Propuesta de Proceso de Transformación de Datos para Proyectos de Explotación de Información

Ezequiel Baldizzoni
Laboratorio de Investigación y Desarrollo en Ingeniería de Explotación de Información
Grupo Investigación en Sistemas de Información
Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús.
Remedios de Escalada, Buenos Aires, Argentina.
ezequiel_baldizzoni@hotmail.com, rgarcia@unla.edu.ar

***Resumen*—La exploración y análisis, en forma automática o semi-automática, de grandes volúmenes de datos para la detección de patrones, se enmarca en los proyecto de explotación de información para lo cual utiliza algoritmos de minería de datos (Data Mining). Dentro de un proyecto normal de explotación de información, existe una de las etapas fundamentales llamada transformación de datos que se ocupa de prepararlos con el propósito de entregar a las etapas posteriores del proceso un conjunto de datos de calidad y de esta forma concluir con resultados mas exactos. Esta etapa normalmente consume alrededor de un 60% del esfuerzo de desarrollo. Dado el gran esfuerzo necesario se propone un proceso de transformación de datos.**

***Índice de Términos*—Proceso, transformación de datos, explotación de información.**

I. Introducción

Se plantea el contexto de la investigación (sección A), se establece su objetivo (sección B), y se resume la estructura general del articulo (sección C).

A. Contexto de la Investigación

Los proyectos de explotación de información son hoy en día una de las herramientas más importantes en las organizaciones y son utilizadas para tomar decisiones de negocio. La minería de datos es un tipo de técnica para extraer información de los datos que las organizaciones fueron almacenando a lo largo de sus vidas. Dado que este tipo de tecnología es de gran ayuda, se puede decir que es importante reducir las dificultades que esta pueda conllevar. También es de destacar que cada una de las etapas de un proyecto de explotación de información consume un esfuerzo distinto según las dificultades por las que atraviesa. Las etapas de un proyecto de explotación de información pueden ser las siguientes:

- Análisis de requerimientos, esta etapa va a cubrir el análisis de las necesidades de la organización y es donde se va a decidir qué datos van a ser necesarios para el resto del proceso.
- Selección del conjunto de datos, tanto en lo que se refiere a las variables objetivo (aquellas que se quiere predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.
- Análisis de los datos, se analizan los datos obtenidos y se decide si son los correctos o si hay que volver a la primera etapa ya que puede ser esta la incorrecta.
- Transformación de los datos, se ejecutara una serie de pasos que mejoraran la calidad de los datos para luego poder ser utilizados por la siguiente etapa.
- Selección y ejecución de las técnicas de minería de datos, según el modelo seleccionado se ejecutaran las distintas herramientas o técnicas de minería de datos.
- Análisis de resultados, Una vez ejecutadas las técnicas de minería de datos se analizan los resultados arrojados en busca de conocimiento.

Los proyectos de explotación de información se encargan de efectuar distintos pasos con el fin de analizar, de forma automática o semi-automática, grandes volúmenes de datos y de esta forma extraer patrones de interés para el negocio que hasta este momento eran desconocidos.

Hay que reconocer que reduciendo el tiempo y el esfuerzo de este tipo de proyectos da como resultado una mayor cantidad de conocimiento adquirido de los datos y un menor costo del proyecto.

La etapa de transformación de los datos consume el 60% del esfuerzo total del proyecto por lo que mejorar el rendimiento en la etapa mencionada, sería de gran ayuda para reducir el esfuerzo total del proceso.

B. Objetivo de la Investigación

Esta investigación tiene como objetivo desarrollar un proceso de transformación de datos para proyectos de explotación de información, las actividades y las técnicas asociadas.

Para llevarlo a cabo, es necesario dividir la etapa de transformación de datos en los proyectos de explotación de información en una cantidad finita de actividades que deben ir ejecutándose secuencialmente en una serie de ciclos que garantizarán la salida a esta etapa.

Se plantea a su vez una serie de técnicas que se utilizaran en cada actividad con el fin de garantizar también que cada entrada/salida sea la correcta.

Cada una de las técnicas se dividirá en distintos pasos los cuales deben ir siendo cumplidos secuencialmente hasta transformar cada entrada de las actividades en salida.

De esta forma se puede decir que el objetivo general de este artículo, es optimizar la tarea de transformación de los datos en proyectos de explotación de información.

C. Estructura del trabajo

En Estado de la Cuestión (sección II) se desarrolla una investigación sobre distintas teorías y técnicas que son concurrentes con los objetivos de este artículo de investigación. Dentro del mismo se presentarán las distintas teorías que encuadran dentro de cada actividad: enriquecer los datos, obtener y ejecutar los casos testigo, determinar y aplicar las estructura de los datos, construir el modelo de entrada y por último la inspección de los datos.

En Descripción del Problema (sección III) se presenta el problema de investigación partiendo de las dificultades que hoy en día poseen las organizaciones al momento de ejecutar proyectos de explotación de información sobre los repositorios de datos que se almacenan desde los sistemas existentes o deprecados. En primer lugar se describe la identificación del problema de investigación, luego se caracteriza el problema abierto y se concluye con un sumario de investigación.

En Solución (sección se IV) se presenta una serie de cuestiones generales sobre la resolución del problema abordado, se describe la propuesta de proceso de transformación de datos para proyectos de explotación de información, la estructura general del proceso y las actividades.

En Conclusiones (sección se V) se presentan los aportes de este artículo de investigación y se destacan las futuras líneas de investigación que se consideran de interés en base al problema abierto que se presenta en este artículo de investigación.

II. Estado de la Cuestión

Se presenta el estado de la cuestión sobre teorías y técnicas que son concurrentes con los objetivos de esta investigación. Dentro del mismo se presentan las distintas teorías que encuadran dentro de cada actividad: enriquecer los datos (sección A), obtener y ejecutar los casos testigo (sección B), determinar y aplicar las estructura de los datos (sección C), construir el modelo de entrada (sección D) y por último la inspección de los datos (sección E).

A. Enriquecer los Datos

En la primera etapa, la de enriquecimiento de los datos, es posible afirmar que no existe una teoría específica que abarque esta tarea, dado que es una tarea de análisis y depende de la formación de la persona encargada del mismo. También, depende en gran medida, del modelo utilizado en el proceso de explotación de información del proyecto.

Por estas razones no se va a desarrollar una teoría específica para esta etapa del proceso.

B. Obtener y Ejecutar los Casos Testigo

Esta es la etapa en la que se decide si los datos cumplen con las características mínimas, para entregar un resultado útil al proyecto de explotación de información que lo ejecuta.

La obtención de los casos testigo se puede lograr mediante entrevistas al personal de la organización que estén involucrados con los datos necesarios o con el proyecto en sí.

Una vez entrevistado al personal involucrado, se generan los casos testigo que serán la guía para decidir qué datos y con qué formato serán necesarios.

Por último es necesario impactar están información en algún tipo de soporte para luego comprobar que los datos cumplen con lo especificado.

Una técnica muy utilizada para este tipo de tareas es la utilización de listas de chequeo (en ingles check list) que consta de una tabla que muestra, al personal que ejecuta las pruebas, una lista con ítems que deben ser analizados para determinar si los datos son los correctos para el resultado final.

Estas listas no tendrán la necesidad de chequear si los datos son de calidad, solo se carga en su interior los ítems que determinen si las variable, atributos, datos, tipos, o lo encontrado, son los correctos para el resultado final (cuentan con las necesidades mínimas).

Una lista de chequeo consta de varias columnas las cuales pueden ser número de identificación del ítem, descripción (alguna descripción de lo que hay que chequear), resultado esperado, resultado encontrado, observaciones y demas. Un ejemplo de lista de chequeo se puede ver en la tabla I.

Una lista de cheque correcta debe detallar uno por uno distintos aspectos que se deben analizar, comprobar o verificar.

TABLA I. Ejemplo de lista de chequeo.

Item	Descripción	Esperado	Encontrado
1	**Columna nombre y apellido**		
1.1	El formato es apellido, nombres	**SI**	
1.2	Los nombres con mayúsculas	**SI**	
1.3	Existe nombre y apellido	**SI**	
2	**Columna Dirección**		
2.1	Separado dirección de numero	**SI**	
2.2	Esta el barrio	**SI**	
2.3	Esta la provincia	**SI**	
2.4	Hay datos perdidos	**NO**	
3	**Columna jefe**		
3.1	Hay datos nulos	**NO**	
4	**Tabla relaciones**		
4.1	**Columna relación**		
4.2.1	Hay datos perdidos	**NO**	
4.2.2	La relación es correcta	**SI**	

C. Determinar y Aplicar la Estructura de los Datos

En esta actividad se debe efectuar un trabajo de análisis y transformación de los datos de tal manera de que si estos provienen de diferentes fuentes, unificarlos para que de varias fuentes quede un solo repositorio. También hay que lograr que a partir de las relaciones existentes entre los objetos de una fuente de datos se obtenga un solo objeto que contenga todos los anteriores. Partiendo de los valores del objeto principal por medio de estas relaciones con los demás objetos, se une a este hasta llegar a tener un solo objeto que contenga todos los datos. Aquí es donde se escoge dejar de lado alguna variable que no sea utilizada en adelante.

Luego de esta tarea es necesario documentar todo criterio de estructuración utilizado para, de aquí en más, no necesitar un análisis extenso y de esta forma facilitar los futuros trabajos.

Posterior a la técnica manual es posible ejecutar una técnica asistida por computadora para mejorar el resultado y siendo monitoreada por un ser humano se obtendrán resultados mucho más valiosos que los obtenidos en un principio con las técnicas manuales.

Una técnica asistida por computadora para esta etapa es la técnica de agrupamiento (en ingles clustering). Agrupamiento es una técnica usada para lograr grupos multidimensionales. Con el fin de determinar que cada grupo utilice el concepto de distancia euclideana. A partir de un elemento calcula esta distancia y según el valor que se obtenga cómo resultado determina si pertenece a algún grupo o no. En este caso se utiliza para hacer un preanálisis de los datos pero también puede ser utilizada para técnicas post análisis.

La Fórmula 1 muestra la distancia euclideana para un espacio bidimensional entre dos puntos P1 y P2, de coordenadas (x1, y1) y (x2, y2) respectivamente.

$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

D. Construir el Modelo de Entrada de Datos

Esta etapa es la más costosa de todas ya que es la que consume más del 80% del esfuerzo de todo el proceso de transformación de datos. Aquí se mencionan las teorías que se utilizan para detectar problemas en los datos y repararlos, modificar sus tipos, el ancho y profundidad. Primero se abordará el tratamiento de los valores nulos o vacíos (sección D.1), luego el tratamiento de duplicados (sección D.2), el tratamiento de valores ruidosos (sección D.3), la normalización (sección D.4), tratamiento de series (sección D.5), reducción del ancho de los datos (sección D.6) y por último la reducción de la profundidad de los datos (sección D.7).

D.1) Tratamiento de los Valores Nulos o Vacíos

La existencia de una cantidad considerable de valores nulos en una variable dificulta el análisis de los datos, ya que usualmente no permite la aplicación de las técnicas existentes que posibilitan el descubrimiento de conocimiento.

Hay una serie de enunciados que pueden agrupar a las distintas causas por las cuales estos valores son extraviados. Ejemplos de estos son:

- Los datos faltantes son parte del dominio de la variable.
- Los datos se han perdido.
- La existencia potencial de desviaciones en el análisis atribuible a las diferencias sistemáticas entre los datos observados y los datos perdidos.
- Problemas en la utilización de hardware disponible.
- Inconsistencia con otros registros de datos que son borrados.
- Datos que nunca fueron ingresados.
- Los datos incompletos.

En la práctica, la eficacia de las técnicas de tratamiento de valores nulos está directamente relacionada con la razón por la cual tuvo su origen el valor perdido. Si existe alguna información acerca de ella, es posible que se encuentre una regla para completar estos valores, por el contrario, si no se cuenta con dicha información, es necesario aplicar técnicas de evaluación de los valores perdidos que encuentren algún patrón que permita ya sea completarlos o descartarlos (en el caso que no afecten el análisis); decisión que depende en gran medida del tipo del valor perdido y la importancia del registro en la base de datos [1]. A continuación se describen las categorías de los datos perdidos (sección D.1.1) y el tratamiento de los datos perdidos (sección D.1.2).

D.1.1) Categorías de Datos Perdidos

Los datos faltantes o perdidos pueden ser categorizados en tres tipos [2]:

- Datos Completamente Ausentes al azar (Missing Completely at random MCAR). Corresponden a variables con datos perdidos que no poseen relación alguna a los valores de otros registros dentro de la misma variable ni a los valores de otras variables. Cuando ocurre esto, las distribuciones de probabilidad de los datos faltantes y de todos los datos son idénticas.
- Datos Ausentes al azar (Missing at Random MAR). Corresponde a variables con datos perdidos que tiene alguna relación con otras variables
- Datos Ausentes No al Azar (Not Missing at random NMAR). Corresponde a variables con datos perdidos donde el mismo dato perdido determina en sí mismo por qué es desconocido.

D.1.2) Tratamiento de los Datos Perdidos

Según [3] se dispone de una serie de técnicas para abordar los problemas de valores nulos o faltantes:

Descartar los registros con datos faltantes: Este método es práctico sólo cuando los datos contienen una relativamente baja cantidad de registros con datos perdidos y cuando el análisis de todos los datos no produce un sesgo importante por no utilizar estos registros. Como métodos para descarte de datos faltantes podeos nombrar a Listwise Deletion que elimina todo aquel registro que contenga datos faltantes en cualquier variable y Pairwise Deletion solo elimina los registros que tengan datos perdidos en las variables poco relevantes o que no son necesarias para el análisis [4]. Según sea el caso se puede plantear la eliminación de variables o de registros:

Eliminación de variable: La eliminación de datos perdidos es una medida poco recomendable debido a la pérdida de información que genera, sin embargo puede ser una buena opción en caso que los datos perdidos sean de naturaleza MCAR y no puedan ser imputados fehacientemente [5].

En el caso que se esté pensando en eliminar una columna se debe tener en consideración la cantidad de datos perdidos que posee en total y si existe o no alguna relación con otra variable. Sería recomendable eliminar la columna en el caso que la cantidad de valores perdidos supere un umbral mínimo que permita análisis (por ejemplo que posea más de la mitad de datos perdidos probablemente no genere información fidedigna), o cuando la variable sea MCAR y no pueda ser deducida de ninguna forma con los datos existentes, por lo que imputarlos generaría un mayor costo de error que por pérdida de información.

Eliminación de registros: Al eliminar registros en primer lugar hay que ser cuidadoso con la proporción de las clases que están evaluando. Si el problema a resolver constituye uno de clases desbalanceadas, la eliminación de un registro del cual hay pocas filas puede ser una gran pérdida de información inclusive si posee gran parte de sus atributos con datos perdidos o vacíos. Por ello se debe tener cuidado con qué tipo de registro se está eliminando y tomar medidas de acuerdo a la información relativa que se pierde con su eliminación. Ahora, se recomienda eliminar registros siempre y cuando posean una gran cantidad de valores perdidos o blancos y correspondan a una pequeña cantidad dentro del total de los datos.

Imputar los datos faltantes: Este método es aplicable cuando la cantidad de atributos con datos faltantes es relativamente pequeña en relación al número de registros que presentan dicha condición. Existen dos grandes tipos de técnicas que pueden ser agrupadas en dos grupos, Imputación Simple (Single Imputation) e Imputación Múltiple (Multiple Imputation).

Imputación Simple (Single Imputation): Es quizás el enfoque más utilizado en la práctica. En este método se estima el dato faltante usando otros datos relacionados que estén disponibles. Esto puede lograrse de varias formas, entre ellas [3]: Imputación por el promedio, Imputación por la moda, Imputación Hot Deck, Imputación por regresión.

Imputación por el promedio: Reemplazar los datos faltantes a través de la imputación por el promedio (en el cual se reemplaza el valor faltante de acuerdo al valor promedio de un grupo apropiadamente definido de valores disponibles).

Imputación simple a través de la imputación por el promedio posee tres limitaciones potenciales [6]:

- Disminuye la variabilidad inherente al conjunto original de datos, particularmente en el caso que el mismo valor promedio sea utilizado para reemplazar varios datos faltantes.
- Es dependiente de las elecciones de grupos de datos.
- Si existen datos fuera de rango (outliers) entre los conjuntos de datos candidatos para obtener el promedio que se empleará para realizar la imputación, este valor puede generar una importante desviación o diversificación en el resultado.

Imputación por la moda: Reemplazar los datos faltantes a través de la imputación por la moda (en el cual se reemplaza el valor faltante de acuerdo a la moda de un grupo apropiadamente definido de valores disponibles). Dado que el promedio es afectado por la presencia de valores fuera de rango, parece natural usar la mediana en vez de la media con el fin de asegurar robustez [Acuña y Rodríguez, 2009].

Imputación Hot Deck (Hot Deck Imputation): En este caso para cada registro que contiene datos perdidos se busca el registro más parecido que no tenga datos perdidos y de esta forma, el dato perdido se imputa con el valor del dato existente en dicho registro [7].

Imputación por Regresión: Utiliza modelos de regresión que a partir de datos de otras variables puede predecir las observaciones faltantes [6].

Imputación Múltiple (Multiple Imputation): Los valores perdidos de cualquier variable son estimados usando los valores existentes en otras variables. Los valores estimados (o imputados) sustituyen a los valores faltantes con lo cual se obtiene un conjunto de datos completo denominado "conjunto de datos imputados". Este proceso es realizado varias veces, produciendo varios conjuntos de datos imputados (de aquí el nombre de Multiple Imputation) [3]. Se realizan análisis estadísticos sobre cada uno de los conjuntos de datos imputados, obteniéndose múltiples resultados. Estos resultados posteriormente son combinados para producir un análisis final.

Reemplazar los datos faltantes con otro valor

Este método tiende a producir serios problemas de inferencia. No se entrará en detalle en este método. Sugerencias para tratar el problema de los datos perdidos [5]:

- No usar imputación por el promedio a menos que el dato sea MCAR.
- Eliminar cuidadosamente verificando antes que es MCAR.
- Imputación simple trabaja bien para datos faltantes MAR, siempre que menos del 10% de ellos sean datos nulos.
- Si se debe usar imputación simple, use EM o Regresión.
- Si las estructuras de varianza en los datos son importantes, no use el método de Eliminación o imputación simple si más del 5% de los datos están perdidos.
- Imputación múltiple opera correctamente para casos sobre el 25% de los datos perdidos.
- Para NMAR, sólo se podría usar imputación múltiple, y con niveles de datos perdidos menores a 25%.
- Siempre que sea factible, usar imputación múltiple debido a sus características.

D.2) Tratamiento de Datos Duplicados

Los duplicados son un problema no menor en los datos de las compañías. Esto se puede deber a errores ortográficos, mal entendimiento de las personas, mal entendimiento de textos que son ingresados a mano, u otros. Un error de este tipo puede ser el ingreso de un usuario con un identificador y un nombre y otro usuario con otro identificador con el mismo nombre (siendo ambos la misma persona) y otro puede ser ingresar nombres, direcciones, documentos, o cualquier campo, con errores (por ejemplo Jorge y Jorje). En esta sección se describen una serie de teorías sobre el tratamiento de duplicados, abordando: detección de duplicados (sección D.2.1) y funciones de similitud (Sección D.2.2).

D.2.1) Detección de Duplicados

El método de detección de duplicados se ejecuta de la siguiente manera [8]:

- Se define un umbral real $\Phi \varepsilon [0,1]$.
- Se compara cada registro de la variable con el resto.
- Si la similitud entre una pareja de registros es mayor o igual que Φ, se asumen duplicados; es decir, se consideran representaciones de una misma entidad real.

D.2.2) Funciones de Similitud

Actualmente existen diversas funciones de similitud, las cuales pueden ser clasificadas en dos categorías: basadas en caracteres y basadas en tokens [9]. Es posible utilizar varios tipos de funciones de similitud:

Funciones de similitud basadas en caracteres: Estas funciones de similitud consideran cada cadena como una secuencia ininterrumpida de caracteres. Hay varios tipos de distancias que aplican a esta teoría.

Distancia de edición: La distancia de edición entre dos cadenas A y B se basa en el conjunto mínimo de operaciones de edición necesarias para transformar A en B (o viceversa) [10]. Esto se puede utilizar cuando existen errores de ortografía dado que la distancia de edición y otras funciones de similitud tienden a fallar identificando cadenas equivalentes que han sido demasiado truncadas.

Distancia de brecha afín: Ofrece una solución al penalizar la inserción/eliminación de k caracteres consecutivos (brecha) con bajo costo, mediante una función afín $\rho(k) = g + h \cdot (k - 1)$, donde g es el costo de iniciar una brecha, k el costo de extenderla un carácter, y h + g [11].

Se suele utilizar cuando hay abreviaciones en las cadenas o cuando hay un gran volumen de datos y además existen prefijos/sufijos sin valor semántico.

Similitud Smith-Waterman: El modelo original de Smith y Waterman [12] define las mismas operaciones de la distancia de edición, y además permite omitir cualquier número de caracteres al principio o final de ambas cadenas. Esto lo hace adecuado para identificar cadenas equivalentes con prefijos/sufijos que, al no tener valor semántico, cuando existe una o más palabras (tokens) que no se encuentran en alguna de las dos cadenas o cuando existen espacios en blanco inútiles.

Similitud de q-grams: Un q-gram, también llamado n-gram, es una subcadena de longitud q [13]. El principio tras esta función de similitud es que, cuando dos cadenas son muy similares, tienen muchos q-grams en común. Se utiliza cuando hay múltiples problemas de similitud o cuando las palabras (tokens) están desordenadas.

Funciones de similitud basadas en tokens: Estas funciones de similitud consideran cada cadena como un conjunto de subcadenas separadas por caracteres especiales, como por ejemplo espacios en blanco, puntos y comas. Esto es, como un conjunto de tokens, y calculan la similitud entre cada pareja de

tokens mediante alguna función de similitud basada en caracteres. Se nombrara la función de coseno TF-IDF que es una función que representa a los tokens como vectores y calcula su distancia mediante los cósenos de sus ángulos [14]. Esta función produce altos valores de similitud para cadenas que comparten muchos tokens poco comunes (con alto poder discriminante). La similitud coseno TF-IDF no es eficiente bajo la presencia de variaciones a nivel de caracteres, como errores ortográficos o variaciones en el orden de los tokens.

D.3) Tratamiento de los Valores Ruidosos

Los valores ruidosos son valores fuera del rango normal de la variable y son detectados mediante funciones especiales. Estos valores por lo general son valores mal ingresados y presentan una distancia con los demás datos del conjunto. Existen una serie de teorías que describen como son usadas para descubrirlos y según sea el caso también existen funciones para tratarlos. Es posible nombrar las funciones a continuación.

Prueba de grubbs: Este método fue planteado por Frank E. Grubbs desde el año 1969 [15] y también es conocido como el método ESD (Extreme Studentized Deviate). La prueba de Grubbs se utiliza para detectar valores atípicos en un conjunto de datos univariantes y se basa en el supuesto de normalidad. Es decir, primero debe verificarse que sus datos pueden aproximarse razonablemente a una distribución normal antes de aplicar la prueba. Es especialmente fácil de seguir y sirve para detectar un valor atípico a la vez [16]. Es muy fácil de usar y funciona bien bajo una variedad de condiciones incluyendo tamaños de muestra muy grandes, recordando que los datos deben provenir de una distribución normal.

Prueba de Dixon: Permite determinar si un valor sospechoso de un conjunto de datos es un outlier. El método define la relación entre la diferencia del mínimo/máximo valor y su vecino más cercano y la diferencia entre el máximo y el mínimo valor aplicado [17]. Los datos deben provenir de una distribución normal. Si se sospecha que una población lognormal subyace en la muestra, la prueba puede ser aplicada al logaritmo de los datos. Antes de realizar el procedimiento es importante definir las hipótesis (si el valor sospechoso se encuentra al inicio o al final del conjunto de datos) y determinar la distribución de la que provienen los datos (normal o lognormal) [18].

Prueba de tukey: El diagrama conocido como diagrama de cajas y bigotes (Box and Whiskers Plot o simplemente BoxPlot) es un gráfico representativo de las distribuciones de un conjunto de datos creado por Tukey en 1977, en cuya construcción se usan cinco medidas descriptivas de los mismos: mediana, primer cuartil (Q1), tercer cuartil (Q3), valor máximo y valor mínimo [19].

Está compuesto por un rectángulo o caja la cual se construye con ayuda del primer y tercer cuartil y representa el 50% de los datos que particularmente están ubicados en la zona central de la distribución, la mediana es la línea que atraviesa la caja, y dos brazos o bigotes son las líneas que se extienden desde la caja hasta los valores más altos y más bajos.

Análisis de valores ruidosos de mahalanobis: El Análisis de Valores ruidosos de Mahalanobis (Mahalanobis Outlier Analysis – MOA), es un método basado en una distancia, llamada distancia de Mahalanobis (DM). Esta distancia es calculada con base en la varianza de cada punto. Ésta describe la distancia entre cada punto de datos y el centro de masa. Cuando un punto se encuentra en el centro de masa, la distancia de Mahalanobis es cero y cuando un punto de datos se encuentra distante del centro de masa, la distancia es mayor a cero. Por lo tanto, los puntos de datos que se encuentran lejos del centro de masa se consideran valores atípicos [20].

Detección de valores ruidosos mediante regresión simple: El análisis de regresión es una importante herramienta estadística que se aplica en la mayoría de las ciencias. De muchas posibles técnicas de regresión, el método de mínimos cuadrados (LS) ha sido generalmente la más adoptada por tradición y facilidad de cálculo. Este método a través de unos cálculos, aproxima un conjunto de datos a un modelo, el cual puede ser lineal, cuadrado, exponencial, entre otros. Es decir, es una técnica de optimización, que intenta encontrar una función que se aproxime lo mejor posible a los datos. La diferencia entre el valor observado y el valor obtenido del modelo de regresión se denominan residuos o suma de cuadrados y el objetivo es tratar de minimizar este valor y así obtener el mejor ajuste [21].

D.4) Normalización

Cuando hablamos de normalizar no estamos hablando de normalización de bases de datos sino de normalización de variables o atributos. Normalizar es transformar una variable aleatoria que tiene alguna distribución en una nueva variable aleatoria con distribución normal o aproximadamente normal.

Existen varias técnicas de normalización [22]: Normalización mínimo – máximo, Normalización a media cero y Normalización de escalado decimal.

Normalización mínimo – máximo: Esta técnica utiliza la fórmula mostrada en la Fórmula 2.

$$v' = \frac{v\ Min_a}{Max_a - Min_a}(RanMax_a - RanMin_a) - RanMin_a \quad (2)$$

Ejecuta una transformación lineal de los datos originales. Con base en los valores mínimo y máximo (Maxa y Mina) de un atributo A y tomando un rango de variación (RanMaxa y RanMina), se calcula un valor de normalización v' con base en el valor v [22].

Normalización a Media Cero: Los valores para un atributo A son normalizados basados en la media y la desviación estándar A (μa y σa). Un valor v de A es normalizado a v' con el cálculo de la Fórmula 3 [22].

Normalización de Escalado Decimal: Normaliza moviendo los puntos decimales de los valores del atributo A. El número de puntos decimales movidos depende del máximo valor absoluto de A, j es el entero más pequeño de Max(|v'|) < 1. Un valor v de A es normalizado a v' con el cálculo de la Fórmula 4 [22].

$$v' = \frac{v - \mu_a}{\sigma_a} \quad (3)$$

$$v' = \frac{v}{10^j} \quad (4)$$

Es de notar, que la normalización puede cambiar los datos originales un poco, especialmente los dos últimos métodos mencionados. También es necesario guardar los parámetros como la media o desviación estándar para uso futuro y que se pueda normalizar de manera uniforme.

D.5) Tratamiento de Series

En el momento de determinar qué hacer con las series de datos es importante entender que las series o sucesiones de

datos son un conjunto de datos que tienen características (patrones) que los relacionan con otros datos formando series o sucesiones. Cada sucesión de datos es un conjunto de registros relacionados en los datos. Estos conjuntos comúnmente están relacionados con una variable tiempo [23].

Un patrón secuencial consiste de una serie de registros que caracterizan a un conjunto. Se puede decir que el problema de encontrar estos patrones es minimizar la intervención del ser humano.

Para la determinación de patrones se puede utilizar un algoritmo propuesto por Quest de IBM, el cual conduce a la solución utilizando una serie de pasos [23].

- Ordenamiento. Convierte la base de datos en sucesiones.
- Ítem. Encontramos el conjunto de todos los ítems L.
- Transformación. Se necesita determinar repetidamente si en un conjunto dado de sucesiones grandes existe una sucesión de clientes. Para hacer esta prueba rápidamente, transformamos cada sucesión de cliente en una representación alternativa. En una sucesión de cliente transformada, cada transacción es reemplazada por el conjunto de todos los ítems contenidos en esa transacción. Si una sucesión de cliente no contiene ningún ítem, esta sucesión es desechada de la base de datos transformada. Sin embargo, todavía contribuye en el conteo total de clientes. una sucesión de cliente se representa ahora por una lista de conjuntos de los ítems.
- Sucesión. Se utiliza el conjunto de ítems para encontrar las sucesiones deseadas.
- Máxima. Encuentra las sucesiones máximas entre el conjunto de sucesiones grandes. En ciertos algoritmos esta fase es combinada con la fase de sucesión para reducir el tiempo al contar las sucesiones no máximas.

D.6) Reducir el Ancho de los datos

En ocasiones puede suceder que ciertas variables o columnas no son necesarias para el propósito del proceso por razones específicas dado que cada uno de sus valores arrojan resultados muy similares. Por esta razón es necesario descartar estas variables con el objetivo de reducir los tiempos de cálculo de los algoritmos ejecutados en la etapa de modelado.

D.7) Reducir la Profundidad de los Datos

Con el fin de reducir la cantidad de registros se pueden definir algunas técnicas estadísticas. En la estadística, la teoría de muestreo, también conocido como estimación estadística, o el método representativo, se ocupa del estudio de los métodos adecuados de selección una muestra representativa de una población, con el fin de estudiar valores estimativos que caractericen a los miembros de una población [24]. Dado que las características estudiadas sólo pueden ser estimadas a partir de la muestra, se calculan intervalos de confianza para dar el rango de valores dentro del cual el valor real caerá, con una probabilidad dada. Hay una cantidad de métodos de muestreo. Algunos métodos parecen ser más adecuado que otros. Algunos métodos de muestreo se describen a continuación.

Muestreo aleatorio simple: Consiste en seleccionar elementos aleatorios, de la población, P, a ser estudiada. El método de selección simple puede ser con reemplazo (SRSWR) o sin remplazo (SRSWOR). Para poblaciones muy grandes, sin embargo, SRSWR y SRSWOR son equivalentes. Para el muestreo simple aleatorio, las probabilidades de inclusión de los elementos pueden o no ser uniforme. Si las probabilidades no son uniformes, se obtiene una muestra aleatoria ponderada [24].

Muestreo aleatorio sistemático: En este caso se elige el primer individuo al azar y el resto viene condicionado por aquél. Este método es muy simple de aplicar en la práctica y tiene la ventaja de que no hace falta disponer de un marco de encuesta elaborado. Puede aplicarse en la mayoría de las situaciones, la única precaución que debe tenerse en cuenta es comprobar que la característica que estudiamos no tenga una periodicidad que coincida con la del muestreo [25].

Muestreo estratificado: Se divide la población en grupos en función de un carácter determinado y después se muestrea cada grupo aleatoriamente, para obtener la parte proporcional de la muestra. Este método se aplica para evitar que por azar algún grupo esté menos representado que el resto [26].

Muestreo aleatorio por conglomerados: Se divide la población en varios grupos de características parecidas entre ellos y luego se analizan completamente algunos de los grupos, descartando los demás. Dentro de cada conglomerado existe una variación importante, pero los distintos conglomerados son parecidos. Requiere una muestra más grande, pero suele simplificar la recogida de muestras. Frecuentemente los conglomerados se aplican a zonas geográficas [26].

E. Inspección de los Datos

En la etapa de inspección de los datos, las teorías involucradas dependen de cada proyecto de explotación de información ya que trata de probar los datos utilizando los algoritmos o procesos de modelado de cada proyecto. Por estas razones no se va a desarrollar una teoría específica para esta etapa del proceso.

III. Descripción del Problema

La presente sección presenta el problema de investigación partiendo de las dificultades que hoy en día poseen las organizaciones al momento de ejecutar proyectos de explotación de información sobre los repositorios de datos que se almacenan desde los sistemas existentes o deprecados.

En primer lugar se el contexto de investigación (sección A), luego se caracteriza el problema abierto (sección B) y se concluye con un sumario de investigación (sección C).

A. Contexto de Investigación

Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes. Pero el éxito de los negocios depende por lo general de la habilidad para ver nuevas tendencias o cambios en los datos.

La aplicación de proyectos de explotación de información sobre los datos puede identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos.

La limpieza de datos dentro de un proyecto de explotación de información es una de las tareas más costosas y se calcula que consume un 60% del total del esfuerzo de ejecución del proyecto [27].

Se puede afirmar que la limpieza de datos, es un trabajo sumamente tedioso y que pocas veces se puede automatizar totalmente, debido al desconocimiento de las combinaciones que se puedan llegar a producir en grandes volúmenes de datos [27].

Con el fin de disminuir el esfuerzo en el trabajo de la limpieza de los datos se presenta como primer problema el de la organización. Al no haber un proceso de limpieza de datos disponible, la tarea de limpieza de datos suele ser desorganizada y además no cuenta con tareas especificas para asegurar la calidad.

Por otra parte, al no contar con un proceso, tampoco existe un circuito de mejora a partir de la gestión del conocimiento sobre la limpieza de los datos. Esto mejoraría sustancialmente el proceso a medida que se ejecutan distintos procesos de limpieza de datos a través del tiempo.

B. Problema Abierto

El problema abierto que se identifica en la presente sección, consiste en que la gran cantidad del esfuerzo que conlleva el proceso de explotación de información, parte de la falta de procesos de limpieza de datos organizados y documentados para lograr el fin, asegurando la calidad de los datos y reduciendo los esfuerzos de la ejecución de esta tarea, larga y tediosa, en todo proyecto de explotación de información.

Por otra parte la falta de documentación de procesos anteriores que ayudarían en gran medida a procesos futuros.

C. Sumario de Investigación

De lo expuesto precedentemente surgen las siguientes preguntas de investigación:

- ¿Se puede plantear un proceso de limpieza de datos divido en actividades que abarque toda las tarea de transformación de datos en proyectos de explotación de información? En caso afirmativo: ¿Cuales son las actividades?
- ¿En caso de existir la posibilidad de dividir en actividades, se puede identificar una serie de técnicas genéricas que se ejecuten en cada una de las actividades? De ser posible: ¿Cuáles son las técnicas y como deben ser ejecutadas dentro de cada actividad?
- ¿En caso de existir estas técnicas, se puede diferenciar una entrada y una salida a cada una de estas de tal forma de diferenciarlas dentro del proceso? De ser posible: ¿Cuáles son las entradas/salidas de cada técnica?
- ¿Se puede documentar cada una de las actividades de tal forma de poder colaborar con la gestión del conocimiento? De ser posible: ¿Como seria el proceso de documentación de cada actividad?

IV. SOLUCIÓN

En esta sección se presenta: Cuestiones generales sobre la solución (sección A), una propuesta de proceso de transformación de datos para proyectos de explotación de información (sección B), la estructura general del proceso (sección C) y sus actividades (sección D).

A. Cuestiones Generales

En función del análisis realizado en la sección III correspondiente a la Descripción del Problema, se considera de interés citar nuevamente el problema abierto que se aborda en este artículo de investigación, recordando que el mismo se focaliza en una de las tareas más costosas y se calcula que consume un 60% del total del esfuerzo de la ejecución de proyectos de explotación de información [27].

Puede afirmarse que esta tarea, es un trabajo sumamente tedioso y que pocas veces se puede automatizar totalmente, debido al desconocimiento de las combinaciones que se puedan llegar a producir en grandes volúmenes de datos [27].

La solución que se propone en este artículo de investigación consiste en el desarrollo de un procesos que ayude a disminuir este esfuerzo con el fin de hacer menos costosos los proyectos de explotación de información y de esta forma poder ejecutarlo con mayor frecuencia dentro de las organizaciones. Por medio de un proceso las tareas de transformación de datos se organizan por tipos y propósitos y se ejecutan una a una hasta lograr resultados de calidad para los pasos posteriores de los proyectos de explotación de información.

B. Propuesta de Proceso de Transformación de Datos Para Proyectos de Explotación de Información

La propuesta de proceso estará formada por una serie de actividades que cumplan funciones específicas y bien divididas.

Cada una de estas actividades tendrá en sí que ejecutar una cierta técnica de tal forma de cumplir con su cometido.

Cada técnica deberá cumplir con una serie de pasos con el fin de transformar los productos de entrada de las actividades en productos de salida.

Cada actividad dependerá de la salida de alguna de las actividades anteriores por lo que su cumplimiento correcto es fundamental.

Cada actividad a su vez aportará de alguna forma a la gestión del conocimiento de la compañía con el fin de disminuir la dificultad de las futuras ejecuciones del proceso.

C. Estructura General del Proceso

Para lograr el objetivo vamos a plantear primeramente un proceso que abarque todo lo necesario. El proceso contará con una serie de actividades fundamentales que son las siguientes:

- Enriquecer los datos.
- Obtener y ejecutar de los casos testigo.
- Determinar y aplicar la estructura de los datos.
- Construir el modelo de entrada de datos.
- Inspeccionar los datos.

Mediante la tabla II se puede observar cómo se distribuyen las técnicas, entradas y salidas de cada actividad del proceso.

Estas actividades están organizadas mediante un circuito organizado de tal forma de que cada una se ejecute en un momento específico dado que estas dependen de la anterior. El circuito propuesto se muestra en la figura 1.

D. Actividades

En esta sección se presenta las actividades del proceso las cuales están dividas en Enriquecer los datos (sección D.1), Obtención y ejecución de los casos testigo (sección D.2), Determinar y aplicar la estructura de los datos (sección D.3), Construir el modelo de entrada de datos (sección D.4) e Inspeccionar los datos (sección D.5).

D.1) Enriquecer los Datos

En esta actividad se reciben los datos que se obtuvieron en momentos previos al proceso actual por ende hay que abstraerse de cómo se obtienen dado que para cada organización esto puede ser más o menos engorroso.

También se obtiene la información sobre el proyecto de explotación de información.

TABLA II. DISTRIBUCIÓN DE TAREAS, ENTRADA Y SALIDA DE CADA ACTIVIDAD.

Actividad	Productos de Entrada		Técnica de transformación	Productos de Salida	
	Entrada	*Representación*		*Salida*	*Representación*
Enriquecer los datos	• Datos posiblemente sucios (DPS) • Información sobre el Proyecto de Explotación de Información (IPEI)	• Archivo plano, SQL, XLS, Access, etc. • Documento formateado	Técnica de Enriquecimiento de los datos (TED)	• Datos Sucios (DS) • Documento de Solución (DO-SO) • Documento de Técnicas de Modelado (DO-TM)	• Archivo plano, SQL, XLS, Access, etc. • Documento formateado • Documento formateado
Obtener y ejecutar de los casos testigo	• `Datos Sucios (DS) • Documento de Solución (DO-SO) • Documento de Técnicas de Modelado (DO-TM)	• Archivo plano, SQL, XLS, Access, etc. • Documento formateado • Documento formateado	Técnica de obtención y ejecución de los casos testigo (TOECT)	• Documento de Listas de Chequeo (DO-LC) • Datos Válidos (DV)	• Documento formateado • Archivo plano, SQL, XLS, Access, etc.
Determinar y aplicar la estructura de los datos	• Datos Válidos (DV) • Documento de Solución (DO-SO) • Documento de Técnicas de Modelado (DO-TM)	• Archivo plano, SQL, XLS, Access, etc. • Documento formateado • Documento formateado	Técnica de Determinación y Aplicación la Estructura de los Datos (TDAED)	• Documento de Integración (DO-IN) • Datos Integrados (DI)	• Archivo SQL. • Documento formateado
Construir el modelo de entrada de datos	• Datos Integrados (DI)	• Archivo SQL.	Técnica de Construir el Modelo de Entrada de Datos (TCMED)	• Documento de Estructuración de los datos (DO-ES) • Datos Estructurados (DE)	• Archivo SQL. • Documento formateado
Inspeccionar los datos	• Datos Estructurados (DE) Documento de Estructuración de los datos (DO-ES) Documento de Integración (DO-IN) Documento de Solución (DO-SO) Documento de Técnicas de Modelado (DO-TM) Documento de Listas de Chequeo (DO-LC)	• Archivo SQL. Documento formateado Documento formateado Documento formateado Documento formateado Documento formateado	Técnica de Inspección de los datos (TIND)	• Datos de calidad (DC)	• Archivo SQL.

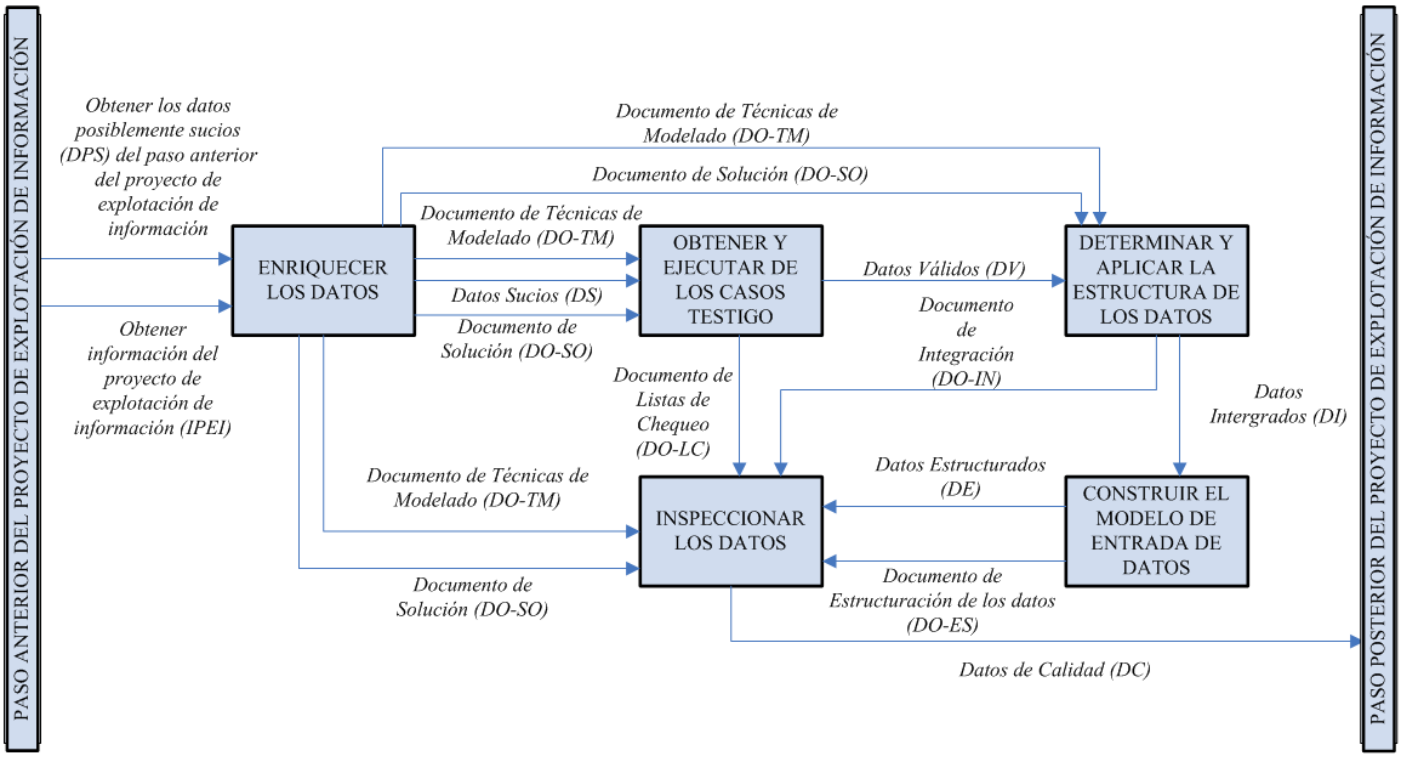

Fig. 1. Diagrama de flujo del proceso completo.

Una vez con los datos en el poder, el primer paso, para la preparación de datos es conocer el problema a resolver, o al menos hacia qué objetivo es necesario llegar. Sin esto resulta imposible afirmar que los datos con los que se cuenta son los correctos para continuar con el proceso. También es necesario conocer la forma en que se debe presentar la información al modelo seleccionado para la explotación de datos.

- ¿Qué solución deben obtener?
- ¿Qué técnica de explotación se utilizarán?

Luego de analizar la información y teniendo respuesta a las preguntas antes generadas, se puede continuar con el proceso actual.

En la tabla III se muestra la serie de pasos que determinan la técnica de enriquecimiento de los datos. Esta técnica se puede describir mediante el gráfico de la figura 2.

TABLA III. TÉCNICA DE ENRIQUECIMIENTO DE LOS DATOS (TED).

Técnica de Enriquecimiento de los datos (TED)	
Entradas:	Datos posiblemente sucios (DPS)
	Información sobre el Proyecto de Explotación de Información (IPEI)
Salidas:	Datos Sucios (DS)
	Documento de Solución (DO-SO)
	Documento de Técnicas de Modelado (DO-TM)
Paso 1.	Conocer el Problema a Resolver
Paso 2.	Analizar Solución a Obtener
Paso 3.	Generación de Documento de Solución
Paso 4.	Analizar Técnicas de Modelado a Utilizar
Paso 5.	Generación de Documento de Técnicas de Modelado

Paso 1: Conocer el Problema a Resolver: Una vez obtenida la información sobre el proyecto de explotación de información se procede a analizar todo lo relacionado con las técnicas de modelado, los datos obtenidos y la forma en que deben presentarse sobre cada una de las técnica de modelado mencionadas.

Paso 2: Analizar Solución a Obtener: En este caso se debe inferir en el resultado que es necesario obtener para poder tener un acercamiento sobre los datos requeridos para el este propósito en especial.

Paso 3: Generación de Documento de Solución: Se genera un documento donde conste todo lo obtenido del paso anterior.

Paso 4: Analizar Técnicas de Modelado a Utilizar: Se analiza las técnicas de modelado a utilizar en los pasos posteriores del proyecto de explotación de información, centrándose en todo lo relacionado con los datos de entrada para cada una, con el fin de detallar cuáles van a ser los requerimientos de los datos de entrada para dichas técnicas.

Paso 5: Generación de Documento de Técnicas de Modelado: En este paso se genera la documentación de toda la información obtenida en el paso anterior.

D.2) Obtención y Ejecución de los Casos Testigo

La obtención de los casos testigo puede convertirse en un proceso muy tedioso dado que esto nos permitirá definir si el modelo al que lo vamos a aplicar es viable o no en relación al conjunto de datos que se obtuvo del paso anterior.

Estos casos son listados de ítems a tener en cuenta de los datos, definen cuales son los atributos a tener en cuenta, formato, tamaño, y demás características.

La ejecución de los casos es simplemente efectuar pruebas sobre los datos y los resultados son los que deciden si estos datos son los correctos para continuar el proceso.

Si se encuentran problemas en los datos se deberá ejecutar nuevamente la actividad de enriquecimiento.

En la tabla IV se muestra la serie de pasos que determinan la técnica de obtención y ejecución de los casos testigo. Esta técnica se puede describir mediante el gráfico de la figura 3.

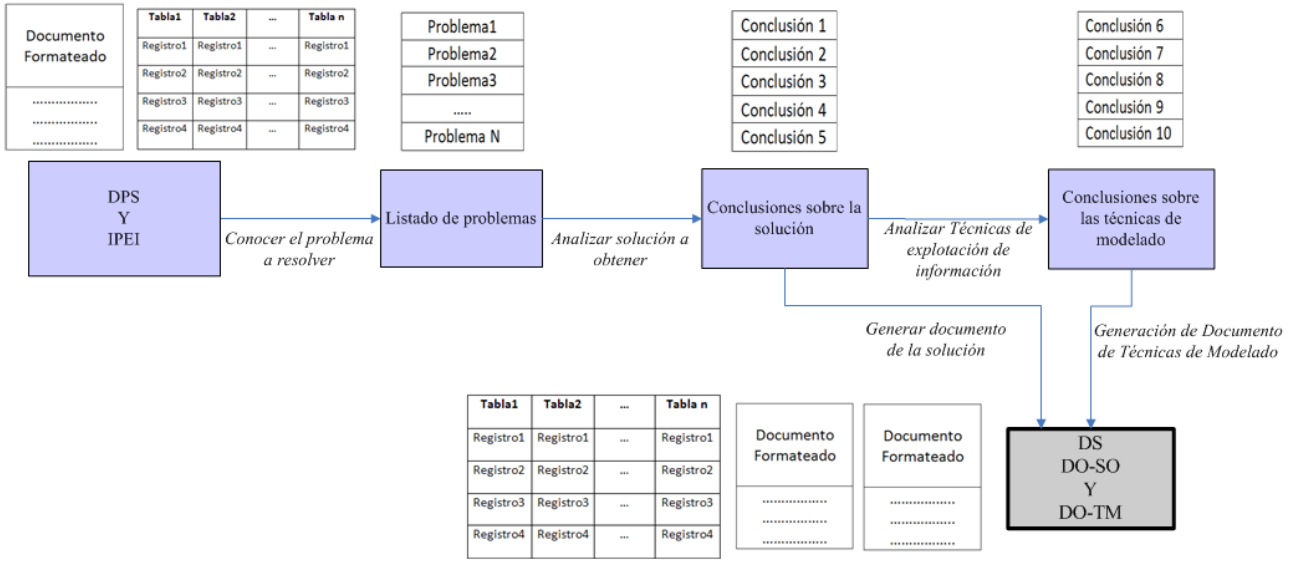

Fig. 2. Diagrama de flujo de la técnica de Enriquecer los datos.

TABLA IV. TÉCNICA DE OBTENCIÓN Y EJECUCIÓN DE LOS CASOS TESTIGO (OECT).

Técnica de Obtención y Ejecución de los Casos Testigo (OECT)	
Entradas:	Datos Sucios (DS)
	Documento de Solución (DO-SO)
	Documento de Técnicas de Modelado (DO-TM)
Salidas:	Documento de Listas de Chequeo (DO-LC)
	Datos Válidos (DV)
Paso 1.	Planteo de los casos testigo.
Paso 2.	Generar Lista de Chequeo.
Paso 3.	Test de los datos.
Paso 4.	Documentar conclusiones.

Paso 1: Planteo de los casos testigo: Partiendo de los documentos DO-SO y DO-TM se genera una serie de ítems que sirven como pautas a tener en cuenta con el fin de determinar la viabilidad del proceso.

Paso 2: Generar Lista de Chequeo: Confeccionar la listas de chequeo partiendo de las pautas obtenidas en el paso anterior

Paso 3: Test de los datos: Se procede a efectuar el test de los datos partiendo de las pautas de las listas de chequeo rellenando los campos de la misma y obtener conclusiones.

Paso 4: Documentar conclusiones: Se pondera los resultados con el fin de hacer una evaluación de viabilidad del proceso con respecto a requerimiento de los datos y a las entradas de las técnicas de modelado. Se da como validados los datos para continuar con el proceso.

D.3) Determinar y Aplicar la Estructura de los Datos

Para poder entender este concepto es necesario definir el término conjunto de datos. El mismo abarca a los datos que serán utilizados por proceso de explotación de información.

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

Esta actividad es necesaria dado que los datos pueden provenir de diferentes fuentes y tener distintos formatos o incluso estar en distintos lugares.

Para esta actividad se pueden utilizar técnicas manuales y asistidas por computadora para lograr una única estructura y a su vez los datos más completos.

En la tabla V se muestra la serie de pasos que determinan la técnica de determinación y aplicación de la estructura de los datos. Esta técnica se puede describir mediante el gráfico de la figura 4.

TABLA V. TÉCNICA DE DETERMINACIÓN Y APLICACIÓN DE LA ESTRUCTURA DE LOS DATOS (TDAED).

Técnica de Determinación y Aplicación de la Estructura de los Datos (TDAED)	
Entradas:	Datos Válidos (DV)
	Documento de Solución (DO-SO)
	Documento de Técnicas de Modelado (DO-TM)
Salidas:	Documento de Integración (DO-IN)
	Datos Integrados (DI)
Paso 1.	Determinar las fuentes de los datos.
Paso 2.	Determinar las relaciones.
Paso 3.	Unificar Tipos de datos.
Paso 4.	Unificar Rangos de variables.
Paso 5.	Generar documento de integración.

Paso 1: Determinar las fuentes de los datos: En este paso se determina de que tipo de fuentes provienen los datos para comenzar con la integración.

Paso 2: Determinar las relaciones: Una vez que se entiende las fuentes se comienza la integración partiendo de las relaciones entre las distintas fuentes de datos para determinar que tablas están relacionadas para concluir con una tabla única.

Paso 3: Unificar Tipos de datos: Si los datos unificados son de distintos tipos hay que unificar criterios y determinar un solo tipo de datos.

Paso 4: Unificar Rangos de variables: Si hay variables discretas se debe unificar los rangos de las mismas hasta obtener un solo rango de variable.

Paso 5: Generar documento de integración: Se reúnen todos los criterios utilizados para la integración y se almacenan en el documento de integración.

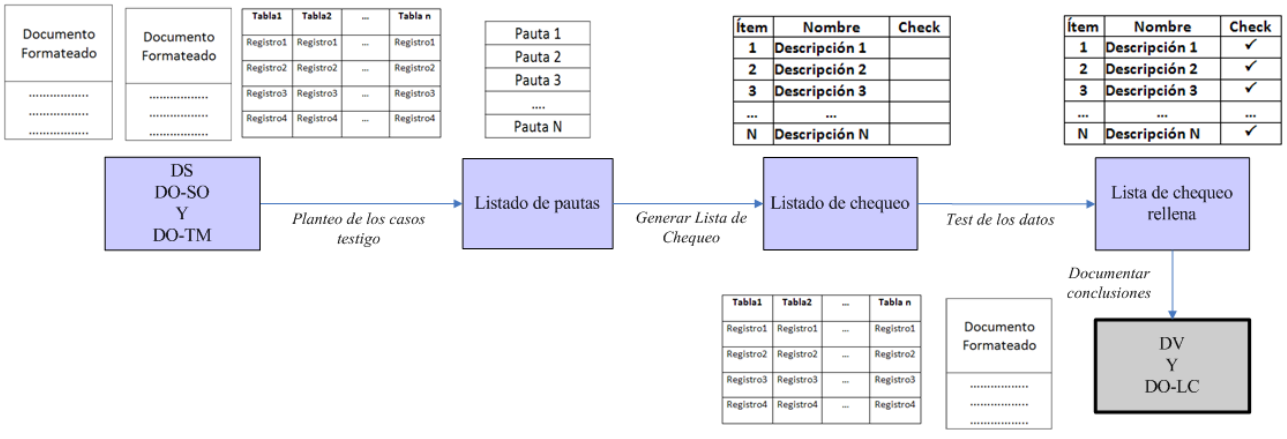

Fig. 3. Diagrama de flujo de la técnica de obtención y ejecución de los casos testigo.

Baldizzoni, E. 2013. *Un Proceso de Transformación de Datos Para Proyectos de Explotación de Información*. Revista Latinoamericana de Ingeniería de Software, 1(2): 57-70, ISSN 2314-2642

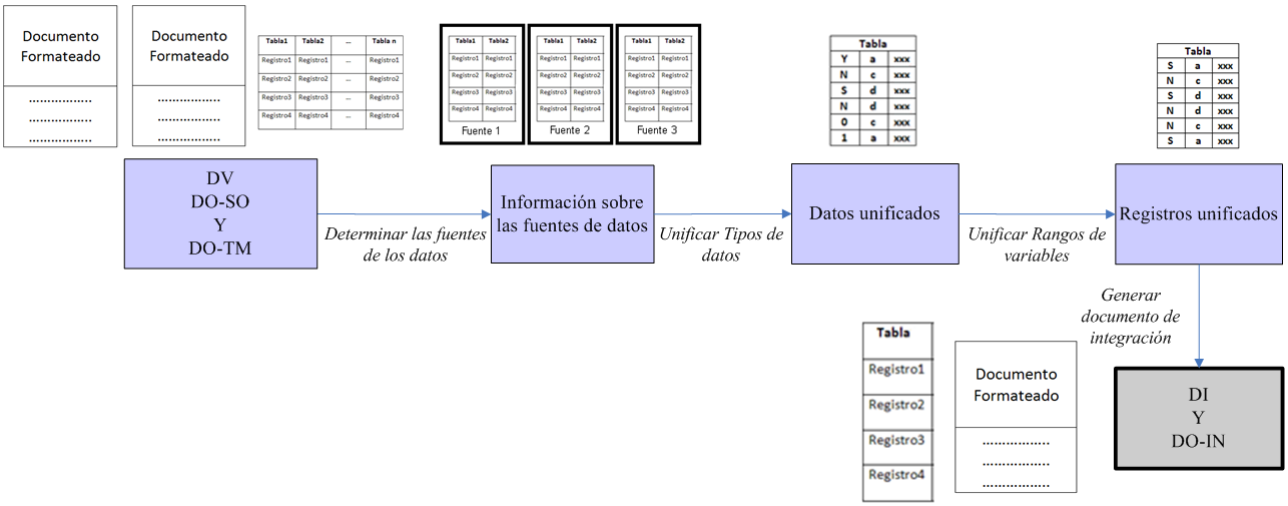

Fig. 4. Diagrama de flujo de la técnica de determinación y aplicación de la estructura de datos.

D.4) Construir el Modelo de Entrada de Datos

Hasta esta actividad el proceso se centra en obtener y conocer los datos disponibles y se han adaptado las diferentes fuentes de datos. Lo que debería suceder en esta actividad es determinarán los procesos que se seguirán para el modelado de los datos, entre los cuales es posible nombrar:

- Tratamiento de los valores nulos o vacíos.
- Eliminación de duplicados.
- Tratamiento de los valores ruidosos,
- Tratamiento de series (las más comunes de tiempo).
- Reducir el ancho de los datos, es decir la cantidad de columnas.
- Reducir la profundidad, la cantidad de registros.

Esta actividad es la que modela los datos de tal forma de darles una calidad suficiente para continuar con la siguiente actividad, que será la de evaluar el resultado. La calidad de los datos es tan importante para el proceso de explotación de información, dado que de no contar con esta el proceso podría tomar caminos diferentes y hasta incluso caminos inexistentes.

En la tabla VI se muestra la serie de pasos que determinan la técnica de construcción del modelo de entrada de datos. Esta técnica se puede describir mediante el gráfico de la figura 5.

Paso 1: Efectuar análisis iniciales: En este paso se obtienen una serie de valores que determinan la calidad de los datos. Cada una de las fases de transformación necesita de estos valores por lo que son de suma importancia para continuar con la técnica. Ejemplos son: Cantidad de datos perdidos por variable, patrón de datos perdidos, valores fuera de rango, cantidad total de registros, cantidad total de variables, tipos de datos de cada variable u otra. También se generan una serie de gráficos que, mediante el análisis, determinan información necesaria para continuar con la técnica.

Paso 2: Ejecutar las distintas fases de transformación: Este paso de dividirá en fases de transformación. Cada fase efectuara uno de los tipos de transformación planteados en esta tesis como ser tratamiento de valores faltantes o nulos, tratamiento de valores fuera de rango, normalización, entre otros.

Paso 3: Generar documento de Estructuración: En este paso se generara un documento donde conste todos los valores obtenidos en el paso 1 y todo lo relacionado con la ejecución del paso 2.

TABLA VI. TÉCNICA DE CONSTRUCCIÓN DEL MODELO DE ENTRADA DE DATOS (TCMED).

Técnica de Construcción del Modelo de Entrada de Datos (TCMED)	
Entradas:	Datos Integrados (DI)
Salidas:	Documento de Estructuración de los datos (DO-ES)
	Datos Estructurados (DE)
Paso 1.	Efectuar análisis iniciales.
Paso 2.	Ejecutar las distintas fases de transformación.
Paso 3.	Generar documento de Estructuración.

D.5) Inspeccionar los Datos

En esta actividad se procede a analizar los datos resultantes de la actividad de construcción del modelo de entrada de los datos, para evaluar si estos datos resultantes son los convenientes para entregar al próximo paso del proyecto de explotación de información, dado que los mismos son los que harán viable el modelo elegido.

En la tabla VII se muestra la serie de pasos que determinan la técnica para inspeccionar los datos.

TABLA VII. TÉCNICA DE INSPECCIÓN DE LOS DATOS (TIND).

Técnica de Inspección de los Datos (TIND)	
Entradas:	Datos Estructurados (DE)
	Documento de Estructuración de los datos (DO-ES)
	Documento de Integración (DO-IN)
	Documento de Solución (DO-SO)
	Documento de Técnicas de Modelado (DO-TM)
	Documento de Listas de Chequeo (DO-LC)
Salidas:	Datos de calidad (DC)
Paso 1.	Efectuar la inspección de los datos.
Paso 2.	Actualizar el repositorio del conocimiento de la compañía.
Paso 3.	Preparar los datos para el siguiente paso del proyecto de EI.

Esta técnica se puede describir mediante el grafico de la figura 6.

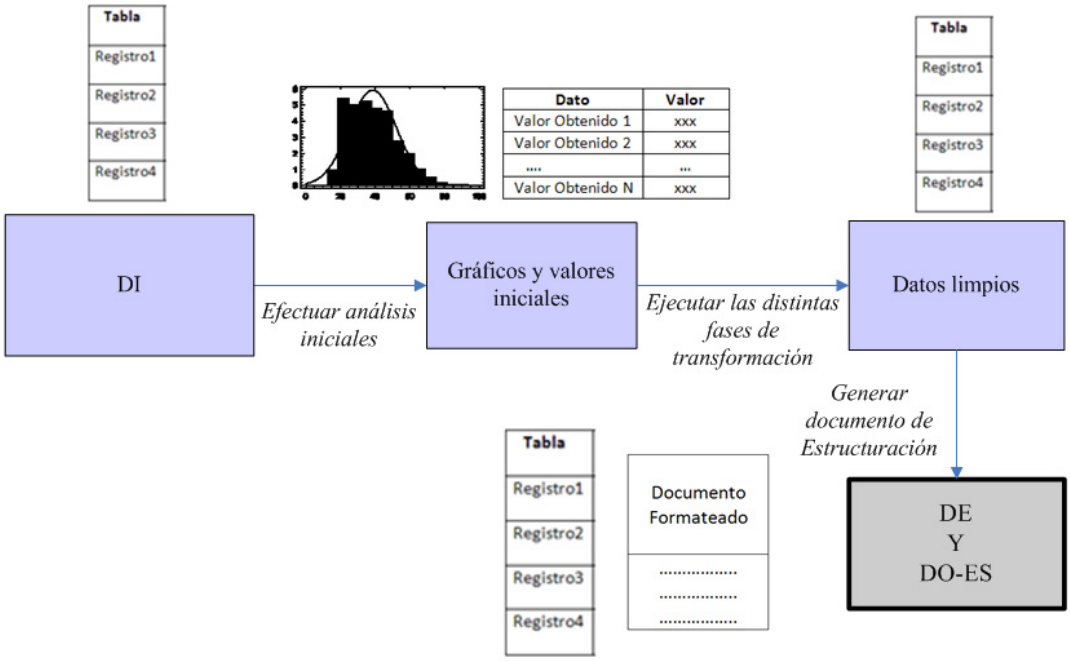

Fig. 5. Diagrama de flujo de la técnica de construcción del modelo de entrada de datos.

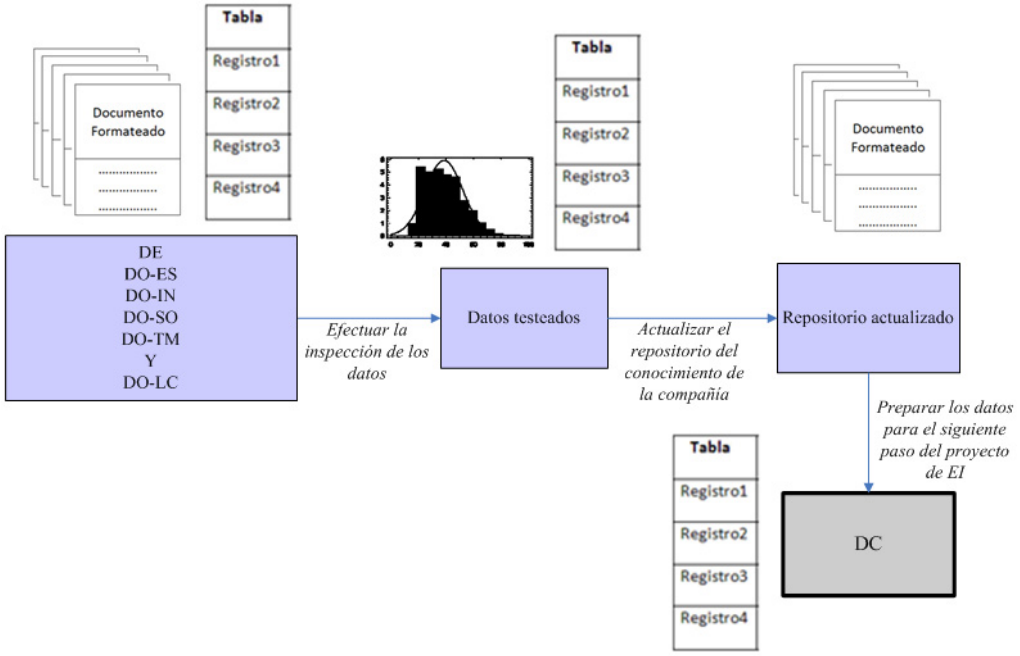

Fig. 6. Diagrama de flujo de la técnica de inspección de los datos

Paso 1: Efectuar la inspección de los datos: Se efectúa una inspección completa de los datos para detectar problemas de calidad. Si es posible se ejecuta los algoritmos de modelado a un muestrario de registros con el fin de detectar problemas en la entrada de dichos algoritmos.

Paso 2: Actualizar el repositorio del conocimiento de la compañía: Se recopila toda la documentación generada en el proceso y se actualiza el repositorio de conocimiento de la compañía con el fin de dar un valor agregado a las próximas ejecuciones.

Paso 3: Preparar los datos para el siguiente paso del proyecto de EI: Se prepara para la entrega de los datos con una calidad óptima al siguiente paso del proyecto de explotación de información.

V. CONCLUSIONES

En esta Sección se presentan los aportes del artículo de investigación (sección A) y se destacan las futuras líneas de investigación que se consideran de interés en base al problema abierto que se presenta en este artículo de investigación (sección B).

A. Aportes del Artículo de Investigación

En este artículo de investigación se ha corroborado que partiendo de una cantidad limitada de datos extraídos de alguna base de datos de interés en condiciones poco optimas para la ejecución de proyectos de explotación de información sobre los mismos, es posible tratarlos de alguna manera en especial con el objetivo de mejorar su calidad a fin de obtener resultados mas certeros posibles.

Inmerso en este contexto y mediante este artículo se propuso:

- Un proceso para la transformación de datos para proyectos de explotación de información dividido en actividades: la primera es la actividad de Enriquecimiento de los Datos, sigue la de Obtener y Ejecutar de los Casos Testigo, continuamos con la de Determinar y Aplicar la Estructura de los Datos, luego Construir el Modelo de Entrada de Datos y por último la Inspeccionar los Datos.
- Dentro de la actividad de enriquecimiento de los datos se propuso, la Técnica de Enriquecimiento de los datos la cual espera como productos de entrada para su ejecución: los Datos posiblemente sucios y también a la Información sobre el Proyecto de Explotación de Información, aporta como productos de salida: los Datos Sucios, el Documento de Solución y el Documento de Técnicas de Modelado.
- Dentro de la actividad de obtener y ejecutar de los casos testigo se propuso la Técnica de obtención y ejecución de los casos testigo la cual espera como productos de entrada para su ejecución: El Documento de Solución y también a el Documento de Técnicas de Modelado y aporta, como producto de salida: El Documento de Listas de Chequeo.
- Dentro de la actividad de determinar y aplicar la estructura de los datos se propuso la Técnica de Determinar y Aplicar la Estructura de los Datos la cual espera como productos de entrada para su ejecución: Los Datos Sucios, el Documento de Solución y también a el Documento de Técnicas de Modelado y aporta, como productos de salida: El Documento de Integración y los Datos Integrados.
- Dentro de la actividad de construir el modelo de entrada de datos se propuso la Técnica de Construir el Modelo de Entrada de Datos la cual espera como producto de entrada para su ejecución: Los Datos Integrados y aporta como productos de salida: El Documento de Estructuración de los datos y los Datos Estructurados.
- Dentro de la actividad de inspeccionar los datos se propuso la Técnica de Inspección de los datos la cual espera como productos de entrada para su ejecución: Los Datos Estructurados, el Documento de Estructuración de los dato, el Documento de Integración, el Documento de Solución, el Documento de Técnicas de Modelado y el Documento de Listas de Chequeo y aporta como producto de salida: Los Datos de calidad.
- La incorporación de la documentación nombrada proporciona información valiosa para el repositorio de conocimiento.

La propuesta de proceso de transformación de datos para proyectos de explotación de información, las actividades y las técnicas asociadas han sido validadas en tres dominios de conocimiento con características bien diferenciadas: El primero trata sobre la necesidad de predicción de clientes de depósitos a largo plazo, el segundo sobre datos de pacientes indios con problemas hepáticos y por último sobre un conjunto de datos para el diagnóstico de cáncer de mama.

B. Futuras Líneas de Investigación

Durante el desarrollo de este artículo de investigación han surgido cuestiones que si bien no son centrales al tema abordado en la misma, constituyen temas concomitantes que (en consideración del investigador) dan lugar a las siguientes líneas de investigación futuras:

En este artículo de investigación se han utilizado teorías de descubrimiento de problemas en los datos que fueron propuestas por otros investigadores las cuales están diversificadas y además podrán surgir en instancias futuras nuevos problemas de calidad en los datos. Además según sea el caso de problema descubierto, se puede utilizar teorías diferentes para solucionar mismos casos de problema. Son de suma importancia la utilización de estas técnicas para proporcionar datos de calidad por lo que surge las siguientes preguntas:

- ¿Cuáles son las teorías necesarias para el descubrimiento de problemas en los datos?
- Según el caso descubierto, ¿Cuál es la teoría que se adapte mejor para la solución del problema en los datos?
- Este proceso propone mejor el rendimiento a la hora de necesitar datos de calidad en proyectos de explotación de información por lo que el aseguramiento de las entradas necesarias y la buena utilización de las salidas depende de los pasos previos y posteriores relacionados al proyecto en si por lo que se pregunta:
- ¿Cuál es la organización necesaria para el proyecto de explotación de información que garantice la buena utilización de este proceso?
- ¿Qué actividades se deben efectuar previamente para garantizar las entradas necesarias para este proceso?
- ¿Qué actividades se deben efectuar posteriormente para garantizar la buena utilización de las salidas de este proceso?
- Si bien el proceso propuesto en este artículo de investigación aporta sistematicidad al proceso de transformación de datos y el mismo ha sido validado en dominios representativos, quedan como temas de trabajo abiertos:
- La validación empírica más amplia del proceso de transformación de datos mediante la técnica de muestras apareadas basadas en grupos experimental y de control.
- La validación empírica de las técnicas propuestas en un conjunto vasto y representativo de dominios de aplicación.

RECONOCIMIENTOS

Este trabajo de investigación ha sido parcialmente financiado por los proyectos 33A105 y 33A167 de la Secretaria de Ciencia y Técnica de la Universidad Nacional de Lanús (UNLa). Además los autores desean agradecer a la Cátedra sobre Tecnologías de Explotación de Información de la Licenciatura en Sistemas de la UNLa que han provisto la información de los proyectos reales utilizados.

REFERENCIAS

[1] Allison, Paul D. Missing Data Techniques for Structural Equation Modeling, CA: Sage Publications. 2001.

[2] Little, R. y Rubin, D. Statistical Analysis with missing data. New York: John Wiley & Sons. 1987.

[3] Farhangfar A., Kurganb L. and Dy J., Impact of imputation of missing values on classification error for discrete data. Pattern Recognition, vol. 41, 2008, 3692 - 3705. 2008.

[4] Jöreskog KG. Structural equation modeling with ordinal variables using LISREL. 2005.

[5] Scheffer Judi, Dealing with Missing Data. Res. Lett. Inf. Math. Sci. 3, 153-160. Disponible desde internet el día 20/02/2013 en http://www.massey.ac.nz/~wwiims/research/letters/. 2002.

[6] Farhangfar, A., Kurgan, L.A. y Pedricks, W. A Novel Framework for Imputation of Missing Values in Databases. 2007.

[7] Nisselson, H., Madow, W., Olkin, I. Incomplete Data in sample Surveys: Treatise. Wonder Book. New York. 1983.

[8] Winkler, W.E. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", Proceedings of the Section on Survey Research Methods, pp. 354-359, 1990.

[9] A.K. Elmagarmid, P.G. Ipeirotis and V.S. Verykios, Duplicate Record Detection: A Survey, IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, 2007.

[10] Levenshtein, V.I. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals", Soviet Physics Doklady, vol. 10, no. 8, pp. 707-710, 1966.

[11] Gotoh, O. An Improved Algorithm for Matching Biological Sequences, Journal of Molecular Biology, vol. 162, no. 3, pp. 705-708, 1982.

[12] Smith, T.F. y Waterman, M.S. "Identification of Common Molecular Subsequences", Journal of Molecular Biology, vol. 147, no. 1, pp. 195-197, 1981.

[13] Yancey, W.E. "Evaluating String Comparator Performance for Record Linkage", Proceedings of the Fifth Australasian Conference on Data mining and Analytics, pp. 23-21, 2006.

[14] Cohen, W.W. Integration of Heterogeneous Databases without Common Domains Using Queries Based on Textual Similarity. En: Proceedings of the SIGMOD International Conference Management of Data SIGMOD (Seattle, Washington, Junio 2-4), 1998.

[15] Grubbs, F. Procedures for Detecting Outlying Observations in Samples, Technometrics, Vol 11, No. 1, pp 1-21. 1969.

[16] Iglewicz, B. y Hoaglin, D. How to detect and handle outliers. American Society for Quality. Statistics Division. 1993.

[17] D. Li y E. Edwards. Automatic Estimation of Dixon's Test for Extreme Values Using a SAS Macro Driven Program. PharmaSug 2001.

[18] Davis, A. y McCuen, R. Storm Water Management for Smart Growth. Springer. 2005.

[19] Tukey, John W . Exploratory Data Analysis. Addison-Wesley. Reading, Mass. : Addison-Wesley Pub. Co. 1977.

[20] Matsumoto, S., Kamei, Y., Monden, A., y Matsumoto, K. Comparison of Outlier Detection Methods in Faultproneness Models. En: Proceedings of the First international Symposium on Empirical Software Engineering and Measurement ESEM 2007 (Madrid, España, Septiembre 20 – 21), 2007.

[21] Rousseeuw, P. y Leroy, A. Robust Regression and Outlier Detection. 3a Ed. New York, John Wiley & Sons, 1996.

[22] Juan A. Botía, Preprocesado de Datos. Departamento de Ingeniería de la Información y las Comunicaciones. Universidad de Murcia. Ingeniería Superior en Informática, UMU. 2010.

[23] Neftalí de Jesús Calderón Méndez. Minería De Datos Una Herramienta Para La Toma De Decisiones. Asesorado por el Ing. Edgar Mauricio Lone Ayala. Guatemala, abril de 2006.

[24] Jerzy Neyman. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society Vol. 97, No. 4, pp. 558-625 Published by: Wiley. Disponible desde internet el día 20/02/2013 en http://www.jstor.org/stable/2342192. 1934.

[25] Jordi Casal, Enric Mateu. Tipos De Muestreo. CReSA. Centre de Recerca en Sanitat Animal / Dep. Sanitat i Anatomia Animals. Universitat Autónoma de Barcelona, 08193-Bellaterra, Barcelona. 2003.

[26] Jordi Casal, Enric. Tipos de muestreo. CReSA. Centre de Recerca en Sanitat Animal / Dep. Sanitat i Anatomia Animals, Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona. 2003.

[27] Merlino, H, Un método de preprocesamiento de datos orientado al uso de explotación de información basado en sistemas inteligentes, Trabajo final especialidad en ingeniería de sistemas expertos, Instituto Tecnológico de Buenos Aires. 2004. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

Ezequiel Baldizzoni. Es Analista Pogramador Universitario y Licenciado en Sistemas por la Universidad Nacional de Lanús. Es Asistente de Docencia en las Asignaturas Ingenieria de Software I y Proyecto de Software. Es Asistente de Investigación en el Grupo de Investigación en Sistemas de Información del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús.