MIGRACION DE SISTEMAS HEREDADOS: UNA METODOLOGÍA DE APOYO BASADA EN EL USO DE HERRAMIENTAS DE KDD (KNOWLEDGE DISCOVERY IN DATABASES).

M. Angélica Caro Gutiérrez
Departamento de Auditoría e Informática,
Universidad del Bio Bio, mcaro@ubiobio.cl

Jorge Bocca
Indecs Ltda.

Daniel Campos
Departamento de Ingeniería de Sistemas
Universidad de Concepción

Resumen

Los Sistemas de Información Heredados (SIH) están presentes en la mayoría de las organizaciones y, en muchos casos, son de importancia estratégica para éstas. Esta clase de sistemas al igual que las organizaciones deben evolucionar y adaptarse a nuevos requerimientos. Frente a esto, una de las soluciones recurrentes es la migración del sistema a uno nuevo que abarque los nuevos requerimientos e incorpore nuevas tecnologías. De acuerdo a distintas investigaciones realizadas en el área [BRODI95][BISBA97], uno de los factores de éxito de un proyecto de migración es el entendimiento del SIH, esto es, poder llegar a entender tanto el modelo de datos como el modelo de negocios que trata de cubrir el SIH.

Este artículo presenta una metodología que, basada en el uso de herramientas de KDD (Knowledge Discovery in Databases), aporta una alternativa real y factible de adquirir el conocimiento necesario de un SIH para llevar a cabo una migración exitosa, presentando además los resultados de la aplicación de dicha metodología en un caso real.

1. Introducción

En la actualidad, la gran mayoría de las organizaciones se ven enfrentadas a una constante necesidad de mantener y mejorar su nivel competitivo y satisfacer los cambiantes requerimientos del mercado. Como directa consecuencia de lo anterior, los Sistemas de Información destinados a apoyar la gestión de dichas organizaciones, se ven enfrentados al cambio. Estos cambios implican, entre otras cosas, la incorporación de nuevas tecnologías de información, nuevos requerimientos de los usuarios y del mercado.

Sin embargo, en la realidad nos encontramos con que no siempre es factible modificar y adaptar todos los sistemas de información para que se puedan incorporar los nuevos requerimientos de la organización que los alberga. Por otro lado, también debemos considerar que muchos de estos sistemas son de misión crítica, e inclusive algunos de ellos funcionan las 24 horas del día. Esto quiere decir, que la detención de dichos sistemas puede traer grandes pérdidas en tiempo, negocios y en la confianza de los clientes de una organización. Esta clase de sistemas de información, son denominados Sistemas de Información Heredados (de ahora en adelante SIH). La problemática asociada a ellos, esto es su evolución, en muchos casos es solucionada a través de la migración, esto es, mover el antiguo sistema a un nuevo ambiente o plataforma que permita que su mantención y adaptación a nuevos requerimientos sea más fácil. Durante el último tiempo han surgido diferentes estrategias para abordar la migración de un SIH, existiendo coincidencia entre ellas [BRODI95],[BISBA97],[GOLD98], respecto de que una tarea fundamental en todo proceso de migración es el entendimiento del SIH. Esto significa conocer y entender cuales son los requisitos que intentaba cubrir el SIH. Lo anterior básicamente porque el nuevo sistema deberá cubrir no sólo los nuevos requisitos de la organización sino también los que cubría el SIH. Principalmente, las fuentes utilizadas para lograr este entendimiento están compuestas por la documentación, el código, los diseñadores del SIH (si están en la empresa), las personas que dan soporte y mantenimiento y los usuarios del SIH. Lamentablemente, muchas veces estas fuentes no existen o están incompletas. Por ejemplo, la documentación no existe o está desactualizada, parte del código fuente no existe y/o el personal no tiene un amplio conocimiento del sistema y sus mantenciones. En consecuencia, inevitablemente surge la interrogante: ¿De qué forma podemos lograr el entendimiento del SIH?

El presente artículo presenta una metodología para apoyar la recuperación de requisitos de un SIH, basada en el uso de herramientas de minería de datos. Mediante esta estrategia se pretende reconstruir algunos aspectos básicos del SIH a migrar, de modo que sea posible entender el modelo de datos del SIH, entender el modelo de negocios que intentaba cubrir el SIH y determinar el nivel de calidad de los datos del SIH. Todos estos elementos son esenciales tanto para la migración de los datos, así como para la migración global del SIH.

Este artículo se organiza como sigue, la sección 2 discute el concepto de Sistemas de Información Heredados y su migración. La sección 3, aborda los conceptos de minería de datos y proceso KDD. La sección 4 presenta la metodología propuesta para la recuperación de requerimientos de un SIH. La sección 5 presenta un caso de aplicación de la metodología. Finalmente, en la sección 6 se presenta las conclusiones y en la sección 7 trabajos futuros.

2. Sistemas heredados

De acuerdo a la definición de Brodie y Stonebraker: "Un Sistema de Información Heredado es cualquier sistema de información que se resiste significativamente a cambios y modificaciones" [BRODI95]. Por otro lado, estos sistemas normalmente son de misión crítica dentro de una organización [BISBA97], esto significa que si alguno de ellos falla o se detiene traerá graves consecuencias en el desempeño de la organización. De acuerdo con Wu et al. [WU97], este tipo de sistemas conforman la columna vertebral del flujo de información en una organización y son el principal vehículo para la consolidación de información acerca del negocio de ésta.

Como las principales características de los SIH ([BRODI95],[BISBA97],[GOLD98]) podemos señalar:

- Típicamente son grandes, con millones de líneas de código,
- son antiguos, más de 8 años desde su construcción,
- están escritos en un lenguaje heredado (COBOL, assembler, etc.),
- se basan en bases de datos heredadas o archivos planos,
- generalmente funcionan en hardware obsoleto que es lento y caro de mantener,
- son autónomos (independientes de otras aplicaciones),
- generalmente, son difíciles de comprender y no existe documentación suficiente o apropiada acerca de ellos,
- su mantención implica un alto costo para la organización y
- que generalmente cumplen una "misión-crítica" dentro de la organización.

Si analizamos las características señaladas anteriormente, podemos detectar varios problemas asociados a este tipo de sistemas. Por ejemplo el alto costo que puede significar realizar mantenciones, esto debido a su gran tamaño (líneas de código) y a que normalmente la documentación es escasa, desactualizada o bien no existe. Otra situación problemática, es que a pesar de que normalmente están soportados por hardware y software obsoletos, estos sistemas son vitales para la organización que los posee y esto significa que se deben asumir las restricciones asociadas a ellos.

Problemática de los Sistemas de Información Heredados

Las organizaciones están en constante evolución. Los negocios y las reglas asociadas a ellos también cambian con cierta frecuencia. Esto enfrenta a las organizaciones a una real necesidad de que sus sistemas de información también evolucionen y es aquí entonces, cuando muchas de ellas se encuentran con problemas como: documentación escasa, desactualizada o inexistente, falta de programas fuentes y personal que no conoce en detalle los SIH.

Bisbal et al. [BISBA99], plantea que las soluciones a esta problemática se encuadran básicamente en 3 categorías: **redesarrollo**, que implica volver a escribir la aplicación existente; **wrapping** (envoltura), que provee una nueva interfaz para el SIH o algún componente de éste, lo que permite mayor accesibilidad desde otras aplicaciones; y **migración,** que mueve al SIH a un nuevo ambiente o plataforma más flexible, reteniendo la funcionalidad y los datos del sistema original. Cada una de estas soluciones tiene un mayor o menor grado de impacto en el sistema y, en consecuencia, en la organización. Por otro lado, la mantención no se aborda como solución para este tipo de sistemas porque se la considera parte del ciclo de vida de cualquier sistema de información, además de su alto costo.

Migración de Sistemas de Información Heredados

En los últimos años ha quedado de manifiesto la necesidad de dar solución al problema de los SIH. Entre las situaciones que han generado esta necesidad, podemos mencionar, entre otras: la constante necesidad de integrar sistemas dentro de una organización, el creciente interés de las distintas organizaciones por habilitar el acceso a sus sistemas, o parte de ellos, a través de internet, el interés por acceder a los beneficios que prometen las nuevas tecnologías (como por ejemplo redes, intranets, bases de datos, etc.) y por último, mejorar el servicio e imagen de la organización ante los clientes y la competencia.

Brodie y Stonebraker [BRODI95], plantean la migración como una solución lógica al problema de los SIH, la cual conllevaría la problemática de reemplazar el hardware y software, incluyendo las interfaces, aplicaciones, y bases de datos que componen la infraestructura del SIH, por un hardware y software nuevos y más modernos. Ellos consideran que la migración de un SIH implica comenzar con un SIH y terminar con un nuevo sistema de información equivalente. El nuevo sistema será significativamente diferente del original, pero deberá contener la funcionalidad elemental y los datos del SIH.

Bisbal et al. [BISBA97,BISBA99], consideran que las migraciones si son exitosas traen mayores beneficios a largo plazo. Por ejemplo, la migración permitirá mayor entendimiento de sistema, facilidad y reducción de costos en mantención. Ellos definen la migración de un SIH como el proceso de mover el SIH a un nuevo ambiente o plataforma que permita que el nuevo sistema de información sea fácilmente mantenido y adaptado a los nuevos requerimientos de negocios de la organización, sin que pierda la funcionalidad del SIH y sin tener que redesarrollarlo completamente.

Un hecho fundamental dentro del proceso de migración, este es que la funcionalidad del SIH debe mantenerse, lo que implicará que quienes tengan a cargo un proyecto de este tipo deberán preocuparse por adquirir el mayor conocimiento posible acerca de dicha funcionalidad. Todas las metodologías para la migración de SIH, revisadas durante esta investigación contemplan entre sus etapas una en la cual se cubre esta necesidad, destacándola como esencial para el éxito de un proyecto de este tipo. Sin embargo, a pesar de este reconocimiento, ninguna la aborda con el debido detalle. Por ejemplo, revisar Metodología Chicken Little y Metodología Butterfly. [BISBA97,BISBA99], [BRODI95].

Junto con lo anterior, es necesario destacar lo compleja que puede ser la tarea de recuperar los requisitos de un SIH, ya que dadas las características de este tipo de sistemas es difícil contar con todos los elementos necesarios para realizar un completo levantamiento de los requerimientos que satisfacen y que deberán ser replicados en el nuevo sistema. También una tarea importante y no abordada por dichas metodologías es la determinación de la calidad de los datos heredados. No olvidemos que los datos deberán ser traspasados al nuevo sistema y si su nivel de calidad es deficiente se deberán tomar medidas correctivas.

En función de todo lo anterior, surge la hipótesis de que la técnica de Minería de Datos, desarrollada para descubrir conocimiento oculto en grandes volúmenes de datos, podría ser de mucha utilidad en un proceso de migración de SIH tanto para el proceso global de migración del sistema así como para la migración de los datos propiamente tal. En particular, en aquellas situaciones en que no se cuenta con mucha documentación sobre el SIH, que no se dispone de programas fuentes y/o que el equipo de migración desconoce totalmente el SIH.

3. Minería de datos

En los últimos años la Minería de Datos (Data Mining) ha atraido la atención de muchos investigadores. Esto principalmente por la diversidad de áreas en las cuales puede ser aplicada, entre otras marketing, medicina, meteorología, y el sector financiero.

La capacidad cada vez mayor de los sistemas computacionales para capturar y almacenar datos (por cierto con gran velocidad) ha generado un explosivo crecimiento de los datos almacenados [AGRAW93]. Este vertiginoso aumento de datos ha provocado que la cantidad de información disponible exceda lo que puede ser procesado y entendido por el ser humano con los medios y técnicas tradicionalmente a su alcance. Por otro lado, existe la percepción de que entre estos datos hay o puede haber, valiosa información que puede ser de gran ayuda para las organizaciones que los han generado.

El descubrimiento de conocimiento en grandes volúmenes de datos se ha denominado **Minería de Datos**, pues lo que se pretende es "excavar" entre los datos para hallar información oculta y que posiblemente sea de gran utilidad en la toma de decisiones [AGRAW93],[ADRIA96][FAYYAD96].

Proceso de descubrimiento de conocimiento en bases de datos (KDD)

En la práctica, la Minería de Datos forma parte de un proceso denominado "proceso KDD" (Knowledge Discovery in Databases). En general, el proceso de KDD se entiende [ADRIA96] como la conjunción de las siguientes etapas:

- **Selección de los datos**
- **Preprocesamiento (limpieza y anriquecimiento de los datos)**
- **Transformación (codificación de los datos)**
- **Minería de Datos (extracción de conocimiento)**
- **Evaluación de los resultados**

Cada una de estas etapas tiene un conjunto de datos de entrada y genera otro conjunto de datos de salida, que son recibidos como entrada por la siguiente etapa. Por otro lado, en cualquier etapa es posible que sea necesario retroceder a una anterior y rehacer algunos procesos. Todo lo anterior, hasta llegar a la obtención de algún conocimiento que el usuario considere útil. La figura 1 nos muestra gráficamente el proceso KDD.

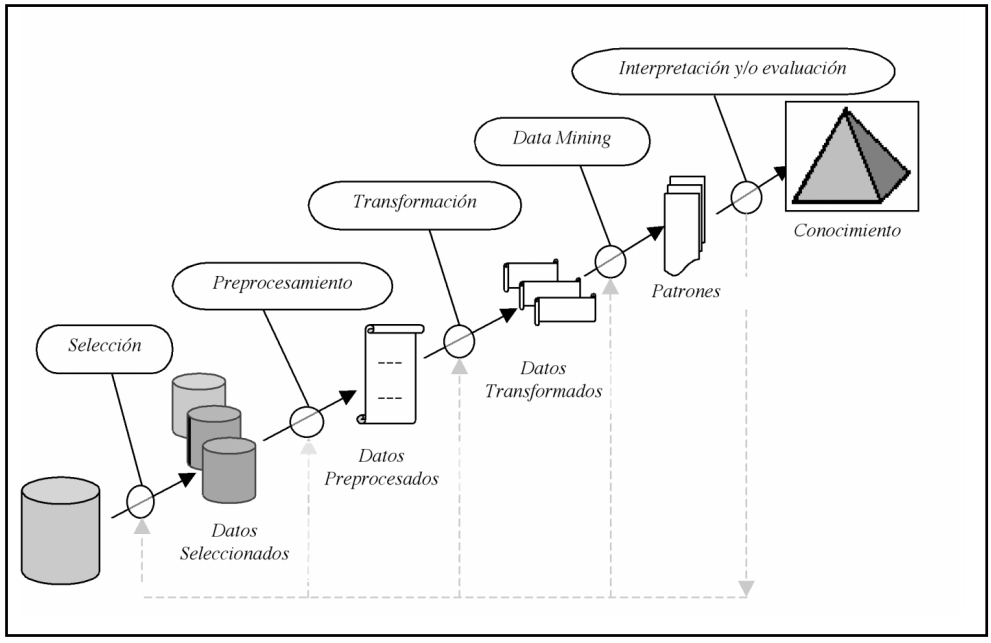

Fig. 1 Proceso de KDD (Knowledge Discovery in Databases)

Técnicas de Minería de Datos

Existen diferentes puntos de vista respecto de qué técnicas pueden ser consideradas parte de la Minería de Datos. Por ejemplo, para algunos, cualquier herramienta

que ayude a extraer más información a partir de un conjunto de datos es útil [ADRIA96]. Bajo este enfoque entonces, también se deberían considerar como herramientas de Minería de datos a las planillas electrónicas, generadores de informes, lenguajes SQL, generadores de gráficos, herramientas de estadística, etc. Otros en cambio, entienden la Minería de Datos como la aplicación de técnicas de inteligencia artificial, incluyendo entre otras modelamiento avanzado e inducción de reglas [AGRAW93] [FAYYAD96].

Considerando ambos enfoques podemos mencionar entre las técnicas más utilizadas para Minería de Datos, las siguientes: Técnicas estadísticas, Visualización, OLAP (Online analytical processing), Aprendizaje basado en ejemplos, Arboles de decisión, Reglas de asociación, Redes neuronales.

Porqué usar una herramienta de KDD en el proceso de entendimiento de un SIH.

Como se ha planteado antes, uno de los problemas centrales en la tarea de lograr el entendimiento de un SIH es que, normalmente, este tipo de sistemas carece de los elementos que se necesitan para cumplir cabalmente dicha tarea, por ejemplo: documentación, personas con conocimiento sobre el sistema, código fuente, modelo de datos.

Sin embargo, a pesar de lo anterior, los datos siempre estarán presente y esa es la característica que ha sido explotada en esta investigación: *si poseemos los datos, entonces que estos sean la fuente que nos provea de conocimiento sobre el SIH*. De acuerdo con esto entonces, el punto es cómo podemos extraer el conocimiento sobre el SIH que suponemos está implícito en los datos. Nuestra respuesta es usando herramientas de KDD cuyo objetivo es, justamente, encontrar conocimiento oculto entre los datos.

Tal como queda de manifiesto en el proceso KDD (ver figura 1), para extraer conocimiento de los datos no es suficiente sólo contar con algoritmos de minería de datos, sino también será necesario preprocesar los datos con el objeto de conocer su estructura, calidad (semántica, sintáctica), seleccionarlos y limpiarlos.

Según [ADRIA96] si nos detenemos a analizar las etapas que componen el proceso de KDD podemos ver que el 80% de éstas, están dedicadas a la preparación de los datos y sólo un 20% a la extracción del conocimiento. Consecuentemente con lo anterior, muchas de las herramientas de minería de datos que existen en el mercado proveen no sólo funciones de minería de datos sino también funciones para la preparación de los datos. Por ejemplo, ver herramientas analizadas en [INT1].

Ambos componentes, son entonces esenciales para la extracción del conocimiento y de acuerdo a nuestra perspectiva en el proceso de entendimiento de un SIH nos permitirían:

- Explorar, conocer y depurar los datos mediante el uso de las funciones de preprocesamiento de datos que posea la herramienta utilizada.

- Determinar la calidad de los datos, esencial tanto para la migración de estos así como para discriminar entre ellos el conjunto a ser analizado con algoritmos de minería de datos.
- Obtener reglas a partir de los datos que nos permitan rescatar el conocimiento necesario (restricciones de integridad, reglas de negocios, relaciones entre datos) para poder reconstruir el modelo de datos y el modelo de negocios que intenta cubrir el SIH.

Entendimiento y recuperación de los requerimientos cubiertos por un SIH.

Una vez que una organización ha decidido migrar un SIH, en particular uno de misión crítica, lo primero que se debe hacer es determinar, en forma exhaustiva, cuales son los actuales requerimientos que este sistema está satisfaciendo, ya que estos junto con los nuevos requerimientos de la organización serán la base del nuevo sistema de información, ver figura 2.

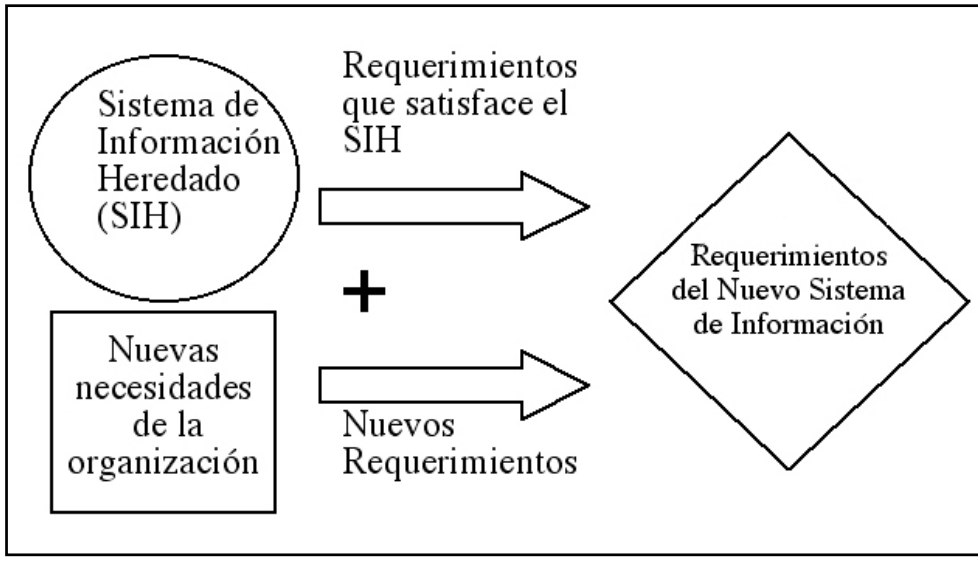

Fig. 2 Determinación de los requerimientos del nuevo sistema de información

Como ya se mencionó antes, esta tarea de recuperación de requerimientos generalmente no es fácil, ya que normalmente los SIH son antiguos y en su vida han ido sufriendo distintas mantenciones (correctivas, adaptativas y/o de perfeccionamiento) [PRESS97], las cuales no siempre son bien documentadas, ni conocidas por todos aquellos que están involucrados en el funcionamiento del sistema, es más, a veces no existe ni la documentación original del sistema, ni códigos fuentes. Adicionalmente, la migración del SIH puede ser encargada a un equipo nuevo, o bien externo a la organización, lo cual implica involucrar a personas que no poseen ningún grado de familiaridad con el sistema, dificultando aún más la reconstrucción de requerimientos.

Lo anterior, nos motiva a pensar que una buena fuente información acerca del SIH son sus datos, ya que, aunque se carezca de otros antecedentes como: interfaces, código fuente, entradas/salidas, etc, que también pueden servir como fuente de información (ver [O'SUL97][WONG95][CHERI94]), los datos siempre estarán presentes.

Si bien es cierto, los datos por sí solos no nos pueden entregar toda la información relativa a los requisitos cubiertos por el SIH, sí nos pueden proporcionar valiosos antecedentes como por ejemplo: reglas del negocio, caracterización de entidades, relaciones entre entidades y antecedentes respecto de la calidad de los datos, elementos importantes para reconstruir el modelo de datos y el modelo de negocios del SIH, ver más adelante caso FONASA.

Junto con los modelos antes mencionados, también será necesario reconocer la sintaxis y semántica asociada a los datos. Para comenzar a operar con la migración de los datos será importante también, definir su nivel de calidad en base a la detección de errores en los datos recibidos. En muchos casos, los datos deberán ser migrados con errores, sin embargo es importante tener presente tal situación, de modo que se pueda establecer, más adelante, mecanismos de mejoramiento de los datos.

Finalmente, todas las tareas antes especificadas deben ser realizadas en conjunto con los usuarios del sistema, quienes deberán ir reconociendo y validando el conocimiento descubierto y por otro lado, tomando conciencia del nivel de calidad de los datos.

A continuación se presenta la metodología propuesta en esta investigación cuyo objetivo es servir de apoyo al proceso de migración de datos de un SIH, y en la cual se pretende abarcar los aspectos, que a nuestro juicio, son los más relevantes de considerar en la migración de datos de un SIH y que también son relevantes para la migración global de un SIH.

4. Metodología de apoyo a la migragración de datos de un SIH

Tal como se expuso hasta ahora, esta metodología pretende cubrir los aspectos esenciales para realizar la migración de un SIH. La estrategia utilizada es el uso de una herramienta de KDD que provea al menos funciones de preprocesamiento de datos y una función de minería de datos de extracción de reglas. Estas condiciones apuntan básicamente a dos consideraciones:

a) Las funciones de preprocesamiento son de gran utilidad para manipular los datos, de modo que se puede facilitar la tarea de determinar cuales son los atributos asociados a una tabla, el conjunto de posibles valores asociados a cada atributo, reconocer la sintaxis y semántica de los datos y determinar el nivel de calidad de los datos analizados.

b) El uso de funciones de extracción de reglas tiene su justificación en que las reglas nos proveen conocimiento acerca de relaciones existentes entre los datos, que a simple vista no es posible identificar y en que la forma de representar este conocimiento es muy fácil de interpretar por el usuario.

La metodología propuesta consta de 7 etapas, ver figura 3, y en la mayor parte de ellas se requiere la intervención del equipo a cargo de la migración y la del usuario del sistema. Se considera como usuario del sistema a un equipo de personas formado por quienes estén a cargo del SIH desde el punto de vista de su mantención, por las personas que realizan el papel de usuario del sistema, y también por quienes serán los usuarios del nuevo sistema.

El papel del usuario es fundamental, ya que será este quien valide en todo momento los resultados obtenidos por el equipo de migración. Quizás algunos de los resultados puedan ser desconocidos para el usuario, sin embargo, es este quien deberá dar la última palabra respecto de su validez para el nuevo sistema.

Etapa 1: Obtención de los datos del SIH y la documentación asociada a ellos.

Etapa 2: Selección de una herramienta de KDD.

Etapa 3: Compatibilización del formato de los archivos o tablas con las entradas para la herramienta KDD y selección de atributos.

Etapa 4: Análisis de los datos utilizando funciones de preprocesamiento.

Etapa 5: Minado de los datos.

Etapa 6: Chequeo de los resultados con el usuario.

Etapa 7: Recomposición del modelo de datos del SIH

Fig. 3 Etapas de la metodología de apoyo a la migración de datos de un SIH

A continuación se entrega un detalle de cada una de las etapas de la metodología:

Etapa 1: Obtención de los datos del SIH y la documentación asociada a ellos.

El objetivo de esta etapa es la obtención de la mayor cantidad de información acerca del SIH. La documentación, aunque no esté completamente actualizada puede darnos alguna idea del contenido de las tablas, los valores que pueden adoptar los atributos, información respecto de reglas de calidad de datos y/o reglas del negocio.

A continuación se deberán seleccionar las tablas a analizar dentro del conjunto de tablas pertenecientes al sistema, es posible encontrarse con algunas que no sea necesario analizar con mayor profundidad, por ejemplo tablas de parámetros o tablas correspondientes a listas invertidas, que pueden ser relativamente sencillas de identificar dentro del grupo de tablas.

Otra importante consideración tiene relación con el tamaño de las muestras de datos con que se va a trabajar. En general, se debe procurar que sean de un tamaño que permita respaldar los resultados obtenidos ya que, cuando un resultado es respaldado por un alto número de ejemplos, nos puede dar mayor seguridad de la validez del conocimiento obtenido, este tamaño también va a depender de la capacidad de procesamiento con que contemos.

Complementariamente, otra estrategia recomendable es dividir las muestras de datos en conjuntos de entrenamiento, los cuales se usarán para la extracción del conocimiento y conjuntos de prueba, con los cuales podamos comprobar los resultados obtenidos a partir de los datos de entrenamiento.

Etapa 2: Selección de una herramienta de KDD.

Esta etapa es muy importante, ya que debemos preocuparnos de seleccionar una herramienta que al menos posea las siguientes características:

- *Distintas alternativas respecto del formato de ingreso de los datos*. Esta característica no es difícil de encontrar ya que la gran mayoría de los productos para KDD proporcionan distintas formas para capturar los datos, entre otras, podemos encontrar: archivos ascii (con formato fijo y variable), planillas electrónicas y conexión a bases de datos. [INT1]
- *Funciones de preprocesamiento y análisis estadístico de los datos*. Estas funciones son fundamentales para la inspección de los datos con el fin de ampliar nuestro conocimiento acerca de ellos y determinar su calidad. No debemos olvidar que para que la etapa de Minería de Datos entregue resultados confiables es imprescindible contar con datos "limpios".
- *Generación de Reglas sobre los datos*, ya sea mediante árboles de decisión o reglas de asociación. La existencia de esta técnica de minería de datos en la herramienta es muy deseable, ya que en general los resultados que se obtienen son muy fáciles de entender y también de explicar.

Es importante señalar en este punto que la adquisición de una herramienta de minería de datos para un proyecto de migración, tiene un doble beneficio, puesto que después de la migración la organización contará con una herramienta de análisis que, cumpliendo su misión original, será de gran utilidad para la gestión global de la organización, no así una herramienta más específica para la migración.

Etapa 3: Compatibilización del formato de los archivos o tablas con las entradas para la herramienta de KDD y selección de atributos.

Una vez que hemos seleccionado la herramienta de KDD y poseemos los datos del sistema debemos compatibilizar su formato con alguno de los formatos para el ingreso de datos que posea la herramienta.

En general, una buena cantidad de los sistemas heredados que se están migrando son aquellos desarrollados en lenguajes de tercera generación como por ejemplo: Cobol, Pascal, o Basic [BRODI95], y donde los archivos, preferentemente, son de tipo ascii y se requiere una buena documentación para conocerlos. En otros casos, donde el sistema a migrar está implementado en algún motor de base de datos más moderno, lo más probable es que se cuente al menos con algún diccionario de datos y otros elementos de autodocumentación que faciliten más el conocimiento de los datos.

Otra consideración a realizar en la compatibilización de los datos tiene relación con la capacidad de la herramienta de KDD para manejar distintos tipos de datos. En general, es deseable usar una que nos permita trabajar con distintos tipos de datos, esto es números y cadenas de caracteres, ya que no deberemos desperdiciar tiempo haciendo conversiones. Existen herramientas que, por ejemplo, sólo permiten el manejo de datos numéricos, por lo cual en este caso será necesario realizar una conversión de los datos antes de incorporarlos al análisis con la herramienta. En aquellas herramientas que permiten la existencia de distintos tipos de datos, se puede dar la situación que nos encontremos con atributos de tipo numérico que tienen un significado simbólico, por ejemplo, hay sistemas antiguos en que el sexo era codificado mediante números, 0 masculino y 1 femenino, en este caso debemos especificarle a la herramienta que trate este atributo en forma simbólica y no de acuerdo a su valor numérico.

Junto con lo anterior, en esta etapa se debe realizar el descarte de algunos atributos que sean considerados irrelevantes para el proceso de inspección y análisis. Por ejemplo:

- Típicamente el atributo que representa la identificación de los registros (como el RUT) puede ser descartado por ser valores únicos. Una excepción a esto se da cuando el atributo siendo único en su valor total, contiene información que puede ser derivada a partir de su descomposición.
- Pueden existir algunos atributos que poseen un único valor para todos los registros, en estos casos no aportan ninguna información.
- También el conocimiento del usuario sobre los datos, puede ser de gran ayuda para descartar atributos de menor importancia.

Una vez hechas estas consideraciones, podemos proceder a realizar los primeros análisis, sin embargo, es posible que en algún momento debamos retroceder a esta etapa porque nos encontremos con algún problema en el formato o especificación de los datos.

Etapa 4: Análisis de los datos utilizando funciones de preprocesamiento.

Los objetivos de esta etapa son: conocer los datos (o bien ampliar nuestro conocimiento inicial acerca de ellos) y chequear su calidad. Ambos podemos lograrlos mediante las funciones de preprocesamiento, ya que éstas nos dan la posibilidad de manipular de distintas formas los datos. Por ejemplo, en el caso Fonasa donde se aplicó esta metodología (ver sección 5) al usar una función de distribución sobre el atributo *sexo de beneficiario* de la muestra analizada nos encontramos con el siguiente resultado:

Valor	%	Ocurrencias
	3.32047	241
Ø	0.041334	3
F	53.3343	3871
M	43.3039	3143

Tabla 1, ejemplo del uso de una función de preprocesamiento

La tabla anterior nos aporta el siguiente conocimiento:

- El sexo en el SIH está codificado alfabéticamente de la siguiente forma: F para el sexo femenino y M para el sexo masculino.
- La muestra posee un poco más del 3% de registros en los cuales el código de sexo es inválido. Si extrapolamos esta situación al archivo original, podría ser un problema bastante considerable respecto de la calidad de los datos.

Como podemos ver en el ejemplo anterior, una simple función de distribución puede ser muy reveladora, ya que nos permite conocer el contenido de un atributo y a partir de éste derivar las reglas sintácticas asociadas a él; y a partir de lo mismo determinar su calidad. Como reacción ante un resultado como este podemos seguir aplicando otras funciones, como por ejemplo una de selección que nos permita aislar los casos (beneficiarios) que están con problemas y analizar por separado el resto de los atributos, de modo que podamos descubrir por ejemplo, una caracterización de los casos con problemas de calidad.

Si se extiende este análisis al resto de los atributos podemos acumular bastante más conocimiento acerca de ellos. En principio este conocimiento es básicamente sintáctico, pero poco a poco y en conjunto con las observaciones del usuario podemos ir agregando un conocimiento semántico sobre ellos.

Otro tipo de funciones que proponemos aplicar, son aquellas que nos permitan relacionar dos atributos de acuerdo a sus valores. Esto puede realizarse a continuación de analizar previamente por separado cada atributo, seleccionar los registros que estén sintácticamente correctos y luego realizar análisis cruzados

de sus valores. Otra posibilidad es cruzarlos sin ningún tipo de discriminación y tratar de descubrir si existe alguna relación entre los registros cuyos valores son sintácticamente incorrectos, por ejemplo: cuando el *atributo A* posee un valor sintácticamente incorrecto el *atributo B* también posee un valor sintácticamente incorrecto.

Al término de esta etapa podemos contar con que ha aumentado nuestro conocimiento acerca de los datos y algunas relaciones sencillas de ellos. Ya no sólo podremos contar con conocimiento de carácter sintáctico, sino que también semántico. Adicionalmente, en esta etapa ya empezamos a tener una clara noción de la calidad sintáctica de los datos, e inclusive podría ser factible tener alguna teoría del origen de algunos de los errores en los datos, por ejemplo: la mayor cantidad de errores se han detectado en los clientes de la sucursal A o bien en los clientes del tipo Y; y por otro lado, esto podría ser un buen punto de partida para desarrollar algún mecanismo para mejorar la calidad de los datos.

En esta etapa al igual que en la siguiente es muy importante contar con la colaboración del usuario, ya que él deberá ir validando cada uno de los resultados y sólo de este modo podremos ir avanzando en forma segura.

Etapa 5: Minado de los datos.

Una vez que se ha realizado el preprocesamiento de los datos, es posible pensar en aplicar alguna técnica de minería de datos sobre ellos. En esta etapa analizaremos una selección de los datos originales que, previamente procesados, se consideran sintácticamente válidos y libres de atributos irrelevantes.

Los datos que no se consideran válidos también pueden ser minados, pero en forma independiente, pudiendo rescatar de ellos, entre otras cosas, conocimiento acerca del origen del problema de calidad en los datos o bien, una nueva regla del negocio que fue pasada por alto al considerar que los datos eran erróneos.

La extracción de reglas básicamente consiste en determinar un atributo, cuyos valores no sean continuos, sobre el cual nos interese conocer reglas y seleccionar todos los atributos que consideramos relacionados con él y que, por lo tanto, pudieran ser parte de la o las reglas a obtener.

En general, una regla puede ser representada mediante una estructura condicional del tipo:

Si <condición> entonces <conclusión>

En particular, en la metodología se sugiere como base una técnica de extracción de reglas, esto principalmente por:

- Los resultados que entrega son fáciles de entender e interpretar.
- Entre los datos de un sistema es inevitable encontrarse con dos tipos de reglas: las ***reglas de negocios*** y por otro lado, ***reglas de integridad***.

Una ***regla de negocios*** es un conjunto de instrucciones que aportan conocimiento al sistema y permiten definir eventos y procedimientos dentro de él configurando cuando se deben hacer las cosas [INT2]. Entre los tipos de reglas de negocios que podemos encontrar tenemos:

- *Regla de Definición*, que corresponde a una característica o propiedad de un objeto dentro del sistema.
- *Regla Tácita,* que define la existencia y dimensión dentro del sistema.
- *Regla Fórmula*, que define un determinando cálculo que debe emplear el sistema.
- *Regla de Validación*, que define los posibles valores válidos para un determinado proceso o variable dentro del sistema.

Por otro lado, estas reglas de negocios son la base para la construcción y definición de las restricciones de integridad (**reglas de integridad**) que el modelo de datos de un sistema debe cumplir [PRESS97]. Entre este tipo de restricciones encontramos:

- *Restricciones de dominio*, que corresponden a las restricciones sobre los tipos de valores que puede tener un atributo.
- *Restricciones de clave y de vínculo*, que en el modelo de datos relacional, se refiere a que una clave es un atributo de un tipo de entidades que debe tener un valor único para cada entidad que pertenezca a dicho tipo en cualquier momento específico.
- *Restricciones generales de integridad semántica*, corresponden a restricciones más complejas que las anteriores y que en general se especifican mediante procedimientos o en forma declarativa. En ambos casos los datos del sistema deben enmarcarse dentro de ellas para ser considerado válidos.

Tal como se dijo en un comienzo, nuestra estrategia fue realizar la minería de datos tomando como técnica fundamental la generación de reglas, sin embargo, esto no restringe complementarla con la aplicación de otras técnicas como por ejemplo *clustering*, con la cual podríamos por ejemplo tipificar clientes y determinar características propias de cada grupo[ADRIA96][AGRAW93].

Etapa 6: Chequeo de los resultados con el usuario. Determinación de las reglas de negocios, tanto de aquellas conocidas como las que el usuario no tenía presentes.

El proceso de extracción de reglas arrojará como resultado una gran cantidad de reglas. El equipo que realiza el minado de los datos en esta etapa ya dispone de

mayor conocimiento acerca de ellos, sin embargo, no el suficiente como para determinar la validez de todas las reglas. Debido a esto será necesario presentar al usuario el conjunto de reglas que se han extraído y comenzar con éste un proceso de chequeo de cada una de ellas.

Es posible que el usuario pueda distinguir claramente las reglas válidas dentro del conjunto presentado, sin embargo, en algunas situaciones(reglas nuevas u olvidadas) esto puede no ser tan simple. En estos casos, resulta importante analizar algunos indicadores estadísticos asociados a las reglas, que reflejan el grado de representatividad de éstas entre los datos. El grupo del migración puede establecer en conjunto con el usuario valores mínimos para estos indicadores, de modo que aquellas reglas que no cumplan con estos requisitos no sean consideradas. Antes de fijar estos valores será necesario estudiar en conjunto con el usuario las distintas reglas y estimar cuál es el grado mínimo de representatividad, ya que puede que existan reglas válidas que sólo se presenten en pocos casos.

Otro factor importante de considerar es el tamaño de la muestra de datos. Idealmente mientras más grande sea la muestra más relevante pueden ser los resultados, sin embargo no debemos olvidar que para poder procesar un gran volumen de datos con técnicas de minería de datos se requiere equipamiento con bastante capacidad de memoria y de procesamiento. Con el fin de sobrellevar esta dificultad, se puede generar un número mayor de muestras, someterlas al mismo procesamiento y luego, hacer un seguimiento de las reglas encontradas en todas las muestras.

Por el momento, no se tiene una estimación precisa respecto de qué porcentaje de datos, como mínimo, deberían procesarse para lograr un mejor y buen entendimiento de un SIH, ya que esto requiere de un estudio mayor en el cual se consideren varios casos. Esta es una tarea que puede ser abordada en un siguiente estudio.

Finalmente, es conveniente tener presente que en cualquier momento puede ser necesario regresar a alguna de las etapas anteriores, y no necesariamente vamos a hacer una ejecución lineal.

Etapa 7: Recomposición del modelo de datos del SIH.

Esta etapa consiste básicamente en elaborar un modelo de datos para el SIH. En base al conocimiento adquirido sobre el sistema, probablemente ya podamos distinguir cuales son las entidades más relevantes que participan en el SIH y cuales son sus relaciones principales.

Tomando como base un MER (Modelo Entidad Relación) podemos representar en el esquema las distintas entidades y unirlas en la medida que éstas se relacionan, también probablemente estaremos en condiciones de asociarles su cardinalidad.

Esta etapa es la culminación del estudio, y el MER generado constituirá nuestra base para la construcción del nuevo modelo de datos que tendrá el sistema al ser migrado. También al llegar a esta etapa habremos acumulado bastante conocimiento sobre el SIH lo que nos brindará la confianza de que la migración se hará sobre bases sólidas.

Una vez que los datos heredados se han migrado, es recomendable establecer alguna estrategia respecto del control de calidad de los datos. Nuestra propuesta es que si los datos siguen siendo inspeccionados incrementalmente con herramientas de minería de datos, podremos llegar a tener un buen sistema de detección de errores así como también de detección de nuevas reglas de negocio. Esta propuesta constituye uno de los trabajos futuros derivados de esta investigación.

5. Caso de estudio: FONASA (Fondo Nacional de Salud).

FONASA es una empresa estatal dedicada a administrar los dineros correspondientes a las cotizaciones por salud de los trabajadores y los subsidios otorgados por el estado a aquellas personas que por alguna causa no cotizan. Todo lo anterior con el fin de brindar a sus beneficiarios la posibilidad de acceder a distintos tipos de prestaciones médicas. Esta empresa abarca un universo de aproximadamente un 70% de la población chilena, el resto de la población cotiza en empresas privadas llamadas Isapres.

Migración del SIH de Control de Beneficiarios

El FONASA contaba con un sistema de información para el Control de Beneficiarios, encargado fundamentalmente de determinar quién es y quién no es beneficiario y en qué categoría se encuentra para optar a las distintas prestaciones médicas ofrecidas.

Las características más relevantes de este Sistema que llevaron a definirlo como SIH eran:

- El sistema se encontraba desarrollado en COBOL.
- Los archivos utilizados eran planos.
- El sistema era de misión crítica para la organización.
- El sistema tenía al menos 8 años desde su construcción.
- El manejo de información histórica era insuficiente.
- Se desconocía el estado de la calidad de los datos.
- Se disponía sólo de documentación general.
- Tecnología incompatible con nuevos requerimientos de la organización.

Este sistema manejaba información sobre aproximadamente un 95% de la población chilena, fueran o no beneficiarios de FONASA. Respecto de los beneficiarios se mantenía la información más relevante que permitiera manejar y controlar en forma adecuada las prestaciones y beneficios que correspondieran a cada uno. El sistema fue creado y administrado varios años por una empresa externa a FONASA.

Luego de un nuevo proceso de licitación, su administración, mantención y actualización, fueron traspasadas a una empresa externa diferente, la cual una vez analizados los nuevos requerimientos de la organización y las características del SIH, determinó que la solución más adecuada era la migración del sistema.

Dado que no se contaba con suficiente información sobre el SIH, se procedió a realizar un proceso de recuperación de los requerimientos del SIH con el fin de obtener un mayor conocimiento acerca de él y traspasar el máximo de su funcionalidad al nuevo sistema a desarrollar. Para esta labor se crearon dos grupos de trabajo independientes: uno que abordó el problema en forma tradicional, esto es, a través de entrevistas con el usuario y estudio de la documentación existente; y otro, con un fin más académico que se dedicó a probar la hipótesis de que inspeccionando los datos con herramientas de KDD, es posible recuperar requerimientos y reglas de negocio, fundamentales para ayudar a reconstruir el SIH y su modelo de datos. A continuación se presentan algunos de los resultados por cada una de las etapas desarrolladas, y que pueden ilustrar el tipo de conocimiento que es posible obtener.

Etapa 1: Obtención de los datos del SIH y la documentación asociada a ellos.

En este proyecto de migración se contaba con los siguientes recursos para iniciar el análisis de los datos del SIH de Control de Beneficiarios:

- **Descripción de tablas**: Esta consistía en una descripción de los nombres de cada campo, tamaño y una breve descripción de su contenido y/o significado.
- **Usuarios con conocimiento del sistema**: A medida que se fueron obteniendo resultados, estos fueron siendo contrastados con el usuario, de modo que éste confirmara o negara las primeras impresiones generadas a partir de los datos.
- **Datos**: Se trabajó principalmente con los datos de la tabla Natural, en ella estaba contenida toda la información de los beneficiarios y no beneficiarios, con un tamaño de registro de 653 caracteres y originalmente de un tamaño de 2 Gbytes, a partir de los cuales se fueron tomando distintas muestras.

Etapa 2: Selección de una herramienta de KDD

La herramienta de KDD utilizada fue Clementine versión 2.0, la cual posee una interfaz gráfica que facilita tanto la operación e interpretación de los resultados. Entre otros formatos de entrada de datos ofrece la posibilidad de trabajar con código ascii, por lo que se pudo trabajar directamente con las muestras de datos del SIH. Provee un conjunto de funciones de preprocesamiento de datos que resultaron muy útiles en el proceso de inspección y chequeo de calidad de los datos. Las técnicas de minería de datos que ofrece son: reglas basadas en árboles de decisión, redes neuronales, y regresión, de éstas se utilizaron sólo reglas y redes neuronales.

Etapa 3: Compatibilización del formato de los archivos o tablas con las entradas para la herramienta KDD y selección de atributos.

En particular, dado que los datos estaban en un archivo plano se utilizó el formato de entrada que hace un tratamiento sobre el archivo como un ascii de formato fijo. Los campos fueron definidos de acuerdo a la especificación de registro entregada por el usuario. Durante la selección de los atributos, aquellos campos que en la documentación aparecían como fuera de uso fueron descartados del análisis.

Entre las tablas del SIH existía una que contenía toda la información de los beneficiarios, su nombre era Natural. A partir de esta tabla se generaron varios conjuntos con muestras de menor tamaño para su análisis con la herramienta KDD. Las primeras muestras con que se trabajó contenían todo tipo de registros (de beneficiarios y no beneficiarios), una vez que se realizaron los primeros análisis, en las muestras sólo se incluyeron registros pertenecientes a beneficiarios.

Etapa 4: Análisis de los datos utilizando funciones de preprocesamiento.

La herramienta de KDD utilizada permitió en un principio explorar los datos, con el objetivo de conocerlos, determinar cuales eran los posibles rangos de valores válidos para cada campo y chequear la calidad de los datos. La información obtenida fue chequeada con el usuario. Los resultados obtenidos en esta primera etapa resultaron fundamentales, ya que efectivamente se logró conocer los datos y determinar la calidad de estos. Algunos de los resultados obtenidos y que pueden ilustrar la utilidad de esta etapa son los siguientes:

Inspección de los datos y chequeo de calidad.

Mediante el uso de una función de distribución se analizó el campo *Nivel del beneficiario*. Este campo corresponde a la ubicación del beneficiario dentro de ciertos rangos, que tienen relación con su nivel de ingresos y que determina el costo que deben cubrir los beneficiarios por las diferentes prestaciones médicas. Los valores válidos para este campo de acuerdo a la documentación son: A, B, C y D. Al aplicar la función de distribución de él se obtuvieron los siguientes resultados:

Valor	%	Ocurrencias
	21.3833	1552
A	28.9474	2101
B	21.6726	1573
C	8.28052	601
D	19.7162	1431

Tabla 2, distribución del atributo *Nivel del beneficiario*

Como podemos observar en la tabla anterior, existe un 21,38% de la muestra que aparece sin valor. Nos preguntamos si tal situación era válida y en qué casos. Más adelante al generar reglas sobre los datos comprobamos que aquellos registros con este campo en blanco pertenecían a personas que no eran beneficiarios, por tanto este valor era correcto. Es importante destacar aquí la diferencia que existe en un análisis estadístico de los datos como una distribución, que sólo nos muestra el estado de los datos, y el análisis mediante reglas que nos permite obtener una explicación de dicho estado.

Chequeo Semántico de los datos

El análisis semántico tiene directa relación con el significado de los datos y de las relaciones entre estos. Dado el desconocimiento inicial sobre los datos y el sistema, las dependencias y relaciones entre los datos se buscaron intuitivamente y no siguiendo algún patrón determinado.

Mediante una función de selección se analizaron los atributos *Rut del beneficiario y Rut del cotizante,* el objetivo de aplicar esta función fue detectar anomalías en la relación existente entre estos campos. Se supone que si un beneficiario es carga, entonces debe tener asociado un rut de cotizante. Como resultado se obtuvo que todos los registros correspondientes a una carga tenían asociado un rut de cotizante.

Otro atributo analizado fue *Código de actividad del empleador*, si un beneficiario tiene un empleador asociado, lo correcto sería que también tuviera el código de la actividad de su empleador. Al aplicar una función de selección para el chequeo de esta condición se detectó un porcentaje no alto de casos que no la cumplen, esto fue justificado por el usuario como problemas con la calidad de los datos respecto del manejo de información histórica.

Lo anterior, constituye sólo una muestra de los análisis realizados y los resultados obtenidos. Esta etapa resultó de gran importancia ya que los antecedentes obtenidos facilitaron la comprensión de los resultados de la siguiente etapa.

Etapa 5: Minado de los datos

Una vez explorados los datos con el fin de conocerlos y chequear su calidad, comenzó el proceso de extracción de conocimiento a partir de ellos. El objetivo era poder obtener reglas del negocio contenidas en ellos que no eran evidentes, conocidas y/o recordadas por el usuario. Para esto se aplicaron las técnicas de inducción de reglas y como apoyo a la anterior, redes neuronales, ambas técnicas provistas por la herramienta utilizada.

Un ejemplo del tipo de regla generada es el siguiente, asociada con el atributo *Sexo*:

if *Tipo_carga* = *"C"* **then** *Sexo* = *"F"*

Producto del análisis sintáctico de los datos, se sabía que cuando el Atributo *Tipo_carga* posee el valor "C", esto es equivalente a decir que el tipo de carga es "cónyuge" y que el valor "F" para el atributo *Sexo* corresponde a sexo "femenino". Entonces, la interpretación lógica de la regla es: *si el tipo de carga es cónyuge entonces el sexo del beneficiario es femenino*. Luego, analizando y chequeando esta regla con el usuario, podemos concluir que es muy lógica y válida, ya que cuando se tiene una carga del tipo cónyuge, nos estamos refiriendo casi en un 100% de los casos a mujeres que son las esposas de los cotizantes.

Etapa 6: Chequeo de los resultados con el usuario.

Una vez que se tuvo el conjunto de reglas asociadas a cada campo analizado, se procedió a la reconstrucción de las reglas más importantes o fuertes. Para esto fue necesario confrontar los resultados anteriores con el usuario. A continuación, se muestran algunas de las reglas más importantes que fueron derivadas del análisis de los datos con una herramienta de KDD.

a) Reglas del atributo Bloqueo:

Este atributo indica si una persona es o no beneficiaria de FONASA.

a.1) Si ***Bloqueo*** = "**N**" entonces "**la persona es beneficiaria**".

a.2) Si ***Bloqueo*** = "**S**" entonces "**la persona es no beneficiaria**".

Al buscar una regla que asocie esta condición con otros campos del registro encontramos, entre otras, las siguientes:

a.3) Si ***Bloqueo*** = "**S**" entonces ***Código*** > **901**

Donde los códigos mayores a 901 indican distintas situaciones por las cuales la persona no es beneficiario.

a.4) Si ***Bloqueo*** = "**N**" entonces ***Código*** <= **901**

Donde los códigos menores e iguales a 901 indican las condiciones en las que una persona es beneficiaria.

El hallazgo de estas reglas fue muy importante en esta experiencia, ya que los juicios iniciales respecto de la calidad de los datos se vieron afectados. Los datos fueron analizados nuevamente y se constató que se cumplían otras reglas que también son válidas e importantes, por ejemplo:

a.5) Si ***Bloqueo*** = "**N**" entonces ***Nivel*** $\in$ {"**A**","**B**","**C**","**D**"}

Esto significa que todo beneficiario debe tener asociado un nivel, esto resulta lógico ya que de este nivel dependerán las condiciones en que cada beneficiario puede acceder a las distintas prestaciones médicas.

b) Reglas del atributo Nivel

Este campo como ya se dijo tiene relación con el nivel de ingresos de los beneficiarios. Por lo tanto, la regla básica para definir el nivel está determinada por los ingresos. Sin embargo al explorar los datos en busca de reglas nos encontramos que en todas las muestras se repetía la siguiente regla para el nivel A:

b.1) Si ***Código*** > **615** entonces ***Nivel*** = "A"

Entonces, dado este resultado y relacionándolo con las reglas del campo *Bloqueo*, podemos determinar que todos los beneficiarios que tienen asociado un código cuyo valor está en el rango de 616 y 901, pertenecen al nivel "A" (donde no se cotiza), situación que resulta ser válida, ya que estos códigos representan a personas indigentes o con alguna situación especial y a las cargas de estos.

c) Reglas del campo Carga

Este campo es un flag que indica si el beneficiario es o no una carga. Al explorar los datos en busca de reglas nos encontramos con los siguientes resultados:

c.1) Si ***Tipo de carga*** = {"H","C","P","N","V"} entonces ***Carga*** = "S"

Esta regla representa la relación lógica entre estos campos, si el beneficiario es una carga debe especificarse también el tipo de carga.

c.2) Si ***Tipo de carga*** = " " entonces ***Carga*** = "N" o ***Carga*** = " "

Más allá de lo que representa en relación al campo *Tipo de carga*, la regla anterior muestra una regla implícita en los datos y que aparentemente fue estableciéndose en forma posterior al diseño del SIH, esto es, el valor de negación en un campo flag puede ser representado mediante blanco o N. Esta misma situación se pudo observar en otros campos tipo flag.

Etapa 7: Recomposición del modelo de datos del SIH.

Con todo el conocimiento recopilado en las etapas anteriores fue factible recuperar el modelo de datos del SIH, a partir del cual, pudo construirse el modelo de datos del nuevo sistema. Así también como todo el modelo de reglas de negocio del sistema, fundamentales ambos tanto para la migración de los datos, así como la de todo el sistema.

Es importante destacar que ninguno de los grupos de investigación que trabajó en este caso, tenía conocimiento previo acerca del SIH o los datos, más bien este fue adquiriéndose en la medida que se obtenían resultados. Este hecho es de gran importancia ya que, tradicionalmente uno de los grandes escollos que tienen los proyectos de migración de SIH es la falta de conocimiento del sistema a migrar por parte de quienes tienen que desarrollarlo.

Como conclusiones de esta experiencia podemos señalar las siguientes:

- Dado que al comienzo del proyecto se tenía muy poca información sobre el SIH, el uso de una herramienta de KDD fue muy útil en la exploración y conocimiento de los datos del sistema Fonasa. Gracias a las distintas funciones que provee la herramienta fue posible conocer antecedentes sobre la sintaxis de los atributos, la calidad de los datos y también conocer algunos aspectos semánticos sobre los datos, los cuales fueron reforzados y complementados con la intervención del usuario.
- Desde el punto de vista del proyecto el establecer el nivel de calidad de los datos en un comienzo fue muy importante, ya que se previno futuros problemas con el usuario respecto de este punto.
- El disponer tempranamente de algunas de las reglas de negocios del SIH, hizo que las reuniones de trabajo con el usuario fueran muy ágiles y participativas.
- En la búsqueda de reglas se encontraron algunas que eran conocidas por el usuario y otras que no, esto permitió capturar la evolución del SIH respecto de las reglas iniciales con que fue creado el sistema.
- En opinión del grupo que tenía a cargo el proyecto de migración de Fonasa (distinto del que trabajó con minería de datos), esta estrategia fue beneficiosa y contribuyó al éxito del proyecto tanto en su desarrollo como en los tiempos empleados, aproximadamente 4 meses.

6. Conclusiones

El presente trabajo aporta un alternativa real y factible para apoyar proyectos de migración de SIH, en particular, en situaciones extremas en las cuales la información acerca del SIH es mínima. Por otro lado, el empleo de una herramienta de KDD en un proceso de migración, le permite a una organización el adquirir una herramienta que puede ser utilizada, también con otros fines de carácter más permanente, como lo es el análisis de los datos en apoyo a la gestión.

Para el equipo de trabajo que lleva a cabo la migración, esta alternativa es bastante favorable, ya que le permitirá contar en forma anticipada con elementos que ayudarán al diálogo con los usuarios. Además, eventualmente, se podrán encontrar con información desconocida respecto de la evolución que ha tenido el sistema heredado en el tiempo, que probablemente no sea fácil de detectar de otro modo. Con estos elementos a su favor, la reconstrucción del modelo de datos y el modelo de negocios del SIH puede ser bastante más sencilla. Por otro lado, la posibilidad de determinar el nivel de calidad de los datos también es muy beneficioso, ya que de algún modo permitirá controlar el nivel de errores con que los datos serán migrados y llegar a establecer mecanismos para mantener un control permanente de su calidad y mejora.

7. Trabajos futuros

Una siguiente etapa de esta investigación, es aplicar la metodología propuesta a otros casos de migración, de manera que esta pueda afinarse y contemplar otras situaciones que no hayan sido consideradas en esta oportunidad. Con un número mayor de casos será posible hacer estimaciones más acertadas respecto de tiempos promedios involucrados en la ejecución de cada etapa de la metodología y el grado de éxito de ésta en distintos tipos de sistemas heredados.

Otra línea de investigación que se puede derivar de la experiencia de este trabajo es la del análisis de la calidad de los datos de los sistemas, que no necesariamente tienen que ser sistemas heredados. En este trabajo se ha comprobado que el uso de herramientas de KDD es muy útil en esta labor, por lo cual podrían perfectamente ser empleadas con este fin. La idea central sería establecer, con mayor formalidad, los tipos de problemas de calidad que pueden ser detectados y cuales no, y qué mecanismos pueden establecerse en un sistema para contar con un control permanente de calidad basado en el uso de este tipo de herramientas.

Finalmente, un tercer aspecto interesante de explorar es el seguimiento de la evolución de las reglas de negocio en un sistema. Como se ha tratado en esta investigación, en la medida que una organización evoluciona las reglas de negocio también lo hacen, por lo tanto los sistemas deben cambiar. Por otro lado, también ocurre que el entorno de una organización evoluciona y esta evolución puede afectar las reglas de negocios del sistema. Ante esta situación, es importante para una organización detectar dichos cambios y reaccionar ante ellos. Entonces la propuesta es establecer mecanismos de detección de estos cambios, de modo que los sistemas puedan reaccionar ante ellos. Todo lo anterior en base a la aplicación de herramientas de KDD y análisis incremental de los datos.

Referencias

1 [ADRIA96] Adriaans P., Zantinge D., **Data Mining**, Addison-Wesley. 1996.

2 [AGRAW93] Agrawal R., Imielinski T, Swami A., **Data Mining: A Performance Perspective**, IEEE Transactions on Knowledge and Data Engineering. December 1993. pp : 914 - 925.

3 [BISBA97] Bisbal J., Lawless D., Richardson R., Wu B., Grimson J., Wade V., and O'Sullivan D. **"An Overview of Legacy Information Systems Migration".** Proceedings of the APSEC'97/ICSC'97: Joint 1997 Asia Pacific Software Engineering Conference and International Computer Science Conference. Hong Kong, pp 529-530, IEEE Computer Society. China. 2-5 December 1997.

4 [BISBA99] Bisbal J., Lawless D., Wu B., Grimson J.. **"Legacy Information Systems: Issues and Directions".** IEEE Software, pp: 103-111. Septiembre/Octubre 1999.

5 [BRODI95] Brodie M., Stonebraker M. **Migrating Legacy Systems: Gateways, Interfaces and the Incremental Approach**. Morgan Kaufman Publishers. 1995.

6 [CHERI94] Cherinka R., Overstreet C. and Sparks R. "**TSME: Maintaining Legacy System with Process Driven Software Engineering Technology** ". The 1912th Computer Systems Group (USAF, langley AFB VA).Enero 31,1994.

7 [GOLD98] Gold N. **The Meaning of "Legacy Systems"..** SABA Project Report: PR-SABA-01 Version 1.1. http://www.duc.ac.uk/CSM/SABA. Enero 27, 1998.

8 [O'SUL97] O'Sullivan D., Richardson R., Grimson J., Wu B., Bisbal J., and Lawless D. "**Application of Case Based Reasoning to Legacy System Migration**". Proceedings of the Fifth German Workshop on Case-Based Reasoning-Foundations, Systems, and Applications, pp 225-234, March , 1997.

9 [PRESS97] Pressman R., **Ingeniería del software, un enfoque práctico**. Mc Graw-Hill. 1997.

10 [FAYYAD96] Fayyad U., Piatetsky-Shapiro G., Smyth P., **The KDD Process for Extracting Useful Knowledge from Volumes of Data,** Comunication of the ACM. Vol. 39, Nº 11, November 1996.

11 [WONG95] Wong K., Tilley S., Muller H. and Storey M. "**Structural Redocumentation: A Case Study** ". IEEE Software, volumen 12, nro. 1. March, 1995.

12 [WU97] Wu B., Lawless D., Bisbal D., Richardson R., Grimson J., Wade V., and O'Sullivan D. "**Legacy System Migration: A Legacy Data Migrating Engine**". Proceedings of the 17th DATASEM '97, Brno, Czech Republic, pp. 129-138, Ed. Czechoslovak Computer Experts. October 12-14, 1997.

Referencias en Internet

[INT1] Productos de minería de datos.

http://www.kdnuggets.com/

http://www.kbs.twi.tudelft.nl/~norbert.

[INT2] Dominios y reglas de negocios.

http://www.geocities.com/SiliconValley/Hills/3243.