UN MÉTODO DE TRANSFORMACION DE DATOS ORIENTADO AL USO DE EXPLOTACIÓN DE INFORMACIÓN

Hernán Merlino, Paola Britos, Jorge Ierache, Eduardo Diez y Ramón García-Martínez

Centro de Ingeniería de Software e Ingeniería del Conocimiento. Escuela de Postgrado. Instituto Tecnológico de Buenos Aires
Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires.

rgm@itba.edu.ar

Abstract: En este trabajo se propone un método de transformación de datos orientado a la explotación de información y se detallan las características necesarias que debe poseer el entorno de trabajo para la automatización del mismo.

Palabras Clave: Transformación de datos. Minería de datos.

Workshop: Ingeniería de Software y Bases de Datos (WISBD).

1. Introducción

La exploración y análisis, en forma automática o semi-automática, de grandes volúmenes de información para la detección de patrones de comportamiento es lo que se denomina minería o explotación de datos, también conocido por su vocablo en inglés data mining. En [Witten *et al*., 2000] se define minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.
Se puede definir el ciclo de vida de la explotación de datos a partir de las siguientes fases: [a] obtención de datos a procesar, [b] transformación de los datos para que pueda ser utilizado, [c] aplicación de la técnica de explotación de datos y [d] evaluación de los resultados obtenidos. La fase transformación de los datos, es la que insume mayor tiempo, llevando aproximadamente el 60% del esfuerzo de desarrollo. En este trabajo se propone un método de transformación de datos y se detallan las características necesarias que debe poseer el entorno de trabajo para la automatización del mismo. En la sección 2 se describe brevemente el ciclo de vida de la explotación de datos y se menciona el principal problema de la preparación de datos. En la sección 3 se detalla el ciclo de vida de la explotación de datos, y los pasos que se encuentran relacionados con la transformación de datos. En la sección 4 se plantean problema abiertos y futuras líneas de investigación.

2. Ciclo de vida de la explotación de datos

En ciertos casos el disparador de un proceso de explotación de datos es la detección de un problema y la necesidad de corregir ese comportamiento anómalo; en otros no es necesario observar nada anormal solo se aplica el proceso de minería para detectar patrones desconocidos. De ser este último el caso aplicado, los resultados obtenidos en la explotación de datos deben ser sometidos a un proceso de validación conocido como minería de reglas de negocio, o en su forma inglesa *business rule mining*, el cual nos permitirá validar o crear una nueva regla de negocio. Las fases del ciclo de vida a seguir se describen en las siguientes subsecciones.

2.1. Obtención de datos a procesar

Suele suceder que este punto siempre parece mucho más sencillo de lo que realmente es, algunos de los problemas que se suelen encontrar es la falta de acceso a los datos, ya sea por razones de seguridad o por no encontrarse disponibles, es decir los datos se encuentran resguardados. Si son cuestiones de seguridad de la información, una vez superada las cuestiones burocráticas, ya estaremos en condiciones de acceder a los mismos. En caso de que los datos se encuentran resguardado el primer problema al que nos enfrentamos, es obtener el espacio suficiente para recuperar los mismos, de estar en alguna base de datos también es necesario obtener los recursos para poder acceder a la misma. Con estos pasos cumplimentados, la próxima tarea a realizar es una primera revisión de los datos obtenidos para conocer sus características.

El proceso de obtención de datos debe acompañarse de un relevamiento con los responsables de las fuentes, que refleje en relación al dominio en estudio, la calidad y completitud, de los registros de datos presente en las bases de datos, estos en función del dominio podrán ser en mayor o menor grado representativo del dominio, se obtendrán resultados limitados si los datos registrados en las bases de datos son incompletos y limitados.

2.2. Transformación de los datos para que pueda ser utilizado

El primer paso para la preparación de datos es conocer el problema a resolver para lo cual se beberán Incluir como actividad preliminar la comprensión del dominio o negocio: el propósito es asegurar el entendimiento del negocio y objetivos del proyecto, o al menos hacia que objetivo queremos llegar, sin esto nos resulta imposible conocer los datos que debemos extraer. Por otra parte debemos conocer la forma en que se debe presentar la información al modelo seleccionado para la explotación de datos, con estas dos precisiones se puede comenzar a recolectar la información y trabajar con ella. Cuando se esta trabajando en explotación de datos, se están utilizando datos que representan hechos de la vida real, esos datos deben se preparados para que las herramientas de explotación puedan trabajar con ellas. La preparación de los mismos no es un proceso automático, por lo cual es necesario aplicar nuestro conocimiento para generar el conjunto de datos necesario para poder aplicar un modelo de explotación. Por lo antes dicho podemos definir como el principal objetivo de la preparación de datos (la vista minable o dataset y su descripción) es tomar información manipularla, transformarla y presentarla para que pueda ser procesada por un modelo de minería de datos.

Para conocer que transformaciones debemos realizar y como la debemos presentar nos debemos hacer dos preguntas fundamentales: ¿Qué solución debemos obtener? y ¿Que técnica de explotación utilizaremos?. La primera cuestión la relacionaremos con las características y cantidad de información que deberemos manipular, y la segunda cuestión, la forma en que se debe presentar la información para la explotación. Con los datos accesibles y hecha la primera revisión de los mismos los pasos comunes en la preparación de datos, se puede definir como:

a. *Enriquecer la información*: Luego de analizar la información y teniendo respuesta a las preguntas antes generadas, se plantea la posibilidad de agregar datos a los ya obtenidos, pues la información con la que se cuenta no cumple con todos los requisitos necesarios para poder generar un conjunto de datos que sea aceptado por el modelo.

b. *Obtener casos testigos*: Esto se puede convertir en un proceso muy tedioso, la obtención de estos casos testigo nos permitirán definir si el modelo al que lo vamos a aplicar es viable o no en relación al conjunto de datos que tenemos.

c. *Determinar la estructura de los datos*: Para poder entender este concepto es necesario definir el término conjunto de datos, este hace referencia a los datos que serán utilizados por

el modelo de minería de datos para encontrar patrones. La estructura de datos hace referencia a la forma en que las variables se relacionan unas con otras en los conjuntos de datos. Es en esta estructura donde se buscarán relaciones y patrones de comportamiento.

d. *Construir el modelo de entrada de datos*: Se puede decir que hasta este paso en lo que nos hemos centrado es en obtener y conocer la información disponible. En este paso lo que se determinarán los procesos que se seguirán para el modelado de los datos, entre los cuales podremos nombrar: [i] normalización, [ii] tratamiento de los valores nulo o vacíos, [iii] detección de series (las mas comunes de tiempo), [iv] reducción del ancho de los datos, es decir la cantidad de columnas y [v] reducción de la profundidad, la cantidad de registros.
e. *Inspeccionar los datos*: Una vez generada todas estas transformaciones, el minero de datos necesitan evaluar el resultado para poder determinar si de las transformaciones hechas al conjunto de datos lo hace viable para que el modelo elegido lo pueda procesar.

2.3. Aplicación de la técnica explotación de datos seleccionada

Luego de realizar todas las transformaciones se procede a modelar los datos en función de la técnica de minería de datos elegida para actuar sobre la vista minable obtenida anteriormente, existen diferentes técnicas a saber: de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéticos, aprendizaje bayesiano, entre otras. Dependiendo de la técnica, se ejecutarán una o varias ejecuciones con uno o varios conjuntos de datos.

2.4. Evaluar los resultados obtenidos

Con los resultados obtenidos de las ejecuciones del modelo, se centra la atención en detectar y poder comprender el resultado de los mismos. Esta tarea no es para nada sencilla e insume gran cantidad de tiempo, esto se debe en muchos casos a la complejidad de los resultados obtenidos. De este análisis es de donde se puede concluir, que los resultados no han sido los esperados, por varios motivos, la técnica no es la correcta para la solución del problema; otra posibilidad es que el conjunto de datos, no haya sido el adecuado, que se deba generar otro conjunto de datos, para validar los resultados obtenidos en la primera modelización; o que el modelo no se ajuste a los requerimientos de negocio, es por estas razones que el último paso sea comenzar con el ciclo nuevamente. El modelado de la explotación de datos es un proceso de aproximación cíclica, el cual se debe ir mejorando a medida que se conoce más de la información con la cual se está trabajando. Es por esto que es necesario reiniciar el ciclo has que la información obtenida satisfaga el requerimiento que la produjo.

3. Método de transformación propuesto

El Método Unificado de Transformación (MUT), es el resultado de la experiencia adquirida en el procesamiento de grandes volúmenes de información sobre distintas plataformas, desde equipos IBM 390 a redes de computadoras personales que poseen alguna de las distintas versiones de Microsoft Windows existente, pasando por el AS 400 y diversas versiones de Unix. En esta categorización se hace necesario agregar la nueva generación de aplicaciones orientadas a sistemas de planeamiento de recursos empresariales, del vocablo en inglés *enterprise resourse plainning (ERP);* esto se debe a que debido a su complejidad, el usuario puede abstraerse del sistema operativo con le cual trabaja su computadora personal y solo operar dentro del entorno que le facilita el ERP, entre estos a modo de paradigma mencionaremos SAP.

3.1. Requerimientos para la aplicación de la metodología

El encargado de la transformación de datos debe tener conocimientos básicos sobre la notación que implemente del lenguaje unificado de modelado, de su vocablo en inglés *unified modeling language*, de aquí en mas UML, específicamente se hace referencia a los casos de usos y los diagramas de secuencia. El explotador de datos conoce los formatos de los archivos con los cuales deberá trabajar, además del formato de salida obtenidos por el proceso de transformación de datos y tiene permiso a los mismos y son accesibles el conjunto de datos de entrada que deberá transformar para poder ingresar los datos al modelo de minería de datos; además se cuenta con espacio suficiente para el proceso de los datos, vale aclarar que esta metodología antepone la agilidad y velocidad de procesamiento en detrimento del espacio de almacenamiento de archivos intermedios. Esta característica del método se basa en que el espacio físico de almacenamiento hoy en día es lo mas accesible y mas barato en comparación con recursos de memoria y procesador.

3.2. Descripción de la metodología

Conociendo las dos preguntas fundamentales para el proceso de transformación, es decir, donde estoy y ha donde quiero llegar, el método recomienda la aproximación gradual al objetivo final basado en un conjunto de pasos que se basan en: análisis de los requerimientos de transformación, modelado de las transformaciones, codificación pruebas, evaluación y nueva iteración. El principal objetivo del método propuesto no es realizar todas las transformaciones en un solo paso, sino que se realizan pequeñas modificaciones a los datos, se realizara una prueba de regresión completa de lo hecho hasta ese momento y una vez evaluada la misma de ser satisfactoria se volverá a reiniciar el ciclo con la próxima transformación a realizar. En las siguientes subsecciones se detallan los pasos mencionados anteriormente.

3.2.1. Fase de análisis de los requerimientos de transformación:

El primer paso que se deberá dar es el de recabar información acerca de que es lo que necesitamos obtener, es decir, conocer el formato de debe tener nuestro conjunto de datos para poder ser ingresado al modelo elegido para la minería de datos. En este paso se aplican las técnicas mas adecuadas que faciliten la extracción y educción (p.ej.: entrevistas), el único requisito al finalizar este paso, es poseer la especificación detallada del formato de datos para el modelo. Cabe mencionar que es posible encontrarnos ante la posibilidad que la misma persona que se encuentra encargada de las transformaciones sea la persona que ha definido el modelo de minería de datos, en tal caso solo se especificará el formato de archivo. Como resultado de la educción de requerimientos, se obtendrá la especificación del formato de archivo (ver Tabla 1) que se presenta modo de ejemplo

Nombre del campo	Tipo	Valores permitidos	Valor por defecto	Obligatorio	Máscara
identificador de usuario	Numérico	[0-9999]		Si	9999
Nombre	Carácter(20)	[A-Z]		Si	
Apellido	Carácter(20)	[A-Z]		Si	
Domicilio	Carácter(50)	[A-Z]		Si	
Número	Numérico	[0-9999]		Si	9999
Código postal	Alfanumérico(8)	[0-9999] [A-Z]	“”	No	Z999ZZZ

Tabla 1. Ejemplo de una especificación de archivo de entrada al modelo

Con el formato de archivo de ingreso al modelo de datos ya especificado, se abren dos cursos de acción: [a] comenzar a recabar la información necesaria para poder detectar el origen de datos para la creación del archivo solicitado ó [b] con el formato de archivo ya especificado, volver sobre el

modelo de minería de datos seleccionado, con la finalidad de detectar los requisitos del conjunto de datos para su uso, es decir, cantidad de registros necesarios para su aplicación, cantidad de conjuntos de datos necesarios para su validación o entrenamiento.
En el primer curso de acción se deberán seguir los siguientes pasos:

a. *Repetir la técnica de entrevista, para detectar el origen de datos:* En esta etapa de entrevistas, lo que se observa es que la cantidad de personas involucradas es mucho mayor de lo que uno a priori puede suponer. Entre las cuestiones a tener en cuenta podemos citar:
   - Se deberá entrevistar al administrador de la base de datos, para conocer la antigüedad de los datos que se encuentran en línea en la base de datos.
   - Otra cuestión a manejar con el administrador es determinar las posibles plataformas donde se encuentran los datos, de ser todas almacenadas en Bases de Datos, cuales y que versiones.
   - Otro punto es solicitarle el diagrama de entidad - relación (DER), para conocer la estructura de las tablas y sus campos.
   - De esto surgen dos implicancias, por un lado, en función de la cantidad que se encuentran en línea, se deberá entrevistar, al encargado del resguardo de los mismos, para conocer desde hace cuanto tiempo se tienen datos resguardados y su posibilidad de acceso. La segunda implicancia es, del análisis del DER, surgirán dudas sobre el origen de los datos esto hará se conserven entrevistas con los responsables de los diversos sistemas. Aquí será necesario realizar entrevistas grupales para resolver las inconsistencias propias de todo modelo de datos.
b. *Acceso a la información:* En la medida que se detecte las fuentes de los datos, se deberán tomar todos los recaudos para poder acceder a los mismos, algunas de las cuestiones que se deberá resolver son:
   - Cuestiones referentes a la seguridad de datos, formalizar los pedidos de acceso a la información
   - Si los datos se encuentran resguardados, es necesario disponer del espacio para su recupero y calcular el tiempo que llevara esta tarea, que puede ser muy significativa.

   En la Tabla 2 se presenta un ejemplo de formato de documento que cuenta el origen de cada dato.

Nombre de Campo	Origen		Responsable	Resguardo	Accesible
identificador de usuario	Base de datos	DBCMP	Administrador de Base de datos	Si	Si
	Tabla	Maestro			
	Campo	Usuario			
	Tipo	Integer			
Nombre	Base de datos	DBCMP	Administrador de Base de datos	Si	Si
	Tabla	Maestro1			
	Campo	NombreUsu			
	Tipo	Carácter(40)			
Apellido	Base de datos	DBCMP	Administrador de Base de datos	Si	Si
	Tabla	Maestro1			
	Campo	ApeUsu			
	Tipo	Carácter(35)			
Domicilio	Base de datos	DBCMP	Administrador de Base de datos	Si	Si
	Tabla	Maestro2			
	Campo	Dom			
	Tipo	Carácter(100)			
Número	Base de datos	DBCMP	Administrador de Base de datos	Si	Si
	Tabla	Maestro2			
	Campo	NumUsu			
	Tipo	Carácter(10)			
Código postal	Planilla Calc.	Códigos		No	Si
	Equipo	Serv01			
	Raiz	D:/			

Tabla 2. Ejemplo de formato de documento que cuenta el origen de cada dato

En el segundo curso de acción se deberán seguir los siguientes pasos:

a. *Repetir la técnica de entrevista, para detectar el origen del conjuntos de datos de pruebas:* En esta etapa de entrevistas es necesario recabar información sobre el modelo de minería a utilizar, esto servirá para poder generar los conjuntos de prueba, como ejemplo de esto se puede citar el caso de la utilización de redes de neuronas para el proceso de minería de datos, el mismo según el tipo de red elegida necesitará, diversos conjuntos de prueba para su entrenamiento.
b. *Análisis del modelo requerido y el disponible:* Con la información recabada en los pasos anteriores Origen de datos y Características del conjunto de datos es necesario generar un hito en el proceso, para esto se debe validar el formato de archivo para el modelo de datos, de este paso pueden surgir las siguientes alternativas.
   - Todos los datos se encuentran disponibles, alternativa mas optimista.
   - Ciertos datos no se encuentran disponibles, por la razón que fuera, tanto la no existencia de los mismos, razones de seguridad, o simplemente por no poderse recuperar del resguardo de datos.

   Ante esta situación se plantea la necesidad de tratar de conseguir los mismos de otro origen, a modo de ejemplo podríamos citar al Instituto de Estadística y Censo, Internet, proveedores, etc. Si la información esta disponible en algún otro origen vuelve a generar los pasos que se han definido en el origen de datos. De no ser así, se plantea una cuestión relacionada con el modelo elegido para la minería de datos. La cuestión a resolver en este punto es la decisión de cambiar el modelo en función de los datos que tenemos.

Regla de escritorio 1: De lo mencionado anteriormente sobre la forma de obtener el formato de archivo y como se ha enunciado en el punto B, la detección de las características del conjunto de datos. Se podría pensar que ambas actividades se pueden realizar a un mismo tiempo, la experiencia en proyectos de minería de datos demuestra que, dentro del proceso de transformación de datos y dentro de ésta una de las etapas mas tediosa es la recolección de datos para formar los conjuntos de datos.

Regla de escritorio 2: La decisión del modelo de minería de datos a utilizar debe estar en función del problema a resolver. En el caso que se nos plantee la situación, antes descripta, no poseemos los datos para proporcionarle al modelo, esto no debe ser considerado como un error a diferencia de esto el proceso de minería de datos, ha detectado el problema en una etapa temprana del desarrollo, que se puede definir como:
"El problema en cuestión esta dado por la falta de información en nuestros sistemas."

En la Figura 1 se esquematiza el proceso de obtención de requisitos para la transformación de datos.

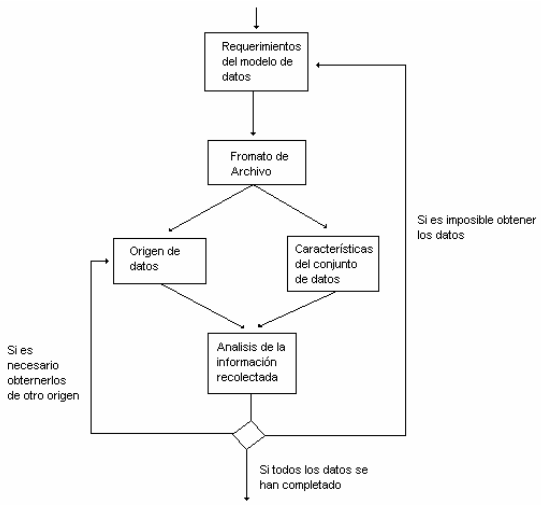

Figura 1. Proceso de obtención de requisitos para la transformación de datos.

3.2.2. Fase de Modelo de las transformaciones

En esta etapa es donde se diseñan las transformaciones necesarias para que los datos tomados del origen lleguen a la estructura requerida por el modelo de minería de datos. Esto lo realizaremos utilizando Casos de Usos. Los casos de uso, del vocablo inglés *use case*, constituyen el concepto central del método OOSE de Ivan Jacobson, uno de los padres de UML. Los casos de uso representan, el medio para describir el carácter funcional de los objetos, son una representación orientada a la funcionalidad del sistema y permiten modelar las expectativas del usuario. Existen tres conceptos fundamentales en el modelado de los casos de uso: los actores que utilizan el sistema, los casos de uso y los escenarios.
Los actores pueden ser de dos tipos: [a] humanos, usuarios de los programas y [b] software, programas que se comunican con nuestro sistema. Desde el punto de vista del sistema exista dos tipos de actores: [a] los actores primarios, que son los que utilizan el sistema y [b] los actores secundarios, que tienen funciones de administración y mantenimiento del mismo.
Los casos de uso representan la utilización del sistema por parte de los actores. Los casos de uso se pueden organizar desde mayor grado de abstracción hasta el detalle que se crea necesario. La representación de los casos de uso puede ser textual (ver Tabla 3) o gráfica (ver Figura 2). Un ejemplo de una representación textual es:

Caso de Uso "Validar valores nulos"
• El controlador ejecuta el programa validar • El programa validar controla la no existencia de nulos • El controlador toma el control nuevamente • El controlador evalúa si se generaron errores

Tabla 3. Ejemplo de una representación textual

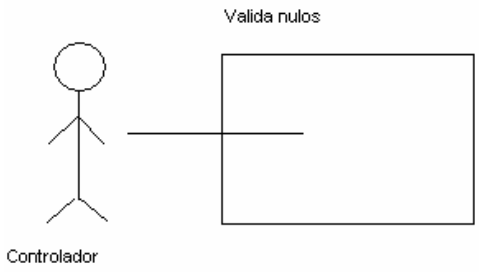

Figura 2. Ejemplo de una representación gráfica

Un escenario es una serie de eventos ordenados en el tiempo, que simulan una ejecución particular del sistema. de manera general, un escenario utiliza dos tipos de conceptos: [a] objetos que normalmente forman parte del sistema y [b] eventos emitidos y recibidos por los objetos implicados en el escenario. Los escenarios permiten experimentar las ejecuciones del sistema, por lo que resultan muy para las pruebas y el mantenimiento. El modelado de las transformaciones tendrán como actor al controlador, que es el encargado de generar los eventos, para que el flujo de los datos tengan las transformaciones necesarias.
Sea el siguiente ejemplo que utilizan casos de uso, escenarios y especificación de requerimientos para el proceso de datos. El caso de uso, donde se modeliza un proceso de transformación de datos, consta de tres operaciones básicas: el primer paso es la validación del formato del archivo de origen, el segundo es el reemplazo de valores nulos por espacios en blanco y por último extraer de la totalidad de los datos disponibles un conjunto de datos, representativo del total. Se puede observar también que el actor de este caso de uso es el controlador de tareas quien es el encargado de invocar a todas las tareas (Figuras 3 y 4).

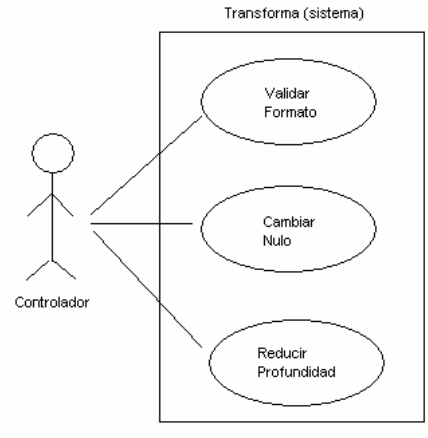

Figura 3. Actor: controlador de tareas quien es el encargado de invocar a todas las tareas

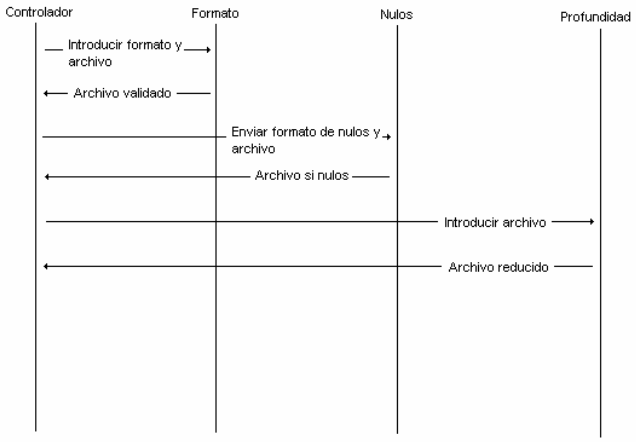

Figura 4. Modelización del escenario sobre el caso de uso.

En el ejemplo tratado, no es necesario profundizar los casos de usos ni el escenario de trabajo, para cerrar el mismo solo hace falta especificar el detalle de los formatos de entrada y salida de cada uno de los pasos involucrados (Tablas 4 y 5).

Nombre de paso	Programa	Entrada	Salida	Observaciones
Validar Formato	formato	Archivo de origen	Archivo validado	Los archivos son de 1 giga. Ver performance.
		Archivo de formato	Archivo de errores	
		Archivo de errores	Archivo de pasos	
		Archivo de pasos		

Tabla 4. especificar el detalle de los formatos de entrada y salida de cada uno de los pasos involucrados

Formato	Nombre campo	Tipo	Longitud	Obligatorio
Archivo origen	Id	Numérico	4	Si
	NomCli	Carácter	20	Si
	ApeCli	Carácter	20	Si
	Dir	Carácter	30	Si

Tabla 5. especificar el detalle de los formatos de entrada y salida del archivo de datos

3.2.3. Fase de Codificación

En este paso se codifican todos los programas que se necesiten para realizar las transformaciones necesarias para el modelo de minería de datos. El encargado de la codificación recibirá, al menos, las especificaciones de los formatos de entrada y salidas. Como el controlador de tareas es independiente del programa que debe ejecutar, se puede usar el lenguaje de programación que se desea, siempre y cuando este pueda ser soportado por la plataforma en la que se desea trabajar. Cabe mencionar que se puede entregar al codificador toda la información que el encargado de las transformaciones crea necesario. Si el lenguaje así lo permite se podría entregar los diagramas de clases necesario para la codificación, así podríamos utilizar uno de los tantos esquemas que nos facilita UML. Volviendo sobre la documentación mínima que se le debe entregar al codificador, en el cambo observaciones de la planilla con los nombres de los archivos que debe recibir y retornar el programa, es muy importante que el codificador sepa cuales son las principales características del archivo no ya de formato que las posee, sino de volumen de información pues ante distintas cantidades de información por procesar, la codificación será muy distinta. El codificador además de realizar el programa se encarga de hacer las pruebas de unidad de los programas que realiza, es pro esto que se le debe también facilitar un archivo de entrada con los datos reales, de ser posible con el volumen de información que en producción se enfrentará. Una vez que se ha finalizado con la codificación, el paso siguiente es la prueba de unidad y de regresión, por parte del encargado de la generación de la secuencia de tareas.

3.2.4. Fase de Pruebas

Esta etapa de prueba no solo se refiere a la comprobación de los programa encargados de la generación de las transformaciones, sino a la construcción del archivo que proveerá la secuencia de pasos al controlador de tareas. Con las primeras transformaciones a realizar, se carga el archivo de formatos del controlador de tareas, es necesario tener en cuenta que no es recomendable agregar varios pasos de una vez, tratando de hacer una prueba, validar la salida y agregar otro paso. Las pruebas que se realizan son:

- *De unidad:* La finalidad de esta prueba es validar que el programa cumpla la función para la cual fue ingresado a las tareas, es necesario poder simular de la manera mas precisa posible una ejecución real. La validación que se hace es en función de la documentación antes desarrollada, se toman los formatos de archivo de entrada y salida, y simplemente se evalúa si los formatos son correctos.
- *De regresión:* En este tipo de pruebas lo que se debe realizar es la validación de los tiempos de procesamiento y recursos necesarios. Para realizar esto es necesario ejecutar la tarea completa hasta el último paso que hemos agregado, es decir hacer una corrida completa de lo que tenemos hasta este momento. Lo que se busca probar es el tiempo de procesamiento, en la sucesión de pasos ejecutados es posible detectar que el tiempo de procesamiento es inaceptable para nuestro sistema, que los recursos utilizados son demasiados, etc.

De la evaluación antes descripta se pueden presentar distintas variantes:

- *Los tiempos son aceptables:* Esta es la posibilidad mas optimista de ser así, lo que se hace es continuar con el agregado de los siguientes pasos, esto puede ser que ya se tenga la especificación de la tarea y la próxima iteración a realizar solo sea agregar un paso mas y rehacer los ciclos de prueba.
- *Es procesamiento es demasiado extenso:* Esto hace que se deba replantear la estrategia de transformaciones a realizar, aquí se debe detectar cual es el paso que más tiempo lleva y modificarlo.
- *Los recursos no son los óptimos:* Este tipo de alternativa se da cuando por ejemplo el espacio de almacenamiento intermedio es demasiado grande y no se dispone de más espacio en disco, esto hace que sea necesario la reformulación de la estrategia a desarrollar.

Sobre los posibles caminos de acción que se puedan seguir en esta opción serán abordados en el próximo paso de la metodología propuesta

3.2.5. Fase de Evaluación:

Con toda la información de las pruebas antes realizadas el encargado de realizar las transformaciones, deberá tomar un camino de acción, como se ha dicho antes, salvo que todo halla sucedido como se esperaba, en el resto de las opciones se deberá modificar algo. La primera alternativa a seguir es una vez detectado el paso, programa, que más recursos o tiempo demora, es tratar de optimizarlo. Otra alternativas no tan costosa es, la posibilidad de ejecutar las tareas en forma paralela, esto se hace agregando un punto de bifurcación en el controlador de tareas y se hace un procesamiento en paralelo; de no poder hacer esto otro camino de acción a seguir es la posibilidad es plantear generar nuevamente el programa en un lenguaje con mejor rendimiento, a modo de ejemplo podemos citar si se ha hecho el programa en un lenguaje como Visual Basic, se lo podría pasar a C/C++, para que su ejecución sea mas optima. Otra alternativa que también podemos elegir es, a semejanza de la normalización de las bases de datos que en una primera instancia se normaliza, y para finalizar se realiza una des-normalización de las tablas para que estas posean una

velocidad de acceso aceptable; se realizaran modificaciones en los programas que integran cada paso, para que se hagan mas de una transformación en un paso, como de la experiencia se ha observado que en cada paso los tiempos de acceso a disco, lectura del archivo y escritura de los mismos, es lo que mas tiempo insume, unir transformaciones puede hacer que se reduzca el tiempo de procesamiento, aunque esto va en detrimento de los reprocesos que se puedan generar, en ciertos casos es la única alternativa mejorar la performance. Esto son algunos de los caminos alternativos que se podrán seguir para la mejora del rendimiento de la tara a ejecutar, en definitiva el encargado de la realización de las transformaciones tendrá la libertad de realizar las modificaciones que desee para poder llevar a buen puerta su trabajo.

3.2.6. Fase de Nueva iteración

De lo dicho hasta el momento se deduce la necesidad de generar nuevas iteraciones con cada paso de la tarea a realizar, este proceso se repite hasta finalizar todas las transformaciones necesarias para satisfacer el Modelo de Minería de Datos.
De la metodología propuesta se desprenden algunas observaciones necesarias de hacer:

- Primero en función del método de trabajo el controlador de tareas es de suma importancia para la realización del trabajo, cuanto mas sofisticado sea el controlador y mas opciones pueda manejar, mejor será la forma que apliquemos la metodología.
- Sobre el uso de un sistema de Monitoreo y Diagnostico para el controlador de tareas, es necesario proveerle una herramienta, en este caso un sistema experto, para que pueda manejar alternativas no contempladas por los programas hechos para generar los pasos de las tareas, podemos citar, cuando parar ante el primer error encontrado en el archivo o después de encontrar cien registros con error, y ante esta situación que se debe hacer enviar los registros erróneos a un archivo temporario para su posterior análisis. Ante estos errores que se debe hacer, en este caso una vez decidido que se ha producido un error el cual es la política que se debe llevar a cabo. En caso de ser necesario dar avisos, entra en juego el subsistema de alarmas, que es el encargado de disparar y controlar todas las alarmas que generara el sistema y el tiempo de respuesta de los mismos. Dentro de las tareas rutinarias en el proceso de transformación llevada a cabo una tarea que es necesaria es la evaluación de las ejecuciones. Esto se trata de generar un seguimiento de los procesos cuando ya se encuentran en producción, donde ya se ha automatizado la tarea y no es controlada por ningún operador humano.
- La experiencia dicta que todos los procesos con el tiempo se van degradando, su tiempo de respuesta empieza a ser peor, el espacio en disco utilizado aumenta, y esto puede llegar a niveles inaceptables, aunque el resultado final es el esperado.

Como se habrá observado en la metodología, cuando se detalla la documentación requerida para la generación de cada paso se especificó un “Archivo de pasos”, el cual no se había tratado, este archivo es el cual nos permitirá evaluar el rendimiento de la tarea, en el mismo se contabilizaran los registros transformados , tiempo de procesamiento y demás información que se crea útil para el análisis del rendimiento en producción .Lo que se realizara se podría denominar como una minería de datos del proceso de transformación de la minería de datos. Es ahí donde la utilización de un sistema experto puede tener mucho valor agregado, pues el mismo se encargará de analizar los tiempos de proceso compararlo con el volumen de información que se ha procesado y determinar un camino de acción a seguir. Algunas situaciones que se pueden detectar con este análisis son, baja en la capacidad de procesamiento en determinados momentos del día, hay que recordar que nuestras pruebas aunque completas, no pueden simular todo el ambiente de producción donde otros procesos están corriendo en paralelo al nuestro y están compitiendo por los recursos del mismo. Aumento de la cantidad de espacio necesario para la ejecución de los procesos, esto se puede producir por la

fragmentación de la información en el disco, además de la perdida de respuesta, como se puede observar cuanto antes se detecten estos problemas menos traumática será su solución.

4. Conclusiones

De lo expuesto podemos destacar que en el proceso de la Minería de Datos, el Modelo de Minería elegido es solo una pequeña parte de la totalidad del proceso. Creer que lo único importante es obtener un buen modelo de minería de datos, es a juicio de los autores uno de los errores más comunes que se encuentran en los proyectos de minería de datos.
Es un problema abierto la generación totalmente automática de las transformaciones de datos para su modelado en minería de datos. Poder llegar a un proceso automático redundaría en una considerable baja de costos y tiempo de implementación de una solución en la minería de datos.
En el contexto de una herramienta para transformación automática de datos debería considerarse: [a] la utilización de un sistema experto embebido en el sistema que ayude con el trabajo del control de tareas y [b] migrar a un sistema experto el subsistema de alarmas embebiéndolo en el controlador de tareas.

5. Referencias

- Adriaans, P (1996). *Data mining.* Addison-Wesley
- Berry, M. (1997) *Data mining techniques.* Wiley
- Booch. G. (1998). *Objects, Components, and Frameworks with UML*. Addison-Wesley
- Booch. G. (1998). *The unified modeling language user guide*. Addison-Wesley
- Booch. G. (1999). *The unified modeling language reference manual*. Addison-Wesley
- Escudero, L. (1977). *Reconocimiento de Patrones*. Paraninfo
- Frakes, W. (1992) *Information Retrieval*. Prentice Hall
- Jacobson, I (1998). *The unified software development process*. Addison-Wesley
- Larman, C. (2002). *UML y Patrones*. Prentice Hall.
- Pyle. D. (1999). *Data preparation for data mining.* Morgan Kaufman
- Turban, E. (1998) *Decision support systems and intelligent systems*. Prentice Hall
- Witten, C. y Frank, H. Clark, P.; Boswell, R. (2000) *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers.