cantidad de bases de datos, lo que resulta en la necesidad de integrar o fusionar diferentes bases de datos [66] , [67] . La fusión plantea varios problemas. Por ejemplo, las tuplas duplicadas deben eliminarse y reemplazarse con una sola tupla que represente la información conjunta de las tuplas duplicadas en la mayor medida ( [68] ), lo que puede violar la integridad referencial. Esta fusión también plantea problemas adicionales, tales como errores tipográficos, falta de estandarización y datos faltantes, por lo que es una tarea considerable. El problema es aún mayor cuando se fusionan fuentes de datos heterogéneas [69] . Por ejemplo, Bilke *et al.*[70]y Naumann *et al.*[71] propuso el sistema HumMer para la integración semiautomática de fuentes de datos heterogéneas. Distingue tres pasos en el proceso de integración de datos: coincidencia de esquemas, detección de duplicados y fusión de datos. En [72] se describen problemas de fusión adicionales: cuando los modelos de base de datos son diferentes (por ejemplo, base de datos relacional frente a base de datos orientada a objetos), diferentes esquemas de datos (es decir, diferentes formas de describir datos o diferentes criterios de codificación), dominios no coincidentes, conflictos semánticos, reconciliación semántica, integración de esquemas, etc.

**C. Datos del sensor**

Recientemente, se han desarrollado diferentes enfoques en la comunicación entre sensores biomédicos y teléfonos inteligentes. La arquitectura del sistema desarrollada por Cobelli *et al.*[73] presenta dos medios de comunicación en el marco de la diabetes tipo 1: el primero es la comunicación entre el sensor del monitor continuo de glucosa (CGM) y el teléfono inteligente, y el segundo es la comunicación entre el teléfono inteligente y la bomba de insulina. La propuesta de [74]enfrenta las características críticas del manejo de la diabetes ya que cada diabético es un caso único con necesidades específicas. A través de algoritmos de minería de datos, su propuesta proporciona asesoramiento individual al usuario diabético. Una contribución interesante de su trabajo es que cuando el número de registros no es suficiente para alcanzar conclusiones útiles, su propuesta utiliza un conjunto de reglas lógicas, definidas a partir de protocolos médicos, directivas. Además, en el marco de las mediciones de colesterol, Oncescu *et al.*[75] presentan un sistema que puede usarse para medir y rastrear los niveles de colesterol directamente en un teléfono inteligente a través de la adquisición y el procesamiento óptimos de la imagen. Además, Yi *et al.*[76]muestra una arquitectura de comunicación entre el teléfono inteligente y diferentes sensores, como un sensor de temperatura y una electrocardiografía (ECG). Fortino y Giampa [77], que presentan la Medición continua de la presión arterial, presentan otros enfoques en el escenario de comunicación entre teléfonos inteligentes y sensores biomédicos . Finalmente, mencione el artículo de [78] que presenta un sistema multisensor para mejorar la precisión de la variabilidad de la frecuencia cardíaca potencialmente peligrosa al considerar la información del contexto del paciente.

Con el objetivo de obtener un conjunto de datos de reconocimiento de la actividad humana, el trabajo de investigación previo desarrollado por Reyes-Ortiz *et al.*[79]presenta el conjunto de experimentos que se llevaron a cabo con un grupo de 30 voluntarios dentro de un rango de edad de 19 a 48 años. El grupo realizó un protocolo de actividades compuesto por seis actividades básicas: tres posturas estáticas (de pie, sentado, acostado) y tres actividades dinámicas (caminar, caminar escaleras abajo y subir escaleras). El experimento también incluyó transiciones posturales que ocurrieron entre las posturas estáticas. Estos son: estar de pie, sentarse a estar de pie, sentarse a estar acostado, estar sentado, estar sentado y estar sentado. Todos los participantes llevaban un teléfono inteligente (Samsung Galaxy S II) en la cintura durante la ejecución del experimento. Capturaron la aceleración lineal 3-axial y la velocidad angular 3-axial a una velocidad constante de 50Hz utilizando el acelerómetro y giroscopio integrados del dispositivo. Las señales del sensor (acelerómetro y giroscopio) se preprocesaron aplicando filtros de ruido y luego se tomaron muestras en ventanas corredizas de ancho fijo de 2.56 segundos y 50% de superposición (128 lecturas / ventana). La señal de aceleración del sensor, que tiene componentes gravitacionales y de movimiento corporal, se separó usando un filtro de paso bajo Butterworth en aceleración corporal y gravedad. Se supone que la fuerza gravitacional solo tiene componentes de baja frecuencia, por lo tanto, se utilizó un filtro con una frecuencia de corte de 0.3 Hz. De cada ventana, se obtuvo un vector de 561 características calculando variables del dominio de tiempo y frecuencia. se separó usando un filtro de paso bajo Butterworth en aceleración corporal y gravedad. Se supone que la fuerza gravitacional solo tiene componentes de baja frecuencia, por lo tanto, se utilizó un filtro con una frecuencia de corte de 0.3 Hz. De cada ventana, se obtuvo un vector de 561 características calculando variables del dominio de tiempo y frecuencia. se separó usando un filtro de paso bajo Butterworth en aceleración corporal y gravedad. Se supone que la fuerza gravitacional solo tiene componentes de baja frecuencia, por lo tanto, se utilizó un filtro con una frecuencia de corte de 0.3 Hz. De cada ventana, se obtuvo un vector de 561 características calculando variables del dominio de tiempo y frecuencia.

Los experimentos se grabaron en video para etiquetar los datos manualmente. El conjunto de datos obtenido se dividió aleatoriamente en dos conjuntos, donde el 70% de los voluntarios fueron seleccionados para generar los datos de entrenamiento y el 30% de los datos de la prueba.

Además, los teléfonos inteligentes también se han utilizado para medir la actividad física y los niveles de vida sedentarios de las personas. Se sabe que la actividad física regular ayuda a prevenir y tratar numerosas enfermedades no transmisibles como la diabetes. Se ha demostrado que las aplicaciones de teléfonos inteligentes aumentan la actividad física en la atención primaria. Es un hecho que hoy en día hay muchas aplicaciones que pueden informar sobre el grado de actividad física de los usuarios. Por ejemplo, las nuevas aplicaciones de estilo de vida saludable pueden medir variables como la velocidad al caminar, correr, subir escaleras y la duración de la actividad física.

**D. Integración de datos**

Como se presentó en [80] y [81] , en el pasado se han desarrollado muchas técnicas sobre la integración de datos de diferentes fuentes de datos heterogéneas:

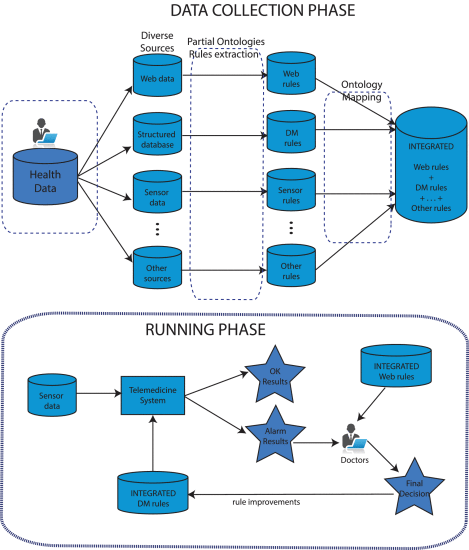
* La mayoría de las investigaciones sobre integración de datos se han centrado en el modelo relacional. En muchos sentidos, el modelo relacional y el lenguaje de consulta del registro de datos son los formalismos más simples y limpios para la representación de datos y consultas, por lo que muchos de los problemas fundamentales se consideraron primero en ese entorno. [82] - [83] [84] [85] [86] [87] .
* XML (recomendaciones del W3C) 3 se ha convertido en el formato predeterminado para la exportación de datos de fuentes de bases de datos y documentos, y se han desarrollado muchas herramientas adicionales para exportar a XML desde fuentes heredadas. Sin embargo, XML no pretende resolver directamente la heterogeneidad semántica o introducir esquemas estándar en ningún dominio [88] - [89] [90] [91] [92] [93] .
* La incertidumbre puede introducirse en múltiples aspectos de la integración de datos: (a) Datos, algunos de los datos pueden extraerse de datos no estructurados y es posible que tengamos dudas acerca de la precisión de la extracción; (b) Las asignaciones de esquemas pueden generarse utilizando técnicas semiautomáticas y es posible que no tengamos los recursos para validar todas estas asignaciones. La incertidumbre sobre las asignaciones de esquemas puede ser muy común en las aplicaciones de integración de datos; (c) Consultas, el sistema puede necesitar ofrecer al usuario una interfaz de búsqueda de palabras clave y debe traducir las consultas de palabras clave en alguna forma estructurada, para que puedan reformularse en las fuentes de datos. El paso de traducción puede generar múltiples consultas estructuradas candidatas y, por lo tanto, habrá incertidumbre sobre la consulta de usuario prevista; (d) Esquema mediado, cuando el dominio del sistema de integración es muy amplio,[94] - [95] [96] [97] [98] [99] .
* El sistema de Representación del conocimiento (KR) almacena el modelo subyacente del dominio que utilizan muchas aplicaciones de inteligencia artificial, como planificadores, robots, procesadores de lenguaje natural y sistemas de juego. Los sistemas KR utilizan técnicas de razonamiento para responder consultas sobre el conocimiento. Algunos aspectos de la integración de datos también pueden verse como un problema de representación del conocimiento. Las fuentes de datos y sus contenidos se prestan a modelos bastante complejos. Determinar las relaciones entre las fuentes de datos, o entre una fuente de datos y un esquema mediado, a menudo requiere un razonamiento sutil. Por estas razones, los investigadores han considerado aplicar técnicas de representación del conocimiento a la integración de datos.[100] - [101] [102] [103] [104] . Finalmente, el trabajo desarrollado por Ganzha *et al.*[105] realiza un rico estudio sobre las ontologías disponibles y listas para ser utilizadas para el desarrollo de aplicaciones interoperables en el entorno de IoT por las interrelaciones entre las diferentes fuentes de datos heterogéneos.

**SECCION III.**

Nuestro enfoque Propuesta de arquitectura

La sección anterior revisa la literatura científica sobre las diferentes fuentes de información y su integración. Debido a la heterogeneidad de estos conjuntos de datos, creemos que en este caso, es necesario abordar el problema con nuevos métodos y enfoques. Las nuevas técnicas son fundamentales, ya que los métodos tradicionales actuales son más inadecuados y restringidos principalmente debido a la situación actual de Big Data. Una de las principales características y ventajas de nuestra arquitectura propuesta es el uso de técnicas de PNL y DM que permiten el procesamiento, la integración y la extracción de información relevante de las diferentes fuentes.

La arquitectura propuesta en este documento se resume en la Figura 1 , que muestra un diagrama que representa la metodología empleada en este estudio.

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral1-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral1-2857499-large.gif)

**FIGURA 1.**

Arquitectura general propuesta. Recolección de datos y fases de ejecución.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Hemos utilizado una arquitectura orientada a la ontología. La ontología central se ha definido como un KB y permite la integración de datos de las diferentes fuentes. Esta integración se realizará identificando los conceptos y relaciones equivalentes (mapeo ontológico) de las diferentes fuentes. En nuestra metodología podemos distinguir dos fases: (1) fase de recopilación de datos y (2) fase de ejecución.

En la fase de recopilación de datos, se recopila la información de varias fuentes de datos. Se utilizan ontologías de dominio especializadas para cada fuente de datos específica. Un proceso, llamado extracción de reglas, se aplica a cada fuente diferente. En el caso de información no estructurada (datos web), las técnicas de PNL se utilizan con el fin de obtener las reglas web. Con respecto a la información estructurada (bases de datos estructuradas), se utilizan técnicas de PNL y DM para obtener reglas de DM. Finalmente, toda la información se integra con el proceso de mapeo ontológico en el que se establecen las equivalencias entre los conceptos de las fuentes de datos. El resultado de la fase de recopilación de datos es un conjunto de diferentes tipos de reglas que hemos denominado reglas integradas (reglas web integradas, reglas DM integradas, etc.). Todas estas reglas se especifican utilizando los mismos conceptos, es decir,

En la fase de ejecución, la información recopilada de los sensores (datos del sensor) y las reglas DM integradas se envían al sistema de telemedicina. Si el sistema detecta alguna medición anormal, envía la alarma correspondiente al equipo médico. El médico analiza la alarma consultando las reglas web integradas. Finalmente, actuarán: (1) rechazando la alarma porque es una excepción, o (2) aceptando la alarma. Se considerará toda esta información para mejorar y refinar las reglas integradas de aprendizaje de DM.

En las siguientes dos subsecciones, los procesos de extracción de reglas y mapeo de ontología mencionados anteriormente se describen en detalle.

**A. El proceso de extracción de reglas**

Podemos distinguir diferentes procesos de extracción según la fuente de información. En este artículo nos centramos en datos no estructurados y estructurados. Por lo tanto, el proceso para extraer las reglas de los datos web y de las bases de datos estructuradas se define a continuación.

**1) Extracción de reglas web**

Con el fin de gestionar la gran cantidad de información no estructurada procesada por el sistema y ayudar a los médicos en su proceso de toma de decisiones, el módulo de PNL realiza una serie de tareas con el objetivo de proporcionar a los médicos información suficiente en términos de cantidad y calidad necesaria para realizar Las decisiones correctas en el tratamiento de los pacientes. Estas tareas combinan diferentes técnicas de PNL como se describe a continuación:

* Selección de documentos. Con respecto a la información extraída de la Web, se seleccionan documentos relacionados con el sector salud. 4 4
* Tarea de recuperación de información (IR). Esta tarea de IR consiste en encontrar aquellos documentos que satisfacen las necesidades de información de los médicos de las grandes colecciones de documentos. Nuestro enfoque utiliza el término técnica de frecuencia de documento inversa de frecuencia (tf-idf) para lograr su objetivo. En esta técnica, cada palabra tiene asignada una medida estadística (peso), que se utiliza para evaluar la importancia de una palabra para una colección de documentos. La importancia aumenta proporcionalmente al número de veces que aparece una palabra en el documento, pero está compensada por la frecuencia de la palabra en la colección. Por lo tanto, de acuerdo con las necesidades de información de los médicos formuladas como un enunciado en lenguaje natural, la tarea de IR recupera aquellos documentos que contienen la mayor cantidad de palabras del enunciado dado en el marco tf-idf explicado.
* Tareas de recuperación y etiquetado semántico basadas en ontología. Después de filtrar las colecciones de documentos, la solicitud de información presentada por los médicos se analiza semánticamente con el objetivo de etiquetar los posibles tipos semánticos esperados en los documentos. Por un lado, esta tarea consiste en utilizar la ontología de dominio como fuente de conocimiento para etiquetar palabras y expresiones contenidas en los documentos recuperados en la tarea de IR previa. 5 5Por otro lado, una vez que las palabras y los enunciados de los resúmenes se han etiquetado semánticamente como entidades, la tarea clasifica estos documentos según la cantidad de tipos semánticos esperados en los documentos. Por lo tanto, con base en esta clasificación, el sistema supone que los documentos principales son candidatos para contener información suficiente en términos de cantidad y calidad que los médicos necesitan para su proceso de toma de decisiones.
* Tarea de extracción de información. Esta tarea realiza un proceso especializado de extracción de información en el dominio médico. Anteriormente, los documentos seleccionados eran sintácticos parciales analizados (solo se identifican frases nominales ( *np* ), frases verbales ( *vp* ) y preposicionales ( *pp* )). Además, el proceso de reconocimiento de la entidad nombrada se llevó a cabo para etiquetar las frases nominales, por ejemplo, GA (albúmina glicosilada) y HbA1c se etiquetan como entidades "proteicas", mientras que DPN (neuropatía periférica diabética) se etiqueta como entidad "enfermedad". .
* Mapeo de ontologías. Las diferentes fuentes de datos tendrán sus propias ontologías o recursos semánticos, mencionados anteriormente como ontologías de dominio. Como se mencionó anteriormente, la información de cada fuente se enriquecerá / asociará semánticamente con la información extraída de su ontología de dominio específica. Además, la ontología central se ha definido como un KB. Esta ontología se utilizará cuando se lleve a cabo la integración de toda la información. Para lograr esta integración, las fuentes de datos se han descrito en términos de la ontología central a través de los conceptos y relaciones equivalentes (mapeo ontológico) entre la ontología central y las ontologías de dominio especializado de las diferentes fuentes. Este proceso se explicará en detalle en la siguiente subsección.
* Extracción de las reglas web integradas. Finalmente, las reglas web integradas que utilizan conceptos básicos se extraen gracias a una serie de patrones predefinidos descritos en términos de los conceptos básicos de ontología. Las reglas contienen información relevante sobre las necesidades de los médicos. Este resultado se proporciona a los médicos especialistas. Con esta información, pueden crear un tratamiento personalizado para el paciente dependiendo de sus características personales con un análisis más preciso de las alarmas producidas por el sistema de telemedicina. Posteriormente, los médicos y expertos en informática utilizarán las reglas web para mejorar las reglas DM utilizadas por el sistema de telemedicina.

**2) Extracción de reglas DM**

Con el objetivo de recopilar las reglas de DM a partir de información estructurada, se llevan a cabo varias tareas de PNL y DM:

* Selección de documentos. Con respecto a las bases de datos estructuradas, podrían utilizarse bases de datos relacionadas con el sector de la salud. 6 6
* Etiquetado semántico basado en ontología. Una vez que se ha seleccionado la base de datos, el objetivo de esta tarea es identificar el concepto semántico de cada campo de la base de datos para posteriormente realizar la integración de diferentes fuentes de datos. Todos los atributos de la base de datos se buscan en la ontología de dominio obteniendo sus conceptos semánticos. Como se mencionó anteriormente, las técnicas WSD deben aplicarse si es necesario para obtener un concepto semántico único.
* Tarea de minería de datos. Como ya se mencionó en la Sección II-B , existen tres técnicas principales de aprendizaje automático. A partir de estos, hemos decidido utilizar técnicas de clasificación porque son las estrategias más adecuadas para resolver el problema del diagnóstico y el tratamiento de enfermedades basadas en los síntomas de los pacientes (nuestro estudio de caso). El resultado de esta tarea para bases de datos estructuradas es la clasificación de datos en clases predefinidas. Dependiendo del problema específico, podemos usar diferentes enfoques de clasificación citados en la Sección II-B . En la experimentación de este documento, hemos utilizado árboles de decisión para obtener un conjunto de reglas que determinan el diagnóstico o el tratamiento de la enfermedad de un paciente.
* Mapeo ontológico y adquisición de las reglas integradas de DM. En esta tarea, las reglas obtenidas después del proceso DM se expresan en conceptos básicos después del mapeo ontológico. Estas reglas se denominan reglas DM integradas y son utilizadas por el sistema de telemedicina con el objetivo de verificar la información recopilada por los sensores (datos del sensor) que detectan cualquier medición anormal.

**B. El proceso de mapeo de ontologías. Integración de datos**

La ontología puede representar relaciones complejas entre las fuentes de datos, el esquema mediado y las consultas. El motor de inferencia puede razonar estas relaciones para decidir si las fuentes de datos son relevantes para las consultas dadas. Además, las ontologías pueden resolver la heterogeneidad semántica (a diferencia del formato XML). Por estas razones, las ontologías se han elegido como un mecanismo integrador de todas las fuentes de datos en nuestra propuesta y, en consecuencia, presentamos una arquitectura orientada a la ontología.

La ontología servirá como el esquema mediado del sistema de integración de datos. Las diferentes fuentes de datos tendrán sus propias ontologías o recursos semánticos, mencionados anteriormente como ontologías de dominio. Por lo tanto, la información de cada fuente se enriquecerá / asociará semánticamente con la información extraída de su ontología de dominio específica.

Además, la ontología central se ha definido como un KB. Esta ontología se utilizará cuando se lleve a cabo la integración de toda la información. Para lograr esta integración, las fuentes de datos se han descrito en términos de la ontología central a través de los conceptos y relaciones equivalentes (mapeo ontológico) entre la ontología central y las ontologías de dominio especializado de las diferentes fuentes.

Con el objetivo de seleccionar la ontología central, nos hemos basado en el concepto de Ontología Universal (UO) presentado por Olivé [108]. Por UO, Olivé significa la especificación formal de todos los conceptos que usamos y compartimos. Esto incluye los conceptos de uso general, los que son particulares de las disciplinas existentes y los específicos de cualquier tipo de actividad humana u organizativa. La UO especifica los conceptos que se aplican a los objetos, a sus relaciones y a las acciones o eventos relacionados con esos objetos. En la UO se distinguen cuatro niveles de conceptos: (1) Modelo conceptual, (2) Fundacional, (3) General y (4) Dominio. Si ordenamos verticalmente los niveles y llenamos cada nivel, el resultado puede interpretarse como una pirámide, llamada pirámide UO, en la que la base es el Nivel de dominio. Cada uno de estos niveles se describe brevemente a continuación:

1. El Nivel del Modelo Conceptual (o Modelo de Ontología) comprende los metatipos y los supertipos directos o indirectos de todos los conceptos en la UO. Los conceptos en este nivel se utilizan para definir el resto de la UO.
2. El nivel fundamental incluye conceptos abstractos que se han propuesto en las ontologías fundamentales. Los conceptos en este nivel no pueden ser instanciados directamente para publicar hechos y, por lo tanto, no son esenciales en la UO propuesta. Sin embargo, pueden ser útiles para aclarar la semántica de otros conceptos, para definir solo una vez el conocimiento que es común a varios conceptos y para propósitos de razonamiento.
3. El nivel general contiene los conceptos para fines generales. Estos conceptos son subtipos o instancias de conceptos en el nivel conceptual del modelo y, posiblemente, de conceptos en el nivel fundamental. Existen varias ontologías que podrían proporcionar una base excelente para construir el Nivel General de la UO. Entre ellos, podríamos mencionar WordNet [109] y CYC [110] .
4. El nivel de dominio contiene los conceptos correspondientes a los idiomas para fines especiales. Por lo tanto, este nivel contiene todas las ontologías de dominio existentes. Dado que hay muchas ontologías de dominio, el nivel de dominio incluye varios millones de conceptos. Lograr un arreglo satisfactorio de estas ontologías es el principal desafío técnico de la UO. Los conceptos en el nivel de dominio son subtipos o instancias de conceptos en el nivel general y, posiblemente, de conceptos en el nivel fundamental.

Según la UO presentada, en nuestra propuesta hemos utilizado WordNet como la ontología central. Es una base perfecta para crear el Nivel General de la UO. WordNet define los sinónimos, los verbos y los adjetivos que pueden ser la fuente de los tipos de entidad y las propiedades de la UO. WordNet 3.0 comprende más de 80,000 synsets sustantivos (conceptos), que incluyen la mayoría de los tipos de entidades que tienen un nombre en inglés (propósito general). Ya hay "redes de palabras" en muchos idiomas, que incluyen enlaces a WordNet en inglés. Comprende también más de 13,000 synsets verbales, que incluyen la mayoría de las propiedades que tienen un nombre en forma de verbo. Finalmente, WordNet comprende también más de 18,000 sintetizadores adjetivos, la mayoría de los cuales pueden considerarse propiedades booleanas.

Con respecto al proceso de mapeo de ontología entre las ontologías de dominio especializado y la ontología central, hay dos tipos de mapeos: vertical y horizontal [108] . Las asignaciones verticales definen las correspondencias entre la ontología de dominio y los conceptos a nivel general (WordNet en nuestro enfoque). Las asignaciones horizontales definen las correspondencias entre la ontología de dominio y las otras ontologías a nivel de dominio. En ambas asignaciones, una correspondencia es una relación entre dos conceptos. En general, puede ser una *equivalencia* (los conceptos son los mismos), un *IsA* (un concepto es un subtipo del otro) o una *desarticulación* (ninguna entidad –o propiedad– puede ser una instancia de ambos conceptos) [111] .

En nuestro enfoque, proponemos el uso de la metodología STROMA [112] (SemanTic Refinement of Ontology MAppings) para determinar las asignaciones de ontología semántica automática tanto vertical como horizontal. Los autores presentan una metodología de dos pasos que aprovecha las capacidades de las herramientas de coincidencia ontológicas de última generación. En un primer paso, aplican una herramienta de coincidencia de vanguardia para determinar un mapeo de ontología inicial con correspondencias de igualdad aproximadas. En el segundo paso, los autores aplican cinco técnicas diferentes (incluidos los enfoques lingüísticos y el uso de diccionarios) para determinar para cada correspondencia su tipo más probable de relación ( *igualdad* , *es-a* –supuesto–, *inversa es-a* , *parte de* y*parte inversa de las* relaciones). Las cinco estrategias implementadas son las siguientes: (1) Estrategia compuesta que procesa las palabras compuestas (encabezado + modificador); (2) Estrategia de conocimiento previo que utiliza recursos o diccionarios lingüísticos (WordNet, OpenThesaurus 7 y partes del Metathesaurus UMLS) 8 para verificar las relaciones; (3) Estrategia de desglose que se utiliza si al menos uno de los dos conceptos en una correspondencia es un desglose (una lista de ítems); (4) Estrategia de estructura que toma en cuenta la estructura explícita de las ontologías; (5) Enlace múltiple es una estrategia específica que saca conclusiones de elementos de esquema que participan en más de una correspondencia.

**SECCION IV.**

Caso de estudio

**A. Descripción general**

Ahora que hemos introducido la arquitectura del sistema ( Figura 1 ), en las siguientes subsecciones explicamos la aplicación de nuestra arquitectura al escenario del entorno sanitario. Específicamente, nos centramos en un sistema de telemedicina para facilitar el diagnóstico y el tratamiento de pacientes con diabetes.

En primer lugar, se presenta el contexto de la enfermedad ( Sección IV-B ). A continuación, la Sección IV-C describe el conjunto de datos utilizado en nuestros experimentos. Después de esto, se explica la fase de recopilación de datos, explícitamente el proceso de extracción de reglas: extracción de reglas web ( Sección IV-D ) y extracción de reglas DM ( Secciones IV-E ). Finalmente, la Sección IV-F presenta la fase de ejecución que incluye la recopilación de datos del sensor y cómo funciona el sistema de telemedicina.

**B. El contexto**

En nuestro estudio de caso, nos centraremos en la diabetes mellitus (DM), comúnmente conocida como diabetes. Es un grupo de enfermedades metabólicas en las que hay niveles altos de azúcar en la sangre durante un período prolongado. Los síntomas de un nivel alto de azúcar en la sangre incluyen micción frecuente, aumento de la sed y aumento del hambre. Si no se trata, la diabetes puede causar muchas complicaciones. Las complicaciones agudas incluyen cetoacidosis diabética y coma hiperosmolar no cetótico y son consecuencia de un control inadecuado de la enfermedad. Las complicaciones graves a largo plazo incluyen enfermedades cardiovasculares, derrames cerebrales, insuficiencia renal crónica, úlceras en los pies y daños en los ojos.

Las principales complicaciones a largo plazo de la diabetes se relacionan con el daño a los vasos sanguíneos. Las complicaciones principales de la diabetes debido al daño en los vasos sanguíneos pequeños incluyen daño a los ojos (retinopatía diabética), riñones (nefropatía diabética) y nervios (neuropatía diabética). Con respecto a las enfermedades "macrovasculares", las complicaciones incluyen accidente cerebrovascular y enfermedad vascular periférica.

Existen tres tipos principales de diabetes mellitus: (1) la DM tipo 1 (“diabetes mellitus insulinodependiente” –IDDM–) resulta del fracaso del páncreas para producir suficiente insulina; (2) La DM tipo 2 (“diabetes mellitus no dependiente de insulina” –NIDDM–) comienza con resistencia a la insulina y produce un defecto progresivo en la secreción de insulina; (3) La diabetes gestacional ocurre cuando las mujeres embarazadas sin antecedentes de diabetes desarrollan un nivel alto de azúcar en la sangre.

En nuestro experimento estudiaremos pacientes con diabetes tipo 1. Se utilizará un sistema de telemedicina para rastrear la enfermedad y el tratamiento de pacientes diabéticos con el objetivo de prevenir las complicaciones mencionadas anteriormente que son típicas de esta enfermedad. Nuestros esfuerzos estarán dirigidos a lograr una medicina personalizada para cada paciente. Con respecto al tratamiento personalizado de enfermedades, se pueden identificar tres escenarios posibles: (1) ayudar a los profesionales de la salud durante el proceso de toma de decisiones en entornos clínicos; (2) proporcionar a los pacientes crónicos y dependientes acceso a información relevante sobre su salud; (3) apoyar la capacitación basada en evidencia para nuevos estudiantes de medicina.

A partir de estos escenarios nos centraremos en lo siguiente:

1. Perspectiva del médico: el tratamiento que los profesionales le dan a un paciente diabético. El objetivo del tratamiento de la diabetes es restaurar los niveles glucémicos normales. En la DM tipo 1 y la diabetes gestacional se aplica un tratamiento para sustituir la insulina o los análogos de la insulina. En la DM tipo 2 se puede aplicar un tratamiento similar o tratamiento con antidiabéticos orales. Para determinar si el tratamiento es adecuado, se realiza una prueba llamada hemoglobina glucosilada, prueba HbA1c. La prueba proporciona niveles promedio de glucosa en sangre 9 durante un período de dos a tres meses. Una persona no diabética 10 tiene unHb A 1 c < 6 % . En pacientes con DM tipo 1, el tratamiento es correcto solo si el resultado de la prueba HbA1c permanece por debajo del valor objetivo.
2. Perspectiva del paciente: la información proporcionada en tiempo real sobre las pruebas de glucosa en sangre y HbA1c le permite al paciente hacer un seguimiento exhaustivo de su enfermedad sin ser hospitalizado. Además, el sistema de telemedicina puede detectar cualquier anomalía en el tratamiento del paciente. Si se produce una anomalía, se enviará una alerta al personal médico. El médico tomará las medidas adecuadas, por ejemplo, cambiar el medicamento o la dosis de un medicamento específico.

**C. Descripción de los datos**

Con el fin de verificar el rendimiento de nuestra arquitectura, hemos utilizado diferentes tipos de fuentes de información. Hemos utilizado datos extraídos de la Web y datos estructurados extraídos de un repositorio de datos en la fase de recopilación de datos. Los datos recopilados de los sensores se han utilizado en la fase de ejecución. Específicamente, hemos experimentado con:

1. Datos web El tipo de información no estructurada y semiestructurada de Internet (incluidas las redes sociales) que se recopila continuamente es ilimitada. Sin embargo, en este ejemplo, vamos a restringir la información al contexto específico de dos sitios web: PubMed y PLOS ONE. PubMed es un excelente lugar para obtener información, ya que comprende más de 24 millones de citas de literatura biomédica de MEDLINE, revistas de ciencias de la vida y libros en línea. La publicación PLOS ONE es una revista multidisciplinaria de acceso abierto. El amplio alcance de PLOS ONE proporciona una plataforma para publicar investigaciones primarias, incluidos estudios interdisciplinarios y de replicación, así como resultados negativos. Presenta informes de investigaciones originales de todas las disciplinas dentro de la ciencia y la medicina.
2. Base de datos estructurada. Hemos utilizado un sitio web altamente referenciado con información estructurada lista para la minería de datos utilizando varios métodos de IA: la base de datos UCI. El depósito de aprendizaje automático de UCI es una colección de bases de datos, teorías de dominio y generadores de datos que son utilizados por la comunidad de aprendizaje automático para el análisis empírico de algoritmos de aprendizaje automático.
3. Datos del sensor. Actualmente, no solo hay muchos dispositivos económicos para recopilar datos del cuerpo humano, sino que los teléfonos inteligentes también pueden recopilar esta información.

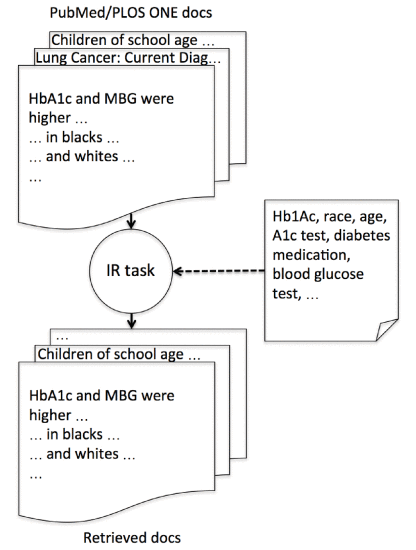
Como se ha mostrado en la Figura 1 , existen otras posibles fuentes de información (los nodos de nuestra arquitectura), pero por razones de limitaciones de espacio de este documento, hemos restringido nuestras fuentes a estos tres nodos muy significativos ya que representan un conjunto completamente operativo.

**D. Extracción de reglas web**

Los pasos para obtener las reglas web, explicados previamente en la Sección III-A.1 , son los siguientes:

* Selección de documentos del sector de la salud (documentos de PubMed y PLOS ONE).
* Tarea de IR para obtener los documentos relevantes para las necesidades del médico (uso de palabras clave como HbA1c, glucosa en sangre, raza, edad, etc.).
* Etiquetado semántico de los conceptos encontrados en los documentos relevantes recuperados (uso de la ontología de dominio, UMLS en nuestro ejemplo).
* Análisis parcial y tarea de IE para identificar las entidades del documento.
* Mapeo de ontologías (UMLS-WordNet) para identificar conceptos equivalentes o compatibles entre diferentes recursos.
* Extracción de reglas web utilizando patrones predefinidos con conceptos de WordNet.

A continuación, explicaremos en detalle los pasos mencionados. Los dos primeros pasos son la selección de documentos y la tarea de IR ( Figura 2 ).

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral2-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral2-2857499-large.gif)

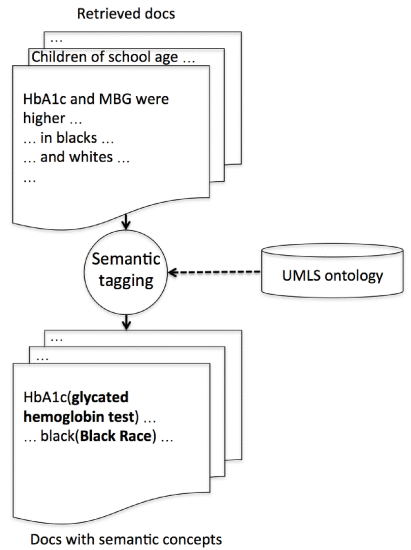
**FIGURA 2.**

Proceso para obtener reglas web: selección de documentos y tarea de IR.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Con respecto a la selección de documentos, hemos elegido documentos relacionados con el sector de la salud (mencionados en la Sección III-A.1 ) de las URL de PubMed y PLOS ONE. Estos documentos son de una amplia variedad de temas y temas. Por lo tanto, es necesario llevar a cabo un filtro para seleccionar solo documentos sobre diabetes. Esto se logra aplicando el siguiente paso, la tarea de IR. Como se mencionó anteriormente en la Sección III-A.1 , el objetivo de esta tarea es recuperar los documentos relevantes para las necesidades del usuario. Los expertos (los médicos) introducen los requisitos del usuario como palabras clave en lenguaje natural relacionadas con el tema principal: diabetes. En la figura 2se muestra un conjunto de palabras clave: HbA1c, raza, edad, prueba A1c, medicamentos para la diabetes, etc. El resultado de este paso es el conjunto de documentos relevantes para las palabras clave. Por ejemplo, en la figura, el documento sobre el cáncer de pulmón se descarta porque no está relacionado con la enfermedad de la diabetes.

El siguiente paso es el etiquetado semántico de los documentos conceptuales ( Figura 3 ). Hemos utilizado el dominio de ontología UMLS. En nuestra propuesta, uno de los objetivos principales es conceptualizar los diferentes términos y expresiones en los textos obtenidos de la Web. Como se vio en la sección de vanguardia anterior, UMLS es el recurso más utilizado por los sistemas de PNL en los diversos subdominios especializados en las áreas de medicina y salud. Para propósitos de conceptualización, consideramos que UMLS es el mejor recurso debido a su amplia variedad de fuentes especializadas en diferentes subdominios de los sectores de medicina y salud.

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral3-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral3-2857499-large.gif)

**FIGURA 3.**

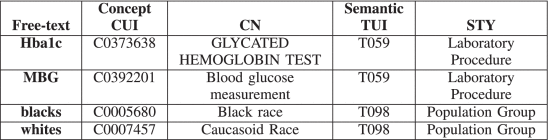
Proceso para obtener reglas web: etiquetado semántico.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

En la Figura 3 se puede observar un ejemplo del marcado semántico. Por ejemplo, del documento de PubMed con el código URL " <https://www.ncbi.nlm.nih.gov/pubmed/26783014> " (visitado el 19 de marzo de 2018) se extrajo el siguiente texto: *HbA1c y MBG fueron más altos ( p* < *0.0001) en negros [10.4% (90.3 mmol / mol), 255 mg / dL] que en blancos [8.9% (73.9 mmol / mol), 198 mg / dL)* .

La Tabla 1 muestra el etiquetado de los conceptos UMLS del texto anterior. Por un lado, la columna CUI identifica de forma exclusiva el concepto, mientras que la columna CN muestra el nombre del concepto, y por otro lado, la columna TUI identifica de forma exclusiva el tipo semántico, mientras que la columna STY describe el nombre del tipo semántico asociado a el concepto.

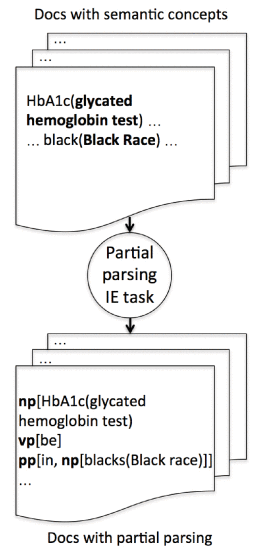
**TABLA 1** Entradas de texto libre asignadas a UMLS

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral.t1-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral.t1-2857499-large.gif)

El resultado de este paso es el etiquetado semántico con conceptos UMLS de los documentos relevantes. Por ejemplo, las palabras HbA1c y negro del documento relevante están etiquetadas con los conceptos UMLS (columnas CN) " *prueba de hemoglobina glucosilada* " y " *raza negra* " respectivamente ( Figura 3 ).

Una vez que se ha realizado el etiquetado semántico de los conceptos, se extrae un conjunto de reglas del análisis parcial del texto en el siguiente paso ( Figura 4 ). De esta manera, cada cláusula que contenga cualquier enlace semántico se convertirá en una regla. Por ejemplo, en la cláusula " *HbA1c y MBG fueron mayores (p* < *0.0001) en negros [10.4% (90.3 mmol / mol), 255 mg / dL] que en blancos [8.9% (73.9 mmol / mol), 198 mg / dL )*", El análisis etiquetará las siguientes frases nominales ( *np* ), frases verbales ( *vp* ) y frases preposicionales ( *pp* ):

* *np* (HbA1c),
* *np* (MBG),
* *vp* (be),
* *pp* (en, *np* (negros, 10.4%, 90.3 mmol / mol, 255 mg / dL)),
* *np* (blancos, 8.9%, 73.9 mmol / mol, 198 mg / dL).

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral4-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral4-2857499-large.gif)

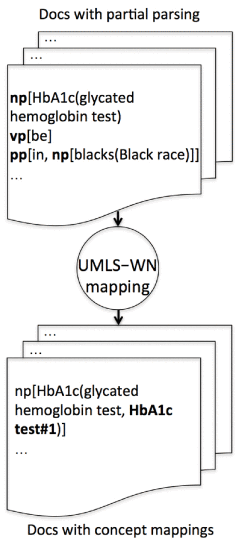
**FIGURA 4.**

Proceso para obtener reglas web: análisis parcial.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Además, después de aplicar la tarea IE, se extrae un conjunto de entidades. En nuestro ejemplo, se identifican tres tipos de entidades que expresan medidas: (i) porcentaje (10.4, 8.9), (ii) mmol / mol (90.3, 73.9) y (iii) mg / dL (255, 198).

En el siguiente paso, se lleva a cabo la asignación de ontología entre los conceptos de la ontología de dominio (UMLS) y la ontología central (WordNet) ( Figura 5 ). Se utiliza la metodología STROMA presentada en la Sección III-B . Por ejemplo, obtenemos que el concepto UMLS "PRUEBA DE HEMOGLOBINA GLICADA" y el concepto de WordNet "Prueba de HbA1c # 1" son equivalentes. El resultado de esta etapa contendrá los conceptos etiquetados con la ontología central que permitirá la integración de datos.

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral5-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral5-2857499-large.gif)

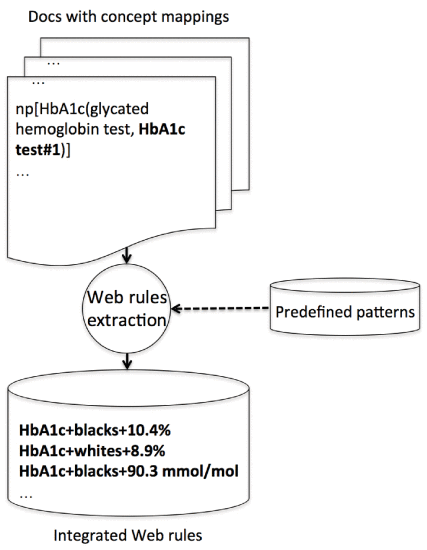
**FIGURA 5.**

Proceso para obtener reglas web después de la asignación de ontología entre los conceptos UMLS y WordNet.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Finalmente, en el último paso, las reglas web integradas que contienen información relevante sobre la diabetes se extraen gracias a una serie de patrones predefinidos que utilizan los conceptos de WordNet ( Figura 6 ). En el ejemplo anterior, se utilizan los siguientes dos patrones relacionados con HbA1c y MBG (glucosa en sangre media):

1. *Concepto de "prueba HbA1c"* + *concepto de "raza"* + + *entidad "medida\_porcentaje"* ∥ *Entidad "measure\_mmol / mol"*
2. *Concepto de "prueba de glucosa en sangre"* + *concepto de "raza"* + + *entidad "measure\_mg / dL"*

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral6-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral6-2857499-large.gif)

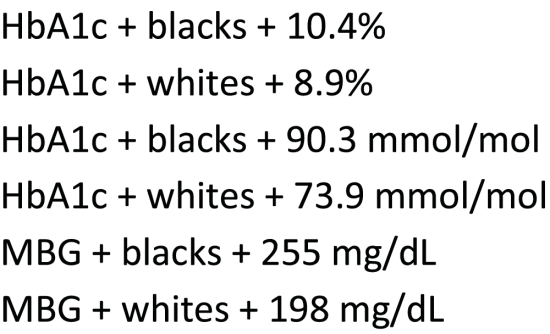
**FIGURA 6.**

Proceso para obtener reglas web: extracción de reglas web utilizando los patrones predefinidos.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Por ejemplo, el primer patrón especifica que se ha detectado en el texto un concepto de "prueba de HbA1c" (relativo a la prueba de HbA1c). Le sigue un concepto de "Raza" (relativo a la raza de la persona) y una entidad que expresa una medida en porcentaje o mmol / mol (el valor del resultado de la prueba HbA1c).

Después de aplicar los patrones antes mencionados, se obtienen las reglas web que se muestran en la Figura 7 .

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral7-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral7-2857499-large.gif)

**FIGURA 7.**

Reglas web obtenidas después de la integración de datos. Influencia de la raza del paciente.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

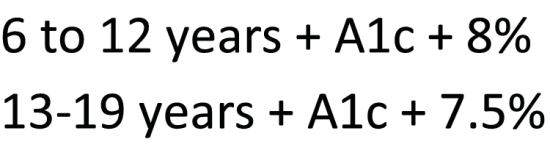
Estas reglas obtenidas de documentos web (reglas web) son muy importantes porque determinan que existe una variación racial del resultado de la prueba de HbA1c en pacientes con afecciones similares. Este resultado se proporciona al médico especialista. Con esta información, el médico puede crear un tratamiento personalizado para el paciente según su raza con un análisis más preciso de las alarmas producidas por el sistema de telemedicina. Finalmente, los médicos y expertos en informática utilizarán más tarde las reglas web para mejorar las reglas DM utilizadas por el sistema de telemedicina.

Otro ejemplo muy ilustrativo fue extraído del documento PLOS ONE con el código URL " <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152332> " (visitado el 19 de marzo de 2018). Se encontró el siguiente texto: *Niños en edad escolar (6 a 12 años)* = *meta A1c de 8% o menos; Adolescentes y adultos jóvenes (entre 13 y 19 años)* = *objetivo A1c de 7.5% o menos.*

En este ejemplo, se utiliza el siguiente patrón:

1. *Entidad "measure\_age"* + *concepto "Test HbA1c"* + *entidad "measure\_percentage"* ∥ *Entidad "measure\_mmol / mol"*

En consecuencia, se obtienen las reglas web que se muestran en la Figura 8 . Con esta información, el médico puede crear un tratamiento personalizado para los pacientes en función de sus edades. Cuando se envía una alarma a un paciente joven, el médico verificará su edad exacta y determinará si el valor de A1c es correcto o anormal.

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral8-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral8-2857499-large.gif)

**FIGURA 8.**

Reglas web obtenidas después de la integración de datos. Influencia de la edad del paciente.

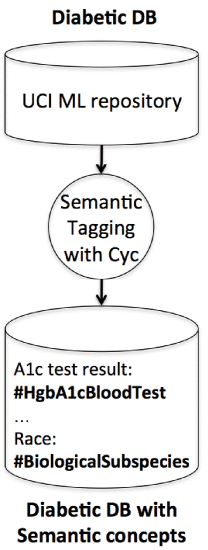
[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

**E. Extracción de reglas DM**

Los pasos para extraer las reglas de DM, presentados en la Sección III-A.2 , son los siguientes:

* Selección de documentos (UCI Machine Learning Repository).
* Etiquetado semántico de los conceptos encontrados en los campos de la base de datos (uso de la ontología de dominio, Cyc en nuestro estudio de caso).
* Aplicación de métodos de IA para obtener las reglas de DM enriquecidas con conceptos semánticos.
* Mapeo de ontologías (Cyc-WordNet) para identificar conceptos equivalentes o compatibles entre diferentes recursos.

A continuación, explicaremos en detalle los pasos mencionados. Los dos primeros pasos son la selección de documentos y el etiquetado semántico ( Figura 9 ).

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral9-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral9-2857499-large.gif)

**FIGURA 9.**

Proceso para obtener reglas DM: selección de documentos y etiquetado semántico.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

En cuanto a la selección de documentos, hemos utilizado los datos estructurados extraídos de pacientes con diabetes utilizados en el estudio desarrollado por Strack *et al.*[113] y se ha extraído de la base de datos de Datos de salud (Cerner Corporation, Kansas City, MO), un almacén de datos nacional que recopila registros clínicos completos en hospitales de los Estados Unidos.

La base de datos contiene datos recopilados sistemáticamente de los registros médicos electrónicos de las instituciones participantes e incluye datos de encuentros (emergencias, pacientes ambulatorios y pacientes hospitalizados), especialidad del proveedor, datos demográficos (edad, sexo y raza), diagnósticos y procedimientos hospitalarios documentados por ICD-9 -Códigos CM, datos de laboratorio, datos de farmacia, mortalidad hospitalaria y características del hospital.

Los datos de Datos de salud que utilizamos fueron un extracto que representa 10 años (1999–2008) de atención clínica en 130 hospitales y redes integradas de entrega en todo Estados Unidos. La base de datos consta de 41 tablas en un esquema de dimensión de hechos y un total de 117 características. La base de datos incluye 74,036,643 encuentros únicos (visitas) que corresponden a 17,880,231 pacientes únicos y 2,889,571 proveedores.

El conjunto de datos se creó en dos pasos. Primero, los encuentros de interés se extrajeron de la base de datos con 55 atributos. Este conjunto de datos está disponible como material complementario disponible en línea, 11 y también está en el Depósito de aprendizaje automático de UCI.

En segundo lugar, los análisis preliminares y el preprocesamiento de los datos se realizaron dando como resultado que solo estas características (atributos) y encuentros que podrían usarse en análisis posteriores se retengan, en otras palabras, características que contienen información suficiente. La lista completa de las características y su descripción se proporciona en [113] . Es importante enfatizar los atributos específicos relacionados con la diabetes: "resultado de la prueba de suero de glucosa" (nivel de glucosa en sangre), "resultado de la prueba A1c", "medicamentos para la diabetes" y "cambio de medicamentos" (indica si hubo un cambio en los medicamentos para la diabetes) –Ya sea dosis o nombre genérico–) entre otros.

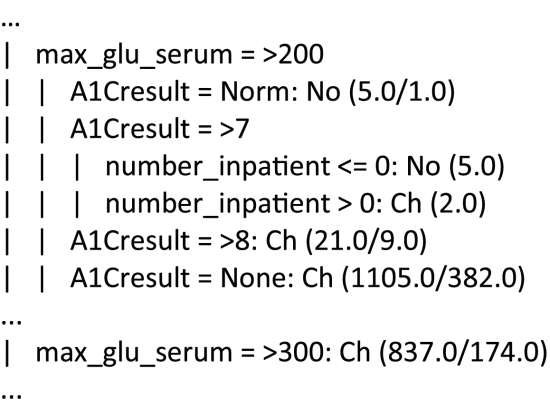
Finalmente, se extrajo la información de la base de datos mencionada para encuentros "diabéticos". De esta forma, se identificaron 101,766 encuentros relacionados con pacientes diabéticos. Esta información fue utilizada en nuestros experimentos.

El siguiente paso es el etiquetado semántico de la "DB diabética" mencionada anteriormente utilizando la ontología de dominio. Hemos utilizado la ontología Cyc 12 en nuestros experimentos porque almacena un gran volumen de información y tiene una gran cantidad de enlaces a otras ontologías / recursos semánticos (incluidos los dominios de salud).

Específicamente, para etiquetar los datos estructurados hemos buscado todos los atributos de la base de datos en Cyc. Por ejemplo, los atributos "Resultado de la prueba A1c" y "Carrera" están etiquetados con los conceptos de Cyc " *# HgbA1cBloodTest* " y " *#BilogicalSubspecies* " respectivamente ( Figura 9 ). El objetivo de esta etapa es identificar el concepto semántico 13 de cada campo de la base de datos para realizar posteriormente la integración de diferentes fuentes de datos.

El siguiente paso consiste en aplicar métodos de Inteligencia Artificial con el objetivo de extraer datos estructurados para descubrir conocimiento oculto. Utilizamos árboles de decisión para obtener un conjunto de reglas de DM extraídas de los datos relacionados con el tratamiento de pacientes con diabetes. Este enfoque ofrece excelentes resultados, así como el beneficio de la visualización fácil y eficiente de los datos. Los experimentos se realizaron con el software Weka [114] . El algoritmo elegido en nuestro estudio es C4.5 [115]. El árbol C4.5 muestra, en forma de árbol, las condiciones para predecir el comportamiento de una variable especificada; además, el conjunto de reglas o condiciones de DM se presentan en forma de texto. En nuestro estudio de caso, redujimos el número original de atributos a 18 (los principales relacionados con la diabetes: "raza", "edad", "resultado de la prueba de suero de glucosa", "resultado de la prueba A1c", "medicamentos para la diabetes", "cambio de medicamentos ", etc.).

Desde el punto de vista del médico y del paciente, una de las cosas más importantes es analizar si el tratamiento del paciente es adecuado o no. Cuando no sea adecuado, el médico debe cambiar los medicamentos o sus dosis para evitar complicaciones de la enfermedad. Este atributo existe en la base de datos y se llama "cambio de medicamentos". Por lo tanto, nuestros experimentos se centran en predecir el valor de esta variable (que usa los valores de *Ch* y *No* para indicar si el medicamento debe cambiarse o no, respectivamente) dependiendo de las condiciones que ocurran. Un fragmento de las reglas DM obtenidas se muestra en la Figura 10 .

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral10-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral10-2857499-large.gif)

**FIGURA 10.**

Extracto de las reglas de DM obtenidas con el algoritmo C4.5 en la "base de datos diabética".

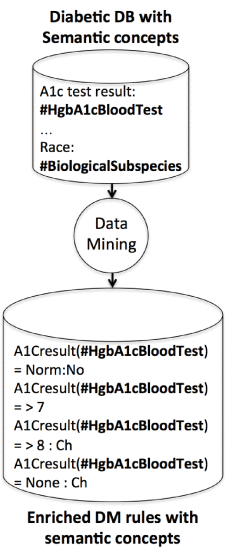
[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Un hecho importante a tener en cuenta es que el sistema aprende a través de las características individuales del paciente (edad, peso, sexo, etc.), que determinan el tratamiento personalizado del paciente. Además, con las reglas de DM aprendidas automáticamente, ya hemos filtrado y eliminado muchas situaciones que no son peligrosas para el paciente. Finalmente, estas reglas de DM son supervisadas por un especialista médico para detectar errores o hacer mejoras en ellos.

En las reglas de DM obtenidas en el experimento, el primer factor que influye en el cambio de la medicación del paciente es el resultado de la prueba de glucosa en sangre, "resultado de la prueba de glucosa en suero". En el DB, esta variable tiene 4 valores: (1) *Norma* (valor normal), (2) *Ninguna* (sin medición de glucosa), (3)> *200* y (4)> *300* .

Nos centramos en las reglas de DM obtenidas cuando la medición de glucosa es anómala (si el valor es *Norma* o *Ninguno* , no se toman medidas inmediatas porque no hay ninguna alarma e, inicialmente, no se envía ninguna advertencia al sistema). Las reglas muestran que si el valor de glucosa es> *200* (es un valor alto), se debe verificar el resultado de la prueba A1c. Dependiendo de su valor y de si el paciente ha sido ingresado previamente en el hospital, se decidirá el cambio de medicación del paciente. Cuando es necesario cambiar el medicamento, se envía una alarma al sistema para su revisión por parte del especialista. Finalmente, las reglas indican que si la medición de glucosa es> *300*(un valor muy alto) se cambiará el medicamento. Esta regla será revisada por un especialista que debe confirmarla o modificarla. 14

El resultado de este paso son las reglas de DM obtenidas con métodos de IA (en nuestro árbol de experimentación C4.5) enriquecidas con conceptos semánticos. En [116] se ha utilizado un enfoque similar para incluir reglas semánticas de C4.5 . En la Figura 11 se pueden observar las reglas enriquecidas de DM. Por ejemplo, en las reglas en las que aparece el atributo A1cresult de la base de datos, se ha agregado su correspondiente concepto semántico Cyc: A1Cresult ( *# HgbA1cBloodTest* ) = Norma: No, A1Cresult ( *# HgbA1cBloodTest* ) => 7, etc.

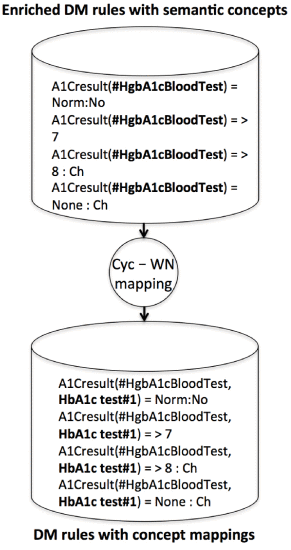
[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral11-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral11-2857499-large.gif)

**FIGURA 11.**

Proceso para obtener reglas de DM enriquecidas con conceptos de Cyc después de Data Mining.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

En el paso final, se lleva a cabo el mapeo de ontología entre los conceptos de la ontología de dominio (Cyc) y la ontología central (WordNet) ( Figura 12 ). Por ejemplo, obtenemos que el concepto de Cyc "# HgbA1cBloodTest" y el concepto de WordNet "HbA1c test # 1" son equivalentes. Esta información se incluye en las reglas que obtienen las reglas de DM enriquecidas con los conceptos semánticos de WordNet: A1Cresult (# HgbA1cBloodTest, *HbA1c test # 1* ) = Norma: No, A1Cresult (# HgbA1cBloodTest, *HbA1c test # 1* ) => 7, etc. El resultado final del proceso de extracción de reglas DM es un conjunto de reglas DM enriquecidas en las que los atributos se etiquetan con sus conceptos semánticos utilizando la ontología central que permite la integración de las diferentes fuentes de datos.

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral12-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral12-2857499-large.gif)

**FIGURA 12.**

Proceso para obtener reglas de DM después de la asignación de ontología entre los conceptos de Cyc y WordNet.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Estas reglas de DM obtenidas automáticamente son muy importantes porque son utilizadas por el sistema de telemedicina. Las reglas de DM determinan el cambio de la medicación del paciente y el envío de la alarma al médico. Podemos resumir que la calidad de vida del paciente ha mejorado considerablemente: (1) el uso de sensores (que realizan pruebas de glucosa en sangre y A1c) permite controlar su estado de salud en tiempo real, sin tener que visitar el hospital; (2) si no hay alarma, no es necesario para una visita al médico o un monitoreo cara a cara; (3) el médico puede "controlar" de forma remota la condición del paciente en tiempo real y "ajustar" su medicación de una manera más eficiente. Además, es importante mencionar que los costos clínicos también se reducen por estas razones mencionadas.

**F. Datos del sensor. Fase de ejecución**

Los resultados de esta etapa para los datos del sensor es la recopilación en tiempo real de datos especializados del paciente mediante sensores. Esto se lleva a cabo en la fase de ejecución. La información recopilada de los sensores se envía al sistema de telemedicina en tiempo real. Si se observan / procesan datos anómalos, el sistema enviará una alarma al equipo médico para que analicen los datos.

En nuestros experimentos, estábamos interesados ​​en monitorear los principales indicadores clave de rendimiento (KPI) de los pacientes con diabetes tipo 1 durante sus actividades físicas diarias. Como se vio en la sección de vanguardia anterior, considerando la gran cantidad de sensores que se pueden monitorear para extraer datos biomédicos de los pacientes, nos hemos centrado en aquellos que brindan más información sobre la diabetes tipo 1. Por lo tanto, utilizaremos un sensor para realizar el control de medición de glucosa en sangre y otro para realizar la prueba A1c o HbA1c en los intervalos de tiempo especificados por el médico. Mediante estas pruebas principales y otros puntajes posibles, como el esfuerzo físico y el tiempo para dormir de los pacientes, el sistema tiene suficiente conocimiento para predecir el comportamiento anormal de esta parte del sistema endocrino en los pacientes y transmite alertas al equipo médico.

Además, nuestra propuesta permite agregar otros sensores que realizan otras mediciones interesantes para este tipo de pacientes. Por ejemplo, el control de la medición del colesterol, la frecuencia cardíaca y los niveles de presión arterial podrían incorporarse al sistema de telemedicina. Toda esta información enviada por los sensores es útil si puede integrarse con las fuentes de información restantes. Por ejemplo, si los sensores envían información sobre la frecuencia cardíaca, las reglas de DM (extraídas de los datos estructurados) deben contener el atributo de "frecuencia cardíaca" o equivalente para llevar a cabo la integración de datos.

Con el fin de monitorear a los pacientes, cada uno de ellos debe tener diferentes sensores en sus cuerpos que rastreen todos estos puntajes y los envíen al servidor, que verifica todos los puntajes para detectar comportamientos anormales. La comunicación entre los diferentes sensores usados ​​por los pacientes y el servidor que verifica y almacena estos puntajes en tiempo real puede realizarse mediante un teléfono inteligente dedicado que use wifi o 4G para enviar estos puntajes al servidor.

Como se mencionó anteriormente, en la evaluación de nuestro enfoque, hemos utilizado la información proporcionada por los sensores relacionados con las pruebas de glucosa en sangre y A1c. Cuando el sistema de telemedicina recibe la información, se verifica con las reglas de DM aprendidas en la fase de recopilación de datos para decidir si hay una alarma (es necesario un cambio en la medicación del paciente).

En la Figura 13 se puede observar un nuevo fragmento de las reglas DM obtenidas. Las reglas muestran que si el resultado de la prueba de suero de glucosa no existe ( *max\_glu\_serum* = *Ninguno* ), se debe verificar el resultado de la prueba A1c. Si la prueba A1c es muy alta (=> *8* ), la medicación del paciente cambiará dependiendo de su edad y raza. Si la edad del paciente es [20-30] y su raza es afroamericana y él / ella ha sido ingresado previamente en el hospital (como máximo una vez), el medicamento debe cambiarse y el sistema produce la alarma. En este momento, los especialistas analizan si se ha producido una excepción o datos anómalos. Verifican las reglas web integradas ( Figura 7 y Figura 8) obtenida en el proceso de extracción de reglas web. Como se mencionó anteriormente, en la Figura 7 , se ha detectado una relación entre HbA1c ( *concepto de "prueba de HbA1c"* ) y la raza del paciente ( *concepto de "raza"* ), estableciendo que en las personas negras el valor objetivo de la prueba de HbA1c es del 10,4%, mientras que en las personas blancas El valor objetivo es 8.9%. Posteriormente, hay una excepción de la regla de DM aprendida en pacientes con *raza* = *afroamericana* . Por lo tanto, la alarma se descarta y la regla DM se mejora con esta excepción para otras decisiones. Por el contrario, si el A1c es => *8*y la edad del paciente es [10-20] la regla de DM dice que no es necesario cambiar el medicamento. Sin embargo, de acuerdo con las reglas web extraídas, Figura 8 , se ha encontrado una relación entre HbA1c ( *concepto de "prueba de HbA1c"* ) y la edad del paciente ( *entidad de "medida\_age"* ). Las reglas especifican que el objetivo A1c del niño de 6 a 12 años podría alcanzar el 8%. Posteriormente, en este caso, se debe realizar el cambio de medicación. Por estas razones, las reglas de DM aprendidas deben mejorarse y refinarse con las reglas extraídas de la Web (en nuestros experimentos extraídos de los documentos PubMed y PLOS ONE).

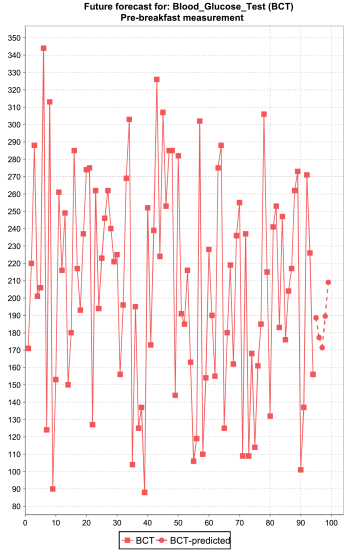
**FIGURA 13.**

Excepciones de las reglas DM obtenidas con el algoritmo C4.5 en la "base de datos diabética".

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Hemos llevado a cabo un segundo experimento, que incluye nueva información que refinará y mejorará el sistema de telemedicina. Para este propósito, hemos utilizado los datos específicos obtenidos de los sensores de prueba de glucosa en sangre de cada paciente. Se han utilizado los archivos de diabetes de 70 pacientes del repositorio UCI ML y el archivo de cada paciente tiene muchos registros con mediciones de glucosa y otros datos tomados a diferentes intervalos. Hay cuatro campos por registro de paciente que contienen la siguiente información: fecha, hora, código descriptivo y valor. El campo del código descriptivo especifica diferentes tipos de información: dosis de insulina regular / NPH / UltraLente, medición de glucosa en sangre antes del desayuno / almuerzo / cena, síntomas de hipoglucemia, ingestión típica de comidas, etc.

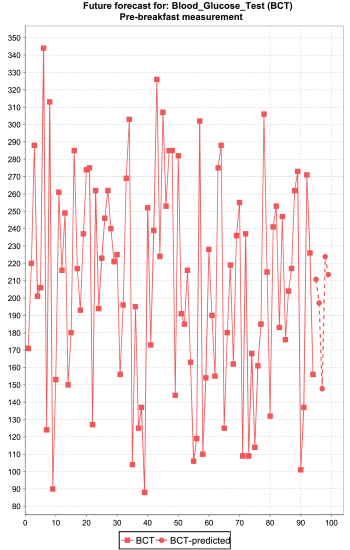
Para cada paciente, hemos seleccionado la información de medición diaria de glucosa en sangre. Específicamente, se han elegido las mediciones de glucosa en sangre antes del desayuno, antes del almuerzo y antes de la cena. El objetivo es analizar la evolución de la glucosa en sangre durante el día para cada paciente específico para facilitar un tratamiento personalizado. En consecuencia, las alarmas personalizadas se activarán cuando se superen los límites establecidos. Utilizamos el software Weka para el análisis de series temporales para pronosticar los pasos futuros a partir de los datos históricos entrantes. En la primera prueba, utilizamos los datos obtenidos todos los días antes del desayuno para el mismo paciente (en el ejemplo utilizamos los datos del paciente 02 durante 94 días consecutivos). Las predicciones de la prueba de glucosa en sangre antes del desayuno para los próximos 5 días utilizando métodos de Regresión lineal (LR) y Máquinas de vectores de soporte (SVM) se muestran enFiguras 14 y 15 . Los resultados de la prueba de glucosa en sangre se muestran en el eje de ordenadas y las diferentes mediciones (en orden secuencial) se muestran en el eje de abscisas.

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral14-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral14-2857499-large.gif)

**FIGURA 14.**

Mediciones diarias de glucosa en sangre antes del desayuno para el paciente 02. Valores pronosticados para los próximos 5 días utilizando LR.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral15-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral15-2857499-large.gif)

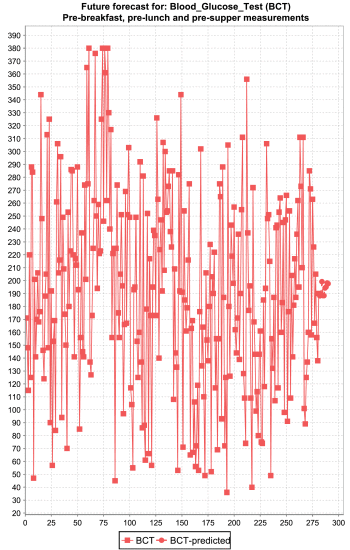
**FIGURA 15.**

Mediciones diarias de glucosa en sangre antes del desayuno para el paciente 02. Valores pronosticados para los próximos 5 días usando SVM.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Como se observa en las Figuras 14 y 15Existen diferencias significativas entre las predicciones que utilizan LR y SVM. En el experimento, para el día 95, LR predice un valor de 188.6 mientras que SVM predice un valor de 210.7. En nuestra propuesta, hemos decidido establecer un rango de predicción obtenido con los diferentes enfoques. Este rango indicará una medición razonable (no anómala) de glucosa en sangre para el día siguiente. Por ejemplo, teniendo en cuenta el historial de mediciones de glucosa en sangre antes del desayuno, para el paciente 02, se establece un rango de [188,6-210,7] para la medición del día 95. De manera similar, se llevarían a cabo predicciones para las medidas de almuerzo y cena. Estas nuevas reglas extraídas de la información diaria y personalizada se incluirán en el sistema de telemedicina con el objetivo de mejorar el sistema de alarma.

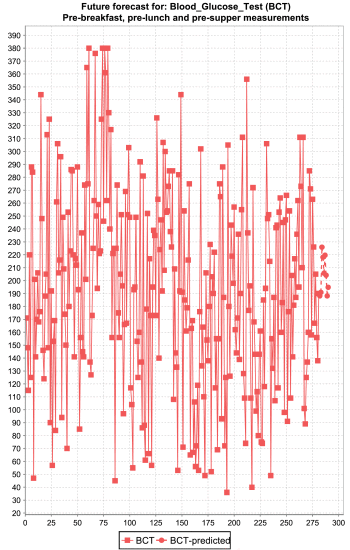
En consecuencia, para mejorar las predicciones de la medición previa al desayuno, hicimos otro experimento. Para esto, también utilizamos las mediciones antes del almuerzo y la cena. La idea es que la medición previa al desayuno no depende exclusivamente de las mediciones obtenidas en el desayuno de los días anteriores, sino que también depende de las mediciones obtenidas en el almuerzo y la cena de los días anteriores. Las predicciones de las pruebas de glucosa en sangre antes del desayuno, antes del almuerzo y antes de la cena para los próximos 3 días (9 predicciones) se muestran en las Figuras 16 y 17 .

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral16-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral16-2857499-large.gif)

**FIGURA 16.**

Mediciones diarias de glucosa en sangre antes del desayuno, antes del almuerzo y antes de la cena para el paciente 02. Valores pronosticados para los próximos 3 días usando LR.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral17-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral17-2857499-large.gif)

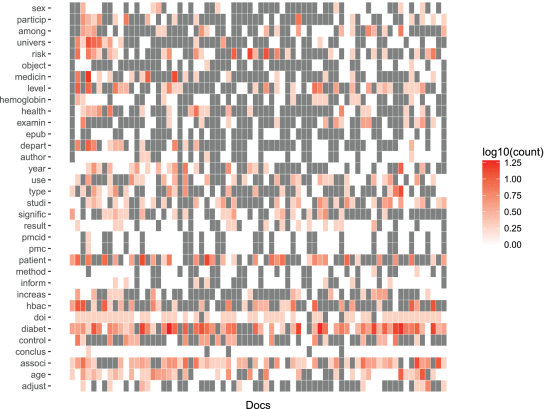
**FIGURA 17.**

Mediciones diarias de glucosa en sangre antes del desayuno, antes del almuerzo y antes de la cena para el paciente 02. Valores pronosticados para los próximos 3 días usando SVM.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

En las Figuras 16 y 17 no hay diferencias significativas entre las predicciones que usan LR y SVM. En el experimento, para la medida 282 (desayuno en el día 95), LR predice un valor de 188.1 mientras que SVM predice un valor de 188.2. En este caso, se establece un rango de [188.1-188.2] para la medición previa al desayuno del día 95. En conclusión, incorporaremos estas reglas más precisas, obtenidas con las tres mediciones diarias, en el sistema de telemedicina.

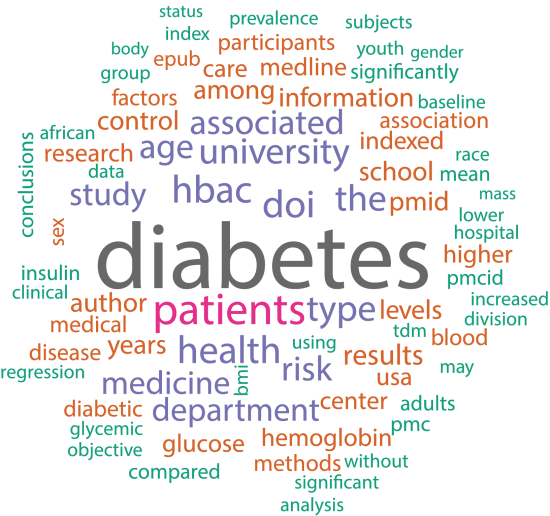
Finalmente, para completar los experimentos, hemos llevado a cabo un análisis para determinar los temas (atributos) de los conceptos más utilizados en los documentos seleccionados en el proceso de extracción de reglas web (documentos PubMed y PLOS ONE). Los atributos principales según su frecuencia en los documentos se representan en la Figura 18 y la Figura 19 . El color o el tamaño de la letra de cada atributo representado en ambas figuras indican su frecuencia. Estos gráficos corroboran los atributos más destacados de los documentos ( *diabetes* , *hbac* , *edad* , *sexo* , etc.) y que se han expresado en las reglas extraídas de la Web.

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral18-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral18-2857499-large.gif)

**FIGURA 18.**

Ejemplo de frecuencias de palabras sobre documentos PubMed y PLOS ONE.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

[[](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral19-2857499-large.gif)](https://ieeexplore.ieee.org/mediastore_new/IEEE/content/media/6287639/8274985/8413084/peral19-2857499-large.gif)

**FIGURA 19.**

Ejemplo de minería de texto visual sobre documentos PubMed y PLOS ONE.

[Ver todo](https://ieeexplore.ieee.org/document/8413084/all-figures)

Gracias a nuestro sistema, todo este proceso se desarrolla en tiempo real, donde los pocos segundos que el equipo médico necesita para tomar la decisión correcta son cruciales para la vida del paciente. Por lo tanto, la principal contribución de nuestro sistema es que proporciona al equipo médico información en tiempo real para tomar las decisiones médicas correctas para prevenir posibles deterioros de los pacientes con diabetes tipo 1.

**SECCION V.**

Conclusión

En este artículo, se han presentado algunas de las tendencias más novedosas en medicina. La literatura de vanguardia ha revelado que los métodos tradicionales no son suficientes, al menos por sí solos, para hacer frente a la situación actual del big data. La arquitectura propuesta en este documento gestiona datos masivos y realiza experimentos eficientes de acuerdo con la complejidad de los escenarios de big data. Se ha definido una arquitectura orientada a la ontología en la que se ha utilizado una ontología central (WordNet) como KB que permite la integración de datos de las diferentes fuentes heterogéneas que utilizan diversas ontologías.

El enfoque se ha aplicado en el campo de la medicina personalizada (estudio, diagnóstico y tratamiento de enfermedades personalizadas para cada paciente). Los métodos de IA se han utilizado con el objetivo de extraer datos en el sector de la salud para descubrir conocimiento oculto en fuentes de datos heterogéneas. Se obtiene un conjunto de reglas aprendidas (utilizando técnicas de minería de datos en datos estructurados, reglas de DM) y sus mejoras (aplicando técnicas de PNL en datos de la Web).

Se ha demostrado que un estudio de caso en el escenario de la diabetes demuestra la validez del modelo propuesto. Se ha presentado un sistema de telemedicina que ayuda al médico a proporcionar tratamiento para la diabetes y tomar decisiones. El sistema permite al médico mejorar las reglas de DM al integrar la información recopilada de documentos web especializados. Con esta nueva información, el médico puede crear un tratamiento personalizado para el paciente (dependiendo de sus características específicas) y analizar las alarmas producidas en el sistema de telemedicina con mayor precisión.

Las principales novedades presentadas son las siguientes: (1) una arquitectura orientada a la ontología que utiliza una ontología central que permite la comunicación entre diferentes fuentes de datos, cada una con su propia ontología; (2) una mejora de los sistemas tradicionales de IA en el tratamiento de la diabetes. El tratamiento personalizado de cada paciente y la mejora de los sistemas tradicionales de IA han sido posibles al incluir diferentes fuentes de datos; y (3) la aplicación de la arquitectura propuesta en el sistema de telemedicina para mejorar su rendimiento.

En un estudio futuro a corto plazo, y como parte de un proyecto en curso, el objetivo es mejorar el rendimiento de otra de las V de Big Data. En particular Velocity, al incorporar un marco flexible para sistemas embebidos en tiempo real. Este proyecto proporciona soluciones para una amplia gama de dispositivos con capacidades muy heterogéneas para las cuales es difícil predecir los tiempos de respuesta [117] . En el trabajo futuro, planeamos incluir nuevas fuentes de datos como las redes sociales. Además, se debe abordar la automatización del proceso para mejorar las reglas de DM, incluidas las reglas obtenidas después de aplicar técnicas de PNL en datos web (reglas web).