WorkShop
EIG2009

WORKSHOP INTERNACIONAL
EIG2009
Departamento de Ingeniería de Sistemas
Universidad de La Frontera – Chile
Diciembre 3 y 4, 2009

RESÚMENES LINGÜÍSTICOS PARA DATOS HISTÓRICOS

Felipe Acevedo, Cecilia Reyes, José Luis Martí

Departamento de Informática, Universidad Técnica Federico Santa María,
Av. Vicuña Mackenna 3939, San Joaquín – Santiago, Chile
felipe.acevedos@alumnos.utfsm.cl, {reyes, jmarti}@inf.utfsm.cl

Resumen. La gran cantidad de datos que las organizaciones han ido acumulando a lo largo del tiempo, puede resultar un problema al momento de generar conocimiento, en particular para apoyar la toma de decisiones. En el último tiempo, y haciendo uso de elementos propios de la lógica difusa, han surgido los resúmenes lingüísticos, que corresponden a breves frases que rescatan las relaciones más relevante entre los datos a estudiar. Este trabajo entrega las bases para poder construir resúmenes lingüísticos sobre datos históricos, estructurados en un formato multidimensional propio de ambientes de *data warehouses* e inteligencia de negocios.
Los resultados obtenidos a la fecha han permitido afianzar la utilidad de los resúmenes lingüísticos, ya que el tomador de decisiones puede entender en su propio idioma, las relaciones más importantes del conjunto de datos a analizar, y en menor tiempo, proponer soluciones de negocios, no importando el rubro en el cual se desenvuelva su organización.

1. Introducción

Durante los últimos años y más en la actualidad, han tomado una vital importancia aquellos sistemas que apoyan la toma de decisiones, debido a la gran cantidad de datos operacionales que cada organización manipula a través del tiempo, la clara necesidad de mejorar el servicio y ofrecer mejores productos a sus clientes, y saber sobrevivir al competitivo mundo que hoy en día se vuelve cada vez más fuerte y agresivo.
El primero de los aspectos mencionados anteriormente guarda relación con bases de datos para la gestión, denominadas *data warehouses* (almacenes de datos) [1], [2], los cuales corresponden a repositorios centrales desde donde la organización obtiene lo necesario para un análisis de sus datos históricos. Un *data warehouse* tiene como característica importante que su contenido se encuentra organizado de manera multidimensional, es decir en forma de un cubo de datos, donde cada casilla contiene los valores resultantes de una operación determinada, y las dimensiones hacen referencia al contexto sobre el cual se explican las medidas anteriores almacenadas en cada registro. La figura 1 muestra un ejemplo de un cubo de datos, para un análisis basados en tres dimensiones (*Route, Source, Time*) [3].

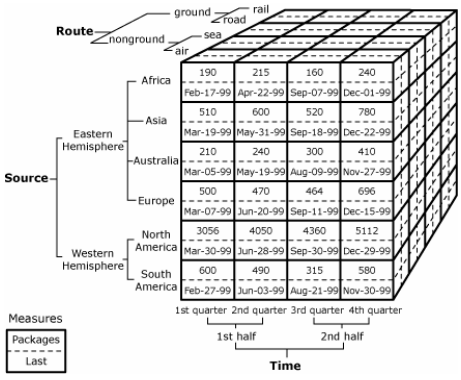

Fig 1. Ejemplo de un cubo de datos.

Al momento de presentar resultados (patrones, tendencias) al usuario o analista del negocio, la estructura multidimensional puede no ser fácil de manejar, en particular cuando el estudio considera varias dimensiones. Por ejemplo, el análisis de ventas podría tomar en cuenta a Cliente, Tiempo, Local de Venta, Producto, Proveedor y Vendedor.

Para ayudar al estudio, existe la posibilidad de trabajar con subconjuntos del cubo (según sea el caso), y a partir de éstos generar resúmenes que tradicionalmente se han

expresado mediante tablas o expresiones matemáticas. No obstante, cuando una persona comunica los mismos datos a otra persona, él o ella lo hace mediante lenguaje natural. Por esta razón, resulta interesante que el resumen sea usando en dicho formato, y que el mismo sistema computacional se encargue de generarlo. En este contexto, tiene sentido hablar de un paradigma de "**computación con palabras**" [4], aplicable a diferentes contextos [5], [6].

Un ejemplo típico de interpretación de un conjunto de datos es la entrega del pronóstico del tiempo. La persona a cargo, al momento de indicar como cual será la temperatura del día siguiente, se limita a entregar sólo los aspectos relevantes; por ejemplo *"en la mañana amanecerá con nublado parcial, con una temperatura del orden de los 15 grados Celsius, para luego variar a despejado con temperaturas altas, que estarán entre los 28 y 30 grados"*. Se aprecia que no es necesario señalar la temperatura hora a hora, sólo resumir los pronósticos de éstas, para generar una frase (resumen lingüístico) que sea comprensible y de fácil uso.

El objetivo del presente trabajo es generar resúmenes lingüísticos para datos que se encuentran en un formato multidimensional, y que sean un real aporte a la toma de decisiones de una organización al concentrar las relaciones más importantes entre variables de interés para el negocio.

2. Resúmenes Lingüísticos

Un resumen lingüístico se define como una oración o frase con un lenguaje muy parecido al habitual, que trata de explicar lo contenido dentro de un grupo de datos. Este conjunto generalmente corresponde a datos numéricos y de un gran volumen, convirtiéndose casi incomprensible en su forma original para el ser humano: luego, es labor del resumen generar el conocimiento más relevante de dichos datos, capaz de apoyar una adecuada toma de decisiones.
Al considerar los siguientes elementos:

- Un conjunto de de objetos ***Y*** en una base de datos, por ejemplo un grupo de trabajadores.
- Un conjunto de atributos ***A,*** característicos de los elementos del conjunto de objetos anterior, como sueldo, edad, sexo, etc.

la definición formal de un resumen lingüístico considera [7]:

- **Un resumidor *P***, que corresponde a un atributo junto con un término lingüístico (predicado difuso), el cual se encuentra definido en el dominio de los atributos de interés de los objetos en estudio.
- **Un cuantificador lingüístico *Q***, por ejemplo "la mayoría", "algunos", "cerca de la mitad".
- **Validez *T* del resumen**, correspondiente a un número dentro del intervalo **[0,1]** que permite evaluar el grado de validez del resumen, por ejemplo 0,7. Por lo general, sólo aquellos resúmenes que poseen un alto valor de *T* se consideran como interesantes de estudiar.
- **Un cualificador *R***, cuya utilización es opcional y responde a uno de los atributos característicos asociados a los objetos en estudio.

Un resumen de tipo I sigue una estructura del tipo ***"Q Y son P"***, como por ejemplo: **la mayoría de los trabajadores ganan bajos sueldos.**
Si se incorpora una propiedad **R** a los objetos **Y**, es decir se extiende el tipo anterior al formato ***"Q Y R son P"***, se obtiene un resumen de tipo II. Por ejemplo: **la mayoría de los trabajadores jóvenes ganan bajos sueldos.**

Por último es importante recalcar que cada uno de los resúmenes lingüísticos debe tener asociado un grado de verdad, que permita determinar cuál resumen es más conveniente que otro.

3. Resúmenes Lingüísticos Tridimensionales

La propuesta de este trabajo consiste en presentar una taxonomía de resúmenes lingüísticos para cubos de datos, los cuales permitan analizar su contenido con diversos niveles de detalles.

3.1 Protoformas Temporales

Como primer paso se plantea una protoforma que incorpore el período de tiempo durante el cual es analizado el comportamiento. De esta manera es posible entregar de manera explícita y certera el intervalo de tiempo involucrado, minimizando errores de interpretación o descartando supuestos. Luego la protoforma de

tipo I tridimensional tendría la forma ***"Q Y se produjeron durante T"***, donde ***Q*** e ***Y*** corresponden a los mismos atributos especificados para las protoformas de los resúmenes tradicionales. Lo novedoso es la incorporación de la dimensión ***Tiempo*** de manera explícita, desde un comienzo, dada su relevancia para el estudio de la historia de una organización; dicho período de tiempo puede ser descrito según la granularidad deseada (días, meses, semanas, horas, etc.). Ejemplos de protoformas de este tipo son:

- La Mayoría de las compras se produjeron durante **las 3 primeras semanas.**
- Pocas ventas se produjeron durante **los 10 primeros días.**
- Muchos reclamos se produjeron durante **el cuarto trimestre.**

Poniéndose en el caso del analista del negocio, es muy probable que tras resultados anteriores, quiera conocer más detalle sobre los objetos que dan origen a los resúmenes iniciales. Un resumen del tipo II tridimensional podría tener el siguiente formato: ***"Q Y R se produjeron durante T"***, extensión con la cual es posible identificar de quién o qué es el comportamiento analizado, permitiendo entregar un resumen claro y preciso. A modo de ejemplos, y extendiéndose los casos anteriores, se tienen:

- La Mayoría de las compras del **Cliente de Tipo A** se produjeron durante las 3 primeras semanas.
- Pocas ventas del **Producto 3** se produjeron durante los 10 primeros días.
- Muchos reclamos por el **Motivo 2** se produjeron durante el cuarto trimestre.

Cabe señalar que tanto en las protoformas propuestas hasta ahora, como en las próximas que se presentan en este trabajo, la presencia de la subfrase "***se produjeron durante***" no corresponde a una estructura fija, ya que puede variar según sea necesario, con el propósito de dar un entendimiento pleno. De manera similar ocurre con lo existente entre los componentes ***Y*** y ***R***, ya que puede aceptar variaciones según sea el contexto del resumen generado y no necesariamente deberá tomar el valor "***de***".

Una nueva extensión es aquella que hace referencia a la mezcla de dimensiones (planos) presentes en cada cubo, con la finalidad de obtener información más específica sobre los datos analizados. Con esto, la protoforma de tipo III tridimensional sigue la estructura: ***"En R', Q Y R se produjeron durante T"***. Con ***R*** se continúa haciendo referencia a aquella dimensión que se está analizando, es decir, con respecto a qué o quién se está generando el resumen lingüístico. Por otra parte, ***R'*** corresponde a un elemento en específico de otra dimensión, lo que permite determinar de qué plano en particular se está haciendo referencia, permitiendo generar una intersección de planos y con ello, extraer información y resúmenes mucho más específicos. La relación entre ***R*** y ***R'*** se visualiza en la figura 2.

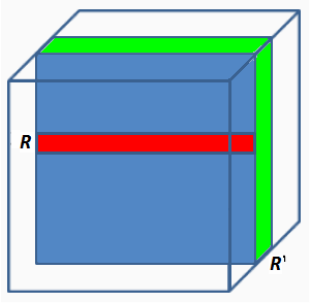

Fig 2. Representación gráfica de Protoforma de Tipo III Tritidimensional

Al extender los ejemplos de resúmenes anteriores a una estructura de tipo III, se tendrían:

- **En la Sucursal 4**, la mayoría de las compras del Cliente de tipo A se produjeron durante las 3 primeras semanas.
- **En la VI Región**, pocas ventas del Producto 3 se produjeron durante los 10 primeros días.
- **En la Oficina Norte**, muchos reclamos por el Motivo 2 se produjeron durante el cuarto trimestre.

Hasta aquí, todas las protoformas planteadas hacen alusión a un comportamiento identificado durante un período determinado de tiempo, quedando excluidos aquellos resúmenes que no lo requieren. Sin embargo, y aunque se trate de datos históricos, igualmente puede ser de interés resumir los datos del cubo, sin tomar en cuenta la dimensión del tipo.

3.2 Protoformas No Temporales

Es posible plantear una nueva protoforma para la entrega de resúmenes lingüísticos, cuya información no sea relativa a un período de tiempo. El formato propuesto, para esta protoforma de tipo IV, es ***"Q Y se produjeron en/por R"***. Algunos casos de ejemplo serían:

- La Mayoría de las compras se produjeron **en la Sucursal 4.**
- Pocas ventas se produjeron **por el Cliente 2.**
- Muchos reclamos se produjeron **en la Zona Oriente.**

Con el mismo razonamiento, es posible modificar la protoforma de tipo IV incorporando otra dimensión (que no sea el Tiempo), indicando claramente a cuál de ellos se hace referencia. La protoforma de tipo V resultante se ajusta al formato ***"Q Y de R' se produjeron en/por R"***.
Al igual que en la protoforma de tipo II, es posible identificar de quién o qué es el comportamiento analizado, permitiendo entregar un resumen más claro. A modo de ejemplo se presentan los siguientes resúmenes de tipo V:

- La Mayoría de las compras del **Cliente de tipo A** se produjeron **en la Sucursal 4**.
- Pocas ventas del **Producto 3** se produjeron **por el Cliente 2**.
- Muchos reclamos por el **Motivo 2** se produjeron **en la Zona Oriente**.

4. Caso de Aplicación

En la mayoría de los estudios y recursos disponibles en la literatura actualmente, con respecto a Resúmenes Lingüísticos se refiere, es posible encontrar como tema principal el análisis del comportamiento de las Ventas en diferentes ámbitos y para diversas organizaciones que lo han creído necesario. También, ciertos aspectos económicos y estadísticos están dentro de los aspectos más comunes o tradicionales que abordan este tipo de estudios, donde es posible encontrar ciertas predicciones entregadas mediante un resumen lingüístico que permita identificar un comportamiento a futuro de un indicador determinado. Finalmente, una práctica un tanto diferente a lo habitual corresponde a la predicción y entrega de resúmenes meteorológicos, en los cuales es posible verificar la temperatura, humedad, vientos, nubosidad, etc. para un próximo día o para sintetizar como se comportó el clima en algún día anterior.
Para demostrar que no sólo dentro de la estadística y de la economía tiene una gran utilidad este tipo de técnica, se analizó una base de datos disponible en [8], donde es posible encontrar la cantidad de nacimientos ocurridos a lo largo y ancho del país de *México*. En el mencionado recurso, la información disponible está clasificada en tres dimensiones:

- ***Entidad Federativa*** o Estado Federado: corresponde a una región de un país, en este caso México, donde cada una de ellas posee leyes propias para regir a las personas que habiten en cada una de las mencionadas porciones de territorio.
- ***Lugar de Nacimiento***: hace referencia al lugar físico en donde ocurrió el nacimiento, en otras palabras, si el alumbramiento ocurrió en un Hospital dependiente del gobierno mexicano o uno de carácter particular, como también está la alternativa de haber ocurrido en el domicilio o simplemente que dicha información corresponda a otro lugar diferente de los anteriores, dejando como última alternativa que el lugar no esté especificado.
- ***Tiempo***: los años involucrados en este estudio hacen referencia desde 1990 hasta 2007, conformando un universo de 18 años.

La figura 3 muestra una representación gráfica del cubo de datos históricos asociados al caso en estudio.

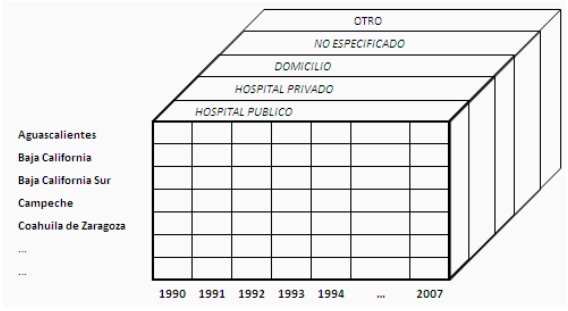

Fig 3. Representación gráfica de Protoforma de Tipo III Tridimensional

Antes de entregar los resultados obtenidos, cabe señalar que para cubos de datos como el utilizado en este trabajo, en donde la información contenida en él

corresponde a índices que en términos generales no poseen mayor variación, los resúmenes lingüísticos tienden a resultar similares entre sí, debido a que quedan distribuidos de una manera uniforme y por lo tanto, el cuantificador obtenido para cada uno de ellos tiende a ser repetido. Por ejemplo, al dividir la dimensión ***Tiempo*** en varios grupos de tres años, es muy probable que el cuantificador obtenido para los resúmenes lingüísticos entregados sea ***"La Minoría"***, ya que la cantidad de nacimientos ocurridos en cada uno de los períodos es pequeña si se compara con el total de lo analizado.

Las combinaciones posibles entre las dimensiones son:

- **Entidad Federativa vs. Tiempo.**
- **Lugar vs. Tiempo.**
- **Entidad Federativa vs Lugar.**

A continuación se presenta como ejemplo, los resultados asociados al primero de dichos casos. Tras el desarrollo de un generador de resúmenes lingüísticos y su aplicación al caso de los nacimientos históricos en México, se obtuvieron resúmenes de tipo I como:

- La minoría de los nacimientos se produjeron en el primer trienio.
- La minoría de los nacimientos se produjeron en el sexto trienio.

considerando que la visión que se tenía de los datos es similar a la presentada en la figura 4.

Fig 4. Visión de los datos para los resúmenes de tipo I, para la combinación Entidad Federativa vs. Tiempo

Al aumentar la necesidad de un mayor detalle de información, por parte del analista del negocio, se podrían obtener resúmenes del tipo II, tales como:

- La minoría de los nacimientos ocurridos en Baja California Sur se produjeron en el primer trienio.
- La minoría de los nacimientos en Colima se produjeron en el quinto trienio.

siendo la visión del análisis equivalente a la descrita por la figura 5.

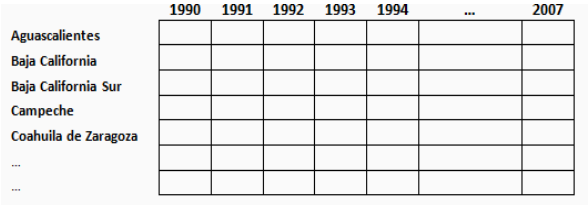

Fig 5. Visión de los datos para los resúmenes de tipo II, para la combinación Entidad Federativa vs. Tiempo

Finalmente, un tercer nivel de detalle permitiría incluir la dimensión que no estaba considerada a la fecha (Lugar), generándose resúmenes de tipo III tales como:

- En Hospital Privado, pocos nacimientos ocurridos en Baja California Sur se produjeron en el segundo trienio
- En Domicilio, casi ninguno de los nacimientos ocurridos en Campeche se produjeron durante el sexto trienio.

ambos bajo una visión del análisis que considera el cubo de datos históricos completo, como muestra la figura 6.

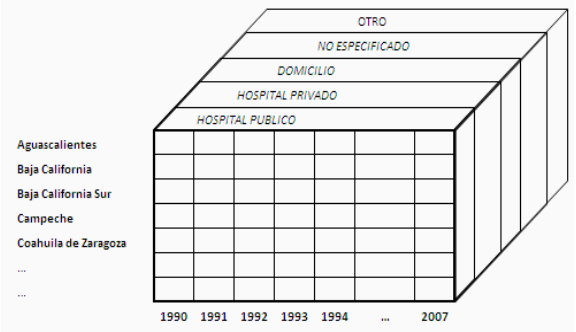

Fig 6. Visión de los datos para los resúmenes de tipo III, para la combinación Entidad Federativa vs. Tiempo

Sólo a modo de comparación con la estructura de los resultados, se muestran a continuación un ejemplo de resumen de tipos I, II y III, respectivamente para la combinación entre Entidad Federativa y Lugar (la visión cada tipo de resumen se muestra entre las figuras 7 a 9):

- La minoría de los nacimientos se produjeron en Hospital Privado.
- Muchos de los nacimientos ocurridos en Guanajato se produjeron en Hospital Público.

- En el 2005, pocos nacimientos ocurridos en Tlaxcala se produjeron en Hospital Privado.

	Hosp. Publico.	Hosp. Privado	Domicilio	No Espec.	Otro
EF					

Fig 7. Visión de los datos para los resúmenes de tipo I, para la combinación Entidad Federativa vs. Lugar

	Hosp. Publico.	Hosp. Privado	Do
Aguascalientes			
Baja California			
Baja California Sur			
Campeche			
Coahuila de Zaragoza			
…			
…			

Fig 8. Visión de los datos para los resúmenes de tipo II, para la combinación Entidad Federativa vs. Lugar

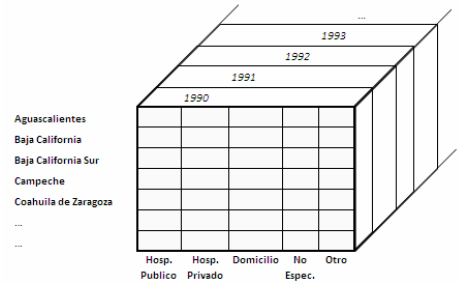

Fig 9. Visión de los datos para los resúmenes de tipo III, para la combinación Entidad Federativa vs. Lugar

Como conclusión principal se destaca que los resúmenes obtenidos reflejan adecuadamente aspectos importantes de los datos, a nivel de las relaciones que los estructuran. Sin la utilización de un generador de resúmenes lingüísticos, la obtención de dichas relaciones tardaría más, y eventualmente podrían no ser identificadas, especialmente cuando el volumen de datos es muy grande.
Por último cabe señalar la importancia que puede tener este tipo de técnicas cuando el negocio está accesando mediante dispositivos móviles, en los cuales las dimensiones de las ventanas de interacción con el usuario son muy limitadas. El contar con un resumen, en lugar del cubo de datos históricos, se convierte más bien en una necesidad.

4. Conclusiones

A través del desarrollo de este trabajo se ha puesto en evidencia las ventajas de los resúmenes lingüísticos, como técnica para rescatar relaciones importantes entre datos históricos, con los cuales la toma de decisiones podría hacerse más simple y rápida, al concentrar el análisis sólo en los puntos de real interés, identificados precisamente por dichos resúmenes.

En un próximo paso, el trabajo se pretende extender a ambientes con más de tres dimensiones, para obtener una generalización de los resultados ya expuestos.

Referencias

[1] M. Jarke, M. Lenzerini, Y. Vassilliu, P. Vassilliadis: *Fundamentals of Data Warehouses*. Primera Edición, Springer (2003).

[2] E. Malinowski, E. Zimányi. *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications (Data-Centric Systems and Applications)*. Springer (2008).

[3] Microsoft Developer Network Home Page. http://i.msdn.microsoft.com/ms175449.523a39d7-6491-4862-8ff2-f810bbdf01d7%28es-es,SQL.90%29.gif.

[4] J. Kacprzyk: Linguistic Summaries of Static and Dynamic Data: Computing with Words and Granularity. *IEEE International Conference on Granular Computing*, pp.4 (2007).

[5] S. Somayajulu, E. Reiter, I. Davy: SumTime-Mousam: Configurable Marine Weather Forecast Generator. *Expert Update*, vol. 6, n 3, pp. 4-10 (2003).

[6] E. D'Avanzo & T. Kuflik: *Linguistic Summaries con Small Screens*. Data Mining VI: Data Mining, Text Mining and their Business Applications (2005).

[7] J. Kacprzyk, S. Zadrożny: Linguistic Database Summaries and their Protoforms: towards Natural Language based Knowlegde Discovery Tools. *Information Sciences*, vol. 173, n 4, pp. 281=304 (2005).

[8] Dirección General de Información en Salud (DGIS). *Base de datos de nacidos vivos registrados 1990-2007*. Secretaría de Salud. http://www.sinais.salud.gob.mx.