UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

Facultad 1

Departamento de Técnicas de Programación

DBAnalyzer 2.1

La visualización científica en el análisis de los datos

Trabajo final presentado en opción al título de

Máster en Informática Aplicada

Autor: Lic. Daynel Marmol Lacal

Tutores: Dr. Ramiro Pérez Vázquez

Dr. Beatríz López Porrero

La Habana, diciembre de 2012

Agradecimientos

En primer lugar a mis tutores el Dr. Ramiro Pérez y la Dr. Beatríz López por ayudarme durante todo el proceso, por sus consejos, enseñanzas y paciencia conmigo.

A mi hermana Damaris, que sin ella todo hubiese sido más difícil.

A todos los que de una forma u otra colaboraron y ayudaron a que todo llegara a feliz término.

Dedicatoria

A mi madre y su insistencia a diario por que todo estuviera listo

A mi esposa, por estar siempre a mi lado

A mi hermano Darién

A mi mismo

Resumen

El presente trabajo muestra el desarrollo de una herramienta para asistir en el proceso de limpieza de datos mediante la representación gráfica de estos. Implementada en el lenguaje de programación Java y con el uso de bibliotecas libres, luego de definir la técnica de visualización a utilizar y verificar algunas de las herramientas que existen con este fin. DBAnalyzer representa de modo gráfico mediante histograma, diagramas de dispersión y gráfico de barras el análisis realizado a bases de datos que se encuentren sobre los sistemas gestores de base de datos PostgreSQL y MySQL. Agilizando de esta manera la búsqueda de errores e inconsistencias en los datos, garantizando que el proceso de limpieza de datos sea más rápido y eficiente

Abstract

This paper shows a tool's development intended to assist data cleaning process by its graphic representation. It was implemented in Java programming language with free library use, after the visualization technique definition and existing tools review. DBAnalyzer shows as bar charts, histograms and scatter plots the analysis result on data bases stored in PostgreSQL and MySQL. Speeding up by this way the errors search and data inconsistency, making data cleaning process faster and efficient.

Tabla de Contenidos

Introducción 1

Capítulo 1 3

1.1 Dimensiones de la visualización 4

1.2 Visualización Científica 7

1.3 Técnicas de Visualización para datos multiparamétricos 8

1.3.1 Técnicas geométricas 8

1.3.2 Técnicas basadas en iconos 11

1.3.3 Técnicas orientadas a píxel 13

1.4 Elementos a tener en cuenta 15

1.5 Herramientas existentes 15

1.5.1 AVS/Express 16

1.5.2 Amira 16

1.5.3 Khoros 17

1.5.4 Vis5D 17

1.5.5 JFreeChart 18

Capítulo 2 20

2.1 Características de la herramienta DBAnalyzer 20

2.1.1 Modelo de Dominio 20

2.1.2 Requerimientos del sistema 21

2.1.3 Modelado del sistema 24

2.2 Diseño de la herramienta DBAnalyzer 26

2.3 Implementación de la herramienta DBAnalyzer 32

Capítulo 3 34

3.1 Ventajas y desventajas de la variante de implementación 34

3.2 Análisis de los datos 35

3.3 Resultados del trabajo con la herramienta DBAnalyzer 36

3.3.1 Base de datos del portal del graduado 37

3.3.2 Base de datos de portal de noticias 45

Conclusiones 53

Recomendaciones 54

Referencias Bibliográficas 55

Bibliografía 58

Introducción

En la actualidad, debido al creciente volumen de información manejada en todos los sectores de la economía y en el día a día de nuestra sociedad, emergen como herramientas imprescindibles los almacenes de datos, los cuales contribuyen de manera efectiva a la toma de decisiones en empresas y organismo. La información de calidad en los mismos asegura que las decisiones que se tomen sean correctas y efectivas. Uno de los procesos que se ha identificado en la creación de un almacén de datos es la limpieza de datos. Una primera etapa de este proceso es el análisis de los datos, el cual brinda qué es necesario limpiar, dónde pueden existir errores potenciales. En la mayoría de los casos la información que se somete a análisis es voluminosa y los reportes textuales no brindan una solución factible al análisis. Surge, entonces, como alternativa complementaria el análisis gráfico de los datos, que proporciona una visual de determinar posibles errores.

Desde hace un tiempo se ha venido desarrollando un trabajo investigativo sobre la limpieza de datos que llevó a la implementación del “DBAnalyzer”, herramienta que ayuda a la detección de posibles errores en los datos de una base de datos determinada, debido que el análisis de forma manual es muy tedioso y puede ser la causa de la introducción de nuevos errores. Esta herramienta permite el análisis de bases de datos almacenadas en los sistemas gestores de bases de datos libres MySQL y PostgreSQL, está implementado en el lenguaje de programación Java. Presenta una interfaz amigable con el usuario de tal manera que se puede realizar el análisis de la tabla que se decida y de la columna que se desee. Los datos inicialmente los brinda por pantalla, pero pueden ser almacenados para su posterior análisis. Evidentemente la herramienta solo brinda información que debe ser procesada conociendo la semántica de los datos, pero estos elementos pueden ser muy útiles en la determinación de la calidad de los datos.

El software en dependencia del tipo de campo que esté analizando brinda diferentes cálculos estadísticos. Estos datos pueden ayudar a encontrar o determinar hasta cierto punto cuando un elemento determinado tiene un valor incorrecto o fuera de rango, o se encuentra repetido más veces de las que debiera.

A lo anterior se le desea adicionar una representación gráfica de los datos, lo cual ayudaría aún más en la determinación de los posibles errores en ellos. A esta representación se le conoce como visualización de datos, que no es más que la transformación de datos o información en imágenes o pinturas. La visualización emplea el aparato sensitivo primario humano, que es la visión, tanto como todo el poder de procesamiento de la mente humana. El resultado debe ser un medio simple y efectivo para comunicar información voluminosa y compleja, en este caso se puede decir que su propósito sería el discernimiento y no la imagen. Las principales ventajas del discernimiento son el descubrimiento, la elaboración de decisiones y la posibilidad de explicar el comportamiento de los

datos, que para este caso sería la detección de posibles errores.

De lo anterior se deriva el objetivo general de esta investigación: incorporar a la herramienta de análisis de datos métodos de visualización que permitan un estudio más efectivo de los datos. Los objetivos específicos:

- Caracterizar los métodos de visualización más conocidos y utilizados.
- Seleccionar métodos de visualización adecuados para el proceso de análisis de datos.
- Adicionar al DBAnalyzer la funcionalidad de visualización de los datos.

La hipótesis de la investigación sería: “la utilización de métodos de visualización de datos mejora la efectividad en el análisis de los datos”.

El documento está estructurado de la siguiente forma:

Capítulo 1: Visualización de la información. Técnicas y herramientas.

Capítulo 2: Implementación de la propuesta.

Capítulo 3: Validación de la herramienta.

Capítulo 1

La disponibilidad de almacenamiento económico y el progreso tecnológico, han llevado a que se hayan creado inmensas bases de datos de negocios, de datos científicos, entre otros tipos de datos. Ante el crecimiento tan vertiginoso en la cantidad de información de estas bases de datos y aún cuando las personas estén acostumbradas a interrogarlas, se hace prácticamente imposible para una persona la tarea de explorarlas para poder extraer conclusiones, tendencias y patrones. En este caso, sin duda los problemas de la consulta y la posterior exploración de las bases de datos son problemas claves. Con el objetivo de colaborar en la solución de los mismos se han desarrollado distintas herramientas de visualización.

Entre las propuestas iniciales para la visualización se encuentran los métodos interactivos basados en técnicas de búsqueda, filtros y facilidades para la construcción de consultas dinámicas que permitan aprender de los datos mediante múltiples consultas.

La visualización de información o visualización científica permite visualizar espacios de información abstracta, tales como datos financieros, información de negocios, colecciones de documentos y concepciones abstractas que pueden también beneficiarse al ser presentadas en forma visual. El problema fundamental radica en mapear abstracciones no espaciales en formas visuales efectivas, para lo cual es crucial descubrir nuevas metáforas visuales y entender qué tareas de análisis soportan.

Ante grandes volúmenes abstractos de información la meta es lograr la cristalización del conocimiento, es decir, permitir a los usuarios obtener la información que necesitan y hacer que esta tenga sentido para que puedan lograrse las decisiones en un tiempo relativamente corto.

Como ejemplos de los objetivos de la visualización de información se pueden enunciar: mostrar tendencias en los datos, detectar discontinuidades en los mismos, identificar fácilmente máximos y mínimos, establecer límites, identificar agrupamiento en los datos, encontrar estructuras en información heterogénea y ver mucha información en una única pantalla pero al mismo tiempo ver un elemento de interés en este contexto, etc. La visualización de información apoya el proceso de producir modelos que puedan ser detectados y abstraídos; puede reducir la búsqueda de datos al agruparlos convenientemente o al relacionar la información visualmente, permite compactar información en un espacio reducido, permite búsquedas jerárquicas mediante la utilización de vistas generales para ubicar áreas de más detalle bajo demanda. La visualización permite la recuperación de modelos de datos y estos modelos sugieren esquemas a un nivel superior. La agregación de datos se revela a través de agrupamiento (clustering) o propiedades visuales comunes.

En conjunto a la visualización de la información surge otro concepto, el de visualización del conocimiento, que estudia el uso de las representaciones visuales para mejorar la creación y

transferencia del conocimiento entre, al menos, dos personas. Por lo que se puede llamar visualización del conocimiento a los medios gráficos que pueden ser utilizados para construir o llevar a un complejo entendimiento. Más allá que un mero traspaso de hechos la visualización del conocimiento aspira a transferir entendimiento, experiencias, actitudes, valores, expectativas, perspectivas, opiniones y predicciones, todo lo anterior de una forma tal que permita a las personas reconstruir, recordar y aplicar las comprensiones de forma correcta.

La visualización de la información es el campo precursor de la visualización del conocimiento, y tanto una como la otra explotan las habilidades del hombre para procesar de manera efectiva representaciones visuales, pero la forma de utilizar esas habilidades difiere en ambos dominios: la visualización de la información intenta explorar grandes volúmenes de datos abstractos (a menudo numéricos) derivando en nuevas comprensiones de estos o sencillamente haciéndolos más accesibles; por otra parte la visualización del conocimiento trata de mejorar la creación y transferencia de conocimiento entre personas, dándoles mejores medios de expresar lo que conocen. Mientras la visualización de la información típicamente facilita la recuperación, el acceso y representación de grandes colecciones de datos (particularmente en la interacción de humanos y computadoras) la visualización del conocimiento tiene como objetivo fundamental aumentar la comunicación de conocimiento intensivo entre los individuos, ejemplo de ello se tiene en la relación de nuevas comprensiones de conceptos ya interiorizados como en el caso de las metáforas visuales.

1.1 Dimensiones de la visualización

¿Cómo escoger el formato de visualización adecuado? Existen muchas posibilidades diferentes de visualizaciones, cada una con sus ventajas y desventajas en dependencia del objetivo a lograr, ninguna es perfecta o ideal para todas las situaciones.

A continuación se presentan una serie de factores a tener en cuenta a la hora de escoger, estos factores se han organizado en forma de dimensiones independientes y abarcadoras, basadas en años de investigación psicológica y organizacional por parte de un equipo especializado [1]:

1. Impacto visual (¿cuán atractiva es la visualización?)
   Esta dimensión está relacionada con las características gráficas de la visualización. El impacto visual es alto si la visualización es atractiva, memorable, emocionalmente evocativa o asemeja un trabajo artístico; es bajo si es muy simple o de baja calidad. Un alto impacto puede ser contraproducente porque puede desviar la atención. Una visualización placentera es aquella que atrae la atención de quien la ve y crea un efecto positivo en los otros aspectos de la visualización, inspira creatividad y da emociones, además de aumentar su valor como elemento recordable.

Escala: 1-genérico/baja calidad 2-básico 3-promedio 4-distintivo 5-asemeja un trabajo artístico

2. Claridad (¿es la visualización fácilmente entendible con un bajo esfuerzo cognoscitivo?)
   Es elevada si el significado de la visualización puede determinarse a primera vista, es bajo si se requiere tiempo y concentración para entenderla. Una baja claridad puede ser necesaria cuando se representen conceptos complejos, si una alta claridad en esos casos puede llevar a una sobre simplificación.
   Está relacionada con la familiaridad de la audiencia con las convenciones, la complejidad y completitud, y la consistencia de los elementos. La dificultad en entender no necesariamente es un elemento negativo, en comparación con la sobre simplificación que puede llevar a un entendimiento erróneo y por tanto a resultados incorrectos.
   Existe una relación entre la claridad y el impacto visual: una visualización diseñada para ser clara no debe incluir decoración excesiva o elementos artísticos.
   Escala: 1-confuso 2-difícil comprensión 3-poco claro 4-bastante claro 5-claro a primera vista

3. Finitud percibida (¿la visualización invita a contribuciones o modificaciones, o muestra un producto terminado, pulido?)
   Esta dimensión es elevada si la visualización representa una ilustración terminada, cuando es baja muestra un trabajo en progreso. La percepción de finitud debe ser consistente con la actual posibilidad de modificación.
   Es influenciada por el medio, el nivel de cambios posibles, y el impacto visual.
   Escala: 1-percibido como “en progreso” 2- percibido como “incompleto” 3-percibido como “modificable” 4- percibido como “terminado” 5- percibido como “pulido”

4. Dirección centrada (¿la visualización dirige la atención a los elementos principales de la discusión?)
   Es elevada cuando la atención de los participantes está dirigida solamente al principal elemento de la discusión. Cuando no hay un centro, o este está dispersado entre varios elementos es baja, lo cual es útil cuando son requeridos pensamientos divergentes. Un gráfico ingenioso que centra en uno o pocos elementos, puede ayudar al grupo a mantener el hilo de la discusión. La proyección se debe tener en cuenta, pues puede desviar la atención del contenido a la forma gráfica. El centro puede ser dispersado cuando se buscan pensamientos divergentes o diferentes alternativas de análisis.
   Un moderado nivel de atención aumenta la claridad, pero en muchos elementos la disminuye causando una competencia por el centro de atención.
   Escala: 1-ninguno 2-en muchos elementos (disperso) 3-en varios elementos 4-en pocos

elementos 5-en un solo elemento principal

5. Fácil entendimiento (¿se generan nuevas formas de entendimiento como resultado de la forma de visualización?)

   Esta dimensión está relacionada al potencial de una visualización para descubrir nuevos patrones o relaciones. Describe la capacidad de la visualización para ayudar a pensar y descubrir procesos. Es el núcleo diferenciador y valor agregado de una visualización sobre el texto. Si permite obtener un nuevo entendimiento solamente al cambiar el tipo de visualización, el centro de atención o las restricciones de representación, tiene un nivel alto, si la visualización no ayuda en obtener un nuevo entendimiento, entonces tiene un nivel bajo.

   Al buscar entendimiento, a menudo se llega a reducir la claridad, debido a formas no convencionales o familiares de visualizar la información. Las formas de visualización que conllevan a generación de inferencia no se desempeñan bien en términos de impacto visual.

   Escalas: 1-ningún entendimiento 2-algún entendimiento 3-varios entendimientos 4-muchos entendimientos evidentes 5- entendimiento sin barreras

6. Modificable (¿los elementos de la visualización pueden ser modificados en respuesta a la dinámica de la discusión?)

   Esta dimensión es alta si cada elemento de la imagen puede ser cambiado fácilmente (movido, borrado o sustituido). Si los elementos son difíciles de cambiar o las modificaciones no son posibles, entonces es baja la dimensión, lo que posibilita la búsqueda de segundas opiniones antes de hacer alguna modificación, mientras que si es alta aumenta la posibilidad de interacción.

   Esta dimensión afecta a la de “apoyo a la interacción de grupo” y a “finitud percibida”, un ambiente altamente modificable impulsa a un mayor nivel de contribución debido a que los cambios pueden ser hechos de forma fácil y hace que las personas tomen riesgos y contribuyan aún más.

   Escalas: 1-imposible 2-difícil 3-posible 4-fácil 5-fácil y rápido

7. Apoyo a la interacción de grupo (¿la visualización ayuda en facilitar o estructurar la interacción de un grupo de personas?)

   Esta dimensión describe la capacidad de mantener la interacción de grupo y discusión en el camino correcto. Si la visualización permite indicadores como la contribución de un grupo de seguimiento (tracking participants’ contribution), repetición y patrones de modificación tanto simultáneos como secuenciales, entonces el apoyo a la interacción de grupo es alto, si no hay indicadores es bajo. Otro de los elementos que provee apoyo a la interacción de grupo es la

referenciabilidad de los elementos en el diagrama (pointing or recalling) que facilita la referencia de los participantes a los elementos de la visualización y documentación que permite regresar a cierto punto en el tiempo (historia) o repetir la interacción completa.

Escalas: 1-ninguna 2- muy limitada 3-alguna 4-extensiva 5-muy extensiva

Por otra parte, la visualización científica (VC) según [2] significa encontrar una representación visual apropiada para un conjunto de datos, que permita mayor efectividad en el análisis y evaluación de los mismos. Simplifica el análisis, comprensión y la comunicación de modelos, conceptos y datos en la ciencia y la ingeniería. La VC ofrece grandes ventajas sobre otros métodos de análisis de datos, permitiendo representar datos de varias dimensiones o variables, logrando visualizar cuatro o más variables al mismo tiempo. A este tipo de visualización se le conoce como visualización de datos multiparamétricos.

La visualización científica [3] transforma los datos numéricos o simbólicos y la información en imágenes geométricas generadas por una computadora. Es una metodología para interpretar a través de una imagen en la computadora tanto datos de mediciones como los generados por modelos computacionales. La investigación y el desarrollo de la VC se han centrado en cuestiones relacionadas con el renderizado de gráficos en tres dimensiones, animaciones de series temporales y visualización interactiva en tiempo real.

1.2 Visualización Científica

La ciencia ha desarrollado diversos métodos para la obtención de información, y uno de ellos se basa en la creación de imágenes a partir de los datos. Este método, conocido como visualización, ha sido utilizado como vía natural para mostrar información [4]. Recientes investigaciones han impulsado en gran medida este campo mediante el uso de la computación.

La visualización de datos en general logra diferentes metas. La naturaleza del objetivo que se obtiene está en relación directa al conocimiento que se tenga sobre los datos iniciales. Los objetivos pueden ser los siguientes [5]:

- Análisis exploratorio.
- Análisis confirmativo.
- Presentación de información.

El análisis exploratorio en la visualización parte de un conjunto de datos sobre el que no se tiene hipótesis y a partir de un proceso interactivo usando una búsqueda no dirigida se llega a la obtención de una imagen que aporta una hipótesis sobre los datos.

El análisis confirmativo comienza con datos sobre los que se tiene a priori una hipótesis. El proceso consiste en la búsqueda de la confirmación de la hipótesis. La presentación de información parte de

hechos que son fijos a priori y que se desean enfatizar y mostrar con extrema calidad.

Son muchas las técnicas de visualización científica existentes. Diversos enfoques se han empleado para agruparlas y clasificarlas. Un enfoque establecido para clasificar las técnicas es a través del tipo de dato sobre los que operan. Por el tipo de datos se entiende el tipo al que pertenecen los atributos o variables. Según este criterio se pueden encontrar las siguientes categorías [4, 5]:

- Técnicas de visualización para datos volumétricos.
- Técnicas de visualización para fluidos.
- Técnicas de visualización para datos multiparamétricos.
- Técnicas de visualización de la información.

Existen diversos enfoques para especificar los datos [6]. Estos enfoques permiten definir una amplia variedad de característica de los datos como son: la dimensionalidad, el nivel de medición y la estructura. En el caso de las dimensiones se tienen los datos multiparamétricos que son aquellos en que el número de variables relacionadas con cada observación es mayor o igual que dos. Estas variables pueden ser cuantitativas o cualitativas y a su vez ordinales o nominales [4-6]. Este trabajo se concentra en las técnicas de visualización para este tipo de datos, debido a la misma estructura de los elementos a representar y el objetivo a lograr con esta representación.

1.3 Técnicas de Visualización para datos multiparamétricos

Existe una multitud de problemas en que cada punto de dato contiene más de un atributo. Estos atributos pueden ser fechas, lugares, precios o valores descriptivos, y pueden tener o no una referencia espacial. A este tipo de datos se les llama multiparamétricos y se encuentran generalmente en aplicaciones de minería de datos, estadística e inteligencia artificial [7].

El objetivo fundamental de los métodos de visualización para datos multiparamétricos es lograr que las representaciones revelen correlaciones o patrones entre los atributos [5, 7, 8]. Con este fin existe actualmente una amplia gama de técnicas de visualización, para las cuales se han creado además diversas mejoras. Las técnicas pueden ser clasificadas en geométricas, basadas en iconos, basadas en píxel y proyecciones [5, 7, 9].

1.3.1 Técnicas geométricas

Las técnicas geométricas son aquellas que utilizan elementos como líneas, puntos o curvas como propiedades visuales para representar los datos. Existe gran número de ellas, como *prosection views* [10-12], *hyper slices* [13], parahistogramas [14] y coordenadas en forma de estrellas, pero hay tres que sobresalen por su generalidad y amplio uso. Estas son los diagramas de dispersión [15], coordenadas paralelas [16] y los gráficos de Andrews [17, 18].

Diagramas de dispersión (en inglés *ScatterPlot*)

El diagrama de dispersión es una técnica simple muy utilizada. Su forma más sencilla se manifiesta cuando los datos tienen solo dos dimensiones. Con dos dimensiones la técnica consiste en trazar un eje de coordenadas y utilizar los valores de las dimensiones como puntos (x,y) de R, resultando un gráfico donde se observan dispersos los puntos de datos.

Por otro lado, visualizar datos de más de dos dimensiones no es obvio, para lograrlo pueden utilizarse proyecciones, que provocan pérdida de información debido a la reducción de la dimensionalidad [4, 5, 7].

Para datos multiparamétricos es muy frecuente utilizar matrices de diagramas de dispersión (ver Figura 1.1). Las matrices resultantes son cuadradas y el elemento (i,j) de la matriz es un diagrama de dispersión de la dimensión i y la j. El diseño evita la pérdida de información, pero en cambio los análisis complejos son engorrosos. Una deficiencia adicional es que la diagonal principal de la matriz es subutilizada. Algunos trabajos actuales están encaminados a aprovechar mejor esta región de la representación [19].

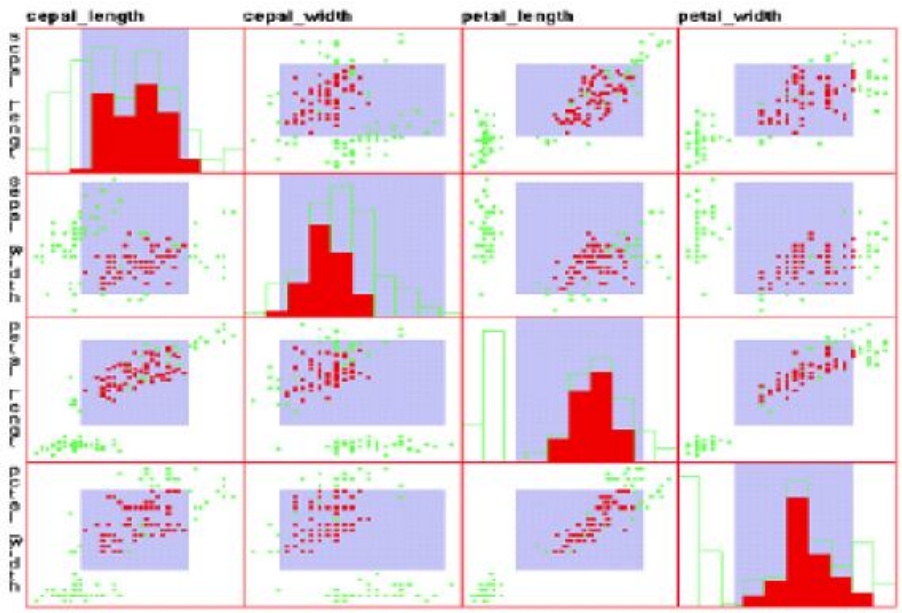


Figura 1.1: Matriz de diagramas de dispersión con la diagonal principal con histogramas.

Coordenadas Paralelas

La técnica de las coordenadas paralelas [16] es un esquema simple de gran generalidad que permite visualizar conjuntos de datos multidimensionales. Esta técnica geométrica es una de las más utilizadas debido a su fácil implementación y los buenos resultados que se obtienen al aplicarla [7]. (Obsérvese Figura 1.2).

Esta técnica usa un sistema de coordenadas como base y consiste en crear un eje de

coordenadas para cada atributo colocándolos paralelamente. El valor de cada dimensión en un determinado punto de datos es marcado en el eje correspondiente. La representación final para un objeto es una línea que recorre las posiciones marcadas en cada dimensión [5, 7].

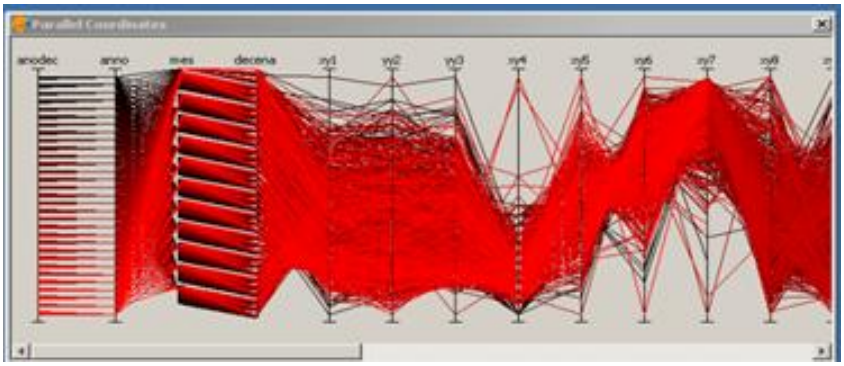


Figura 1.2: Coordenadas paralelas.

El color de las líneas que representan los objetos puede ser elegido por varios criterios. El más simple es utilizar un color constante para todos los objetos. Un criterio que maximiza la calidad de la imagen es seleccionar una dimensión para que sea el color del objeto, de tal forma que puntos con diferentes valores en el atributo de color serán mostrados con diferentes tonos y los similares serán mostrados con tonos equivalentes [5, 7].

Gráfico de Andrews

Una idea similar para representar datos multiparamétricos es el gráfico de Andrews (Obsérvese la Figura 1.3). En esta técnica cada observación es representada por una función f(t) que se evalúa en el intervalo **[0,1]**. Cada función es una serie de Fourier, cuyos coeficientes se igualan a los valores de las dimensiones para cada observación [17, 18].

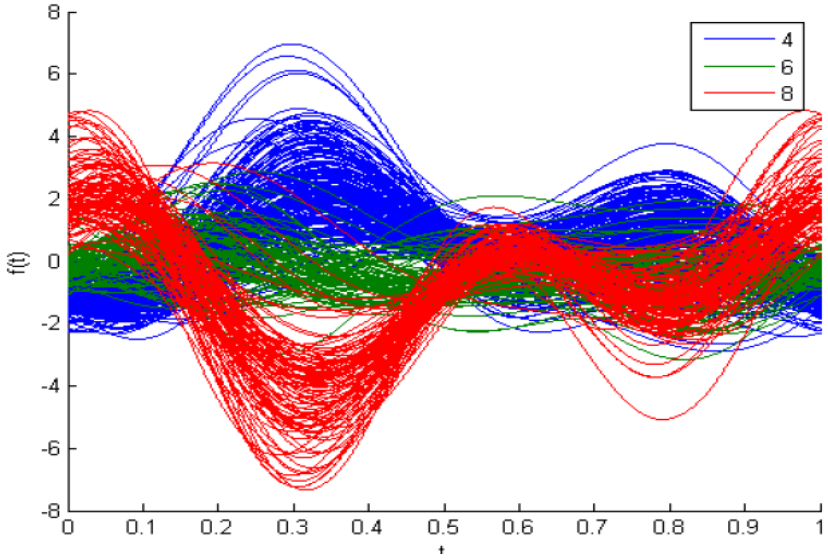


Figura 1.3: Gráfico de Andrews

Esta técnica permite identificar con facilidad diferencias entre grupos de observación, ya que por lo general observaciones pertenecientes a un mismo grupo presentan una forma de la función similar. Los análisis sobre variables individuales resultan en cambio mucho más engorrosos [17]. La virtud fundamental de la técnica es que puede representar conjuntos de datos de un tamaño relativamente grande y además con un número de dimensiones elevado.

1.3.2 Técnicas basadas en iconos

Las técnicas basadas en iconos tienen dos parámetros que la caracterizan: el primero es el tipo de figura que representará cada observación, o sea, la forma del icono; el segundo parámetro es la forma en que se definirá la posición de cada icono en la imagen [5, 20].

Estas técnicas no sufren de pérdida de información. Se logra evitar la pérdida de información al realizar una proyección de las dimensiones a diferentes atributos de un icono [5, 9].

Entre los métodos para crear iconos están los rostros de Chernoff (*Chernoff Face*) y los campos de estrellas (*StarField*). Además, suelen crearse editores de iconos para aplicaciones específicas [5, 9, 21]. Por otro lado, la solución más popular para la colocación de los iconos en la imagen está basada en el uso de proyecciones [20].

Campo de estrellas

El campo de estrellas utiliza un algoritmo para componer los iconos, lo que le confiere cierta generalidad. En la forma básica el método utiliza dos dimensiones como coordenadas de posición en un eje imaginario [5, 21]. El resto de las dimensiones son normalizadas al intervalo **[0,1]**. Estas coordenadas son el punto de inicio en el dibujado del icono. Las dimensiones restantes se expresan a partir de líneas que parten del punto inicial y cuya longitud está determinada por el valor del atributo. Estas líneas o rayos que representan las diferentes dimensiones están dispuestos entre sí con igual distancia angular, lo que genera una figura de estrella. Frecuentemente los extremos de las líneas son conectados entre sí. Esta variación elimina la silueta de estrella y crea una figura cerrada que suele presentar más claramente las características del objeto [8]. En la Figura 1.4 puede observarse un ejemplo.

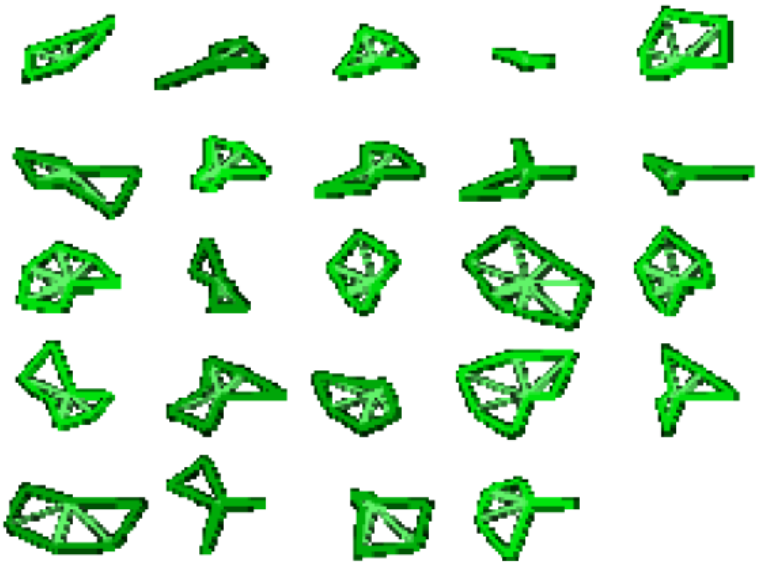


Figura 1.4: Campo de estrellas

Todos los puntos del conjunto de datos pueden mostrarse con el mismo color, pero resulta muy conveniente utilizar esta característica para codificar algún atributo de interés. Igualmente pueden usarse otros rasgos de la figura para codificar otras informaciones, como por ejemplo, la calidad de los datos [22].

Una cuestión de particular importancia en la técnica es la estrategia de posicionamiento del icono. En el proceso de posicionamiento de los iconos pueden usarse los datos de ciertas dimensiones, que en el caso más simple utiliza dos dimensiones y en caso de un número mayor de dimensiones requiere el uso de proyecciones [20].

Se ha mencionado que la visualización de un conjunto de datos de gran tamaño resulta un reto para técnicas geométricas y basadas en iconos. Al graficarlos suele surgir cierto desorden, que está originado por el tamaño de la figura que representa una observación simple. A partir de esta idea resulta lógico concluir, que minimizando el espacio que ocupa un solo punto de datos en la imagen se mejoraría la percepción visual [4, 21, 23].

Las técnicas basadas en píxel son las más eficientes cuando el número de dimensiones es grande y cuando crece el número de registros. Esto se debe a que utilizan un píxel para representar cada atributo de una observación. Los retos fundamentales en estos métodos son la elección del color para cada elemento y el modo de posicionamiento de los píxeles [7, 24].

En este esquema el asunto principal es como colocar los píxeles en la imagen. Este tipo de técnicas utilizan diferentes modos de posicionamiento para lograr diferentes objetivos. Colocar los píxeles en la forma apropiada ofrece la posibilidad de observar información sobre correlaciones, dependencias y regiones trascendentales. Dos de los modos de posicionamiento

de los píxeles son: los Patrones Recursivos y los Segmentos Circulares [4].

La técnica de los patrones recursivos se basa en un posicionamiento recursivo general de atrás hacia delante de los píxeles. Está particularmente dirigida a representar conjunto de datos con un orden natural de acuerdo a un atributo, propiedad que la convierte en una opción para problemas de series de tiempo [5, 7, 24].

Los segmentos de círculo utilizan como imagen base un círculo que es dividido en segmentos iguales a partir del origen. Cada segmento corresponde a un atributo del conjunto de datos. Dentro de cada segmento el valor del atributo para cada registro de datos se representa con un píxel simple. La colocación de los píxeles comienza en el centro de la circunferencia y continúa hacia fuera dibujando sobre una línea ortogonal al segmento [4, 25].

Debe observarse que los únicos atributos visuales de estas técnicas son la localización de los píxeles y la intensidad del color. De ello resulta que la elección del color sea un proceso de vital importancia, que requiere escalas de colores elegidas cuidadosamente para cada uno de los atributos [7, 24, 25].

1.3.3 Técnicas orientadas a píxel

Se ha mencionado que la visualización de un conjunto de datos de gran tamaño resulta un reto para técnicas geométricas y basadas en iconos. Al graficarlos suele surgir cierto desorden, que está originado por el tamaño de la figura que representa una observación simple. A partir de esta idea resulta lógico concluir, que minimizando el espacio que ocupa un solo punto de datos en la imagen se mejoraría la percepción visual [4, 21, 23].

Las técnicas basadas en píxel son las más eficientes cuando el número de dimensiones es grande y cuando crece el número de registros. Esto se debe a que utilizan un píxel para representar cada atributo de una observación. Los retos fundamentales en estos métodos son la elección del color para cada elemento y el modo de posicionamiento de los píxeles [7, 24].

En este esquema el asunto principal es como colocar los píxeles en la imagen. Este tipo de técnicas utilizan diferentes modos de posicionamiento para lograr diferentes objetivos. Colocar los píxeles en la forma apropiada ofrece la posibilidad de observar información sobre correlaciones, dependencias y regiones trascendentales. Dos de los modos de posicionamiento de los píxeles son: los Patrones Recursivos y los Segmentos Circulares [4].

La técnica de los patrones recursivos se basa en un posicionamiento recursivo general de atrás hacia delante de los píxeles (Obsérvese Figura 1.5). Está particularmente dirigida a representar conjunto de datos con un orden natural de acuerdo a un atributo, propiedad que la convierte en una opción para problemas de series de tiempo [5, 7, 24].

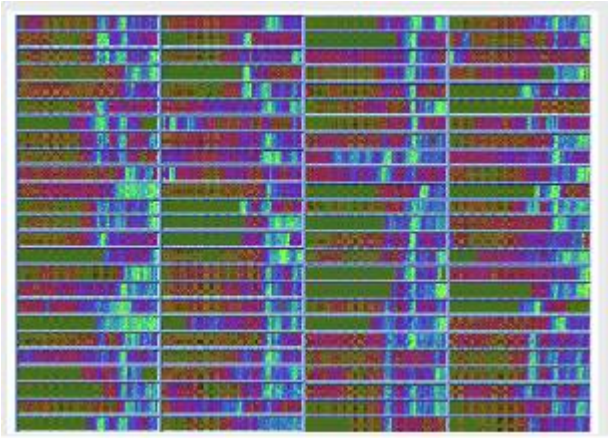


Figura 1.5: Patrones recursivos

Los segmentos de círculo utilizan como imagen base un círculo que es dividido en segmentos iguales a partir del origen. Cada segmento corresponde a un atributo del conjunto de datos. Dentro de cada segmento el valor del atributo para cada registro de datos se representa con un píxel simple. La colocación de los píxeles comienza en el centro de la circunferencia y continúa hacia fuera dibujando sobre una línea ortogonal al segmento [4, 25] (Obsérvese Figura 1.6).

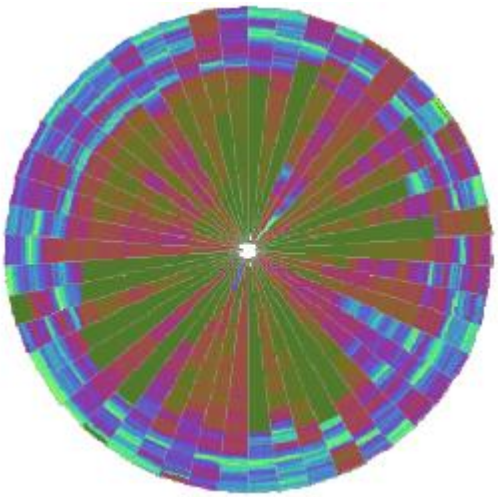


Figura 1.6: Segmentos de círculo

Debe observarse que los únicos atributos visuales de estas técnicas son la localización de los píxeles y la intensidad del color. De ello resulta que la elección del color sea un proceso de vital importancia, que requiere escalas de colores elegidas cuidadosamente para cada uno de los atributos [7, 24, 25].

Después de caracterizar las diferentes técnicas de visualización asociadas a datos multiparamétricos, y partiendo de las características de los elementos a analizar y de los resultados finales que se desean mostrar, la técnica que mejor se ajusta a la problemática a resolver es la geométrica, particularmente los diagramas de dispersión, aunque es necesario

acotar que no se utilizará esa técnica solamente, pues se añadirán elementos de otras técnicas que permitirán una mejor claridad en la imagen resultante. Por lo que además de los diagramas de dispersión se utilizarán gráficos de barras e histogramas, que a pesar de estar entre las representaciones gráficas más simples son de las más eficaces para mostrar con claridad los elementos con dos dimensiones (valor y cantidad de repeticiones).

1.4 Elementos a tener en cuenta

Una de las razones en las que radica la potencia de la visualización está en el hecho de que hay una serie de operaciones de identificación y reconocimiento que el cerebro realiza de forma "automática" sin necesidad de centrar nuestra atención, por lo que manejar los objetos cuyo procesado es "pre-atentivo" puede marcar la diferencia en una interfaz de usuario, sin dejar de lado lo que se conoce como "distractores" [26].

Según [27], la lista de características que se procesan de forma pre-atentiva se pueden agrupar en cuatro categorías básicas:

- **Color**: tanto la diferencia de tonalidad como en intensidad.
- **Forma**: diferente orientación, longitud, ancho, agrupación espacial, etc.
- **Movimiento**: un objeto que se mueve dentro de un fondo o que parpadea.
- **Localización espacial**: localización 2D, profundidad estereoscópica o concavidad/convexidad (proveniente del sombreado).

1.5 Herramientas existentes

Debido al auge de la visualización en las diferentes ramas de la ciencia y en conjunto con el desarrollo de la informática y sus aplicaciones, muchas empresas a nivel mundial se han dedicado a la implementación de herramientas y paquetes especializados en el área, tanto privativos como libres. Dentro de los sistemas de visualización privativos se destacan herramientas como AVS/Express [28] y Amira [29], dentro de los libres se destacan Khoros [30] y Vis5D [31], implementados en diferentes versiones del lenguaje C y disponibles para la mayoría de las distintas plataformas existentes.

Todas estas son herramientas de visualización de propósito general, poseen módulos para realizar los diferentes tipos de técnicas de Visualización Científica (gráficos en 2D, 3D, entre otros) y para determinadas áreas, como la biología molecular, la neurociencia, las ciencias de la tierra, etc.

1.5.1 AVS/Express

AVS/Express es una herramienta de software propietario diseñada para científicos, investigadores y otros profesionales técnicos que requieren de un amplio conjunto de funcionalidades de visualización de datos y análisis. Proporciona una tecnología de última generación para la creación de aplicaciones de gráficos avanzados, procesado de imagen, visualización de datos y presentación técnica. Está desarrollado por Advanced Visual Systems Inc. (AVS). Posee versiones para trabajar tanto en las diferentes versiones de Windows como en las de GNU/Linux y en Mac OS.

La última versión de AVS/Express es la 7.2 y se comercializa en diferentes ediciones:

- AVS/Express Visualization Edition

  Incluye un subconjunto de los módulos de visualización de AVS/Express Professional Edition. En concreto incorpora aquellos módulos específicos para la visualización de datos, procesado de imagen y diseño de interfaces de usuario, y como Express Professional, permite un completo desarrollo C/C++.

- AVS/Express Professional Edition (o Developer Edition)

  Es un entorno de desarrollo de aplicaciones de visualización, que permite el rápido prototipaje y construcción de aplicaciones comerciales y científicas que requieren de funciones gráficas y de visualización interactivas. Utiliza el lenguaje de programación C/C++ que facilita la construcción de aplicaciones con necesidades de gestión extensiva de datos.

- AVS/Express Multipipe Edition

1.5.2 Amira

Es un sistema de software extensible para la visualización científica, análisis de datos y representación de datos en 3D y 4D. Desarrollado y comercializado por Visage Imaging GmBH, Berlin en cooperación con Zuse Institute Berlin (ZIB). Presenta una interfaz de usuario flexible y su arquitectura modular la convierten en una herramienta universal para el procesamiento y análisis de datos en varias modalidades.

Tiene gran cantidad de opciones para importar imágenes en disímiles formatos de los utilizados por equipos médicos en diversas ramas. Para los desarrolladores permite la creación de módulos nuevos y personalizados, la implementación de lectura y escritura en nuevos formatos y programación en el lenguaje C++.

Como producto comercial Amira debe ser comprado en forma de licencia o subscripción académica. Posee una muy completa versión de evaluación por tiempo limitado que está

disponible en línea. Posee versiones para trabajar tanto en las diferentes versiones de Windows como en las de GNU/Linux y en Mac OS. La última versión disponible es Amira 5.4.2 que fue lanzada en marzo de 2012.

En ocasiones Amira es erróneamente confundido con un producto llamado Avizo, este último es una derivación del primero pues su implementación tiene como base la versión de Amira del año 2009, desarrollado por la compañía VSG (Visalization Sciences Group). Avizo es una aplicación de software comercial de propósito general utilizada para la visualización y análisis de datos científicos e industriales. Posee versiones para trabajar tanto en las diferentes versiones de Windows como en las de GNU/Linux y en Mac OS. La última versión disponible es Avizo 7.0 que fue lanzada en diciembre de 2011.

1.5.3 Khoros

En los comienzos del desarrollo de esta herramienta, su propietario era la empresa Khoral Research, Inc., que establecía una serie de requisitos para la adquisición de la herramienta, y la licencia de distribución menos restrictiva era la de libre acceso que permitía obtener la herramienta sin pagar, pero con derechos de distribución limitados y el trabajo derivado solamente era solamente para uso interno de la entidad. Siendo las demás licencias restrictivas y pagando anualmente un impuesto. Pudiendo utilizarse sobre sistemas GNU/Linux y BSDI entre otras plataformas dependiendo del hardware. Está desarrollado en C++ y diferentes versiones de C según la plataforma. En 1994 fue comprado por Accusoft Corporation y pasó a llamarse VisiQuest, actualmente forma parte del paquete de herramientas llamado ImageGear, perdiendo el anterior nombre. Disponible en cinco ediciones, en dependencia del ambiente de desarrollo.

- ImageGear for .NET
- ImageGear Professional
- ImageGear for Silverlight
- ImageGear for Java
- ImageGear Medical

Siendo en estos momentos enteramente privativo.

1.5.4 Vis5D

Es un sistema de visualización en 3D usado fundamentalmente para visualizar en 3D simulaciones del tiempo. Fue el primer sistema en producir imágenes animadas 3D completamente interactivas de conjuntos de datos volumétricos dinámicos en el tiempo, y el primer sistema de visualización en

3D de código abierto. Su licencia es GNU GPL. Puede ejecutarse sobre sistemas GNU/Linux y sobre Windows (cumpliendo determinadas restricciones de software).

Vis5D brinda opciones de manipulación de memoria, para que grandes conjuntos de datos puedan ser visualizados en espacios de tiempo individuales sin la necesidad de calcular los gráficos sobre la secuencia completa de tiempo de la simulación. Además brinda una API facilitando a los programadores de otros sistemas la incorporación de nuevas funcionalidades.

Vis5D+ surge como un repositorio central de versiones mejoradas y desarrollo sobre Vis5D, un programa de visualización volumétrico de código abierto basado en OpenGL para conjuntos de datos científicos en 3+ dimensiones.

1.5.5 JFreeChart

No es un sistema en sí, sino una biblioteca que los programadores pueden utilizar en sus programas para la construcción de variados tipos de gráficas (gráficos de barra, de pastel, diagramas de dispersión, gráficas de tiempo, etc.). Está implementado en el lenguaje Java y posee una licencia LGPL (GNU Lesser General Public Licence), lo que permite que el software derivado pueda licenciarse de forma libre o privativa y se distribuye gratuitamente conjuntamente con su documentación, comercializándose únicamente por el autor una muy completa guía del programador que ayudaría a un entendimiento más completo de la biblioteca y una mejor explotación de las potencialidades que presenta. Tiene canales de retroalimentación para el reporte de errores y un sistema para aceptar mejoras en el código interno y contribuciones de programadores.

Permite además exportar las gráficas a diferentes formatos (jpeg, png, pdf, entre otros) y a la hora de su uso la personalización de las gráficas a mostrar cambiando la orientación de los ejes, el tamaño de las unidades de medida, entre otras muchas funcionalidades. Producto del lenguaje en el que fue implementado puede utilizarse tanto en aplicaciones web como en aplicaciones de escritorio.

En noviembre de 2011 fue liberada la última versión 1.0.14, siendo la versión anterior (la 1.0.13) la más popular hasta el momento con cerca de medio millón de descargas.

Ha sido ampliamente utilizado por numerosas compañías, por lo que forma parte de gran cantidad de productos disponibles en la red.

Luego de analizar un pequeño grupo de las herramientas existentes y considerando las características actuales del DBAnalyzer, al que se le adicionarán nuevas funcionalidades, específicamente la de visualizar los reportes de posibles errores, la biblioteca JFreeChart se ajusta perfectamente, pues no hace falta cambiar absolutamente nada en la estructura actual del sistema al

estar implementados en el mismo lenguaje de programación. La otra gran ventaja es que tienen licencias afines y se distribuye de forma gratuita.

El resto de las herramientas analizadas a pesar de ser muy eficientes y poseer gran uso y aplicación, tienen el gran inconveniente del costo monetario asociado a su utilización y el lenguaje en el que fueron implementadas, lo que limitaría la portabilidad a varias plataformas del software resultante.

Capítulo 2

Como parte del proyecto de investigación se procede a la implementación de una nueva versión de la herramienta DBAnalyzer. En el presente capítulo se tiene como objetivo mostrar los elementos necesarios para adicionar la funcionalidad al DBAnalyzer de visualización de los datos, y se mostrarán las tecnologías utilizadas durante el desarrollo, así como los principales artefactos generados durante el diseño e implementación de la herramienta propuesta.

2.1 Características de la herramienta DBAnalyzer

La herramienta DBAnalyzer fue concebida desde sus inicios de tal forma que el analizar diferentes bases de datos no se tornara en una tarea tediosa y además en hacer una búsqueda y detección efectiva de posibles errores en las bases de datos analizadas. En su primera versión solamente se generaba un reporte en modo texto, este reporte mostraba los datos estadísticos de cada elemento, es decir, de los diferentes campos de las tablas seleccionadas a analizar. Si el campo analizado es de tipo entero, real o moneda la herramienta calcula el valor máximo, el mínimo, la media, la moda, la cantidad de ceros, cuantos elementos vacíos tiene, la cardinalidad (cantidad de valores diferentes), la desviación estándar y los posibles valores que pueden estar fuera de rango, cálculo que se hace teniendo en cuenta el valor de la media y la desviación estándar. Si el campo es de tipo cadena determina la cantidad de valores diferentes que hay (cardinalidad), lista cada uno de los valores del campo junto con su por ciento de ocurrencia y la cantidad de veces que aparece, además de cuantos elementos vacíos y nulos tiene. Si es de tipo fecha determina cuantas fechas distintas hay y cuantos elementos vacíos y nulos tiene dicho campo. Producto de la cantidad de datos que se pudieran almacenar, estos reportes en modo texto no resaltan ningún valor en específico y solamente se veían los datos estadísticos referentes al campo.

Esta nueva versión permitirá generar un reporte en modo gráfico que en dependencia del tipo de campo que se analice resalte los posibles valores con problemas, además de mostrar visualmente y con diferenciación de colores cada uno de los valores con la cantidad de veces que se repite, dependiendo además del tipo de gráfica que se muestre.

2.1.1 Modelo de Dominio

No existe un negocio definido para esta investigación, pues está orientada al desarrollo de una herramienta que permite el análisis de bases de datos con el fin de identificar los posibles errores que tenga en los diferentes campos. Por lo antes explicado es que se realiza un modelo de dominio para identificar y describir los componentes del sistema a implementar.

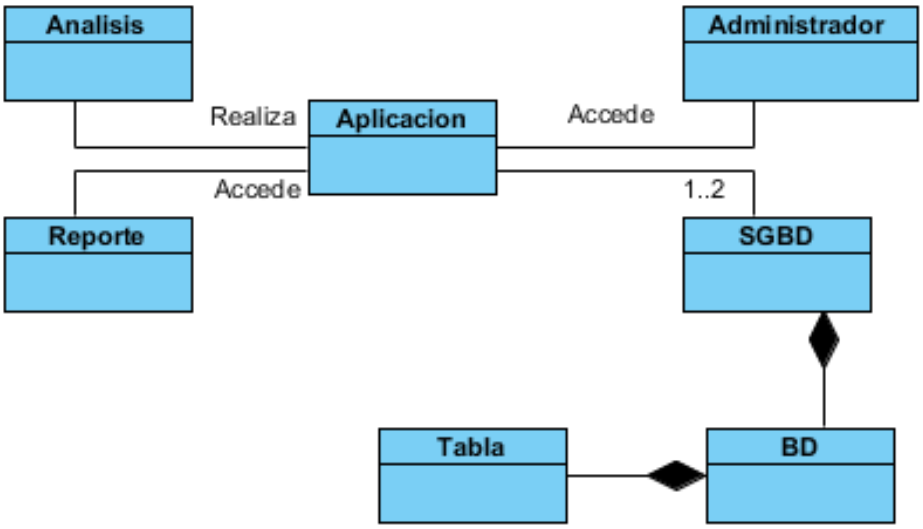


Figura 2.1 Modelo de Dominio de la herramienta DBAnalyzer.

En la Figura 2.1 se muestra la relación existente entre las entidades o clases, es posible apreciar como el administrador (usuario que manipulará la herramienta) accede a la aplicación, que debe conectarse a un sistema gestor de base de datos (SGDB) compuesto por una o muchas bases de datos y estas a su vez contienen una o muchas tablas. La aplicación realiza el análisis de los elementos correspondientes generándose un reporte que luego debe mostrarse en modo texto y gráfico.

2.1.2 Requerimientos del sistema

Los requerimientos funcionales y no funcionales fueron identificados partiendo de la versión anterior de la herramienta y las nuevas necesidades a adicionar. A continuación se muestran de manera general.

Requerimientos Funcionales

ADMINISTRADOR

RF 1 Permitir conectarse a las Bases de Datos

Descripción: Permitir la conexión al SGBD que puede ser PostgreSQL o MySQL luego que el administrador introduzca la dirección (URL) del servidor y nombre de la BD.

RF 2 Autenticar usuario

Descripción: Teniendo ya establecido los datos de la conexión el sistema debe permitir las diferentes variantes de conexión, las cuales son: la autenticación mediante nombre de usuario y contraseña, utilizando usuario solamente y mediante usuario anónimo, siempre y cuando el gestor de base de datos lo permita.

RF 3 Analizar tablas

Descripción: Luego de establecer la conexión se muestran todas las tablas que componen la base de datos a analizar, el sistema debe permitir seleccionar las tablas deseadas, pudiendo ser una de ellas, varias (en cualquier orden), o todas las tablas. Luego de escogerlas se procede al análisis de todos los campos de las tablas seleccionadas, que no es más que almacenar en una lista los valores que tengan cada uno de los campos.

RF 3.1 Mostrar tablas de la BD.

RF 3.2 Permitir escoger una tabla.

RF 3.3 Permitir escoger varias tablas.

RF 3.6 Permitir escoger todas las tablas.

RF 3.7 Permitir analizar una tabla.

RF 3.8 Permitir analizar varias tablas.

RF 3.9 Permitir analizar todas las tablas.

RF 4 Gestionar reporte

Descripción: Luego de analizar las tablas seleccionadas el sistema debe generar y guardar el reporte de estas en dos variantes, modo texto y gráfico, luego debe permitir buscarlos y mostrarlos, para en el caso de los reportes gráficos permitir guardar como archivo pdf una gráfica específica de reporte.

RF 4.1 Generar reporte modo texto.

RF 4.2 Buscar reporte modo texto.

RF 4.3 Mostrar reporte modo texto.

RF 4.4 Generar reporte modo gráfico.

RF 4.5 Buscar reporte modo gráfico.

RF 4.6 Mostrar reporte modo gráfico.

RF 4.7 Guardar gráfica de reporte como archivo pdf.

Requerimientos no funcionales

- **Apariencia o Interfaz Externa**

    - El diseño de la interfaz deberá ser claro y preciso para guiar al usuario en los procesos a realizar. Para una mejor visualización debe utilizarse una resolución de 1024x768 píxeles.

    Debe permitir:

- Diferenciar cuando autenticar utilizando usuario y contraseña y cuando autenticar con usuario anónimo.
- Mostrar en ventanas diferentes los procesos de conexión a base de datos, mostrar reporte texto y mostrar reporte gráfico, estando el primero en la principal y el resto accesibles mediante menús.
- Una vez que la conexión se establezca y se muestren las tablas de la base de datos, hacer de varias formas una misma función, de manera que esta se pueda hacer más rápido, es decir, la acción de seleccionar varios elementos de uno en uno o todos a la vez con un solo clic.
- Utilizar una barra de progreso para el proceso de análisis de las bases de datos y el guardado de los reportes en disco.
- Abrir el reporte texto en una interfaz que no permita la modificación de los datos contenidos en este.
- Mostrar los resultados del reporte gráfico utilizando una estructura jerárquica (similar a la usada por los sistemas operativos para mostrar directorios y archivos), de forma tal que se sepa la tabla a que pertenece cada campo.
- Utilizar solo gráficos de barra, diagramas de dispersión e histogramas, por las características de los elementos a mostrar.
- Seleccionar desde un menú el tipo de gráfica que se desea utilizar para visualizar los resultados.
- Seleccionar desde un menú diferentes temas para visualizar las gráficas según las preferencias del usuario.
- Mostrar una leyenda en las gráficas para saber el significado de cada elemento utilizado.
- Mostrar en una nueva ventana los mensajes de error cada vez que estos ocurran.

- **Portabilidad**
  - El sistema debe ser multiplataforma.
- **Seguridad**
  - El sistema no almacenará ni mostrará las contraseñas proporcionadas mediante la autenticación a las bases de datos.
  - La conexión a las bases de datos debe hacerse en modo de solo lectura.

- Abrir los descriptores de archivos solamente cuando se utilicen y cerrarlos apenas dejen de usarse estos.
- Mostrar diálogos de confirmación cada vez que se vaya a escribir por primera vez en un archivo.

- **Confiabilidad**
  - Una vez que se conecte a una base de datos y muestre sus tablas no cambiar los datos de la conexión a no ser que se quiera realizar el análisis a otra base de datos.
- **Legales**
  - Las herramientas de desarrollo y bibliotecas a utilizar deben ser libres o de código abierto.
- **Restricciones de la implementación**
  - Se programará en lenguaje Java, en la herramienta NetBeans.
  - Solo permitirá conexiones a BD en PostgreSQL y MySQL.
- **Ayuda**
  - El sistema contará con un breve manual de ayuda al usuario.
- **Hardware recomendado**
  - 2 GB o más de capacidad en el disco duro.
  - Microprocesador de 1.0 GHz o superior.
  - 512 MB de memoria RAM o superior (256 MB mínimo).
  - Tarjeta de red de 100 Mbps o superior.

2.1.3 Modelado del sistema

Actores del Sistema

A continuación se realiza una breve descripción de los actores que intervienen en el sistema DBAnalyzer.

Actor	Descripción
ADMINISTRADOR	Usuario encargado de ejecutar la aplicación y de realizar todas las operaciones necesarias para obtener los resultados esperados.

Modelo de Casos de Uso del sistema

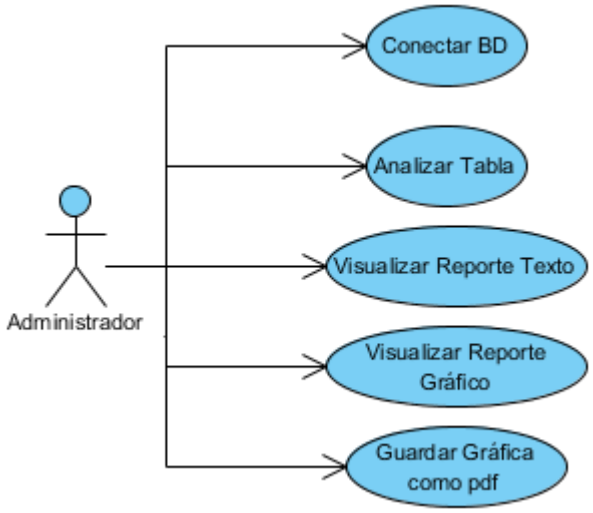


Figura 2.2 Diagrama de casos de uso herramienta DBAnalyzer.

Descripción de los Casos de Uso

CUS Conectar BD

Caso de Uso	**Conectar BD**
Actores	**Administrador**
Resumen	El caso de uso inicia cuando el administrador escoge el gestor de base de datos al que se va a conectar e introduce los datos correspondientes a esta conexión.
Precondiciones	El administrador debe conocer la dirección del servidor, el nombre de la base de datos a la que se va a conectar y el usuario y contraseña para establecer la conexión.
Referencias	**RF 1, RF 2**
Prioridad	**Crítica**

CUS Analizar Tabla

Caso de Uso	**Analizar Tabla**
Actores	**Administrador**
Resumen	El caso de uso inicia cuando se muestran las tablas por la que está compuesta la base de datos a la que se conectó y el administrador procede a seleccionar aquellas tablas a las que va a realizar el análisis.
Precondiciones	La conexión con el servidor de base de datos debe haberse establecido
Referencias	**RF 3, RF 3.1, RF 3.2, RF 3.3, RF 3.6, RF 3.7, RF 3.8, RF 3.9**
Prioridad	**Crítica**

CUS Visualizar Reporte Texto

Caso de Uso	**Visualizar Reporte Texto**
Actores	**Administrador**
Resumen	El caso de uso inicia una vez que el administrador haya realizado el análisis de las tablas deseadas y haya guardado en disco el archivo conteniendo el reporte en modo texto.
Precondiciones	El análisis de las tablas deseadas debe haberse realizado.
Referencias	**RF 4, RF 4.1, RF 4.2, RF 4.3**
Prioridad	**Crítica**

CUS Visualizar Reporte Gráfico

Caso de Uso	**Visualizar Reporte Gráfico**
Actores	**Administrador**
Resumen	El caso de uso inicia una vez que el administrador haya realizado el análisis de las tablas deseadas y haya guardado en disco el archivo conteniendo el reporte en modo gráfico.
Precondiciones	El análisis de las tablas deseadas debe haberse realizado.
Referencias	**RF 4, RF 4.4, RF 4.5, RF 4.6**
Prioridad	**Crítica**

CUS Guardar Gráfica como pdf

Caso de Uso	**Guardar Gráfica como pdf**
Actores	**Administrador**
Resumen	El caso de uso inicia una vez que el administrador esté visualizando el reporte gráfico.
Precondiciones	Debe estarse realizando la visualización del reporte gráfico.
Referencias	**RF 4, RF 4.7**
Prioridad	**Auxiliar**

2.2 Diseño de la herramienta DBAnalyzer

Para implementar las funcionalidades necesarias es preciso identificar y construir los artefactos del diseño correspondientes a los casos de uso de la herramienta.

Diagramas de colaboración

Los diagramas de interacción se utilizan para modelar los aspectos dinámicos de un sistema, lo que trae consigo modelar instancias concretas o prototípicas de clases interfaces, componentes y nodos junto con los mensajes enviados entre ellos, todo esto se denomina flujo de control. Estas

operaciones se llevan a cabo en el contexto de un escenario que ilustra un comportamiento de dicho sistema. Existen dos tipos de diagramas de interacción en UML, los Diagramas de Colaboración (dimensión estructural) y los Diagramas de Secuencia (dimensión temporal). Los diagramas de colaboración destacan la organización de los objetos que participan en una interacción, se construyen colocando en primer lugar los objetos que participan en la colaboración como nodos del grafo, a continuación se representan los enlaces que conectan esos objetos como arcos del grafo y por último, estos enlaces se adornan con los mensajes que envían y reciben los objetos. Poseen dos características que los distinguen de los diagramas de secuencia, una de ellas es el camino, el cual indica cómo se enlaza un objeto a otro y la otra es el número de secuencia el que indica la ordenación temporal de un mensaje o el orden en que estos se llevan a cabo [32].

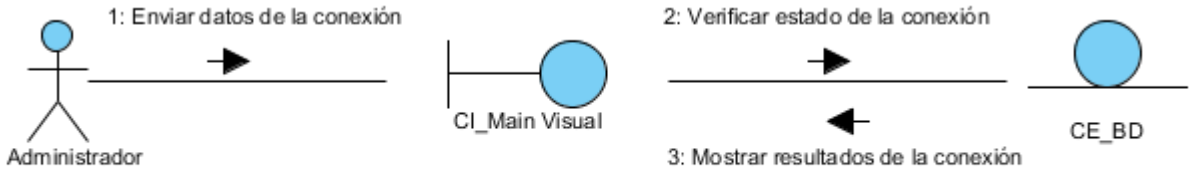


Figura 2.3 Diagrama de Colaboración del CUS Conectar BD.

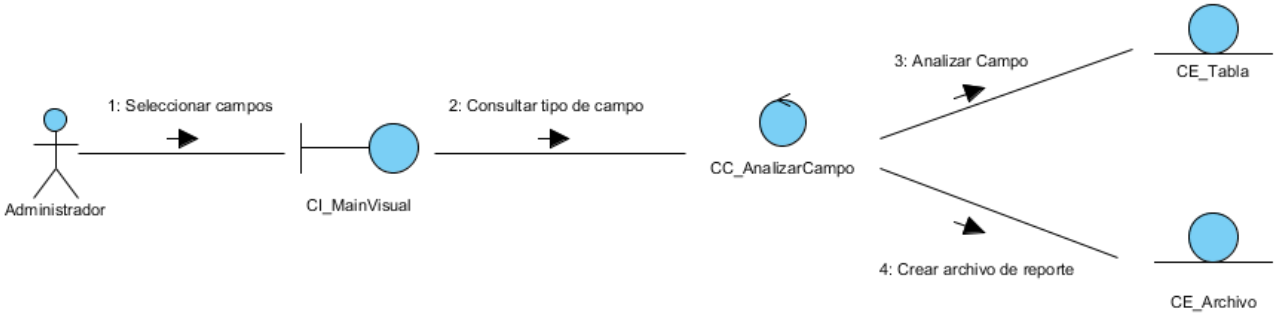


Figura 2.4 Diagrama de Colaboración del CUS Analizar Tabla.

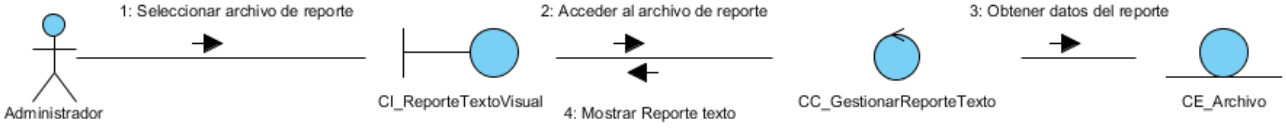


Figura 2.5 Diagrama de Colaboración del CUS Visualizar Reporte Texto.

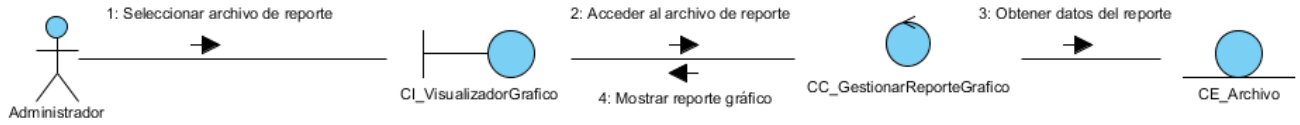


Figura 2.6 Diagrama de Colaboración del CUS Visualizar Reporte Gráfico.

Diagrama de clases

Un diagrama de clases del diseño es un diagrama que describe gráficamente las especificaciones de las clases existentes en un producto de software así como de las interfaces involucradas. Estos diagramas contienen información útil para el usuario como por son por ejemplo clases, asociaciones y atributos, interfaces con sus operaciones y constantes, métodos, información sobre los tipos de atributos, navegabilidad y dependencia existente. A diferencia de un modelo conceptual, un diagrama de clases del diseño contiene las definiciones de las entidades del software en vez de conceptos del mundo real. Pero como en UML no está bien definido el concepto de diagrama de clases diseño, este se sirve de un término genérico denominado Diagrama de Clases [32].

El diagrama de clases perteneciente a la herramienta DBAnalyzer, está compuesto por tres paquetes de clases, el paquete “interfazgrafica” que contiene las clases referentes a las interfaces de la aplicación, el paquete “dbanalyzer” con las clases referentes al análisis de los diferentes campos de las bases de datos y la generación de reporte en modo texto y el paquete “reportegrafico” que contiene la estructura de como guardar un objeto con el reporte gráfico, así como la generación de las diferentes gráficas. Los dos primeros paquetes mencionados fueron implementados completamente en la versión anterior, por lo que solamente se mostrará el diagrama de las nuevas clases implementadas (Figura 2.7) y su explicación.

A la hora de implementar la nueva funcionalidad para la herramienta, en principio siempre se tuvo en cuenta la separación entre los elementos pertenecientes a la interfaz gráfica y la lógica, esto para garantizar ante cualquier solicitud de cambio no tener que reprogramar todo y evitar así la repetición innecesaria de código fuente. Partiendo de lo anterior se implementó lo siguiente:

ChartInfoGranLista tiene como atributos:

- BaseDatos de tipo cadena que representa el nombre de la base de datos.
- ListaTablas que va a ser una lista de objetos de tipo CharInfoLista que contiene la información de cada una de las tablas de la base de datos.

ChartInfoLista tiene como atributos:

- Tabla de tipo cadena que representa el nombre de la tabla.

- ListaCampos que va a ser una lista de objetos de tipo ChartInfo que contiene la información de cada uno de los campos que tiene la tabla.

ChartInfo tiene como atributos:

- DatosGrafica de tipo Grafica, con la información necesaria para generar la gráfica correspondiente al campo.
- NombreCampo de tipo cadena que representa el nombre del campo.
- TipoDato de tipo cadena que representa el tipo de dato que contiene y puede tener uno de los valores siguientes: “Entero”, “Real” o “Cadena”.
- CantidadVacios de tipo entero con la cantidad de elementos vacíos que tenga el campo.
- CantidadNulos de tipo entero con la cantidad de elementos nulos que tenga el campo.
- Cardinalidad de tipo entero con la cantidad de elementos diferentes que tenga el campo.
- CantidadTotalElementos de tipo entero con la cantidad total de elementos que tenga el campo.

Grafica tiene como atributos:

- ArregloEntero es una lista de enteros y se va a utilizar para almacenar los valores en caso que el campo a analizar sea de tipo “Entero”.
- ArregloReal es una lista de reales y se va a utilizar para almacenar los valores en caso que el campo a analizar sea de tipo “Real”.
- ArregloCadena es una lista de cadenas y se va a utilizar para almacenar los valores en caso que el campo a analizar sea de tipo “Cadena”.
- ArregloCantidad es una lista de enteros, donde en cada posición de esta lista va a estar representada la cantidad de veces que se repita el valor correspondiente en esa misma posición de la lista ArregloEntero, ArregloReal o ArregloCadena según sea el caso.
- media es de tipo real y va a representar la media en caso que el campo sea numérico.
- DStd es de tipo real y va a representar la desviación estándar de los valores en caso que el campo sea numérico.

Además de las mencionadas anteriormente fueron implementadas las clases:

- GraficoBarras
- DiagramaDispersion
- Histograma

Cada una con los métodos necesarios para la visualización de los datos mediante gráficos de barras, diagramas de dispersión e histogramas respectivamente, siempre teniendo en cuenta el

tipo de dato a visualizar.

Esta variante de implementación presenta las ventajas:

- Al generar el reporte y luego cargando el archivo resultante se pueden visualizar las gráficas, solamente se necesita la conexión con el servidor a la hora de la generación del reporte gráfico.
- El tiempo en el que la conexión con el servidor está establecida es el mínimo posible.
- El archivo generado con el reporte gráfico ocupa poco espacio en disco (en dependencia directa con el volumen de datos analizados).
- Permite poder modificar cualquier otra parte del código fuente de la aplicación sin tener que reanalizar las bases de datos para generar un nuevo reporte.
- En tiempo de ejecución se puede cambiar el tipo de gráfica con el que se visualiza determinado campo sin necesidad de realizar más de un análisis a la base de datos.
- No es necesario analizar la base de datos completa, pudiendo seleccionar solamente las tablas de mayor interés al usuario.

Y como desventajas:

- Si cambian los datos de algún campo en la base de datos se debe realizar otro análisis para tener los datos actualizados.
- Generar las gráficas en tiempo de ejecución puede traer como consecuencia que para un campo con gran cantidad de elementos se sobrecargue el procesador a la hora de tratar de visualizar el reporte de dicho campo.

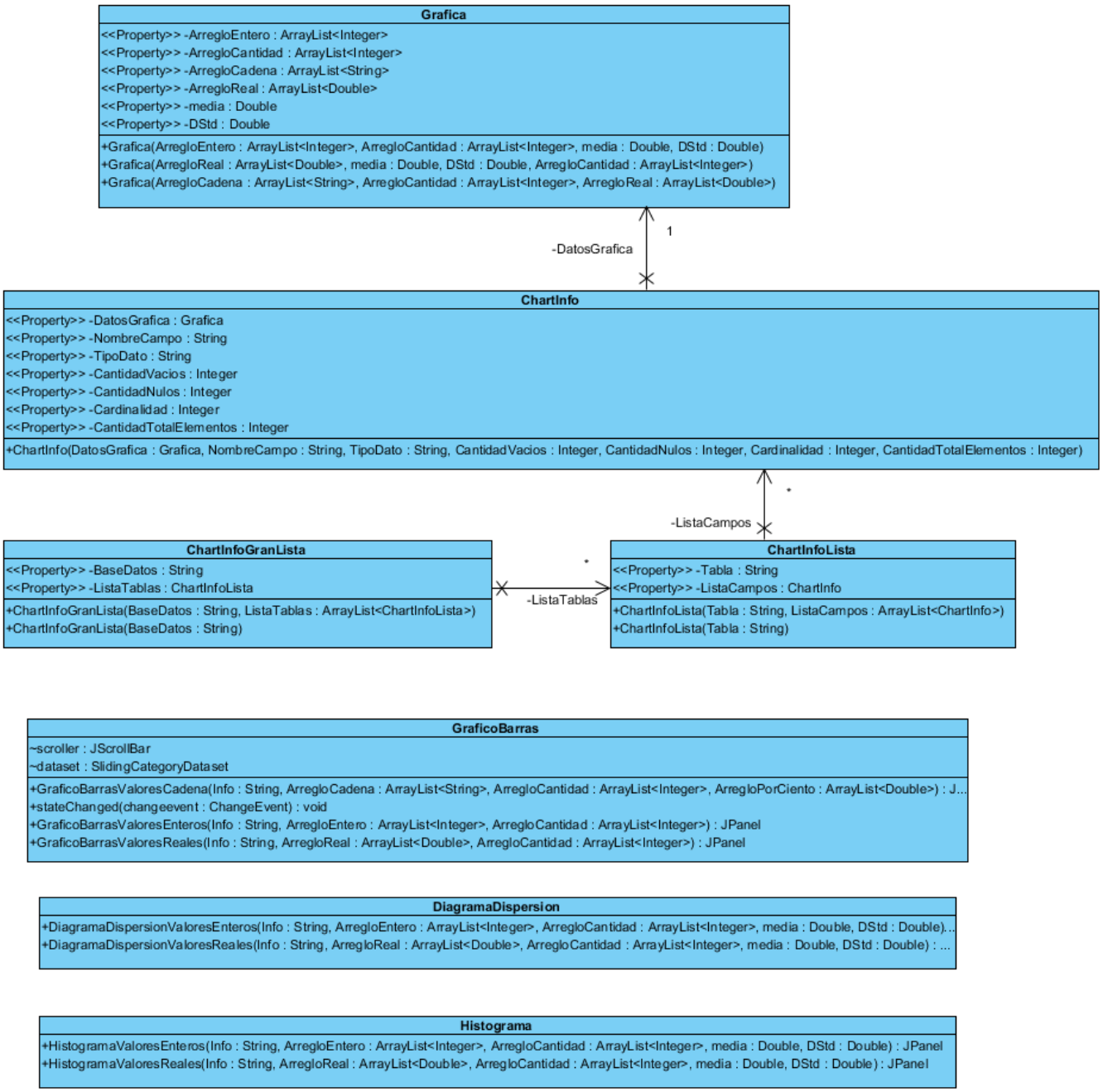


Figura 2.7 Diagrama de clases del paquete “reportegrafico”.

Diagrama de componentes

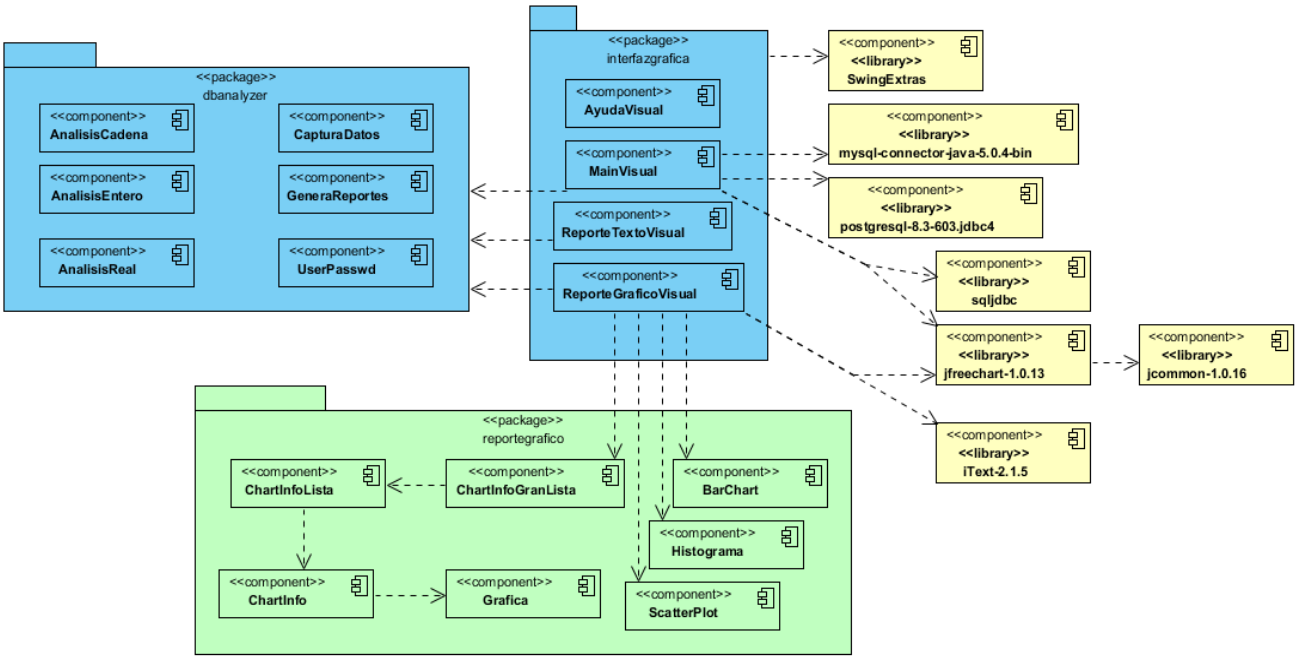


Figura 2.8 Diagrama de componentes.

2.3 Implementación de la herramienta DBAnalyzer

Para la implementación de esta nueva versión del DBAnalyzer se utilizaron una serie de tecnologías y herramientas las cuales se describirán a continuación.

Lenguaje unificado de modelado: UML (en inglés: Unified Modeling Languaje) es un lenguaje utilizado para visualizar, especificar, construir y documentar los artefactos de un sistema que involucra una gran cantidad de software [33], es de los más utilizados ya que ofrece un estándar para describir un plano o modelo del sistema en el cual se incluyen los conceptos de proceso de negocio así como funciones del sistema. UML en su funcionamiento emplea los siguientes tipos de diagramas [32]: de estructura estática (clases, objetos y casos de uso), de comportamiento (interacción, estado, actividad) y de implementación (componentes, despliegue).

Herramienta CASE (por sus siglas en inglés: Computer Aided Software Engineering): Visual Paradigm es una suite de trabajo (entiéndase por suite: varias herramientas dentro de una sola) fácil de utilizar, la cual cuenta con herramientas para el diseño de diagramas y los artefactos que se generan durante los ciclos de desarrollo de software. Permite además la integración con otros IDEs de desarrollo como NetBeans y Eclipse. En la UCI se cuenta con la licencia para esta suite, lo cual facilita las condiciones de uso [34].

Lenguaje de programación: Java es un lenguaje de programación orientado a objetos, libre y portable. La portabilidad viene dada por el hecho de que el compilador del lenguaje genera un código binario (bytecode) el cual es interpretado por la Máquina Virtual de Java (JVM por sus

siglas en inglés: Java Virtual Machine), por lo que un programa escrito en Windows puede ser interpretado en un entorno libre, con solo disponer de la máquina virtual para dichos entornos y viceversa.

IDE de desarrollo: Netbeans es un entorno de desarrollo libre y de código abierto que soporta varios lenguajes, entre ellos Java [35]. Está desarrollado en este mismo lenguaje y facilita el desarrollo de aplicaciones.

Capítulo 3

Una vez concluido el proceso de diseño e implementación de la herramienta DBAnalyzer en su versión 2.1, se procede a realizar el análisis correspondiente sobre la factibilidad de su uso, pasando por las ventajas y desventajas de la variante de implementación utilizada.

3.1 Ventajas y desventajas de la variante de implementación

Dentro del epígrafe 2.2 perteneciente al capítulo anterior, se mencionaron un conjunto de ventajas y desventajas de la implementación de la herramienta, pero aquellas vienen dadas desde el punto de vista de diseño, siendo las mismas independientemente del lenguaje de programación a utilizar, las que se relacionan a continuación fueron determinadas producto del uso de esta.

Ventajas de la herramienta:

- Es multiplataforma, puede ejecutarse tanto sobre sistemas Windows como GNU/Linux, teniendo solamente como prerrequisito el tener la máquina virtual de Java instalada para la plataforma en cuestión.
- Las gráficas poseen opciones de acercamiento mediante la selección de un área determinada con el mouse, permitiendo ver con mayor precisión un conjunto de datos determinado.
- En las gráficas en el caso de los valores numéricos se hace una diferenciación con colores diferentes entre los elementos válidos (de azul) y los posibles elementos con errores (de rojo).
- En las gráficas para los valores de tipo cadena, estas se ordenan alfabéticamente, permitiendo así aparte de ver la repetición de cada una, elementos que pudieran haber sido mal escritos (con alguna letra diferente) quedarían cercanos, pudiendo identificarse de forma fácil (excepto posiblemente en los casos que la letra diferente sea la primera).

Desventajas:

- Al analizar bases de datos muy voluminosas, producto de los datos almacenados en memoria puede producirse el error de falta de espacio en la pila de ejecución del Java (java.lang.OutOfMemoryError: Java heap space), para esto se pueden tener dos soluciones alternativas, incrementar el tamaño de la pila de ejecución y realizar el análisis de la base de datos en varias iteraciones.
- Al mostrar el reporte gráfico de un campo con una cardinalidad muy grande, por ejemplo 12000 elementos o más, la gráfica en su conjunto no es muy clara (producto de la misma resolución de pantalla), por lo que el análisis de ese campo deberá realizarse por etapas.
- Si los valores mínimo y máximo en campos numéricos están muy separados, entonces si la

gráfica seleccionada es por ejemplo un histograma, las barras de este serán muy finas, en caso de ser un diagrama de dispersión los puntos serán muy pequeños. Esta dificultad la compensa el hecho de poder seleccionar una parte de la gráfica y hacerle un acercamiento.

Destacando además la forma en que se puede realizar el análisis de los diferentes tipos de datos mediante la utilización de la herramienta toda vez que se tenga el archivo generado con el objeto serializado del contenido de la base de datos. Al mostrar una gráfica determinada (de un campo de alguna de las tablas que contenga), se pueden ver los datos asociados a un valor, solamente al dar clic sobre el elemento correspondiente en la gráfica, y se verán las tuplas donde este aparezca. Valorando así en el entorno donde está si realmente es un posible error.

3.2 Análisis de los datos

Los metadatos guardados en el esquema de una base de datos son insuficientes para asegurar la calidad de los datos de dicha base. En la mayoría de los casos, solo algunas reglas de integridad son declaradas y almacenadas en el esquema. Por ello es importante analizar las instancias actuales para obtener metadatos reales (reingeniería) y patrones inusuales de los datos. Los metadatos ayudan a encontrar problemas en la calidad de los datos.

Existen dos formas de realizar análisis de los datos: realizando un perfil de los datos y utilizando minería de datos [36].

El perfil de los datos centra el análisis de las instancias en atributos individuales. Se deriva información tal como: tipo de dato, longitud, rango de valores, valores discretos y su frecuencia, varianza, unicidad, ocurrencia de nulos, patrones típicos de cadena. Todos estos elementos pueden dar una visión exacta de los datos, de su calidad. [37]

La minería de datos, por su parte, ayuda a descubrir patrones de datos en conjuntos grandes de datos. Se utilizan modelos descriptivos de minería de datos, los cuales incluyen clustering, sumarización, descubrimiento de asociaciones, descubrimiento de secuencias. Se pueden derivar dependencias funcionales, algunas reglas de negocio, las cuales pueden servir para completar datos ausentes, descubrir datos ilegales, identificar duplicados, etc.

A partir de estas ideas se construyó una herramienta computacional que es capaz de realizar el perfil de datos almacenados en una base.

El perfil de los datos puede ser importante para detectar problemas presentes. En la Tabla 3.1 se dan algunos ejemplos:

Problemas	**Metadato**	**Ejemplos/heurísticas**
Valores	cardinalidad	Ej. Si cardinalidad(sexo)>2 indica problemas.

ilegales	max, min	Los valores max, min no deben estar fuera del rango permitido.
	varianza, desviación	La varianza y desviación no deben ser mayores que el umbral.
Errores ortográficos	valores de atributos	Ordenando los valores de un atributo, hace que datos con errores ortográficos queden cerca y puedan detectarse.
Valores ausentes	valores nulos	Por ciento / número de valores nulos.
	valores por defecto	La presencia de valores por defecto puede indicar realmente un valor nulo.
Duplicados	cardinalidad+unicidad	La cardinalidad del atributo debe coincidir con el número de filas.
	valores de atributos	Ordenar los valores por el número de ocurrencias. Más de una ocurrencia indica duplicados.

Tabla 3.1 Perfil de datos.

Esta herramienta realiza un análisis de cada uno de los atributos de la base y determina:

Para todos los datos

- Tipo de dato.
- Cantidad de valores ausentes.
- Cardinalidad.

Para datos numéricos

- Valores máximo y mínimo.
- Valor medio, moda, desviación estándar.
- Contar ceros (pueden indicar valores ausentes).

Para datos tipo cadena

- Contar cadenas vacías (pueden indicar valores ausentes).
- Valores y cantidad de repeticiones de cada uno.

3.3 Resultados del trabajo con la herramienta DBAnalyzer

Al tener disponible la actual versión del DBAnalyzer se realizó el análisis de una serie de bases de datos con vistas a verificar la eficacia de la herramienta. Para ello se va a desglosar este análisis según los tipos de campos analizados junto a los resultados correspondientes. Entre las bases de datos analizadas se encuentra la del sitio del graduado de la Universidad de las Ciencias Informáticas (UCI), la del portal de noticias “Entérate”, entre otras más pertenecientes a proyectos productivos.

3.3.1 Base de datos del portal del graduado

La base de datos del portal del graduado, se encuentra sobre el sistema gestor de base de datos PostgreSQL en su versión 9.1, está compuesta por un total de 118 tablas y luego de analizarla con el DBAnalyzer se encontraron los siguientes detalles:

Campos numéricos de tipo entero

En el campo “field_efemeride_dia_value” de la tabla “content_type_efemeride”, donde los valores numéricos que almacena se corresponden con los días del mes, hay 8 valores en cero (ver Figura 3.1).

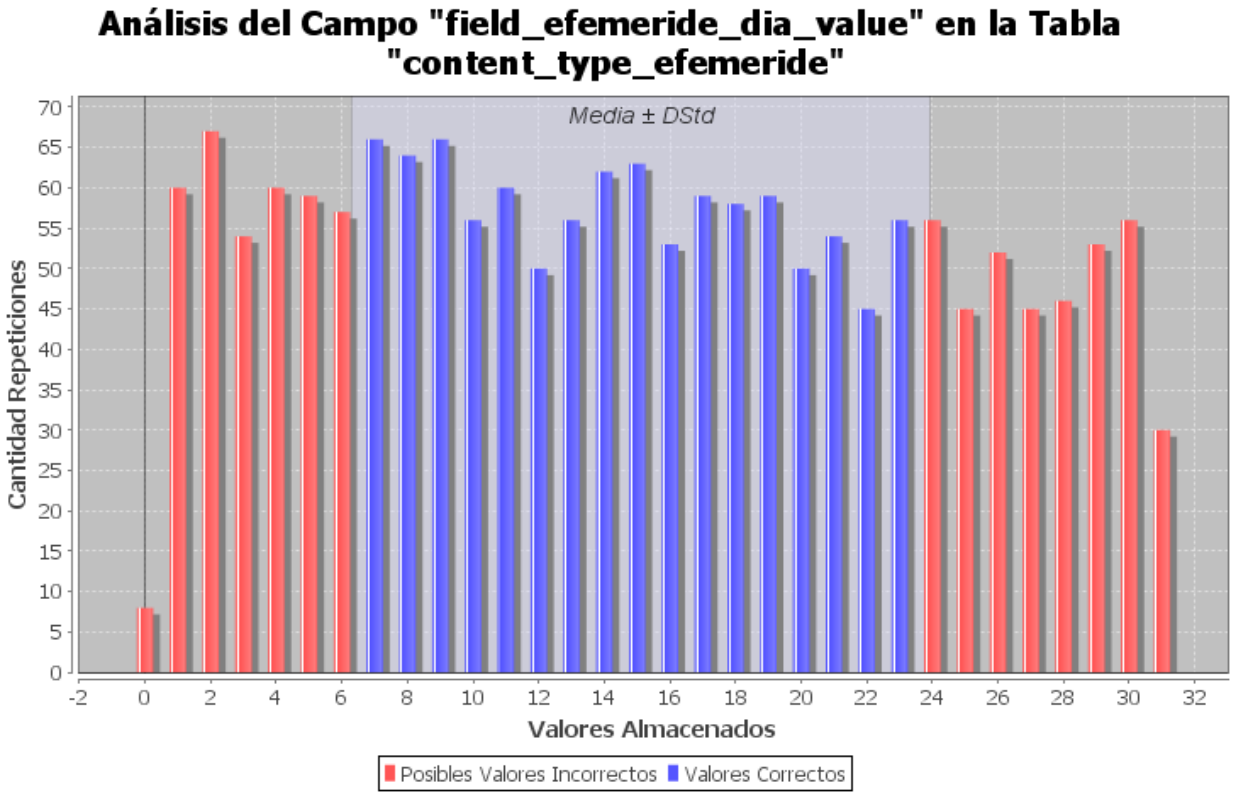
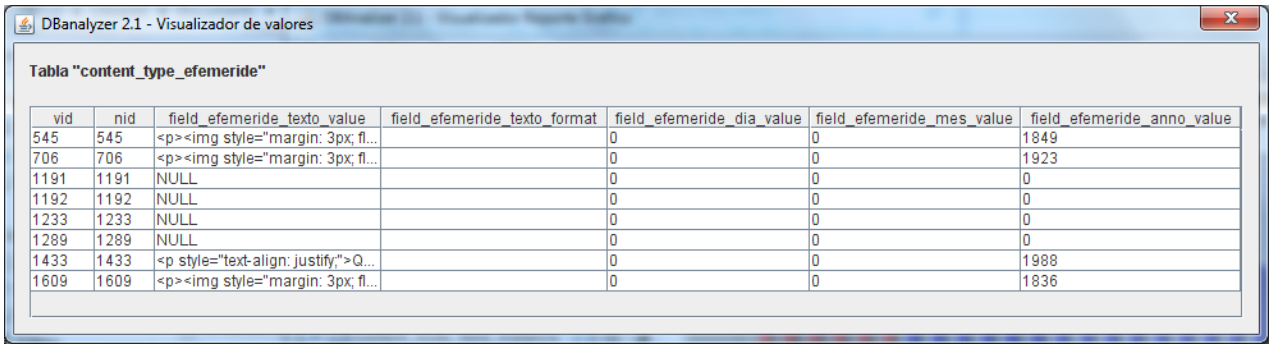


Figura 3.1 Histograma con los valores almacenados en “field_efemeride_dias_value”.



vid	nid	field_efemeride_texto_value	field_efemeride_texto_format	field_efemeride_dia_value	field_efemeride_mes_value	field_efemeride_anno_value
545	545	<p><img style="margin: 3px; fl...		0	0	1849
706	706	<p><img style="margin: 3px; fl...		0	0	1923
1191	1191	NULL		0	0	0
1192	1192	NULL		0	0	0
1233	1233	NULL		0	0	0
1289	1289	NULL		0	0	0
1433	1433	<p style="text-align: justify;">Q...		0	0	1988
1609	1609	<p><img style="margin: 3px; fl...		0	0	1836

Figura 3.2 Tuplas donde “field_efemeride_dia _value” tiene valor cero.

Adicionalmente y como se explicó antes, al dar clic en alguno de los valores mostrados en la

gráfica, se ve el elemento en cuestión todas las veces que aparezca junto a los elementos a los que se encuentra asociado en la tabla. Teniendo de esta forma una visión más completa del marco en el que se encuentra y valorando mejor si realmente pudiera llegar a ser un valor incorrecto.

En la Figura 3.2 se muestra la tabla “content_type_efemeride” con todas las tuplas donde “field_efemeride_dia_value” tiene valor cero. Observando el resto de los campos, se observa que “field_efemeride_mes_value” que representa el valor del mes (un número entre 1 y 12) también tiene sus valores en cero (esto no garantiza que no existan más valores de este campo que también estén en cero, ver Figura 3.4), y 4 de las tuplas tiene el valor de “field_efemeride_texto_value” en “NULL”.

La conclusión de este análisis indica que los valores de “field_efemeride_dia_value” y “field_efemeride_mes_value” que estén en cero se deben cambiar, y que las tuplas donde “field_efemeride_texto_value” tiene valor “NULL” deben ser eliminadas, o reemplazadas por efemérides válidas.

En la figura 3.3 se muestra el campo “field_efemeride_mes_value” de la tabla “content_type_efemeride”, donde los valores numéricos que almacena se corresponden con los meses del año, hay 10 valores en cero. Al verificar estos datos en la Figura 3.4, y compararlos con los de la Figura 3.2, aparecen 2 nuevas tuplas (de las 8 que tenían el día en cero, todas tenían el mes en cero) a las que habría que ponerles un valor de mes válido para la efeméride que representan, pues el valor del día y del año ya lo tienen definido.

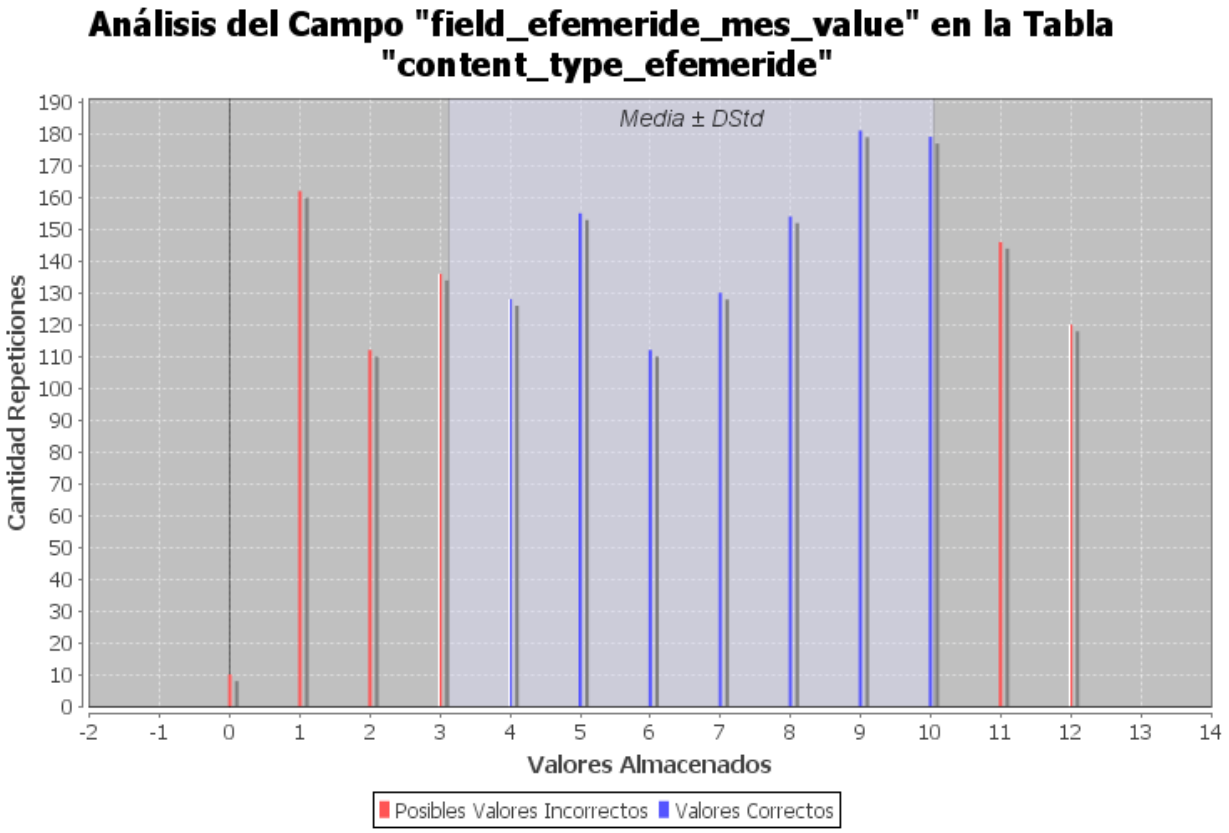


Figura 3.3 Histograma con los valores almacenados en “field_efemeride_mes_value”.

DBanalyzer 2.1 - Visualizador de valores

Tabla "content_type_efemeride"

vid	nid	field_efemeride_texto_value	field_efemeride_texto_format	field_efemeride_dia_value	field_efemeride_mes_value	field_efemeride_anno_value
545	545	<p><img style="margin: 3px...		0	0	1849
706	706	<p><img style="margin: 3px...		0	0	1923
675	675	<p style="text-align: justify;">...		29	0	1911
1191	1191	NULL		0	0	0
1192	1192	NULL		0	0	0
1233	1233	NULL		0	0	0
1289	1289	NULL		0	0	0
1433	1433	<p style="text-align: justify;">...		0	0	1988
1609	1609	<p><img style="margin: 3px...		0	0	1836
1707	1707	<p><!--[if gte mso 9]><xml> ...		14	0	1954

Figura 3.4 Tuplas donde “field_efemeride_mes _value” tiene valor cero.

En el campo “filesize” de la tabla “files”, donde los valores que se almacenan corresponden al tamaño en disco de los archivos que forman parte de las noticias, que pueden ser imágenes, documentos Word, pdf, etc. En la Figura 3.5 se aprecian los valores que tiene almacenado, entre ellos se escogió primeramente al mayor valor que aparece un total de 12 veces (Figura 3.6), este valor pertenece a un mismo archivo, pues el campo “filename” tiene el mismo valor en todas las tuplas, al igual que el tamaño, por lo que todo parece indicar que es el mismo archivo, aunque el campo “filepath” que es donde está físicamente sea diferente. Otro ejemplo lo constituye también el valor que se repite 34 veces en la Figura 3.5, que al verificar los datos

asociados a este valor numérico mostrados en la Figura 3.7, donde ocurre exactamente el mismo fenómeno acabado de explicar, al parecer el mismo archivo lo que en esta ocasión colocado 34 veces.

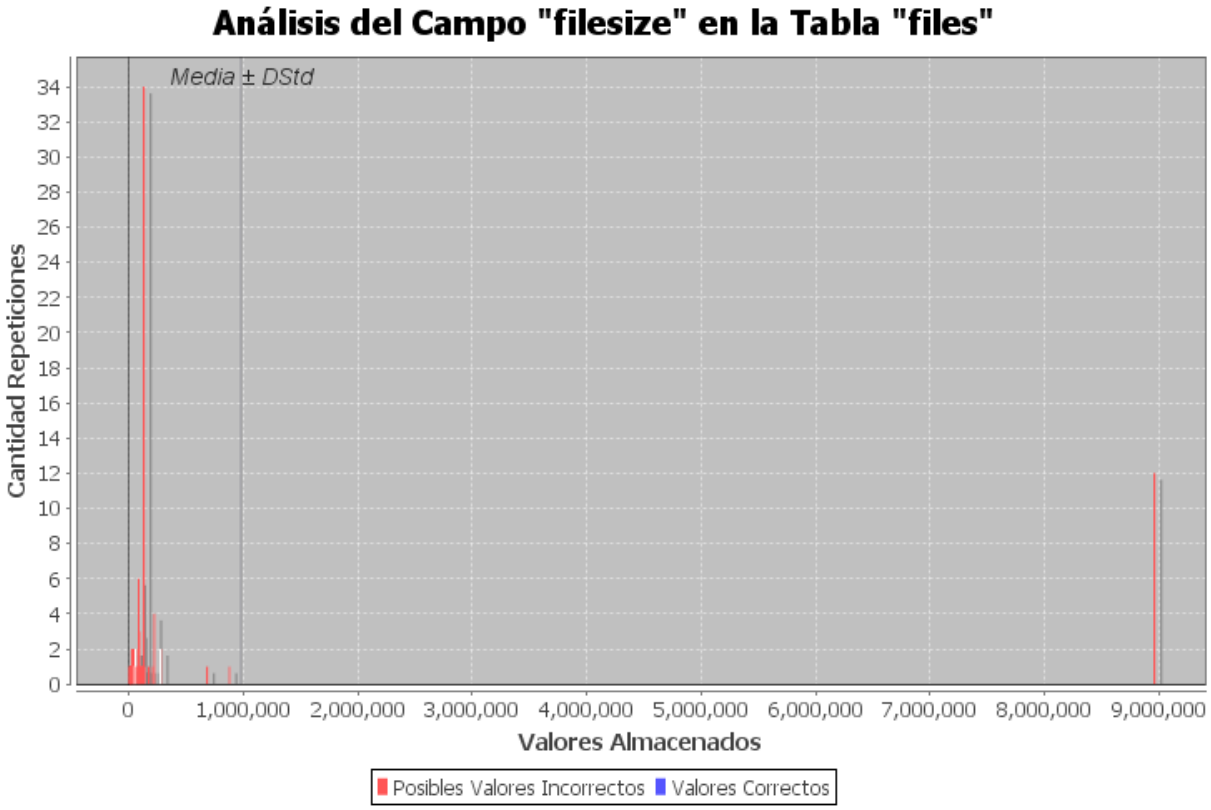
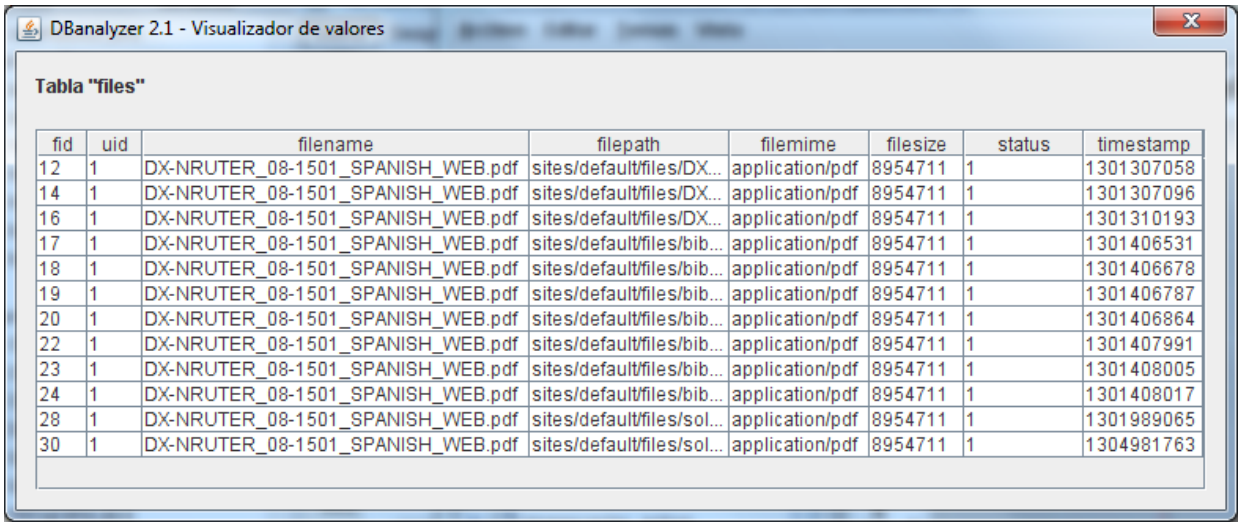
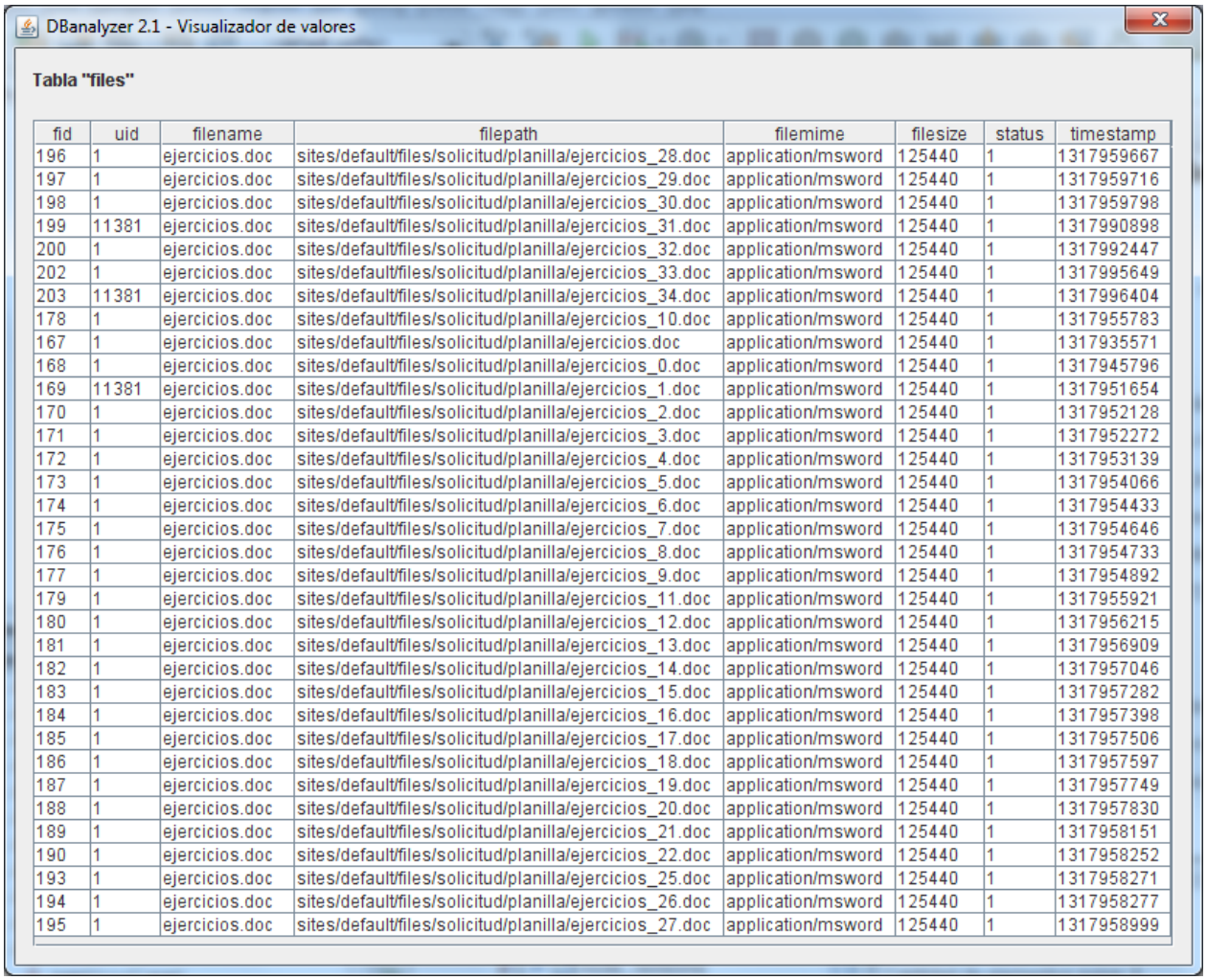


Figura 3.5 Histograma con los valores almacenados en “filesize”.



fid	uid	filename	filepath	filemime	filesize	status	timestamp
12	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/DX...	application/pdf	8954711	1	1301307058
14	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/DX...	application/pdf	8954711	1	1301307096
16	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/DX...	application/pdf	8954711	1	1301310193
17	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/bib...	application/pdf	8954711	1	1301406531
18	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/bib...	application/pdf	8954711	1	1301406678
19	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/bib...	application/pdf	8954711	1	1301406787
20	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/bib...	application/pdf	8954711	1	1301406864
22	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/bib...	application/pdf	8954711	1	1301407991
23	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/bib...	application/pdf	8954711	1	1301408005
24	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/bib...	application/pdf	8954711	1	1301408017
28	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/sol...	application/pdf	8954711	1	1301989065
30	1	DX-NRUTER_08-1501_SPANISH_WEB.pdf	sites/default/files/sol...	application/pdf	8954711	1	1304981763

Figura 3.6 Tuplas donde “filesize” tiene valor 8954711.



fid	uid	filename	filepath	filemime	filesize	status	timestamp
196	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_28.doc	application/msword	125440	1	1317959667
197	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_29.doc	application/msword	125440	1	1317959716
198	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_30.doc	application/msword	125440	1	1317959798
199	11381	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_31.doc	application/msword	125440	1	1317990898
200	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_32.doc	application/msword	125440	1	1317992447
202	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_33.doc	application/msword	125440	1	1317995649
203	11381	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_34.doc	application/msword	125440	1	1317996404
178	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_10.doc	application/msword	125440	1	1317955783
167	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios.doc	application/msword	125440	1	1317935571
168	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_0.doc	application/msword	125440	1	1317945796
169	11381	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_1.doc	application/msword	125440	1	1317951654
170	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_2.doc	application/msword	125440	1	1317952128
171	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_3.doc	application/msword	125440	1	1317952272
172	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_4.doc	application/msword	125440	1	1317953139
173	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_5.doc	application/msword	125440	1	1317954066
174	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_6.doc	application/msword	125440	1	1317954433
175	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_7.doc	application/msword	125440	1	1317954646
176	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_8.doc	application/msword	125440	1	1317954733
177	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_9.doc	application/msword	125440	1	1317954892
179	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_11.doc	application/msword	125440	1	1317955921
180	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_12.doc	application/msword	125440	1	1317956215
181	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_13.doc	application/msword	125440	1	1317956909
182	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_14.doc	application/msword	125440	1	1317957046
183	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_15.doc	application/msword	125440	1	1317957282
184	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_16.doc	application/msword	125440	1	1317957398
185	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_17.doc	application/msword	125440	1	1317957506
186	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_18.doc	application/msword	125440	1	1317957597
187	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_19.doc	application/msword	125440	1	1317957749
188	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_20.doc	application/msword	125440	1	1317957830
189	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_21.doc	application/msword	125440	1	1317958151
190	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_22.doc	application/msword	125440	1	1317958252
193	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_25.doc	application/msword	125440	1	1317958271
194	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_26.doc	application/msword	125440	1	1317958277
195	1	ejercicios.doc	sites/default/files/solicitud/planilla/ejercicios_27.doc	application/msword	125440	1	1317958999

Figura 3.7 Tuplas donde “filesize” tiene valor 125440.

Campos de tipo cadena

En el campo “field_efemeride_texto_value” de la tabla “content_type_efemeride”, donde los valores que se almacenan son el contenido de la efeméride en cuestión, se puede ver que hay un elemento que se repite 2 veces (Figura 3.8).

Figura 3.8 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.

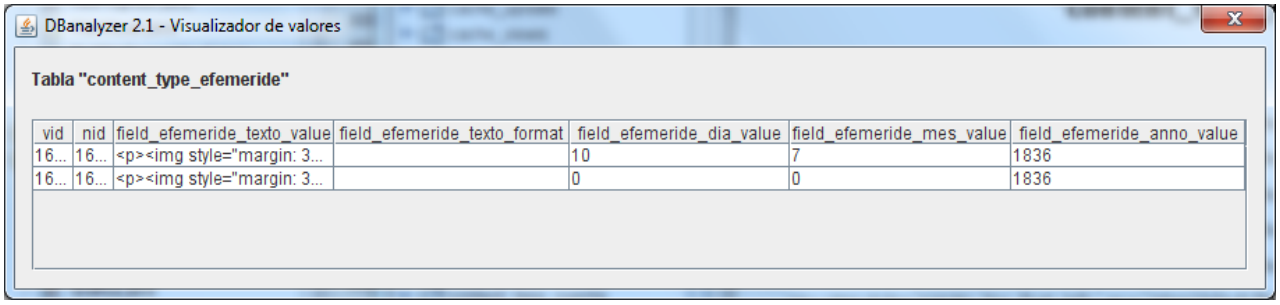


Figura 3.9 Tuplas donde “field_efemeride_texto_value” tiene el mismo valor.

Observando la Figura 3.9, donde se muestran las tuplas donde “field_efemeride_texto_value” tiene el mismo valor se puede notar que una de las efemérides tiene valores correctos, mientras la otra tiene los que corresponden al día y mes en cero. Indicando esto que está repetida y por lo tanto la segunda tupla debe eliminarse.

Algo parecido ocurre en la Figura 3.10, lo que al examinar los datos correspondientes a la cadena con dos repeticiones en la Figura 3.11, se observa que en este caso excepto el identificador, los demás valores son iguales, por lo que la efeméride está duplicada, debiéndose eliminar una de ellas.

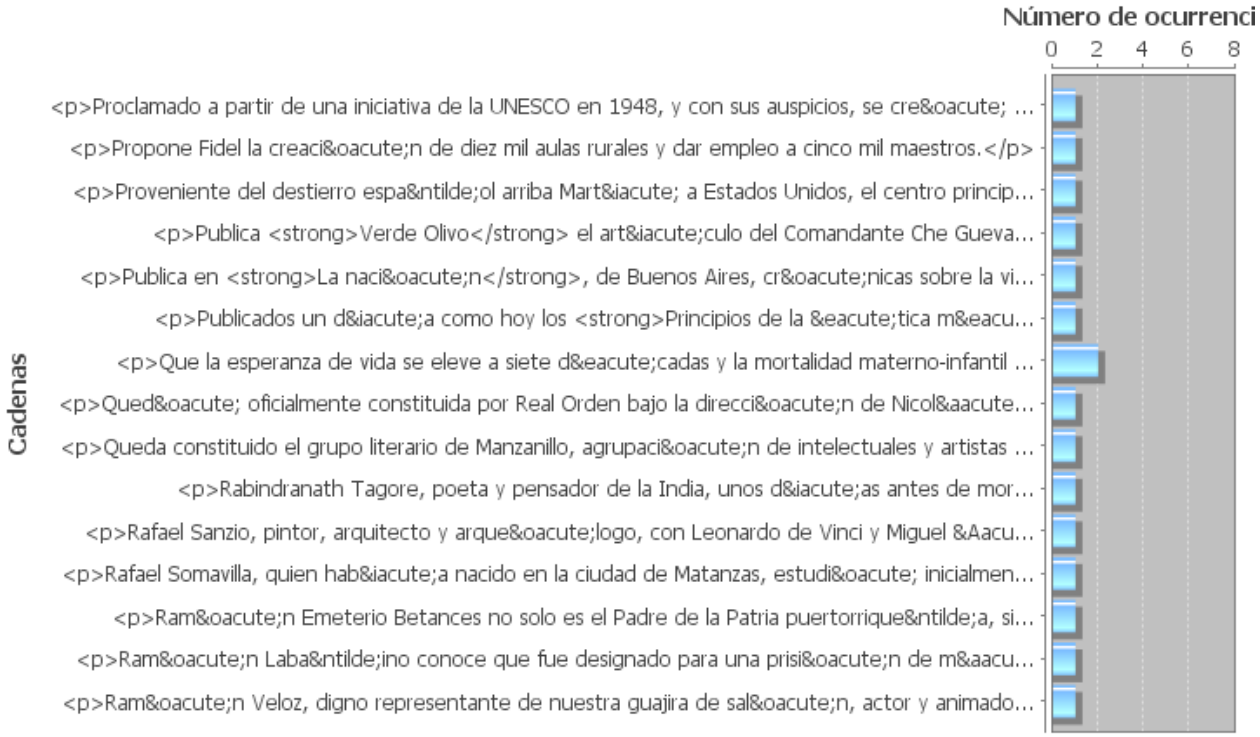


Figura 3.10 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.

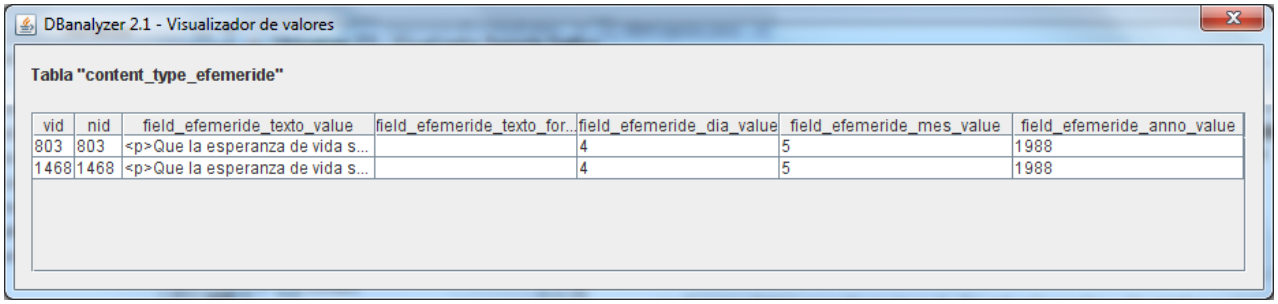


Figura 3.11 Tuplas donde “field_efemeride_texto_value” tiene el mismo valor.

En el campo “nombre” de la tabla “municipio”, en la Figura 3.12 se puede apreciar la repetición de uno de sus elementos, que representa el nombre de uno de los municipios del país, en este caso el de Artemisa, y como se aprecia en la Figura 3.13, el nombre del municipio Artemisa que pertenece a la provincia con el mismo nombre aparece repetido, por lo que debe eliminarse una de sus apariciones, siendo un elemento duplicado.

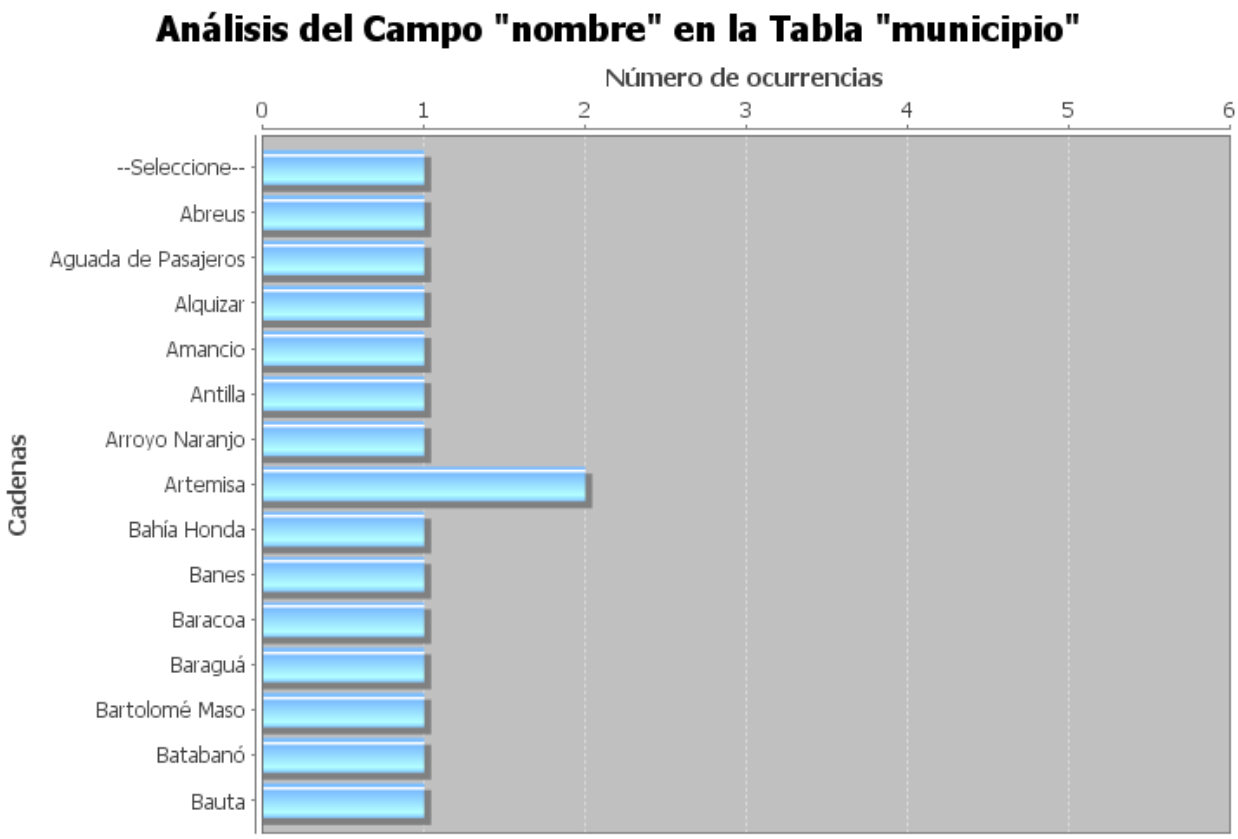


Figura 3.12 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.

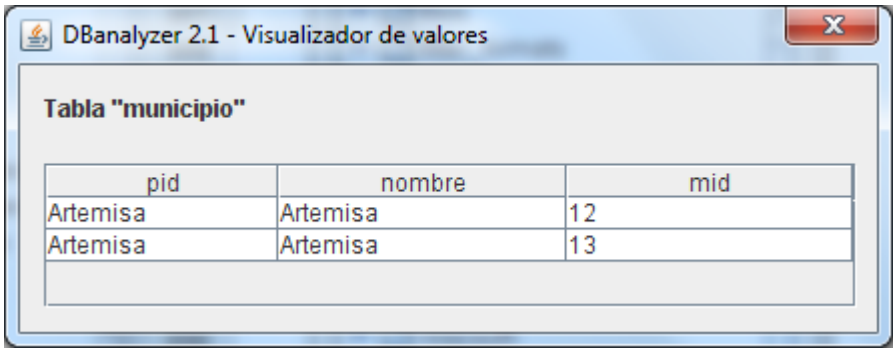


Figura 3.13 Tuplas donde tanto “pid” como “nombre” tienen el mismo valor.

En el campo “filename” de la tabla “files” (la cual se estuvo analizando en las Figuras 3.10, 3.11 y 3.12) aparece el nombre de un archivo repetido 34 veces (Figura 3.14), al verificar los datos en la tabla asociados a este fichero (Figura 3.12), se coincide en el análisis previo que se había realizado.

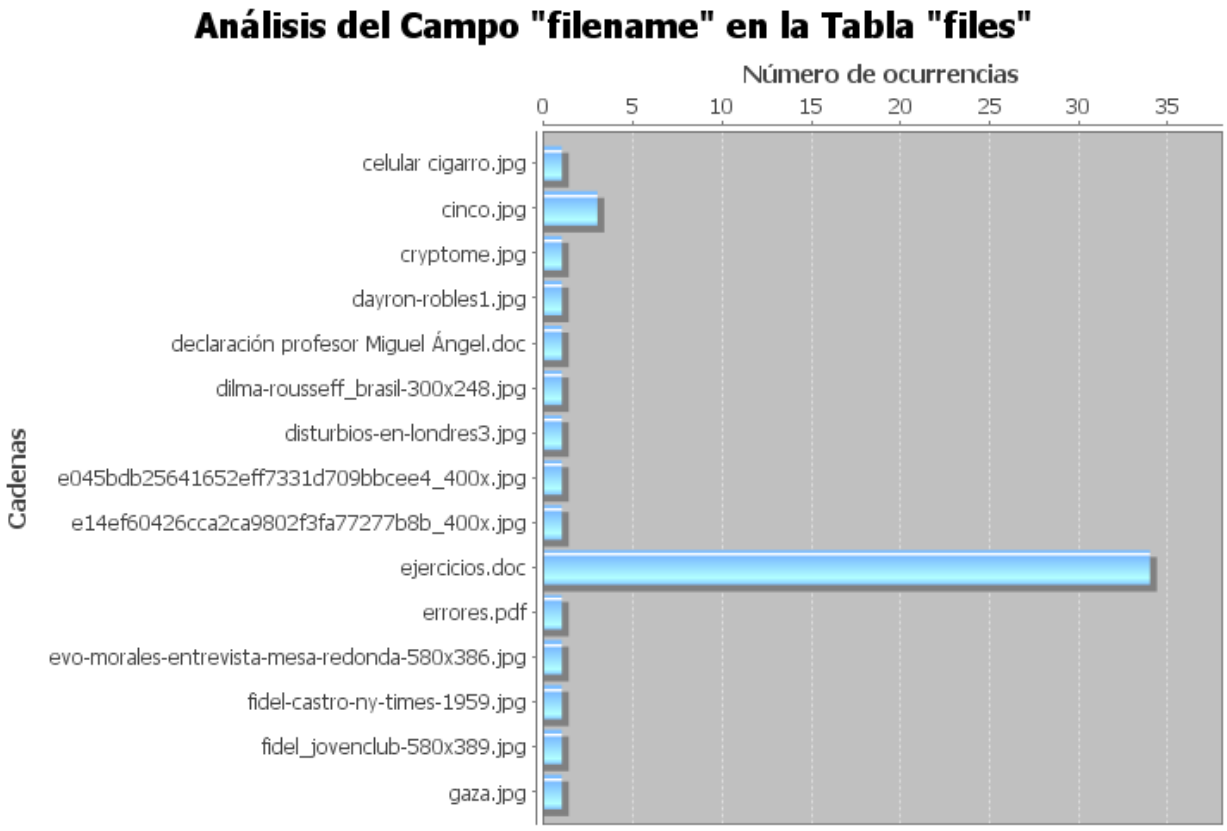


Figura 3.14 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.

3.3.2 Base de datos de portal de noticias

La base de datos del portal de noticias, se encuentra sobre el sistema gestor de base de datos MySQL en su versión 5.1.22, está compuesta por un total de 60 tablas y luego de analizarla con el DBAnalyzer se encontraron los siguientes detalles:

Campos numéricos de tipo entero

En el campo “subscribe_ip” de la tabla “pd_posts_notification_emails” (ver Figura 3.15), se puede apreciar una dispersión muy grande de los valores, pues de un total de 13 elementos con una cardinalidad de 11, el menor elemento es 0 y el mayor 171247608.

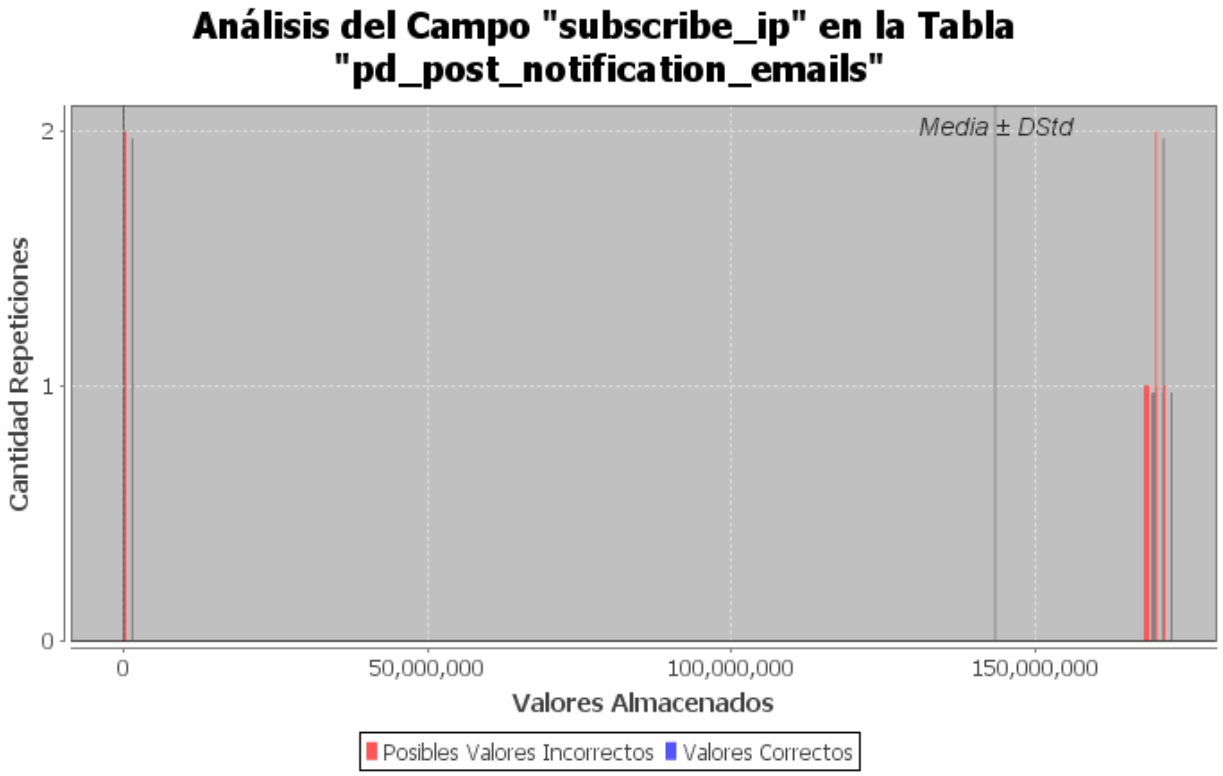


Figura 3.15 Histograma con los valores almacenados en “subscribe_ip”.

En el campo “term_id” de la tabla “pd_terms”, mostrado en la Figura 3.16, se puede apreciar una marcada dispersión en los valores que tiene representado, con una cardinalidad de 366 elementos y valores entre 1 y 456822. Lo curioso de este campo es que representa un valor llave y es de tipo incremental (según se pudo apreciar en la estructura de la base de datos) y los elementos están concentrados en dos fragmentos principales, uno con los valores entre 1 y 133 (Figura 3.17) y otro con los valores entre 456566 y 456822 (Figura 3.18), por lo que el espacio comprendido entre el 134 y 456565 no tiene ningún valor.

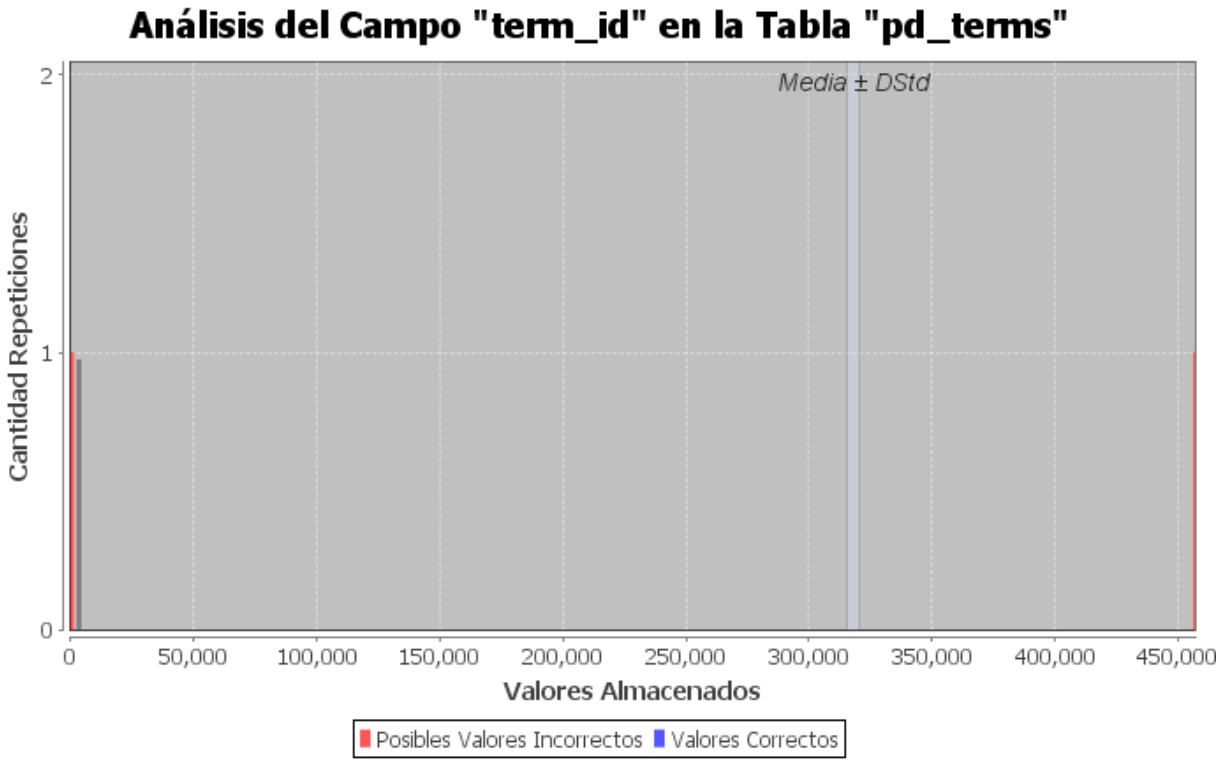


Figura 3.16 Histograma con los valores almacenados en “term_id”.

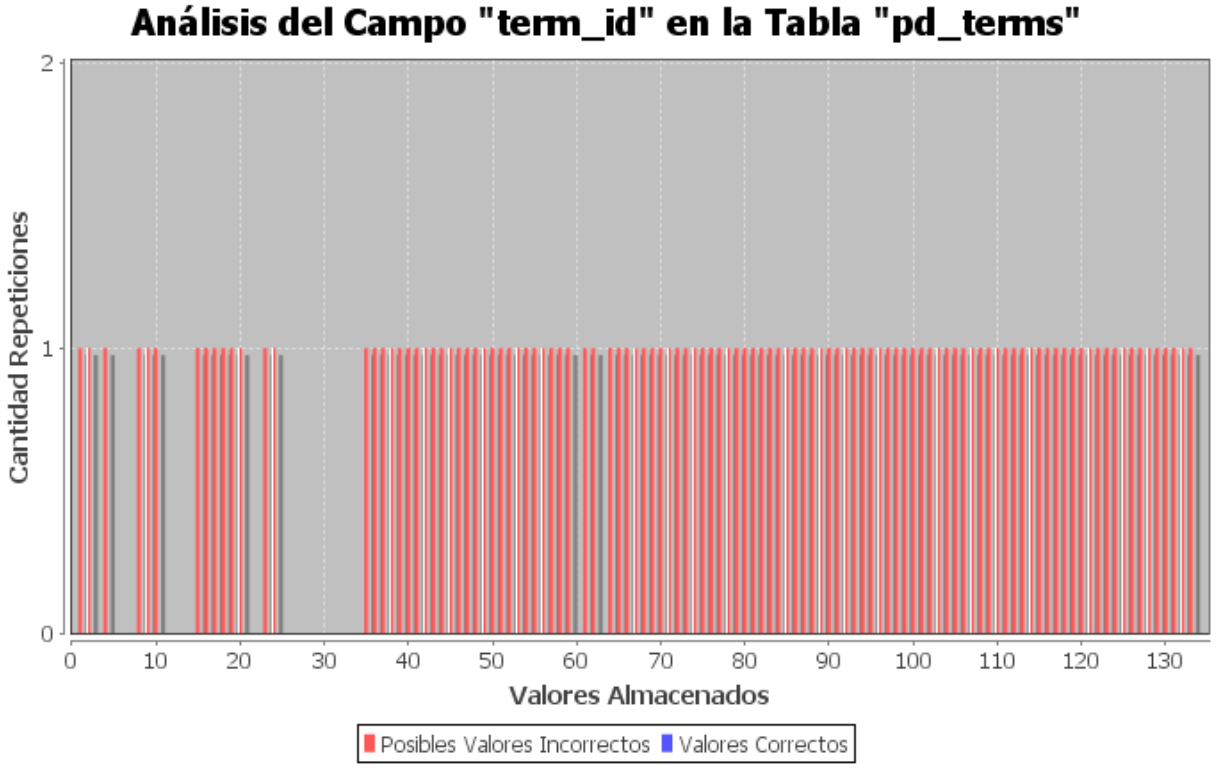


Figura 3.17 Histograma con el acercamiento al primer fragmento de los valores almacenados en “term_id”.

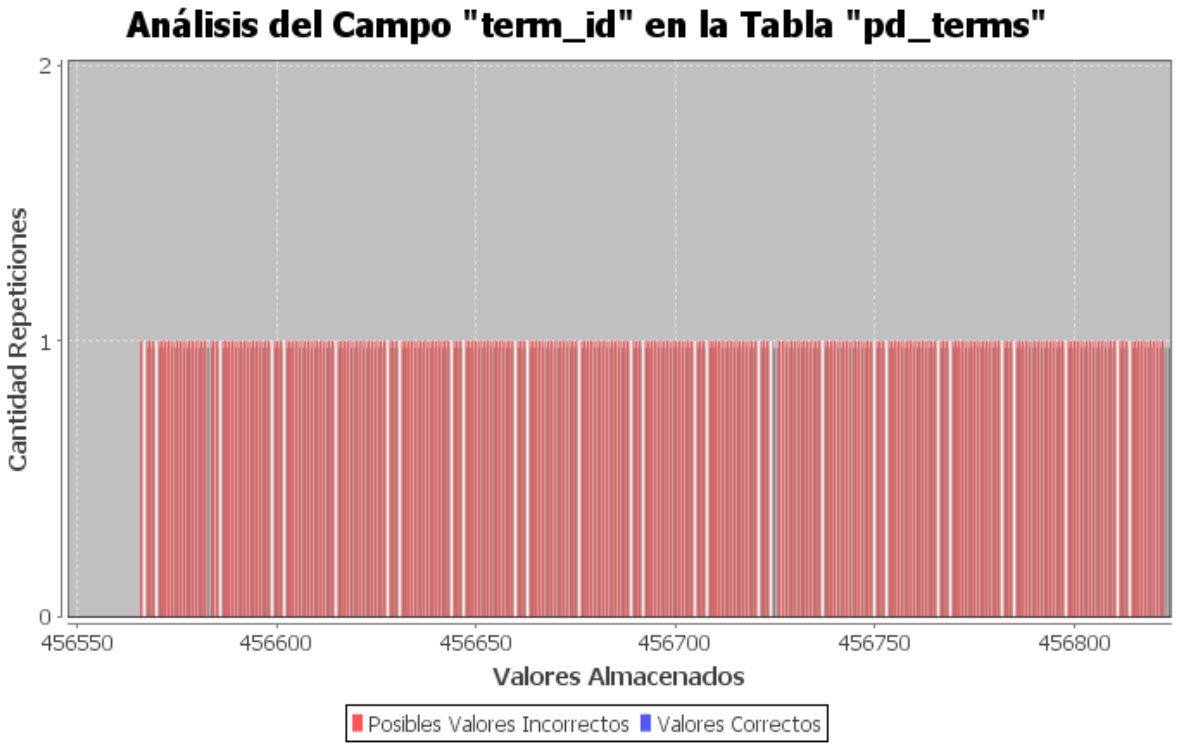


Figura 3.18 Histograma con el acercamiento al fragmento final de los valores almacenados en “term_id”.

Campos de tipo cadena

En el campo “post_title” de la tabla “pd_posts”, mostrado en la Figura 3.19, que almacena los títulos de los posts hechos, y tiene un total de 3038 elementos con una cardinalidad de 2187, se puede apreciar como uno en particular tiene 28 repeticiones. En la tabla correspondiente a los valores asociados a este título en cuestión (ver Figura 3.20) se identifica un patrón, en el cual según el campo “post_type” que muestra el tipo de post que representa cada elemento, uno en particular tiene el valor “post” y el resto “revision”, mirando además el campo “post_status” se aprecia que el primero mencionado tiene valor “publish” y el resto “inherit” y en el campo “post_name” pasa lo mismo, uno con su nombre particular, y el resto con un nombre autogenerado formado por el identificador del post que le dio origen y seguido de la cadena “revision” con un numeral al final que indica la versión de la revisión. En este caso particular se tendría el post original, 26 revisiones y un guardado automático. Si este patrón de publicación de los posts se repite con frecuencia esta tabla almacenará demasiados valores que no son publicados en el sitio. Esto eventualmente hará que el acceso a las páginas que si representan los posts sean lentos, debido al mismo proceso de búsqueda y por la gran cantidad de elementos. Lo antes mencionado no es un error en sí, pero conllevaría a disminuir el tiempo de respuesta del sitio, aquí una posible solución pudiera ser que los elementos que no van a ser publicados (los del proceso de edición y los no aceptados) estén en una tabla diferente para que los tengan a modo de estadística, pero que no influya con los accesos a los elementos

válidos.

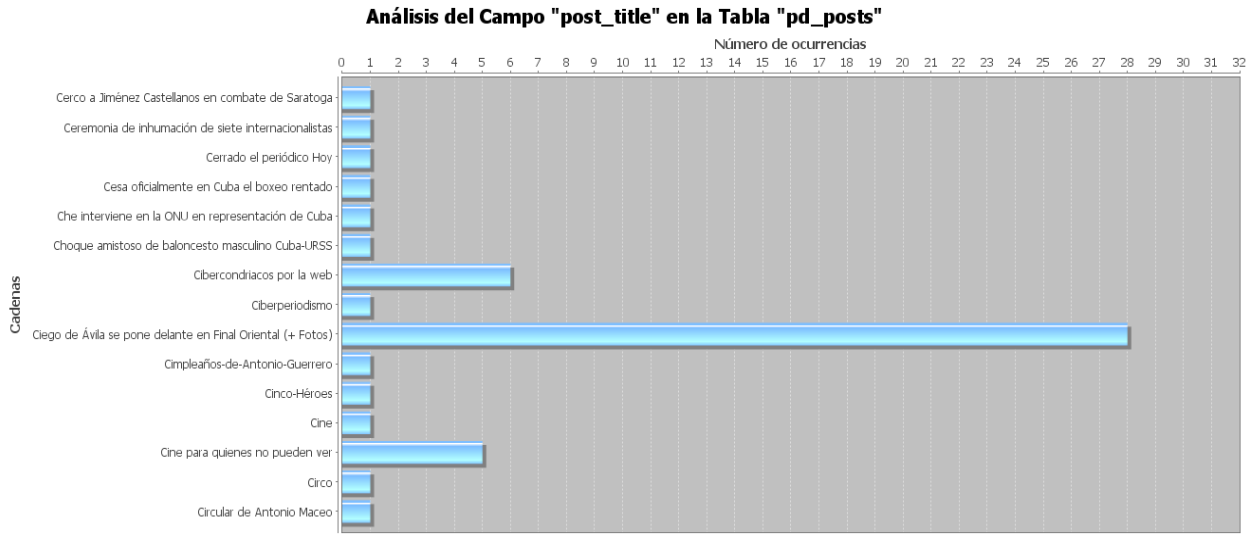
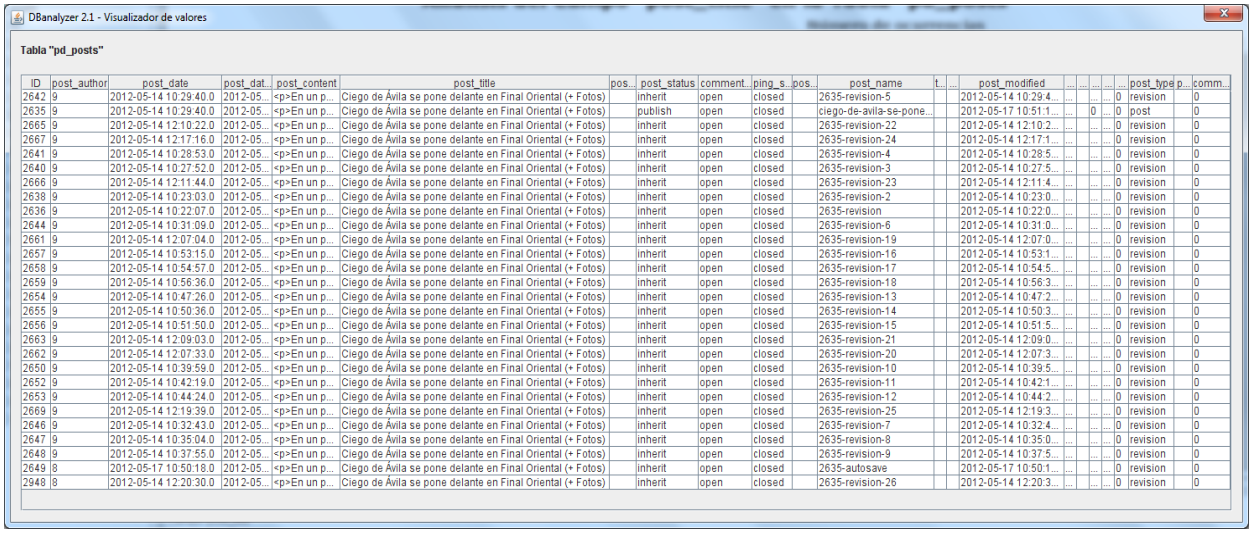


Figura 3.19 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.



ID	post_author	post_date	post_dat...	post_content	post_title	pos...	post_status	comment...	ping_s...	pos...	post_name	t...	post_modified	...	...	...	...	...	post_type	p...	comm...
2642	9	2012-05-14 10:29:40.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-5		2012-05-14 10:29:4...	...			...	0	revision		0
2635	9	2012-05-14 10:29:40.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		publish	open	closed		ciego-de-avila-se-pone...		2012-05-17 10:51:1...	...		0	...	0	post		0
2665	9	2012-05-14 12:10:22.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-22		2012-05-14 12:10:2...	...			...	0	revision		0
2667	9	2012-05-14 12:17:16.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-24		2012-05-14 12:17:1...	...			...	0	revision		0
2641	9	2012-05-14 10:28:53.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-4		2012-05-14 10:28:5...	...			...	0	revision		0
2640	9	2012-05-14 10:27:52.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-3		2012-05-14 10:27:5...	...			...	0	revision		0
2666	9	2012-05-14 12:11:44.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-23		2012-05-14 12:11:4...	...			...	0	revision		0
2638	9	2012-05-14 10:23:03.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-2		2012-05-14 10:23:0...	...			...	0	revision		0
2636	9	2012-05-14 10:22:07.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision		2012-05-14 10:22:0...	...			...	0	revision		0
2644	9	2012-05-14 10:31:09.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-6		2012-05-14 10:31:0...	...			...	0	revision		0
2661	9	2012-05-14 12:07:04.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-19		2012-05-14 12:07:0...	...			...	0	revision		0
2657	9	2012-05-14 10:53:15.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-16		2012-05-14 10:53:1...	...			...	0	revision		0
2658	9	2012-05-14 10:54:57.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-17		2012-05-14 10:54:5...	...			...	0	revision		0
2659	9	2012-05-14 10:56:36.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-18		2012-05-14 10:56:3...	...			...	0	revision		0
2654	9	2012-05-14 10:47:26.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-13		2012-05-14 10:47:2...	...			...	0	revision		0
2655	9	2012-05-14 10:50:36.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-14		2012-05-14 10:50:3...	...			...	0	revision		0
2656	9	2012-05-14 10:51:50.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-15		2012-05-14 10:51:5...	...			...	0	revision		0
2663	9	2012-05-14 12:09:03.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-21		2012-05-14 12:09:0...	...			...	0	revision		0
2662	9	2012-05-14 12:07:33.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-20		2012-05-14 12:07:3...	...			...	0	revision		0
2650	9	2012-05-14 10:39:59.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-10		2012-05-14 10:39:5...	...			...	0	revision		0
2652	9	2012-05-14 10:42:19.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-11		2012-05-14 10:42:1...	...			...	0	revision		0
2653	9	2012-05-14 10:44:24.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-12		2012-05-14 10:44:2...	...			...	0	revision		0
2669	9	2012-05-14 12:19:39.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-25		2012-05-14 12:19:3...	...			...	0	revision		0
2646	9	2012-05-14 10:32:43.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-7		2012-05-14 10:32:4...	...			...	0	revision		0
2647	9	2012-05-14 10:35:04.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-8		2012-05-14 10:35:0...	...			...	0	revision		0
2648	9	2012-05-14 10:37:55.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-9		2012-05-14 10:37:5...	...			...	0	revision		0
2649	8	2012-05-17 10:50:18.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-autosave		2012-05-17 10:50:1...	...			...	0	revision		0
2948	8	2012-05-14 12:20:30.0	2012-05...	<p>En un p...	Ciego de Ávila se pone delante en Final Oriental (+ Fotos)		inherit	open	closed		2635-revision-26		2012-05-14 12:20:3...	...			...	0	revision		0

Figura 3.20 Tuplas donde “post_title” tiene el mismo valor.

Siguiendo la tónica de lo mencionado en el análisis del campo anterior, la Figura 3.21 muestra el campo “post_content” de la tabla “pd_post”, específicamente el contenido del post al que se hacía referencia, donde los 28 posts están distribuidos de la siguiente forma por la diferencia del texto que contienen, 1 con 18 repeticiones, 2 con 2 repeticiones y 6 con 1 repetición. Este campo tiene un total de 2578 elementos y una cardinalidad de 2145, que pudiera reducirse considerablemente con la recomendación planteada en el análisis anterior.

Mediante los análisis anteriores, y comprobando contra los archivos con los reportes en modo texto se evidencia la sustancial ventaja de las gráficas, llevando esto a una detección más rápida a posibles errores, tanto de diseño de las bases de datos como errores en el contenido de las tablas.

Figura 3.21 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.

En el campo “polla_answers” de la tabla “pd_pollsa” se muestran los valores que pueden tomar las encuestas que se realizan en el sitio, allí se observan dos repeticiones de la cadena “Bien” pues está referido a dos encuestas diferentes, también las cadenas “Mal” y “Mañ”, esta última mal escrita (ver Figura 3.22).

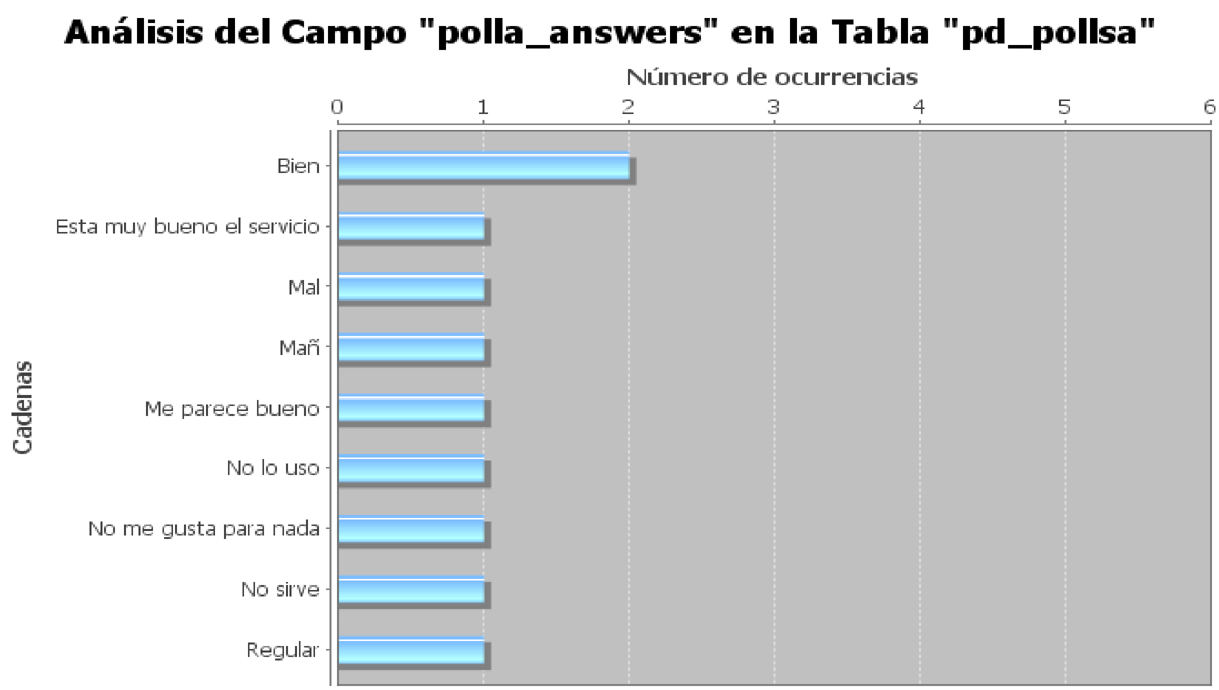


Figura 3.22 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.

En el campo “name” de la tabla “pd_terms” que hace referencia a temas que se han tocado en diversas secciones del sitio, la irregularidad que más se detecta es la de términos mal escritos, ejemplo de ello puede apreciarse en la Figura 3.23 donde aparece el término “olipmpicos” y “olímpicos”, en la Figura 3.24 con los términos “Fidel” y “Fidel Castro”, y en la Figura 3.25 con “Despagne” y “Despaigne” y con “Educacion” y “Educación”. Pudiendo cada una de las parejas

de términos anteriores unificarse, pues hacen referencia a noticias o posts diferentes donde cada uno de ellos se mencionan.

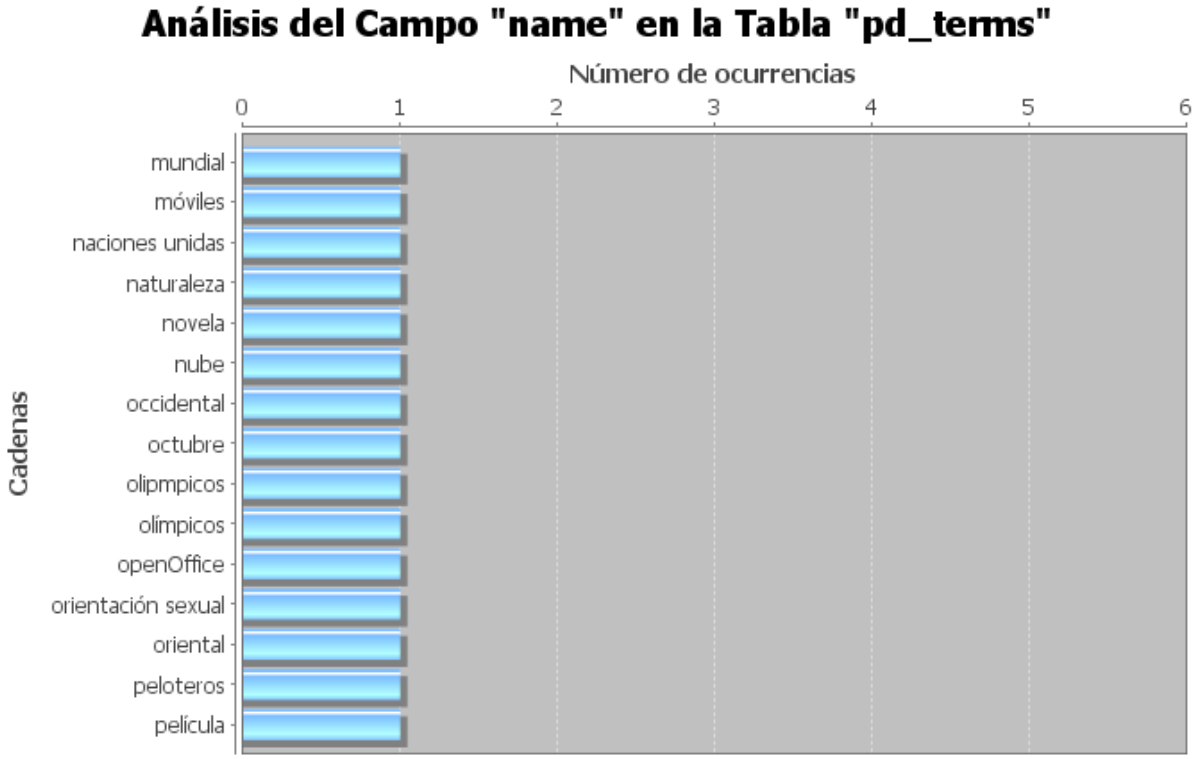


Figura 3.23 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.

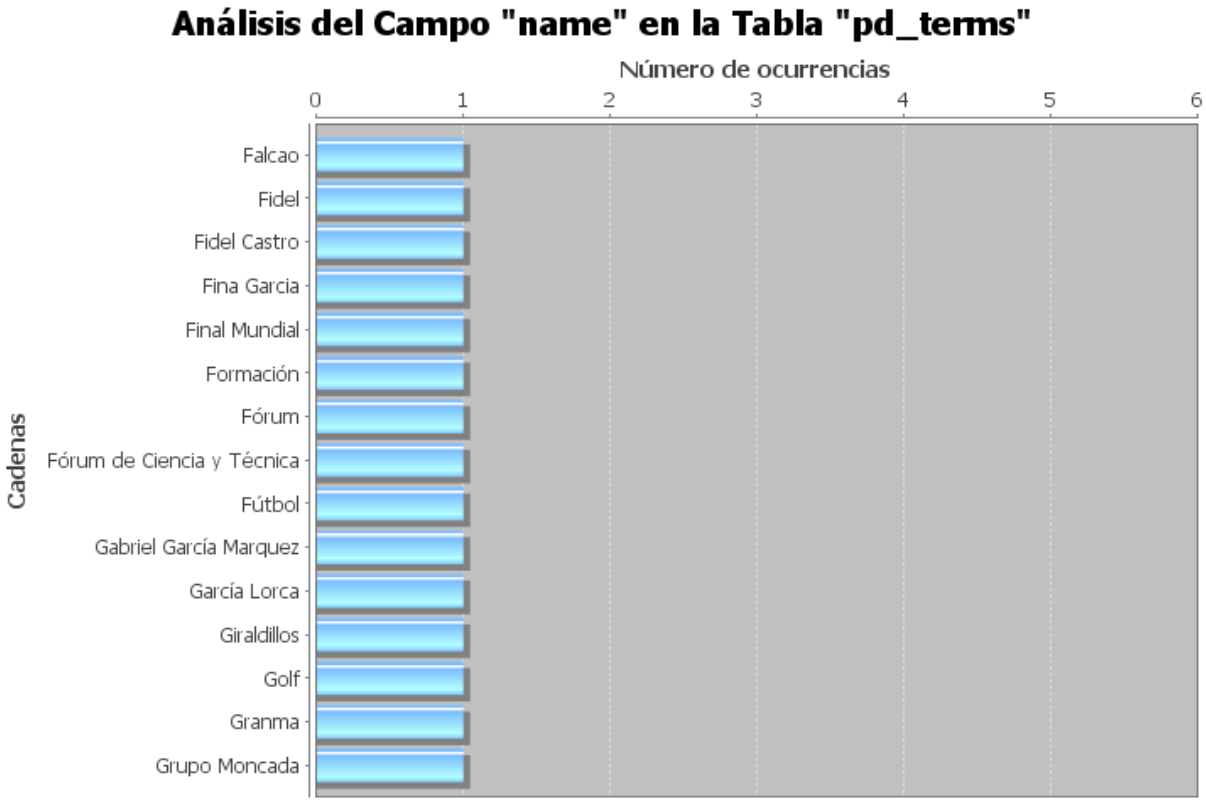


Figura 3.24 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.

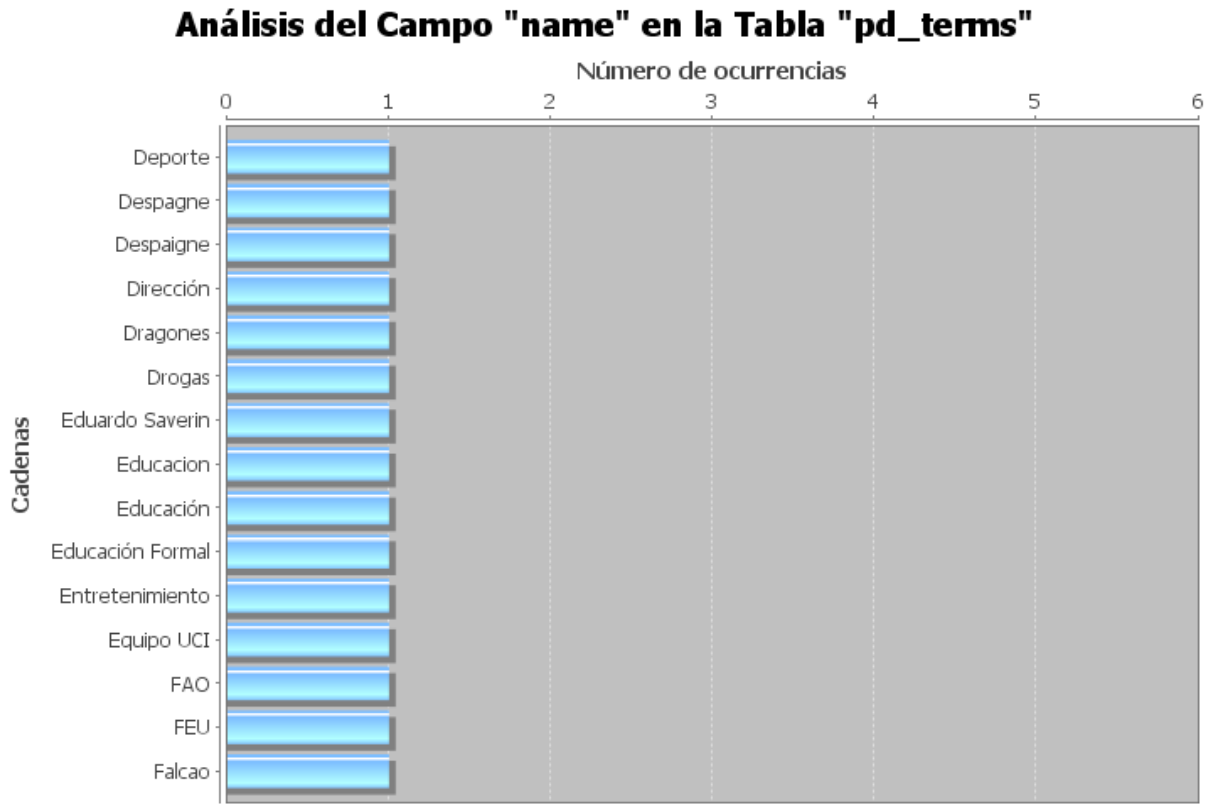


Figura 3.25 Gráfico de barras mostrando las cadenas con su cantidad de repeticiones.

Luego del análisis a las gráficas anteriores, brindadas por la herramienta implementada, deben hacerse algunas acotaciones, la primera viene dada que en el caso de los campos numéricos solo se mostraron los datos utilizando histogramas, pudiendo además representarse como diagramas de dispersión y gráficos de barras. Donde los histogramas y los diagramas de dispersión son las representaciones que mejor representan a los datos numéricos por mostrarlos en orden y hacer un tratamiento visual con los elementos que contenga en dependencia de los cálculos estadísticos realizados con tal efecto. Para los campos de tipo cadena los gráficos de barras son la mejor forma de representarlos y en el caso de la herramienta es como único se visualizarán.

Este análisis ha permitido encontrar una serie de elementos dentro de los valores en diversos campos dentro de las bases de datos analizadas, que han ayudado a mejorar la calidad de los datos contenidos. Siendo esto posible gracias a la representación gráfica de ellos.

Conclusiones

Con el desarrollo del presente trabajo se verifica que los objetivos fueron cumplidos y se arribó a las siguientes conclusiones:

1. Con la implementación de la nueva versión de la herramienta DBAnalyzer se posibilita una mejor detección de los errores mediante los reportes gráficos.
2. Los diferentes gráficos generados por el análisis muestran de una forma más clara los datos existentes que mediante el reporte en modo texto.
3. Con la utilización de la nueva versión de la herramienta, se pueden detectar errores existentes de forma más rápida, por lo que el proceso de limpieza de datos podrá hacerse en un menor espacio de tiempo y con un mejor resultado.

Recomendaciones

Se recomienda:

1. Estudiar otros mecanismos para almacenar los reportes de forma tal que no se produzca el desbordamiento de memoria en los casos de bases de datos muy pobladas.
2. Introducir la herramienta en los departamentos productivos para mejorar la calidad de los productos que trabajen con bases de datos.
3. Adicionarle nuevos tipos de gráficas de forma tal que permitan una mejor interpretación de los datos.
4. Adicionarle técnicas de inteligencia artificial que permitan la realización de un reporte automático luego del análisis, con los campos con posibles problemas, sin la intervención de un especialista.

Referencias Bibliográficas

[1] BRESCIANI, Sabrina, BLACKWELL, Alan F. and EPPLER, Martin, 2008. A Collaborative Dimensions Framework: Understanding the mediating role of conceptual visualizations in collaborative knowledge work. In: 41st Hawaii International Conference on System Sciences (HICCS 08). 2008. pp. 180–189.

[2] MORELL, Alberto and PÉREZ, Carlos, 2006. Biblioteca de módulos de visualización de fluidos para OpenDX. 2006.

[3] RYNE, Theresa-Marie and MACEACHREN, Alan, 2004. Visualizing Geospatial Data. In: ACM SIGGRAPH 2004 Course #30. 2004.

[4] HANSEN, Charles D. and JOHNSON, Chris R., 2005. The visualization handbook. S.l.: Elsevier. ISBN 0-12-387582-X. 2005.

[5] THEISEL, Holger, 2000. Scientific Visualization. 2000.

[6] PÉREZ-RISQUET, Carlos and ORTEGA-CAMACHO, J.C., 2005. Modelación de datos para la visualización científica. In: Compumat 2005. La Habana. 2005.

[7] KEIM, D.A., 2002. Information Visualization and Visual Data Mining. In: IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS 7 NO. 1. 2002.

[8] EICK, Stephen G., 2000. Visualizing Multi-Dimensional Data. In: ACM SIGGRAPH. 2000.

[9] SALGADO-MILÁN, E., 2003. Visualization Techniques. Germany. Rostock. 2003.

[10] BERGERON, R.D. and GRINSTEIN, G., 1989. A Reference Model for the Visualization of Multidimensional Data. In: Proceedings Eurographics ’89. Hamburg. 1989. p. 393–399.

[11] BRODLIE, K.W., CARPENTER, L.A., GALLOP, J.R., HUBBOLD, R.J., MUMFOLD, A.M., OSLAND, C.D. and QUARENDON, P., 1992. Scientific Visualization. S.l.: Springer-Verlag. 1992.

[12] SCHUMANN, H. and MÜLLER, W., 2000. Visualisierung Grundlagen und allgemeine Methoden. S.l.: Springer-Verlag. 2000.

[13] WIJK, J.J.v. and LIERE, R.D.v., 1993. Hyperslice. In: IEEE Visualization ’93. IEEE Computer Society Press. Los Alamitos. 1993. pp. 119–125.

[14] ONG, H.L. and LEE, H.Y., 1996. Software report WInViz -a visual data analysis tool. In: Computation & Graphics 20(1). 1996. pp. 83–84.

[15] CLEVELAND, W.S., 1993. Visualizing Data. In: Hobart Press, Summit New Jersey. 1993.

[16] INSELBERG, A. and DIMSDALE, B., 1990. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In: Visualization 90. San Francisco. 1990. p. 361–370.

[17] MATLAB, 2004. Ayuda del Matlab 7.0.

[18] ANDREWS, D.F., 1972. Plots of high dimensional data. Biometric. 1972.

[19] CUI, Q., WARD, M.O. and RUNDENSTEINER, E.A., 2005. Enhancing Scatterplot Matrices for Data with Ordering or Spatial Attributes. 2005.

[20] WARD, M.O., 2002. A Taxonomy of Glyp Placement Strategies for Multidimensional Data Visualization. Computer Science Department. Worcester Polytechnic Institute. 2002.

[21] ANDREWS, K., 2005. Information Visualisation.

[22] XIE, Z., HUANG, S., WARD, M.O. and RUNDENSTEINER, E.A., 2006. Exploratory Visualization of Multivariate Data with Variable Quality. In: IEEE. 2006

[23] YANG, J., PATRO, A., HUANG, S., MEHTA, N. and WARD, M.O., 2003. Value and Relation Display for Interactive Exploration of High Dimensional Datasets. 2003.

[24] KEIM, D.A., 2000. Designing Pixel-Oriented Visualization Techniques: Theory and Applications. In: IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS 6. 2000.

[25] ANKERST, M., KEIM, D.A. and KRIEGEL, P., 1996. “Circle Segments”: A Technique for Visually Exploring Large Multidimensional DataSet. In: Visualization. San Francisco. 1996.

[26] HEALEY, C.G. and ENNS, J.T, 2012. Attention and Visual Memory in Visualization and Computer Graphics. In: IEEE Transactions on Visualization and Computer Graphics 18. 2012. pp. 1170–1188.

[27] WARE, C., 2004. Information Visualization. Perception for Design. San Francisco: Morgan Kaufmann. 2004.

[28] AVS, 2009. AVS Applications Guide. Advanced Visual Systems Inc.

[29] AMIRA, 2008. Amira Visualize Analyze Present.

[30] YOUNG, M., ARGIRO, D. and KUBICA, S., 1995. Cantata: Visual programming environment for the Khoros system. In: Computer Graphics 29. Vol. 2, pp. 22–24.

[31] VIS5D, 1998. Vis5D Version 5.0.

[32] LARMAN, C., 2004. UML y Patrones. Introducción al análisis y diseño orientado a objetos.

[33] JACOBSON, Ivar, BOOCH, Grady and RUMBAUGH, James, 2000. El Proceso Unificado de Desarrollo de software. S.l.: Adison-Wesley. 2000.

[34] Increase productivity and enhance communication and collaboration efficiency by using UML. In: [online]. [Accessed 20 September 2012 a]. Available from: http://www.visual-paradigm.com.

[35] Java Help Center. In: [online]. [Accessed 20 September 2012 b]. Available from: http://www.java.com/en/download/help/index.xml.

[36] RAHM, E. and DO, H.H., 2000. Data Cleaning: Problems and Current Approaches. In: Data Engineering Bulletin 23. 2000. pp. 3–13.

[37] KIMBALL, R., 2000. Is Your Data Correct? In: Intelligent Enterprise. pp. 22.

Bibliografía

[1] BLACKWELL, Alan, PHAAL, R., EPPLER, Martin and CRILLY, N., 2008. Strategy Roadmaps: New Forms, New Practices. In: Fifth International Conference on the Theory and Application of Diagrams. S.l.: s.n. 2008.

[2] BRESCIANI, Sabrina, BLACKWELL, Alan F. and EPPLER, Martin, 2008. Choosing visualisations for collaborative work and meetings: A guide to usability dimensions. In: DCRR-007. 2008.

[3] BRESCIANI, Sabrina and EPPLER, Martin, 2007. Usability of Diagrams for Group Knowledge Work: Toward an Analytic Description. In: IKNOW07. 2007.

[4] BRESCIANI, Sabrina and EPPLER, Martin, 2008. The Risks of Visualization. A Classification of Disadvantages Associated with Graphic Representations of Information. In: ICA Working Paper # 1. 2008.

[5] BRODLIE, K.W. and NOOR, Nurul Mohd, 2007. Visualization Notations, Models and Taxonomies. In: Theory and Practice of Computer Graphics. 2007.

[6] BURKHARD, Remo A. and EPPLER, Martin, 2007. Visual Representations in Knowledge Management: framework and cases. In: Journal of Knowledge Management. 2007. Vol. 4, no. 11, pp. 112–122.

[7] BURKHARD, Remo Aslak, ANDRIENKO, Gennady, ANDRIENKO, Natalia, DYKES, Jason, KOUTAMANIS, Alexander, KIENREICH, Wolfgang, PHAAL, Robert, BLACKWELL, Alan, EPPLER, Martin and HUANG, Jeffrey, 2007. Visualization Summit 2007: ten research goals for 2010. In: Information Visualization. 2007. Vol. 6, pp. 169–188.

[8] CHI, Ed H., 2000. A Taxonomy of Visualization Techniques using the Data State Reference Model. In: 2000.

[9] EPPLER, Martin, 2006. A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. In: Information Visualization. 2006. no. 5, pp. 202–210.

[10] EPPLER, Martin, 2007. Toward a visual turn in collaboration analysis? In: Building Research & Information 35. 2007. Vol. 5, pp. 584–587.

[11] EPPLER, Martin, 2008. A Process-based Classification of Knowledge Maps. In: Journal of Knowledge and Process Management. 2008. Vol. 15, no. 1, pp. 59–71.

[12] EPPLER, Martin and BRESCIANI, Sabrina, 2008. The perils of visualization: a review of the dysfunctional effects of images for communication based on information visualization studies. In: DGPuK. 2008.

[13] EPPLER, Martin and BURKHARD, Remo A., 2004. Knowledge Visualization. Towards a New Discipline and its Fields of Application. In: ICA Working Paper # 2. 2004.

[14] EPPLER, Martin and GE, J., 2008. Communicating with Diagrams: How Intuitive and Cross-cultural are Business Graphics? In: Euro Asia Journal of Management. 2008. Vol. 18, no. 35, pp. 3–22.

[15] EPPLER, Martin, MENGIS, Jeanne and BRESCIANI, Sabrina, 2008. Seven Types of Visual Ambiguity: On the Merits and Risks of Multiple Interpretations of Collaborative Visualizations. In: Proceedings of the 12th International Conference on Information Visualization IV08. London. 2008.

[16] EWENSTEIN, B. and WHITE, J.K., 2007. Visual representations as artefacts of knowing'. In: Building Research & Information 35. 2007. Vol. 1, pp. 81–89.

[17] HELLERSTEIN, Joseph M., 2008. Quantitative Data Cleaning for Large Databases. In: UC Berkeley. 2008.

[18] KEIM, Daniel A., MANSMANN, Florian, SCHNEIDEWIND, Jörn and ZIEGLER, Hartmut, 2006. Challenges in Visual Data Analysis. In: 2006.

[19] LENGLER, Ralph and EPPLER, Martin, 2007. Towards a Periodic Table of Visualization Methods for Management. In: IASTED Proceedings of the Conference on Graphics and Visualization in Engineering. 2007.

[20] MARMOL-LACAL, Daynel, 2005. Determinación de una taxonomía de errores en los sistemas operacionales de nuestro entorno. UCLV.

[21] MENGIS, Jeanne and EPPLER, Martin, 2006. Seeing versus Arguing The Moderating Role of Collaborative Visualization in Team Knowledge Integration. In: Journal of Universal Knowledge Management. 2006. Vol. 1, no. 3, pp. 151–162.

[22] PFITZNER, Darius, HOBBS, Vaughan and POWERS, David, 2001. A Unified Taxonomic Framework for Information Visualization. In: 2nd Australian Institute of Computer Ethics Conference. 2001.

[23] QIAN, Yu and ZHANG, Kang, 2005. The Role of Visualization in Effective Data Cleaning. In: SAC'05. 2005.

[24] TORY, Melanie and MÖLLER, Torsten, 2002. A Model-Based Visualization Taxonomy. School of Computing Science. Simon Fraser University.

[25] TORY, Melanie and MÖLLER, Torsten, 2004. Rethinking Visualization: A High-Level Taxonomy. In: InfoVis04. 2004.

[26] VALDES-HERNANDEZ, Yenisel, 2008. DBAnalyzer 2.0, sistema para analizar bases de datos

libres. UCI.

[27] WENZEL, Sigrid, BERNHARD, Jochen and JESSEN, Ulrich, 2003. A taxonomy of visualization techniques for simulation in production and logistics. In: Proceedings of the 2003 Winter Simulation Conference. 2003. pp. 729–736.

[28] WHITE, J.K. and EWENSTEIN, B., 2007. Visual practices and the objects used in design. In: Building Research & Information 35. 2007. Vol. 1, pp. 18–27.

[29] WING, Winnie and CHAN, Yi, 2006. A Survey on Multivariate Data Visualization. S.l. Hong Kong University of Science and Technology.