Un Modelo para Soporte de la Investigación Asistido por Técnicas de Visión Artificial

Jorge Kamlofsky, María Lorena Bergamini

[1] CAETI - Universidad Abierta Interamericana

Av. Montes de Oca 725 – Buenos Aires – Argentina

{Jorge.Kamlofsky, Maria.Bergamini}@uai.edu.ar

Resumen

Muchos trabajos de investigación de diversas disciplinas basan sus conclusiones en el resultado de aplicar algoritmos y/o técnicas de minería de datos sobre grandes volúmenes de datos. Sin embargo, el proceso de obtención de las bases de datos de la propia investigación no es una tarea trivial. En este trabajo se presenta un modelo asistido por técnicas de visión artificial que permite analizar encuestas digitales y en papel, normaliza la carga de los resultados a la base de datos, posibilita su guarda en un único almacén, permitiendo y facilitando el posterior análisis integral y extracción de conocimiento subyacente mediante la implementación de algoritmos o técnicas de minería de datos.

1. Introducción

1.1. Trabajos relacionados

El corazón de gran parte de los trabajos de investigación consiste en el análisis de un gran caudal de datos obtenidos. Para ello, se suelen aplicar algoritmos o técnicas de minería de datos, analizar propiedades estadísticas o indicadores de diversos parámetros que forman parte de un conjunto de datos. Muchos bancos de datos son públicos y accesibles gratuitamente, mientras que otros son pagos. En general, presentan gran variedad, cantidad y calidad de datos. Por ejemplo, muchas investigaciones biomédicas, parten de búsquedas en el banco de datos médicos MedlinePlus[1] producido por la biblioteca Nacional de Medicina de EEUU [1]. Sin embargo, cuando el investigador decide encarar un nuevo proyecto, basado en ideas innovadoras, los bancos públicos de datos muchas veces no coinciden con las necesidades del investigador. En ese momento, el investigador decide recurrir a la generación de datos a los fines de satisfacer las necesidades de información del trabajo de investigación. Y se encuentra con que esto no es un trabajo trivial. Es clave el diseño de la investigación y el diseño del modelo de datos, a fines de poder ser usado y analizado mediante las técnicas disponibles.

Durante el diseño de la investigación, una tarea requerida es la definición del universo de estudio o población. A partir de ello se puede definir el tamaño de la muestra del experimento. En ese momento, surge el interrogante acerca de cómo obtener datos no sesgados, con el tamaño de la muestra requerido. Puede hacerse uso de herramientas de marketing, útiles a muchas disciplinas. Se han usado herramientas de marketing, por ejemplo, para mejorar la eficiencia de la salud pública, presentadas en [2]. Una de ellas es el marketing directo. El marketing directo es un sistema interactivo de marketing que utiliza uno o más medios publicitarios para conseguir una respuesta medible [3]. De este modo, es posible la obtención de datos relevantes, a escala necesaria.

Realizar encuestas web enviadas por correo electrónico a bases de datos de población segmentada útil a los fines del investigador, es una opción muy atractiva ya que soluciona gran parte de los inconvenientes mencionados previamente: Población sin sesgo, posibilidad de envío de la encuesta a una gran cantidad de destinatarios, y carga efectiva de las respuestas. Sin embargo, enviar gran cantidad de correos electrónicos no

[1] http://www.nlm.nih.gov/medlineplus/

asegura obtener gran cantidad de respuestas. Mientras más segmentadas las bases, menor la cantidad efectiva de correos electrónicos disponibles. Pero si el tema de la encuesta está dentro del interés del segmento de la muestra, la proporción de respuestas suele ser mayor.

La eficacia del uso del correo electrónico en campañas de marketing fue ampliamente tratado en algunos trabajos [4]. La correcta segmentación de la lista de correos electrónicos de los destinatarios, así como la pertinencia del mensaje aumenta la eficacia de respuesta de las campañas [5]. También se ha constatado que los consumidores suelen aceptar los mensajes que tienen contenidos relevantes para ellos [6]. La calidad técnica y la vistosidad gráfica del mensaje, dependen también de la cantidad y calidad de los elementos gráficos del arte final. Según el estudio realizado por Chittenden [5] el número de imágenes mejora las tasas de respuesta de las campañas de correo electrónico. El remitente y el asunto es lo primero que se ve en la bandeja de entrada. De ello depende que el mensaje se abra o no. En otros [7] se realizan recomendaciones acerca del correcto diseño y formato de e-mails que optimizan su aceptación. Los costos, características, ventajas y desventajas de los principales servicios más conocidos de envío masivo de correos electrónicos se presentan en un trabajo de Manchón y Fernandez [8].

Y luego surge un nuevo interrogante: ¿cómo cargar todo ese volumen de datos en una base de datos?

El uso de encuestas o formularios para la recolección de datos vía web ofrece una serie de ventajas por sobre las técnicas tradicionales (formularios de papel): ofrece menor dependencia en personal y equipamiento de ingreso de datos. Otra ventaja es la universalidad tecnológica: no requiere ninguna aplicación extra más allá del navegador de Internet, presentes en todos los dispositivos electrónicos de conexión a la web (PCs, tablets, celulares, etc.), con cualquier sistema operativo. También tiene bajo costo de implementación y permite la carga directa de las respuestas en una base de datos, listas para su posterior análisis por medio del uso de programas estadísticos. El formato web de las encuestas, además, incrementa la confiabilidad de los datos ingresados, disminuyendo la pérdida de datos por encuestas incompletas o datos erróneos, ya que permite el desarrollo de herramientas de control de campos de carga obligatorios y validación del tipo de dato ingresado, evitando errores de tipeo [9].

Experiencias más recientes muestran que se ha aumentado la tasa de respuestas de las encuestas web siendo ésta similar a la tasa de respuesta de las encuestas en papel. En parte, esto puede deberse a que el uso de Internet se ha hecho parte de la vida cotidiana permitiendo que esta no tan nueva tecnología se pueda utilizar para la recolección de datos vía web [9].

Los métodos tradicionales de recolección de datos basados exclusivamente en el papel tienen limitaciones inherentes y riesgos como la pérdida de datos entre su recolección y carga. Cualquier recolección de datos se vuelca luego a un formato de datos electrónicos: planilla de cálculo o base de datos, para su posterior utilización y análisis. Gracias a que el error humano de carga es inevitable, se necesita realizar una doble carga para la validación de los datos, lo que aumenta los costos de un proyecto de investigación [10]. La recolección de datos históricamente ha sido realizada por formularios pre-impresos, entrevistas telefónicas o personales. Esta metodología ha demostrado ser costosa y consume mucho tiempo, pudiendo condicionar la realización de la investigación. Por otro lado el tiempo entre la distribución del cuestionario y el análisis estadístico puede ser enlentecido por la necesidad de carga de datos posteriores [11].

A pesar de lo anterior, ésta suele ser la mejor (y muchas veces la única) opción en situaciones donde hay gran cantidad de potenciales encuestados (gracias a lo cual hay grandes posibilidades de obtener buena cantidad de datos), pero a su vez, no es posible presentar un formulario electrónico en un kiosco o terminal que contenga la encuesta. Podemos mencionar: salas de espera de centros médicos u oficinas públicas, estaciones de transporte público, bancos y entidades financieras, ingreso a centros educativos, etc.

Cargar automáticamente los datos provenientes de estos "formularios analógicos" soluciona los problemas relativos al tiempo, eficacia y costo de la carga de las respuestas de los encuestados al sistema. Usando técnicas de procesamiento de imágenes y visión artificial descriptas en [12], [13] y [14] entre otras, se pueden digitalizar las respuestas que los encuestados transcriben en los formularios de papel.

En este trabajo se presenta un modelo que permite obtener un conjunto de datos normalizados y con el formato adecuado para su análisis partiendo de datos provenientes de distintos orígenes: formularios electrónicos, y formularios en papel. Los distintos tipos de formulario se cargan a una única base de datos de transacciones (o sistemas OLTP: procesamiento de transacciones en línea) siguiendo el modelo relacional hoy vigente presentado en [15]. A los fines de optimizar su funcionamiento, la base se construye siguiendo las tres formas normales [16] lo que permite acceso a datos en forma rápida y estructurada. Los sistemas OLTP son sistemas de origen para los almacenes de datos o data warehouse (OLAP). Un data warehouse es un repositorio de datos históricos, integrados, especialmente organizados y referidos a un tema. Usar esta tecnología permite organizar los datos facilitando que los mismos se transformen en información. Es la única forma de evitar ser rico en datos y pobre en información [17]. Sobre el

almacén de datos pueden implementarse algoritmos de minería de datos [18] y así obtenerse conocimiento nuevo. En [17] se presenta un modelo de evolución del soporte de decisión basado en cuatro categorías:

- Las consultas estándar: Es el método de análisis más difundido.
- El análisis multidimensional: Ofrece diferentes perspectivas de los datos mediante el uso de dimensiones o puntos de vista como ser: tiempo, ubicación y producto.
- Segmentación: Permite dividir un conjunto de datos en segmentos. Luego, cada segmento puede analizárselo como dimensión.
- Descubrimiento del conocimiento: Usando poderosas técnicas y algoritmos se pueden encontrar patrones que permitan obtener conocimiento nuevo, oculto dentro de los datos.

1.2. Motivación y alcance

Este trabajo surgió como respuesta a diversos requerimientos, principalmente provenientes de organizaciones o instituciones relacionadas con la investigación médica entre las que se pueden mencionar: el sector de "Políticas de Salud" de la Sociedad Argentina de Cardiología y el CAESIS (Centro de Altos Estudios en Ciencias de la Salud) dependiente de la Universidad Abierta Interamericana.

Mientras que por un lado se busca la informatización de encuestas usando la tecnología de Internet para la recolección de datos, por otro lado, frecuentemente se presentan situaciones donde hay gran cantidad de posibles encuestados y no se dispone de infraestructura tecnológica acorde. En esos casos, resulta ideal disponer de formularios impresos en papel.

La integración eficaz de encuestas en papel y de origen electrónico dentro de un mismo modelo de datos permite disponer de mayor cantidad de datos, útiles no sólo para investigaciones médicas sino para investigaciones de otras disciplinas.

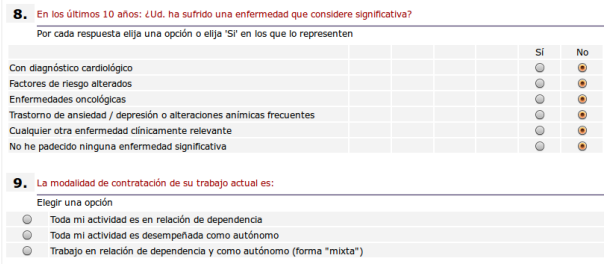
1.3. Esquema del trabajo

En la sección 2 presenta los detalles técnicos del sistema de software y del modelo de datos. La sección 3 muestra la implementación de las técnicas de procesamiento de imágenes y visión artificial para la carga automática de datos provenientes del papel al modelo de datos. La sección 4 presenta la implementación de un caso de estudio y sus resultados. Finalmente, en la sección 5 se presentan las conclusiones.

2. Descripción del Sistema

2.1. Presentación general del sistema

Consiste en un sistema de encuestas que permite la recolección de datos a partir de formularios analógicos (en papel) y digitales (en formularios web), concentra la carga en una base de datos única que permite el análisis posterior, unificado. Cada formulario consta de un conjunto de preguntas y opciones de respuesta. La figura 1 muestra un tramo de una encuesta en formato digital y en papel.



(a)

8 ¿Sufrió alguna enfermedad significativa en los últimos 10 años?	
☐ Con diagnóstico cardiológico	☐ Factores de riesgo alterados
☐ Enfermedades oncológicas	☐ Otra enfermedad relevante
☐ Trastorno de ansiedad /depresión o alteraciones anímicas frecuentes	☐ Ninguna enfermedad relevante
9 Su modalidad de contratación:	
☐ Rel. de dependencia ☐ Autónomo	☐ Mixto: rel. dependencia y autónomo

(b)

Figura 1: (a) Una porción de una encuesta en formato web. (b) La misma porción de la encuesta en formato papel.

En la figura 1 puede observarse que las mismas opciones presentes en el formulario electrónico lo están también en el formulario en papel.

Con los formularios web, la carga al sistema se produce en el momento de presionar el botón "Submit" del formulario. Con los formularios en papel, se escanea la totalidad de los formularios, generándose un archivo de imagen por encuestado-respuesta. La carga al sistema OLTP se produce luego del análisis de cada una de las imágenes con herramientas de procesamiento de imágenes.

Con origen en el sistema OLTP, el proceso de ETL (Extract, Transform, Load) permite que los datos se guarden adecuadamente en el almacén de datos des-normalizado: el data-mart, pequeño datawarehouse o cubo OLAP. Sobre estos almacenes o estructuras de almacenamiento de datos se puede realizar cómodamente el análisis de datos usando una amplia diversidad de las técnicas de diversa complejidad: desde simples consultas, hasta complejas técnicas de minería de datos.

2.2. El sistema OLTP: el modelo transaccional de datos

La figura 2 muestra el diseño lógico mediante un diagrama *patas de gallo* del modelo transaccional de los datos u OLTP. En el mismo, se observan las tablas de "preguntas", de "respuestas" y del "encuestado".

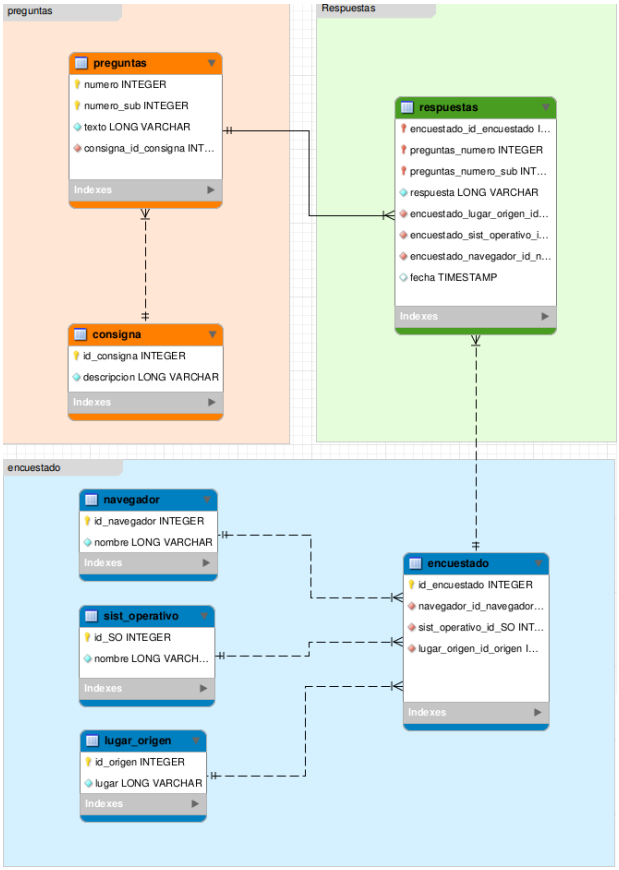


Figura 2: Diseño lógico de la base de datos transaccional.

En las respuestas obtenidas por web, el navegador y el sistema operativo para la tabla "encuestado" se obtienen con $_SERVER['HTTP_USER_AGENT'] de php. El lugar de origen se obtiene con $_SERVER['REMOTE_ADDR']. En caso de respuestas en papel, el lugar de origen se hace referencia al lugar donde se realizó la encuesta. La tabla "respuestas" incluye fecha y hora realizada la respuesta.

2.3. El sistema OLAP: El data-mart

La tecnología de data warehousing facilita el análisis muti-dimensional. Permite obtener en línea respuestas a consultas complejas. Estos sistemas son llamados OLAP (del inglés: procesamiento analítico en línea).

Un data-mart es un data-warehouse más pequeño, orientado a un tema. Son también estructuras multi-dimensionales. Una estructura multi-dimensional consiste en una gran tabla con columnas, donde cada columna se corresponde con una dimensión. Se obtienen a partir de los sistemas OLTP luego de realizarse operaciones ETL.

La figura 3 muestra un ejemplo una estructura de datos multidimensional típica en data-marts llamada "estrella".

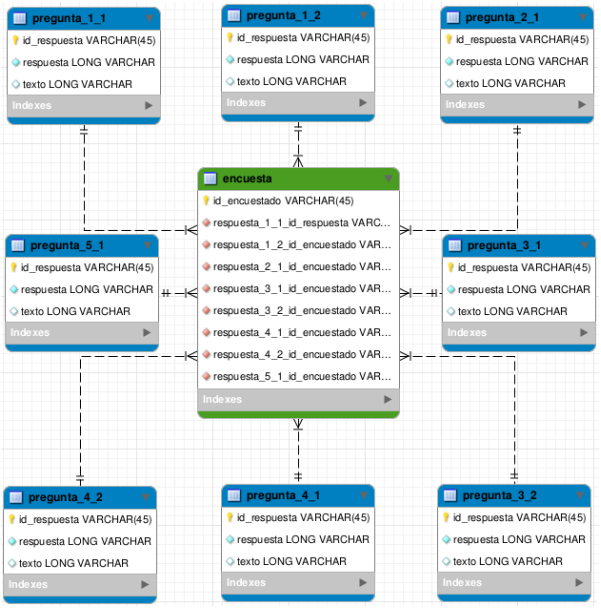


Figura 3: Una configuración de datos "estrella".

Este es un ejemplo de un data-mart de un sistema de encuestas. El núcleo o centro de la estrella tiene las respuestas de cada encuestado. Cada "punta de la estrella" representa una dimensión. En el sistema de encuestas, cada dimensión se corresponde con cada una de las preguntas.

También pueden usarse estructuras llamadas "copo de nieve" [17], en las que cada dimensión vuelve a ramificarse. En el caso de las encuestas, una dimensión corresponde a un tema dentro del cual se agrupan las preguntas.

3. La carga automática de datos provenientes de formularios en papel

3.1. Presentación del problema y esbozo de su solución.

El procedimiento de adquisición de datos a partir de encuestas contestadas en papel, consiste en la lectura del estado de casilleros o sectores específicos ubicados en posiciones predefinidas dentro de la imagen.

Se presenta el inconveniente que los dispositivos de adquisición automática de imágenes (escáneres) rara vez capturan a las imágenes exactamente en la misma posición y orientación, algo también observado en la manipulación imperfecta de hojas por parte de los operadores en escáneres manuales. Entonces, obtener acierto entre un casillero ubicado en un lugar predeterminado con su respectivo sector dentro de la imagen a partir de esto, sería pura coincidencia.

Para poder tener una referencia precisa de la orientación o inclinación del papel, se incluye en los formularios dos marcas en posiciones específicas. En particular, en este trabajo se ubicaron en el borde superior a una distancia fija de los bordes del papel.

Dichas marcas tienen la función de calibrar la ubicación de cada imagen escaneada dentro del marco de cada página. La calibración se realiza luego de la detección de las marcas.

La detección de las marcas se logra tras la comparación entre los patrones de forma de los distintos objetos obtenidos con el patrón de las marcas. Tras la detección, se obtiene la ubicación de las marcas dentro de la imagen adquirida, y con ello, se calcula el desplazamiento y cambio de orientación.

3.2. La detección de las marcas: uso de técnicas de visión artificial

La visión artificial comprende técnicas de procesamiento de imágenes para identificar un objeto dentro de una imagen digital a partir de patrones característicos. Existen distintas técnicas que permiten separar a los objetos del fondo de una imagen [12]. En este trabajo se utiliza el método de segmentación por umbralizado que asigna un valor booleano a los píxeles de acuerdo a si su valor de gris supera o no a un número definido como umbral.

Luego de separados los objetos del fondo, se pretende hallar patrones característicos de los objetos. Se destacan dos estrategias para el análisis de patrones: el análisis de sus bordes o de toda la región que comprende al objeto. El borde de una figura es de gran interés en el análisis de objetos dentro de imágenes, ya que muchas características geométricas (convexidad, tamaño, agujeros, etc.) pueden estudiarse analizando su borde. Además, el borde conforma la frontera entre el objeto digital y el fondo de la imagen. Un algoritmo simple para la obtención del borde de un objeto digital fue presentado por A. Rozenfeld en [19].

El proceso de adquisición y digitalización de una imagen introduce ruido que tiene evidente influencia en los bordes dificultando su estudio. La poligonalización de bordes reduce sus efectos y permite disminuir notoriamente la cantidad de puntos sin perder aspectos relevantes de la forma [20].

En [13] se presentó un método de representación de formas a fin de ser aplicado para el reconocimiento confiable de objetos, en forma eficiente: *el patrón de giro*. Se basa en aproximar inicialmente el borde de un objeto por un polígono y obtener luego una representación de la evolución de curvatura y concavidad del borde rectificado a lo largo de la curva.

Se considera una curva poligonal de n vértices ordenados $v_1, v_2, \ldots, v_n$. Sea $\kappa(i)$ el giro del polígono en el vértice v_i: el ángulo formado por los segmentos orientados $v_{i-1}v_i$ y v_iv_{i+1}, se define $\kappa_{acum}(i)$ al ángulo de giro acumulado desde el primer vértice hasta el vértice i.

Claramente, $\kappa_{acum}(n) = -2\pi$ si se recorre el polígono en sentido horario. Sea $l(i)$ la suma de las longitudes de los segmentos del polígono desde v_1v_2 hasta v_iv_{i+1} y L la longitud total del polígono. Entonces, $\lambda(i) = l(i)/L$ es la longitud normalizada. El patrón de giro de un polígono es la curva lineal por tramos en el plano longitud de arco-curvatura l-κ, cuyos vértices son $\beta(i) = (\lambda(i), \kappa_{acum}(i))$, siendo El punto inicial de esta curva se toma con $\lambda(0) = 0$, $\kappa_{acum}(0) = 0$. También puede pensarse esta curva como la gráfica lineal por tramos de la función $\kappa_{acum}(\lambda)$ con $\lambda \in [0,1]$, donde ocurre que $\kappa_{acum}(0) = 0$ y $\kappa_{acum}(1) = -2\pi$.

El análisis de la orientación de las formas, permite identificar puntos de inicio adecuados. Así, la representación constituye un patrón que describe la forma, independientemente de la posición, rotación y escala.

Basándose en esta representación normalizada se define una distancia entre formas. Llamamos distancia entre dos patrones κ_1 y κ_2 al valor:

$$D(\kappa_1, \kappa_2) = \int_0^1 (\kappa_1(\lambda) - \kappa_2(\lambda))^2 \, d\lambda$$

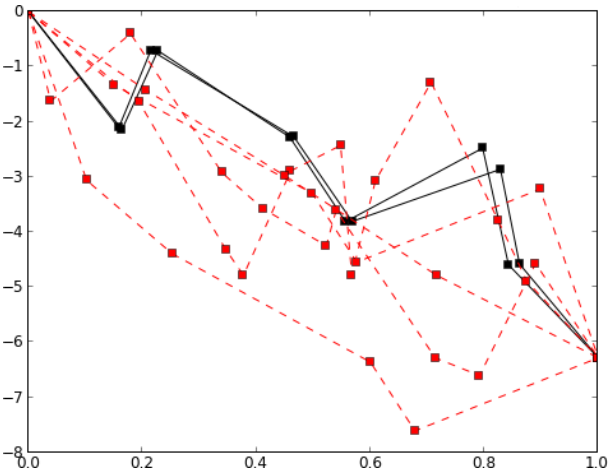
Encontrar un objeto (llamado "modelo") dentro de una imagen, entonces, consiste en comparar sus patrones de curvatura y medir la distancia entre ellos y el patrón modelo. La localización es positiva si la distancia obtenida no supera un umbral predefinido.

La figura 4 muestra un ejemplo del patrón de giro aplicado a la identificación de dos marcas dentro de una imagen digital. El patrón modelo de las marcas se encuentra almacenado en la base de datos del sistema. Para identificar a un objeto de la figura 4 (a) como uno de los elementos almacenados, se mide uno a uno la distancia entre el patrón modelo y el patrón de cada objeto de la imagen. Su resultado se lo compara con un umbral predefinido. Y se habrá reconocido al objeto, si la distancia medida está por debajo del umbral. En la figura 4 (b), se observan los patrones de giro de todos los objetos. En particular, los patrones en trazo negro continuo indican que se ha identificado al objeto modelo entre los distintos objetos.

3.3. La calibración de la imagen: re-ubicación de los píxeles de la imagen

Mediante una aplicación de reconocimiento de formas se detecta la existencia de las formas de las marcas dentro de la imagen. Se devuelve también, la ubicación de dichas marcas. Luego, mediante comparación con la ubicación de cada marca en la imagen modelo, puede obtenerse el vector de traslación, el factor de escala y el ángulo de rotación con los que el escáner transformó la imagen.

(a)



(b)

Figura 4: (a) Una imagen conteniendo dos marcas: flechas hacia arriba, y un texto. (b) Los patrones de todos los objetos. Entre ellos, en negro, los patrones de las marcas.

Sean las marcas A y B. Sean $A_0 = (x_{A0}, y_{A0})$ y $B_0 = (x_{B0}, y_{B0})$ las coordenadas del punto extremo superior de cada marca en la imagen modelo y $A_1 = (x_{A1}, y_{A1})$ y $B_1 = (x_{B1}, y_{B1})$ sus correspondientes ubicaciones en la imagen escaneada. Sean t el vector de traslación, e el factor de escala y α el ángulo de rotación. t se obtiene haciendo: $t = (x_{A1} - x_{A0}, y_{A1} - y_{A0})$. e se calcula mediante el cociente de las distancias euclídeas entre las marcas: $e = dist(A_1, B_1) / dist(A_0, B_0)$. α es el ángulo de inclinación de la recta que une los puntos extremos de las marcas en la imagen escaneada, pues las marcas en la imagen modelo se encuentran alineadas a 0º. Se puede entonces calcular $\alpha = \text{arctg}((y_{B1} - y_{A1})/(x_{B1} - x_{A1}))$.

El vector de traslación t, el factor de escala e y el ángulo de rotación α obtenidos caracterizan la transformación rígida hecha durante el escaneo de la imagen. La calibración de la imagen consistirá entonces en revertir estas transformaciones no deseadas realizando las transformaciones inversas a las realizadas durante el proceso de escaneo de la hoja de respuestas.

En el texto de Lengyel [21], se presentan varias herramientas matemáticas útiles para realizar esta tarea. Entre ellas, se eligió el uso de coordenadas homogéneas, que permiten realizar las operaciones de escalado, rotación y traslación simultáneamente luego de calcular el producto matricial entre la matriz X' de todos los puntos (expresados como vectores fila extendidos hasta la coordenada w) y la matriz de la transformación inversa F. Siendo $F = \{((1/e)\cos\alpha, -(1/e)\text{sen}\alpha, 0, 0); ((1/e)\text{sen}\alpha, (1/e)\cos\alpha, 0, 0); (0, 0, 1, 0); (x_{A0} - x_{A1}, y_{A0} - y_{A1}, 0, 1)\}$, la matriz X de los puntos de la imagen calibrados puede obtenerse calculando $X = X'.F$

Hay casos donde la calibración no puede realizarse: cuando la transformación realizada en el proceso de escaneo es tal que alguna respuesta queda fuera del área de adquisición. Esto se puede calcular conociendo el ángulo de rotación y el vector de traslación, y es dependiente del diseño de la encuesta.

3.4. Lectura de resultados: transferencia de las respuestas desde el papel a la base de datos

Responder a una pregunta de las encuestas en papel consiste en marcar con tinta, o grafito espacios destinados a tal fin. En cada encuesta, se leen estos espacios y se determina si están marcados o no.

Formalmente, sean $R_i = (x_i, y_i)$ las coordenadas del centro de la casilla de la respuesta i en la imagen modelo. Se asocia a esa casilla una ventana digital dada por $D_i = [x_i-\gamma, x_i+\gamma\ ]x[\ y_i-\delta, y_i+\delta]$, donde 2γ es el ancho y 2δ es el alto de la ventana, elegidos de acuerdo al diseño gráfico de las encuestas.

Para cada respuesta, se define el nivel de la ventana asociada: $N_{Di} =$, donde v_j es el valor del nivel de gris leído por el escáner en el pixel p_i.

Para leer las respuestas, se leen los niveles de las ventanas de todas las respuestas, sobre la imagen ya normalizada (transformada según se explicó en la sección anterior). Si el nivel de la ventana supera un umbral predefinido, la respuesta no fue seleccionada por el encuestado. Y viceversa: un nivel por debajo del umbral indica que la respuesta fue seleccionada (en imágenes RGB Monocromáticas: Negro=0 y Blanco=1). Luego, se incorpora dicha respuesta al sistema transaccional.

4. Caso de Estudio.

4.1. Presentación

En Octubre de 2013, el sector "Políticas de Salud" de la Sociedad Argentina de Cardiología lanzó una encuesta a sus socios y clientes (mayormente cardiólogos), con la finalidad de conocer la situación de los cardiólogos en Argentina [22].

La encuesta se realizó en formato web y se enviaron invitaciones a socios y a quienes sin ser socios, hubieran realizado capacitaciones en la sociedad. Las invitaciones se incluyeron en el newsletter semanal en cuatro oportunidades desde su lanzamiento hasta Marzo de 2014; enviándose aproximadamente 6500 correos electrónicos a su base de clientes (socios y no socios) en cada oportunidad.

Durante los días 16 y 17 de Mayo de 2014 se realizó en la ciudad de Junín el I Congreso Multidisciplinario de Cardiología [23]. En dicho evento se reunió a aproximadamente 700 personas. Para ello (sin conocer a-priori la cantidad de asistentes) se imprimieron 400 ejemplares de la encuesta en papel.

4.2. Tecnología utilizada

La aplicación web de la encuesta y las herramientas de ETL se desarrollaron en PHP[2]. El motor de la base de datos usada es Mysql[3]. Para el modelado de la base de datos se usó Mysql wordbench[4].

La encuesta en papel se transcribió en un procesador de textos con la consigna que la totalidad de la encuesta debe ocupar una hoja. Se imprimió en doble faz: en el frente, el título de la encuesta y el objeto de la misma. A continuación, las instrucciones. En la página siguiente, se colocó la totalidad de las preguntas y sus opciones. Para escanear los documentos en papel se usó un escáner Epson[5] con alimentador automático de documentos (ADF en inglés). La aplicación para tratamiento de los archivos de imagen se desarrolló en Python[6].

4.3. Resultados

4.3.1. Acerca del origen de las respuestas. Al día 30 de Mayo de 2014 se cuentan 236 respuestas por web y 28 respuestas en papel.

Las 236 respuestas hechas en la web fueron resultado de aproximadamente 26000 correos electrónicos no nominados enviados desde Octubre de 2013. Esto da un ratio de 0,9% respuestas por persona-envío.

Las 28 respuestas en papel se obtuvieron en un único evento donde participaron aproximadamente 700 personas. Lo cual arroja un ratio de 4% de respuestas por persona-asistente.

4.3.2. Las respuestas y el modelo de evolución. La implementación del modelo detallado en este trabajo aplicado al caso de estudio, ha permitido extraer información relevante a partir de los datos. Aquí se presentan algunas respuestas obtenidas siguiendo el modelo de evolución del soporte de decisión de cuatro categorías presentado en [17].

Consultas estándar:

Al 90% de los cardiólogos les preocupan los posibles juicios por mala praxis. La percepción de los encuestados acerca de su situación actual: Muy mala 15%, Mala 32%, Regular 31%, Buena 17%, Muy buena 4%.

Análisis multidimensional:

Desagregación de la sub-dimensión “Cardiólogos” (93% del total) de la dimensión “Especialidad médica” según Subespecialidad.

Unidad Coronaria: 31%, Ecodoppler Cardíaco - Vascular: 24%, Ninguna: 20%, Cardiología del Deporte: 7%, Electrofisiología: 6%, Hemodinamia: 5%, Investigación: 4%, Medicina Nuclear: 2%, Electrocardiografía: 1%, Otra: 0%.

Segmentación:

Segmentación de la dimensión “Género”: Masculino 75%, Femenino 25%.

Segmentación de la dimensión “Lugar donde desempeña sus actividades”: Ciudad de Buenos Aires y Conurbano 58%, resto de provincia de Buenos Aires 15%, interior de Argentina 26%, fuera de Argentina 1%.

Minería de datos:

Modelo de clasificación mediante árboles de decisión “ADTrees”:

2 http://php.net/
3 http://www.mysql.com/
4 http://www.mysql.com/products/workbench/
5 http://www.epson.com/
6 https://www.python.org/

```
=== Run information ===
Scheme:weka.classifiers.trees.ADTree -B 10 -E -3
 [list of attributes omitted]
=== Classifier model (full training set) ===
Alternating decision tree:
: -0.9
|  (1)perc_su_situac_retiro_mala = 0: 0.219
|  |  (2)perc_su_situac_retiro_regular = 0: 0.467
|  |  |  (3)perc_su_situac_retiro_buena = 0: 0.839
|  |  |  |  (4)perc_su_situac_retiro_muy_buena = 0: 3.572
|  |  |  |  (4)perc_su_situac_retiro_muy_buena = 1: -1.914
|  |  |  (3)perc_su_situac_retiro_buena = 1: -1.545
|  |  (2)perc_su_situac_retiro_regular = 1: -1.635
|  (1)perc_su_situac_retiro_mala = 1: -1.54
|  (5)subesp_Electrofisiologia = 0: -0.827
|  (5)subesp_Electrofisiologia = 1: 0.111
|  (6)act.prof_H-privado_>60hs = 0: -0.507
|  |  (7)perc_su_situac_10anos_mala = 0: -0.438
|  |  (7)perc_su_situac_10anos_mala = 1: 0.024
|  (6)act.prof_H-privado_>60hs = 1: 0.126
Legend: -ve = 0, +ve = 1
Tree size (total number of nodes): 22
Leaves (number of predictor nodes): 15
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         100       %
Incorrectly Classified Instances         0       %
=== Confusion Matrix ===
   a   b   <-- classified as
 132   0 |   a = 0
   0  21 |   b = 1
```

Figura 5: Implementación del árbol ADTree sobre un atributo de la encuesta.

En la figura 5 se muestra una impresión de pantalla de los resultados arrojados por el algoritmo ADTree implementado en Weka[7] sobre el nodo "Percepción de su situación el momento de su retiro: Muy Mala".

5. Conclusiones y Trabajos Futuros

La automatización de la carga de encuestas en papel usando reconocimiento automático de imágenes es una respuesta rápida y económica frente a situaciones en las que no se dispone de infraestructura para su implementación electrónica.

La carga de las respuestas en un modelo de datos coherente con un sistema de encuestas electrónico facilita su integración y su posterior análisis.

Los acelerados avances en el campo de visión artificial y reconocimiento de imágenes permiten disponer de más y mejores prestaciones que pueden solucionar una mayor diversidad de requerimientos. Y aquí se presenta un nuevo caso. Los investigadores pueden tener opciones rápidas y económicas para obtener datos frente a distintas situaciones, dándoles flexibilidad en el momento de diseño de su investigación.

Futuros trabajos podrán mostrar la interpretación automática de textos manuscritos que podrán incluirse en cajas de texto y permitir la obtención de más información de los encuestados.

Agradecimientos

Se agradece a la Sociedad Argentina de Cardiología por permitir hacer uso de una de sus encuestas la cual fue parcialmente tomada como caso de estudio de este trabajo. El agradecimiento se extiende especialmente a los doctores: Adriana Salazar, Marcelo Boscaro, Carlos Boissonnet y María Inés Sosa Liprandi, del área de "Políticas de Salud" de dicha sociedad quienes trabajaron activamente en varias etapas de la presentación del caso de estudio aquí presentado.

7 http://www.cs.waikato.ac.nz/ml/weka/

Referencias

[1] Ospina, E; Reveiz Herault, L y Cardona, A. "Uso de bases de datos bibliográficas por investigadores biomédicos latinoamericanos hispanoparlantes: estudio transversal". *Rev Panam Salud Pública*, 2005, vol. 17, no 4, pp. 230-236.

[2] Beerli-Palacio, A; Martín-Santana, J, y Porta, M. "El marketing como herramienta para incrementar la eficacia de los planes de salud pública". *Informe SESPAS 2008*. Gaceta Sanitaria 22, 2008, pp. 27-36.

[3] Alet, J. "*Marketing directo e integrado"*. Barcelona: Gestión 2000, 1994.

[4] Serrano Ramos, M y Muñiz Velázquez, J. "La eficacia de la publicidad directa e interactiva a través del correo electrónico". *Congreso AE-IC*, Malaga, 2010.

[5] Chittenden, L and Rettie, R. "An evaluation of email marketing and factors affecting response". *Journal of Targeting, Measurement and Analysis for Marketing*, 11 (3), 2003, pp. 203-217.

[6] Carroll, A et al. "Consumer perceptions and attitudes toward SMS advertising: recent evidence from New Zealand." *International Journal of Advertising* 26.1, 2006, pp 79-98.

[7] Asociación de Industriales del Calzado de Elche. "Estrategia de Difusión y Marketing On-Line para el Sector Calzado de la comunidad de Valencia", *PCEV*, 2011.

[8] Manchón Pedroza, P y Benítez Fernández, Y. "MAGIC SMTP". *Tesis de grado. Escuela de Ingeniería, Universidad de Barcelona*, 2013.

[9] Otero, P. "¿Es una metodología válida la recolección de datos vía Web?". *Archivos argentinos de pediatría* 106.5, 2008, pp. 390-391.

[10] Weber B, Yarandi H, Rowe M and Weber J. "A comparison study: paper-based versus web-based data collection and management*". Appl Nurs Res*, 2005;18(3):182-5

[11] Bälter, O, and Bälter, K. "Demands on web survey tools for epidemiological research". *European journal of epidemiology* 20.2, 2005, pp. 137-139.

[12] Gonzalez, R; and Woods, R. "*Digital Image Processing*". Ed: Addison-Wesley publishing company, 1993.

[13] Kamlofsky, J y Bergamini, M. "Patrón de Evolución Discreta de Curvatura y Concavidad para Reconocimiento de Formas". *CONAIISI*, 2013.

[14] Rosenfeld, A, and Kak, A. *"Digital picture processing"*. Vol. 1. Elsevier, 2014.

[15] Codd, E. "*Relational completeness of data base sublanguages*". IBM Corporation, 1972.

[16] Morteo, F; Bocalandro, N; Cascon, H; Descalzo, C; De Rosa, K y Krauthamer, D. "*Fundamentos de diseño y modelado de datos*". Buenos Aires: Ediciones Cooperativas, ISBN: 978-987-1246-51-9, 2007.

[17] Dyche, J. "*E-Data: Turning data into information with data warehousing*". Addison-Wesley Professional, 2000.
[18] Hernández Orallo, J, Ramírez Quintana, M, y Ferri Ramirez, C. "*Introducción a la Minería de Datos*". Editorial Pearson Educación SA, Madrid, 2004.
[19] Rosenfeld, A. "Digital topology". *American Mathematical Monthly*, 1979, pp 621-630.
[20] Kamlofsky, J, y Bergamini, M. "Aproximación de formas de objetos digitales por polígonos." *VII Jornadas Argentinas de Robótica*. 2012.
[21] Lengyel, E. "*Matemáticas para Videojuegos en 3D*". Segunda Edición, Cengage Learning, 2011.
[22] Sociedad Argentina de Cardiología. "*Encuesta SAC 201303: Area Políticas de Salud*", 2013, http://www.encuestas.sac.org.ar/201303/ [Consulta: 20 de Julio de 2014].