Sistema de Soporte a la Toma de Decisiones basado en datawarehouse para pacientes diabéticos

M. E. Llorente [1], A. Sigura [1,2], A. J. Hadad [1,2], B. Drozdowicz [1,2]

[1]Facultad Ciencia y Tecnología, Universidad Autónoma de Entre Ríos

[2]Facultad Ingeniería, Universidad Nacional de Entre Ríos

Ruta 11, Oro Verde, Entre Ríos, Argentina

mellorente@arnet.com.ar, bdrozdo@santafe-conicet.gov.ar

Resumen

El proyecto propone el diseño y desarrollo de una Datawarehouse (DW) para el apoyo a la toma de decisiones de profesionales médicos que atienden pacientes diabéticos. Estos pacientes requieren de un control y seguimiento continuo de su estado a través de chequeos periódicos. Por otro lado también tienen una mayor probabilidad de diversas complicaciones, tales como problemas cardíacos, de cicatrización, oftalmológicos, entre otros. Como consecuencia de estas situaciones se generan una gran cantidad de datos e información (señales, imágenes, etc.), que generalmente se encuentran en forma no vinculada en diferentes ámbitos y formas, esto genera dificultades para realizar una interpretación integral de la evolución de estos pacientes a lo largo de períodos prolongados de tiempo. La integración y procesamiento de dicha información mejora las decisiones relacionadas a diagnósticos y tratamientos. Considerando las diferentes fuentes de información se analizarán y definirán metodologías ETL (Extraction-Transformation-Loading) apropiadas para la identificación de información relevante del lado de las fuentes y sus aspectos semánticos (ontologías). También se analizan los procesos de extracción, limpieza e integración en formatos comunes, para su posterior carga en la DW, orientando a facilitar procesos de minería de datos.

Palabras clave: Datawarehouse, OLAP, Pacientes Diabéticos, Metodologías ETL, Granularidad

Contexto

El presente trabajo describe un Proyecto de Investigación Plurianual (PIDP) denominado "*Sistema de Soporte a la Toma de Decisiones basado en datawarehouse para pacientes diabéticos*". Dicho proyecto es desarrollado en la Facultad de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos (FCYT - UADER).

Introducción

La necesidad de lograr una información integrada se ha convertido en una prioridad para los niveles de la toma de decisiones de una organización o proceso. La tecnología brinda herramientas para tratar esta problemática, por ejemplo la potenciaciónde las Bases de Datos ha permitido una administración más segura y ágil de la información que a menudo procede de varias fuentes u orígenes de datos.

Para estos casos se han creado almacenes de datos o datawarehouses (DW) que permiten satisfacer necesidades particulares.

La mayor capacidad de procesamiento de las herramientas y la aplicación de técnicas analíticas, ha dado como resultado la posibilidad de almacenar y procesar datos en DW en forma diferente a lo que sucede con las BD (Bases de Datos) relacionales tradicionales [1].

Estas BD están basadas en el concepto de Procesamiento Analítico On Line (OLAP), almacenan información proveniente de diferentes orígenes, con un concepto integrador y orientado a la toma de decisión; lo que plantea

un concepto diferente al hasta entonces utilizado Procesamiento de Transacciones On Line (OLTP), como es la administración de la información proveniente de los distintos procesos presentes en las diferentes fuentes de información [2,3].

W.H. Inmon define DW como: "un conjunto de datos orientado a temas, integrado, no volátil, variante en el tiempo, como soporte a la toma de decisiones" [1].

Las herramientas de Soporte de Decisión se refieren a las aplicaciones que se emplean para manipular y analizar los datos del DW y luego presentar los resultados. Estas herramientas se utilizan para diferentes tareas (verificación y descubrimiento) y bajo diferentes enfoques de análisis (informático, analítico y minería de datos). Para tareas de verificación, se propone una hipótesis de trabajo e intenta comprobarla a través de la información existente en el DW, descubriendo ó resaltando a través de procesos de *Minería de Datos*, posibles asociaciones entre los datos existentes en el DW. [4,5].

El modelo de datos multidimensional es una buena opción para las tecnologías OLAP y de soporte a la toma de decisiones. Frente a las multibases de datos, que dan acceso a bases de datos inconexas y en general heterogéneas, un DW es con frecuencia un conjunto de datos integrados provenientes de fuentes diversas, procesados para su almacenamiento en un modelo multidimensional.

Esta variedad en la fuente de información, cada una de ellas con grandes diferencias en su propósito, calidad de información, objetos descriptos, temporalidad, clasificación y tipificación, impone un análisis y un trabajo exhaustivo para compatibilizar la heterogeneidad planteada.

A diferencia de la mayoría de las bases de datos transaccionales, los DW suelen mantener series de tiempo y análisis de tendencia, que necesitan más datos históricos de los que contienen generalmente las bases de datos transaccionales. La información del DW es menos detallada (de grano grueso) y se actualiza de acuerdo a una política elegida con cuidado, y que es generalmente incremental. Como los DW están libres de las restricciones del entorno transaccional, se mejora la eficiencia del procesamiento de consultas.

Datawarehouses en el ámbito médico

El almacenamiento de grandes cantidades de datos en datawarehouses, es una parte central en el desarrollo de sistemas de información que den soporte al proceso de toma de decisiones y actividades de investigación en el ámbito médico, como por ejemplo la práctica de medicina basada en la evidencia [18]. Esto esta motivado por la necesidad de las instituciones del cuidado de la salud, de relacionar la mayor cantidad de información sobre el paciente con el objetivo de mejorar el proceso de atención al mismo y reducir los retardos y costos asociados.

En lo que refiere al modelado de los datos, el modelo conceptual clásico de los componentes de los datawarehouses, no permite el procesamiento de datos segmentados en el tiempo, como por ejemplo la administración de drogas. Esto ocurre porque estos datos poseen una granularidad menor que la granularidad básica presente en el modelo conceptual clásico. En general las tareas de modelado para problemática dentro del ámbito médico requiere el preprocesamiento de datos, a fin de contextualizar su contenido informativo para facilitar su interpretación en un proceso de toma de decisiones. Dicho preprocesamiento puede involucrar la aplicación de técnicas para abstracción temporal, generación de índices, extracción de características relevantes, identificación de estados, etc. [12-17].

En [18] se propone ampliar el modelo conceptual clásico a través de XML. En el caso de la administración de drogas, se modela la información asociada como un atributo complejo de XML. Para esto define los elementos *AdminDose* (granularidad básica del DW) y *dose* como un "acto" de consumo de la droga en un determinado instante de tiempo. El

elemento *dose* a su vez tiene asociado atributos tales como tiempo de consumo, unidad de tiempo, cantidad y dosis.

Así como en el caso de las drogas, en áreas tales como unidades de cuidados intensivos (ICU), surge la necesidad de almacenar información de menor nivel de granularidad, tales como la evolución (resumida), por ejemplo, de un paciente diabético durante su estadía en ICU y los incidentes ocurridos en ese lapso[19].

Líneas de investigación y desarrollo

1. Estructuras de datos representativas del dominio de análisis.

2. Métodos ETL para fuentes de información de referencia.

3. Relaciones que vinculen la semántica de las distintas fuentes de información (ontologías).

4. Definición de la información pasible de ser analizada a través de procesos de Data Mining

5. Granularidad y dimensiones del modelo.

6. Metadatos y reglas de negocios.

Resultados y Objetivos

Los objetivos del proyecto están orientados a la obtención de métodos y modelos para sistemas de información, que permitan a los profesionales médicos asociados a pacientes diabéticos, visualizar en forma integral los estados y eventos vinculados a estos pacientes, en diferentes escalas de tiempo y a partir de diversas fuentes de información. Esto permitirá a los profesionales tomar mejores decisiones y por lo tanto mejorar la calidad de vida de los pacientes. La acumulación de estos datos en una DW ofrece una base apropiada para un proceso de Minería de Datos, el cual permite la extracción de conocimiento útil, no trivial, y previamente no existente.

Más precisamente este proyecto está orientado a estudiar la viabilidad y aplicación de metodologías de DW, para la integración de múltiples dominios de información asociados al paciente diabético, y que puede ser utilizado para evaluar la eficacia de la gestión de la enfermedad diabética.

Teniendo en cuenta que el proyecto se encuentra en su etapa inicial de desarrollo hasta este momento no se han obtenido resultados publicables. Sin embargo, esbozaremos a continuación las estrategias de avance previstas hasta este momento.

En relación a las líneas de investigación *1* y *2*, para nuestro caso las fuentes de información a considerar son:

- Historias Clínicas (Consultório, Eventos, Medicamentos).
- Contenido de las HC de Diabéticos
- Datos de Laboratorio. Tipos de análisis y/o estudios. Contenido.
- Datos Recogidos por el paciente en su casa (Ej.: Toma de datos de niveles de glucosa en sangre a través de dispositivos portátiles)
- Base de Datos de Imágenes. Caso de referencia: Retinopatías Diabéticas. Análisis Evolutivo de la patología, registración de imágenes, etc.
- Internaciones: Sala Común – Terapia Intensiva (ICU). Caso de Referencia: Problemas Circulatorios. Señales y planillas de datos del paciente durante las internaciones. Tablas y Relaciones. Abstracción temporal de las señales
- Cirugía
- Base de Datos de Farmacia

También se identifica la importancia de cada fuente de información y el tiempo de actualización de los datos. En este sentido se considera que una de las medidas más importantes para el paciente diabético es el HbAa1c%, el cual indica el nivel de azúcar en sangre de largo plazo y brinda un buen indicador del estado del paciente durante los

meses recientes. Esta medida es tomada aproximadamente cada 3 meses.
Además para el paciente diabético, un estilo de vida saludable es más importante que para un paciente normal, y literalmente puede marcar la diferencia entre la vida y una muerte prematura. Para monitorear el estilo de vida se consideran varios factores. Estos incluyen peso, hábito de fumar, ingesta de alcohol y hábitos de ejercicio. Estos factores no son medidos de manera regular, pero pueden ser considerados válidos desde una registración a otra.
En relación a la línea de investigación *5*, en este contexto se considero que un DW clínico debe dar soporte al análisis de datos en varios niveles. El más bajo es el nivel del paciente, donde los datos sobre el paciente individual se pueden visualizar y analizar, por ejemplo, para encontrar un patrón en el desarrollo de una enfermedad vinculado al mismo. Este nivel de análisis se centra en dar al paciente en particular el mejor tratamiento posible, y por tanto es importante para la práctica de la atención médica.
El siguiente es el nivel de grupo de pacientes, donde los datos sobre el mismo son analizados, por ejemplo, cuando tengan características clínicas asociadas. Una aplicación de este nivel es la gestión de la calidad clínica, donde los tratamientos y los resultados son analizados y comparados con las normas, a fin de identificar cómo el proceso de atención se puede mejorar. Este nivel también es de interés en la investigación médica, por lo que es importante desde el punto de vista más científico.
Un nivel más amplio sería el de una empresa/institución de salud, donde clínicos, administradores y especialistas en epidemiología, combinan datos para investigar la calidad y la rentabilidad global de los servicios proporcionados. Este nivel de análisis se centra en el rendimiento general de la empresa/institución y es importante desde una perspectiva de gestión.
Otro aspecto de esta línea de investigación es considerar la granularidad temporal de cada base de datos, considerando que finalmente estos datos en la DW deben quedar consistentes.
Por otro lado es posible que parte de los datos crudos no sean utilizados en el funcionamiento del DW, pero en el diseño se debe tener en cuenta el potencial total de estos datos. Puede suceder que algunos requerimientos puedan surgir avanzado el proyecto y que requiera de datos que al principio no se consideraron. Si para el diseño del DW se conocen los modelos completos de las diferentes fuentes, resulta más fácil cambiar el modelo del DW y adaptar los procedimientos de recolección para incluir los nuevos datos requeridos.
También en esta línea se debe considerar que varias dimensiones y sus atributos modelados en un ambiente de DW pueden incluir, reglas y restricciones, mapeos, formatos, tipos de medidas, instrumentos, dimensiones, atributos, campos, tratamientos, pacientes, prácticas médicas, prescripciones médicas, resultados de pruebas, períodos, unidades, niveles de glucosa, el tipo de la glucosa, resultados de oftalmología, la ingesta de alimentos, tipos de alimentos, higiene de los alimentos normales, y el nivel de colesterol.
En relación a las líneas de investigación *2* y *6*, se considera el aspecto complementario de las diferentes fuentes de información, tendiendo a identificar cualquier tipo de inconsistencias. Por lo tanto será necesario desarrollar filtros y detectores de inconsistencias.
Este aspecto influye en la calidad de los datos de los registros de pacientes que, por ejemplo, pueden tener errores tipográficos, valores perdidos o información incorrecta de los pacientes. También los registros pueden estar duplicados, y no poseen a menudo un formato compatible para el modelado mediante ontologías. En este sentido se considera que el enfoque del DW reconcilia las diferencias de formato y esquemas de codificación, permitiendo la exploración y el descubrimiento de patrones. La visualización se utiliza para mejorar la comprensión de la representación de los registros de la evolución del paciente diabético.

En relación a la línea de investigación *3*, es importante la construcción de una ontología que permita mapear los términos más relevantes con el objetivo propuesto para la DW. Esto es necesario pues en una estructura tan distribuida de las fuentes de información, la denominación de los atributos y de sus valores pueden ser diferentes dependiendo del entorno que les dio origen. El esquema central y la arquitectura de la DW deben estar basados en esta ontología.

Un enfoque de este tipo permite abordar problemas relacionados con la falta de conectividad, de comunicación y la interacción entre los profesionales médicos, nutricionistas, farmacéuticos y las organizaciones de bienestar social involucrada en el control de la diabetes. Un apropiado modelado y almacenamiento de los datos permitirá la descripción y la integración de múltiples fuentes dentro y fuera de las organizaciones.

El servicio de un sistema de información que utiliza una ontología sanitaria es proporcionar una representación normalizada para los datos de salud.

El desarrollo de un sistema de información consta de una plataforma de almacenamiento y uso de una ontología de salud, en la que se describan los conceptos y sus relaciones derivados del corpus de conocimiento de dominio específico y la vinculación con la normalización de los sistemas terminológicos.

En relación a la línea de investigación *4*, uno de los objetivos es realizar "data mining" en el DW. Se debe tener esto en cuenta para definir la estructura de los datos, así se deben analizar métodos que permitan transformar datos en forma de series temporales e información secuencial en estructuras más adecuadas para la minería de datos. La minería datos podría utilizarse para predecir la evolución de la diabetes o encontrar condiciones de alto riesgo, para proponer nuevos procedimientos, mejorar los resultados de los diferentes estados de la diabetes o para identificar casos "outliers" con buena calidad de atención.

Formación de Recursos Humanos

El equipo de trabajo esta conformado por especialistas del área informática y de bioingeniería. Integrantes del equipo tienen formación de postgrado tanto en el área de sistemas de información como en el área biomédica, así como también experiencia en el ámbito profesional en lo que refiere al desarrollo de sistemas.

Referencias

[1] Fundamentos de Sistemas de Bases de Datos, Elmasri-Navathe.3ra Edición, Addison Wesley, 2000

[3] The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses. Decision Support Systems, Vol 27, Issue 3, 1999, Pages 289-301. Anindya Datta, Helen Thomas

[4] A survey on summarizability issues in multidimensional modeling. Data & Knowledge Engineering, Volume 68, Issue 12, December 2009, Pages 1452-1469. Jose-Norberto Mazón, Jens Lechtenbörger, Juan Trujillo

[5] The Development of Health Care Data Warehouses to Support Data Mining. Clinics in Laboratory Medicine, Volume 28, Issue 1, March 2008, Pages 55-71. Jason A. Lyman, Kenneth Scully, James H. Harrison Jr.

[12] "Temporal Abstraction for the Analysis of Intensive Care Information" A. Hadad, D. Evin, B. Drozdowicz, O. Chiotti. Journal of Physics: Conference Series. Volume 90, 2007. ISSN: 1742-6596

[13]"A Comparative Analysis of Preprocessing Techniques in Colour Retinal Images",Adrián Salvatelli, Gustavo Bizai, Gisela Barbosa, Bartolomé Drozdowicz, Claudio Delrieux. Journal of Physics-Conferences Series (JPCS). Noviembre de 2007.

[14]"Implementación y aplicación de algoritmos Retinex al preprocesamiento de imágenes de retinografía color". Autores: N. Londoño, G. Bizai, B. Drozdowicz. Revista Ingeniería Biomédica. Volumen 3. Paginas 36-43. ISSN 1909 – 9762.

[15]"Modelos de seguimiento para la supervisión de procesos complejos en aplicaciones biomédicas". Autor: Alejandro Hadad. Encuentro Internacional de Investigación en Ingeniería de Sistemas e Informática. Tunja, Colombia, 6 al 8 de Octubre de 2010.

[18] Marko Banek, A. Min Tjoa, Nevena Stolba: Integrating Different Grain Levels in a Medical Data Warehouse Federation. DaWaK 2006: 185-194

[19] Pitarch et al. Context-Aware Generalization for

Cube Measures. DOLAP'10, October 30, 2010, Canada