Análisis de fuentes de información para el proceso de diseño de un datawarehouse sobre pacientes diabéticos

M. E. Llorente[1], A. Sigura[1,2], J. Besso[1], E. Mangia[1] ,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

Para el desarrollo de un modelo de datos para un Datawarehouse (DW) se tuvo en cuenta la evolución en las metodologías de diseño de los Modelos Multidimensionales, en consecuencia es posible definir que las mismas están basadas en dos aspectos fundamentales: Requerimientos y Fuentes de Información existentes.
En este trabajo se describen las diversas fuentes de información y sus estructuras de datos a tener en cuenta para el diseño y desarrollo de un DW para el apoyo a las decisiones de profesionales médicos que atienden pacientes diabéticos.
Analizando las fuentes de información existentes, en una primera clasificación, se distinguen datos estructurados y no estructurados. Se denomina estructurada a toda aquella fuente de información que tenga un diseño lógico y conceptual, ejemplo una base de datos de historia clínica. Es decir que exista información sobre el dato en sí mismo.
Dentro del conjunto de las fuentes no estructuradas se incluyen aquellas que necesitan un procesamiento previo para contextualizar los datos contenidos y convertirlos en información. En este grupo se encuentran las señales fisiológicas, imágenes y el texto libre, los cuales indefectiblemente necesitan un preprocesamiento, para formar parte de una estructura orientada al análisis, como es un DW.

Palabras clave: Datawarehouse, Pacientes Diabéticos, Estructuras de Datos, Metodologías ETL, Herramienta de Desarrollo

Contexto

El presente trabajo se inserta en 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

Este trabajo está orientado a describir parte del proceso para un primer modelo multidimensional del DW a ser utilizado en el PIDP mencionado.

El diseño del mismo está basado en un enfoque de tres niveles de análisis, a través de un proceso iterativo [1].

Para este desarrollo el más bajo es el nivel del paciente individual, donde los datos sobre el mismo 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 una enfermedad particular asociada. El tercer nivel es el relacionado con una empresa/institución de salud, donde profesionales clínicos, administradores y especialistas en epidemiología, combinan datos para investigar la calidad, efectividad y eficiencia global de los servicios proporcionados.

En esta etapa del modelado se considerarán solamente los datos relacionado con el primer nivel, es decir el del paciente individual.

Para el desarrollo del modelo se tuvo en cuenta la evolución en las metodologías de diseño de los Modelos Multidimensionales desde el año 1998 a la actualidad, y analizando las propuestas de diferentes autores de referencia [2], es posible definir que las metodologías están basadas en dos aspectos fundamentales:

- Requerimientos
- Fuentes de Información existentes

En el trabajo presentado en [1] se realizó una primera propuesta de modelado, considerando solamente los Requerimientos basados en Casos de Uso, dentro de un proceso iterativo.

Esto fue necesario teniendo en cuenta la diversidad de las fuentes de información y la variabilidad de las características y condiciones de los pacientes diabéticos. Por este motivo generalmente los usuarios médicos de un DW clínico, como el propuesto en este trabajo, no tienen perfectamente definidos como van a analizar los datos y por lo tanto resulta casi imposible comenzar su diseño conociendo todos los requerimientos a priori [3]. En consecuencia en un proyecto de DW clínico puede resultar necesario implementar para su diseño un análisis del tipo iterativo. Un objetivo de estos DW es ser lo suficientemente flexibles para tratar con estos cambios, esto conlleva a considerar un proceso de diseño iterativo diferente al proceso incremental, aún cuando este último sea el más convencional.

Por su parte el presente trabajo se enfoca en el segundo de los aspectos indicados anteriormente, las fuentes de información existentes [1,4], en lo que refiere a su estructura de origen y necesidades de transformación en cada caso. Estos aspectos forman parte de las líneas de investigación descriptas en [4] para el PIDP.

En un DW clínico se deben involucrar procesos para el procesamiento de imágenes, señales y datos, como ser: registración, extracción y cuantificación de características. También se requiere el modelado de datos de multimedia en variadas formas- texto libre, reportes estructurados, imágenes en 2 o 3 D, capacidad de hacer zoom a las imágenes, señales, datos espectrales, gráficos, video, publicaciones escaneadas, en lugar de datos en formato texto como la mayor parte de las DW empresariales. Deben tener capacidad para buscar cualquier formato de información en forma cualitativa o cuantitativa y no solamente texto estructurado o números en registros.

Los DW clínicos enfatizan la preparación y adquisición de datos con protocolos predefinidos. Además proveen herramientas analíticas y estadísticas para soportar procesos de verificación o de Minería de Datos.

Líneas de investigación y desarrollo

1. Estructuras de datos representativas del dominio de análisis.

2. Métodos ETL para fuentes de información de referencia.

Resultados y Objetivos

Analizando las fuentes de información existentes, en una primera clasificación, es posible distinguir datos estructurados y no estructurados. En este contexto se denomina estructurado a toda aquella fuente de información que tenga un diseño lógico y conceptual, como por ejemplo una base de datos de Historias Clínicas. Es decir que exista información sobre el dato en sí mismo.
Dentro del conjunto de las fuentes no estructuradas se incluyen aquellas que necesitan un procesamiento previo para contextualizar los datos contenidos y convertirlos en información. En este grupo se encuentran las señales fisiológicas, imágenes y el texto libre, los cuales indefectiblemente necesitan un preprocesamiento, para formar parte de una estructura orientada al análisis, como lo es un DW. Un resumen de esta clasificación se presenta en la Figura 1.

Esta diferenciación entre fuentes de información tiene consecuencias en ciertos aspectos de la implementación, como ser la formulación de las ETL que proveerán de datos al DW.

Los requerimientos de información identificados anteriormente proporcionarán las bases para realizar el diseño y la modelización del Modelo Multidimensional del DW. En esta fase se identificarán las fuentes de los datos (sistema operacional, fuentes externas) y las transformaciones necesarias para, a partir de dichas fuentes, obtener el modelo lógico de datos del DW. Este modelo estará formado por entidades y relaciones que permitirán resolver las necesidades del proceso de atención del paciente. El modelo lógico se traducirá posteriormente en el modelo físico de datos que se almacenará en el DW y que definirá la arquitectura de almacenamiento del mismo adaptándose al tipo de uso que se realice.

Teniendo en cuenta este análisis se realizará el desarrollo de un primer prototipo utilizando la herramienta provista por IBM (IBM InfoSphere Platform), para la cual la institución tiene un convenio de uso académico de sus desarrollos [5].

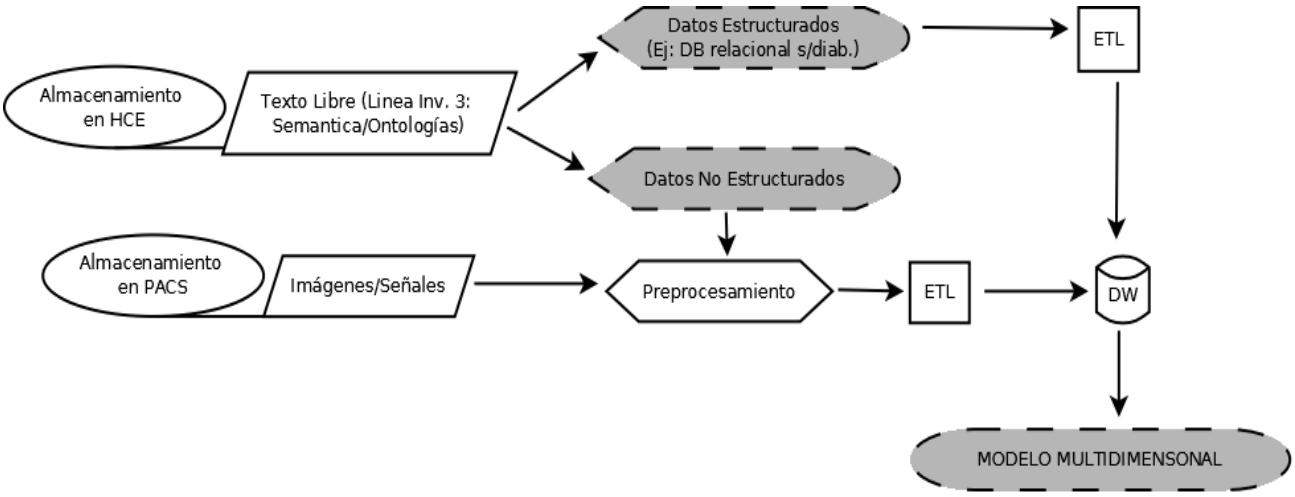

Figura 1 - Clasificación de las fuentes de información

Considerando las diferentes estructuras de datos que generan información se procederá a la definición de las ETL correspondientes y los preprocesamientos previos necesarios.

La herramienta indicada anteriormente tiene capacidad para extraer datos automáticamente de diferentes fuentes u orígenes, como las mencionadas a continuación:

- Base de datos relacionales (MySQL, MS SQLServer, Progress, DB2, Informix, Oracle, Teradata).
- Archivos de texto.
- Extracción desde planillas de cálculo, archivos separados por punto y coma (.csv).
- Servicios web
- Archivos XML

Los orígenes de datos que puede manipular la herramienta incluyen archivos indexados, archivos secuenciales, bases de datos relacionales, orígenes de datos externos, aplicaciones y colas de mensajes. Ello puede implicar algunas de las transformaciones siguientes:

- Conversiones de tipos de datos y de formato de serie y numérico.
- Derivaciones y cálculos que apliquen algoritmos y normas a los datos.
- Comprobaciones y aplicación de datos de referencia para validar identificadores. Este proceso se utiliza para crear un depósito de datos normalizado.
- Conversión de datos de referencia de orígenes dispares a un conjunto de referencia común, creando coherencia entre estos sistemas. Esta técnica se utiliza para crear un conjunto de datos maestro (o dimensiones conformadas).
- Agregaciones para generación de informes y análisis.
- Creación de bases de datos analíticas o de informes, como por ejemplo cubos o depósitos de datos. Este proceso implica desnormalizar datos en estructuras tales como esquemas de estrella o de copo de nieve para mejorar el rendimiento y la facilidad de uso para los usuarios.

Las fuentes de información que se consideran en el dominio de atención de pacientes diabéticos son:

- Historias Clínicas (Consultorio, Eventos, Medicamentos). Para este trabajo se considera que las historias clínicas están almacenadas en una Base de Datos Relacional (BDR).

- Datos de Laboratorio o Tipos de análisis y/o estudios. Contenido. Tablas y Relaciones. **BDR**

- Datos Recogidos por el paciente en su casa (Ej: Toma de datos de niveles de glucosa en sangre a través de dispositivos portátiles). **Archivos de Texto Estructurado**

- Base de Datos de Imágenes en formato DICOM. La imagen contenida no tiene ningún procesamiento lo cual se considera un dato no estructurado. Para que pueda ser interpretado por la herramienta, se debe generar un módulo de preprocesamiento que brinde la información asociada en un modelo de datos. Ejemplos: Imágenes de Fondo de Ojo con presencia de Retinopatías Diabéticas. Análisis Evolutivo de la patología, registración de imágenes, etc.[6-9]

- Base de Datos de señales fisiológicas. Caso similar al de las imágenes. Ejemplos: señales de electrocardiograma, presión arterial, saturación de oxigeno, etc. Análisis de estabilidad hemodinámica, patrones temporales de arritmias, índices de severidad, etc. [10-12]

- Base de Datos de Farmacia. Medicamentos. **BDR + Planillas de cálculo**

- Información complementaria accesible vía **servicios web**, como ser aspectos regulatorios (ANMAT) o información de otras instituciones (Colegios Médicos, Obras Sociales, etc.).

Este relevamiento ha permitido considerar los aspectos más relevantes del dominio, sin embargo con el avance del proyecto pueden surgir otras consideraciones sobre esta temática.

Como trabajo futuro, lo descripto en este trabajo relacionado con las diferentes estructuras de datos con las que va a tener interacción el DW, se complementará con el proceso iterativo de diseño de la DW basado en Casos de Uso [1].

Formación de Recursos Humanos

El equipo de trabajo está 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] Proceso de Diseño basado en Casos de Uso para un Datawarehouse Clínico. M. E. Llorente, Aldo Daniel Sigura, Javier Besso, Alejandro Hadad, Bartolomé Drozdowicz. CACIC 2012

[2] A survey of Multidimensional Modeling Methodologies. Oscar Romero, Alberto Abelló. International Journal of Data Warehousing & Mining, 5(2), 1-23, April-June 2009

[3] Wong S, Hoo K, Knowlton R., et al. Design an applications of a multimodality image datawarehouse framework. J Am Med Inform Assoc. 2002; 9: 239-254

[4] Sistema de soporte a la toma de decisiones basado en datawarehouse para pacientes diabéticos. M. E. Llorente, Aldo Daniel Sigura, Alejandro Hadad, Bartolomé Drozdowicz. WICC 2012

[5] Sitio Oficial de IBM Infosphere DataStage: http://www-01.ibm.com/software/data/infosphere/datastage/

[6] “Implementación y aplicación de algoritmos Retinex al preprocesamiento de imágenes de retinografía color”, N. Londoño, G. Bizai, B. Drozdowicz, Revista Ingeniería Biomédica, ISSN 1909-9762, volumen 3, número 6, julio-diciembre 2009, págs. 36-46.

[7] "Algorithms evaluation for fundus images enhancement", Braem V., Marcos M., Bizai G., Drozdowicz B, Salvatelli A., Journal of Physics: ConferenceSeries 332 (2011) 012035.

[8] “Analysis and Implementation of Methodologies for the Monitoring of Changes in Eye Fundus Images”, A. Gelroth, D. Rodríguez, A. Salvatelli, B. Drozdowicz, G. Bizai, Journal of Physics: Conference Series (JPCS), 23 December 2011, Vol. 332 (2011) 012036, con referato doi:10.1088/1742-6596/332/1/012036

[9] “Monitoring of Changes in Fundus Image”, Authors: R. M. Torres, A. Gelroth, D. Rodriguez, A. Salvatelli, B. Drozdowicz, G. Bizai. ARVO/ISIE Imaging Conference Saturday, May 5, 2012, Grand B Ballroom, Greater Fort Lauderdale/Broward-Poster, http://www.arvo.org/eweb/startpage.aspx?site=isie

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

[11] Hypotension States’ Prediction by using the Hidden Markov Models. Diego Evin, Alejandro Hadad, Mauro Martina, Bartolomé Drozdowicz Revista Facultad de Ingeniería, UPTC, 2011, vol. 20, No. 30, pp 55-63, ISSN 0121–1129.

[12] Hadad, Alejandro Javier; Solano, Agustin Ezequiel y Drozdowicz, Bartolomé (2012). “Prototipo para la comparación de patrones temporales secuenciales de arritmias cardíacas”. En: Ventana Informática No. 26 (ene.-jun., 2012). Manizales (Colombia): Facultad de Ciencias e Ingeniería, Universidad de Manizales. pp 29-43 ISSN: 0123-9678