Resúmenes Lingüísticos Multidimensionales basados en Segmentación de Datos

Christopher Pope, Cecilia Reyes, José Luis Martí*

Resumen

El objetivo de este trabajo es generar resúmenes lingüísticos multidimensionales en base a los resultados de un algoritmo de segmentación. Los resúmenes lingüísticos multidimensionales corresponden a frases en lenguaje natural que describen los datos y sus dimensiones. Para este trabajo se utilizan K-Means como algoritmo de segmentación, un cubo de datos como base para realizar la segmentación de datos y la generación de resúmenes lingüísticos multidimensionales, y protoformas también multidimensionales basadas en lo propuesto por L. Zadeh y R. Yager, para estructurar los resultados al usuario

Palabras clave:

Lógica Difusa, Minería de Datos, Resúmenes Lingüísticos.

Abstract

The aim of this work is to generate multidimensional linguistic summaries based on the results of a segmentation algorithm. Multidimensional linguistic summaries correspond to natural language phrases that describe the data and its dimensions. For this work are used as K-Means segmentation algorithm, a data cube as a basis for the segmentation of data and multidimensional linguistic summarization, and also protoforms multidimensional based on those proposed by L. Zadeh and R. Yager to structure the results to the user.

Keywords:

Fuzzy Logic, Data Mining, Linguistic Summaries.

*Departamento de Informática, Universidad Técnica Santa María – Santiago, Chile
christopher.pope@alumnos.inf.utfsm.cl, reyes@inf.utfsm.cl, jmarti@inf.utfsm.cl

Los resúmenes lingüísticos consisten en una herramienta para la entrega de información en un lenguaje natural, al alcance de cualquier persona con capacidad para comprender el lenguaje y contexto en el que se trabaja. Esta herramienta expresa características de los datos sobre los que está establecida, utilizando adjetivos y calificativos propios del dominio o caso en estudio.

Un resumen lingüístico es una frase o sentencia que explica cómo están conformados los datos. Cada resumen tiene asociado un valor de verdad, el que determina cuán verdadero o falso es el resumen en relación con los datos que lo respaldan. En el presente trabajo se propone un método para generar resúmenes lingüísticos multidimensionales, es decir, resúmenes que describan múltiples dimensiones de los datos.

I. Introducción

Basándose en un cubo de datos, se obtendrán resúmenes lingüísticos que describan ciertas dimensiones en base a otras del mismo cubo. De esta manera los resúmenes entregarán una visión global de los datos que componen el cubo.

Previo a generar los resúmenes lingüísticos, se realizará una segmentación de los datos, utilizando el algoritmo de minería de datos K-Means. El resultado de la ejecución de este algoritmo sobre el cubo de datos será utilizado como base para la generación de dos tipos de resúmenes lingüísticos: el primero, encargado de describir los segmentos uno a uno, y el segundo, de describir todos los segmentos como uno solo. Finalmente, se utilizará un caso de prueba para demostrar el uso del proceso propuesto como también del algoritmo de segmentación, protoformas y cálculo de valor de verdad.

II. Resúmenes lingüísticos

Un resumen lingüístico se define como una herramienta intuitiva y con una forma cercana a los humanos para la extracción de información desde una base de datos [1]. En palabras más simples, un resumen lingüístico es una frase que entrega información sobre los datos presentes, usando palabras del lenguaje natural, tales como "alto", "bajo", "mucho" y "poco", lo cual puede resultar más fácil de comprender que un valor estadístico o un gráfico. Dado que el lenguaje natural es el que se utiliza para comunicarnos, es muy deseable que un sistema "inteligente" sea capaz de entregar resultados en este lenguaje. Por ejemplo, si se trabaja con una base de datos de trabajadores, un resumen lingüístico podría ser "muchos trabajadores son jóvenes".

Como fue explicado por Yager [1], a partir de las siguientes definiciones:

- Un conjunto D={d1,…,di,…,dn} de entidades que se manifiestan sobre un conjunto de atributos.
- Un conjunto A={a1,…,ai,…,an} de atributos pertenecientes al conjunto D.

un resumen lingüístico está compuesto por:

- Un descriptor S, definido como una expresión lingüística semánticamente representada como un conjunto difuso.

- Un cuantificador Q para la cantidad, definido como un cuantificador lingüístico, por ejemplo "muchos".

- Una medida de la validez o la verdad T, definida como el valor de verdad sobre un resumen lingüístico.

Los resúmenes lingüísticos cumplen con ciertas estructuras, llamadas protoformas (formas prototípicas) [2] [3], las cuales definen la forma que presentaran los resúmenes. La literatura ha definido algunos tipos de protoformas, de las cuales destacan las llamadas Tipo 1 (T1) y Tipo 2 (T2), cuyas estructuras se muestran en la Tabla I. Como se puede ver, la diferencia entre ambas surge por la presencia de un elemento denominado calificador (R), que corresponde a una propiedad que se aplica sobre las entidades de un resumen T1; por ejemplo, "muchos trabajadores altos son jóvenes".

TABLA I

TIPOS DE PROTOFORMAS

Protoforma	Estructura	Ejemplo
T1	Q di son S	Muchos trabajadores son jóvenes
T2	Q R di son S	Muchos trabajadores nocturnos son jóvenes

Es posible crear muchos resúmenes lingüísticos, dadas las múltiples combinaciones de atributos con sus respectivos cuantificadores y calificadores.

Una característica muy importante de un resumen es su valor de verdad, simbolizado normalmente como T, que entrega un valor que valida la veracidad del resumen. El valor de verdad T es un valor numérico entre 0 y 1, que entrega la validez del resumen frente al conjunto de datos D. Mientras mayor sea el grado de verdad, es decir más cercano a 1, más "verdadero" es el resumen lingüístico.

Se definen dos maneras de calcular el grado de verdad; en el caso de la asociada al tipo T1, la expresión asociada es [1]:

$$T(D,\{Q,S\}) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_S(d_i)\right) \quad (1)$$

Para el caso de la protoforma de tipo T2, la fórmula asociada es:

$$T(D,\{Q,R,S\}) = \mu_Q\left(\frac{\sum_{i=1}^{n}(\mu_S(d_i)\wedge\mu_R(d_i))}{\sum_{i=1}^{n}\mu_R(d_i)}\right) \quad (2)$$

Donde el operador simboliza el mínimo entre dos valores. La función $\mu Q(di)$ representa la pertenencia del cuantificador Q en relación al conjunto de datos observados.

Diversos trabajos de investigación se han realizado a la fecha en esta área, siendo muy relevante lo publicado por Zadeh [1] sobre las "disposiciones" o proposiciones que contienen cuantificadores difusos, y Yager [2] quien crea el concepto de resumen lingüístico como una herramienta para la extracción de conocimiento desde una base de datos. En este último caso se considera al resumen lingüístico más bien como una técnica de minería de datos, ya que entrega información que no es evidente a simple vista.

Kacprzyk [4], en el año 1999, implementa en Access un algoritmo interactivo con el usuario para la obtención de resúmenes lingüísticos. Kacprzyk y Zadrozny [3] [5] explican las protoformas, establecidas por Zadeh y las aplican para la obtención de información desde una base de datos. Kacprzyk, Wilbik y Zadrozny [5] [6] [7] [8], desde el año 2005 a la fecha, aplican los resúmenes lingüísticos para la obtención de información basada en el análisis de series de tiempo.

Finalmente, en cuanto a aplicaciones se tienen casos, tales como el uso de resúmenes lingüísticos para el análisis del tráfico en redes en tiempo real [9].

III. Protoformas multidimensionales

La primera parte de este trabajo consiste en la definición de protoformas para resúmenes basados en múltiples variables, o dimensiones si se trata de datos rescatados desde un cubo. Dado que dichas protoformas se obtienen del resultado de la aplicación de un algoritmo de segmentación (lo que se explicará más adelante), es preciso diseñar una protoforma en la cual se exponga el tamaño del segmento con respecto al total de los datos y cuál es el centro del mismo, simbolizada como:

$$TO_{Seg}(D,(X,YY,Ri)) \quad (3)$$

donde:

- D corresponde al dominio.
- X representa el número (identificador) del segmento.
- YY es el porcentaje de los datos del segmento con respecto al total de datos.
- R_i simboliza al descriptor i

Por lo que la expresión se puede leer como *"El segmento X representa al YY% de los datos, y se encuentra centrado en los D ocurridos R_1, R_2 y R_3"*.

Esta protoforma es llamada Tipo 0 Segmentos. Cada uno de los Ri asociados representa una de las características o dimensiones del centroide

del segmento, pudiendo tener tantos R_i como dimensiones en estudio se tengan. Gracias a esta protoforma es posible obtener resúmenes lingüísticos que entreguen una descripción de cada uno de los segmentos creados. Estos resúmenes contienen un componente preciso (entregado por YY) y otro difuso (entregado por los R_i). Un ejemplo concreto de su aplicación podría ser el siguiente: "El segmento 4 representa el 6,7% de los datos, y se encuentra centrado en los nacimientos ocurridos en Zacatecas, durante el Año 2003 y en un hospital público".

Dicha protoforma tiene asociada una expresión matemática para la obtención de su valor de verdad, dado que tiene un componente difuso. Este cálculo del valor de verdad está centrado en los R_i, dado que éstos son el punto difuso de la expresión, y corresponde al promedio de la pertenencia de cada uno de los descriptores R_i en relación con la posición del centroide del segmento. El símbolo Desv%R_i corresponde a la desviación porcentual de cada dimensión para el segmento. En otras palabras, entrega el "radio de apertura" del segmento, donde un valor cercano a 0 dice que los datos para esa dimensión en ese segmento están todos cercanamente ubicados. Así, la expresión para el cálculo de verdad de un resumen del tipo T_0 Segmentos es:

$$\mathrm{T}(D,(X,YY,R_i)) = \frac{\sum_{i=1}^{N} \mu_{R_i}(d_i) * (1 - Desv\%_{R_i})}{N} \qquad (4)$$

Donde:

- $\mu R_i(d_i)$ es el valor de pertenencia del descriptor R_i.
- $Desv\%R_i$ es a la desviación estándar del descriptor R_i.
- N representa la cantidad de dimensiones del problema.

El segundo tipo de protoformas a plantear está basada en la expuesta por Yager [2], y considera la misma estructura básica original: cuantificador, descriptor y resumidor. Para manejar la multidimensionalidad se ha expandido la cantidad de descriptores y resumidores, siendo posible agregar múltiples descriptores y múltiples resumidores. Al permitir esto, se hizo necesario incorporar conectores entre ellos, por lo que también se agregaron los conectores para descriptores y para resumidores, como se muestra en la expresión siguiente, donde los conectores corresponden a conjunciones o disjunciones. Esta protoforma se le denomina **Tipo 2 Multidimensional**, y se rige por la siguiente estructura:

$$T2_{MD}(D,(R_i,S_j)) = Q\ D\ R_i\ \{C_{R_i}\ R_{i+1}\}^*\ son\ S_j\ \{C_{S_j}\ S_{j+1}\}^* \qquad (5)$$

Siendo:

- Q es el cuantificador.
- Ri corresponde a un descriptor.

- C representa a un conector lógico (y, o).
- Si es un resumidor.

Un ejemplo de un resumen lingüístico a generar utilizando esta protoforma puede ser: "Casi todos los nacimientos durante el Año 2003 y durante el Año 1997, fueron en un hospital público, y en Chihuahua o en el Distrito Federal."

La protoforma T2 Multidimensional tiene asociada una expresión general para obtener su valor de verdad, que corresponde a una extensión del cálculo de verdad asociado a la protoforma T2 expuesta por Yager [2].

Para obtener la desviación porcentual (6) se obtiene la desviación estándar para la dimensión del segmento en cuestión, luego este valor es dividido por el tamaño de la escala completa de la dimensión.

$$Desv\%_{Si} = DesvEstDim_{Si} / (2(MaxDim_i - MinDim_i) \quad (6)$$

donde:

- DesvEstDimSi corresponde a la desviación estándar de la dimensión i dentro del segmento S

- MaxDimi, MinDimi son los valores extremos del dominio de la dimensión i

IV. Proceso PROGREL

El proceso de generación de resúmenes lingüísticos (PROGREL de ahora en adelante) consiste básicamente en el análisis de los datos, en este caso un cubo de datos, para luego generar los resúmenes pertinentes y verificarlos contra el análisis obtenido anteriormente. PROGREL considera varios pasos a realizar, siendo siempre un proceso lineal y no paralelo. En la Figura 1, se presenta una estructura general de PROGREL, la cual está conformada por las siguientes partes:

- BDN: que corresponde a la base de datos del Negocio, almacenada como un cubo de datos.
- MINDAT: módulo de Minería de Datos, responsable de la generación de los segmentos de datos.
- GRES: encargado de la generación de resúmenes lingüísticos.
- BDRyS: la base de datos de Resúmenes y Segmentación.

En la misma 1 se expone cual es el flujo de los datos entre las capas de persistencia (BDN y BDRyS) y las capas de procesamiento (MINDAT y GRES). Este flujo de los datos permite apreciar la fuerte relación que debe existir entre los componentes BDRyS y GRES.

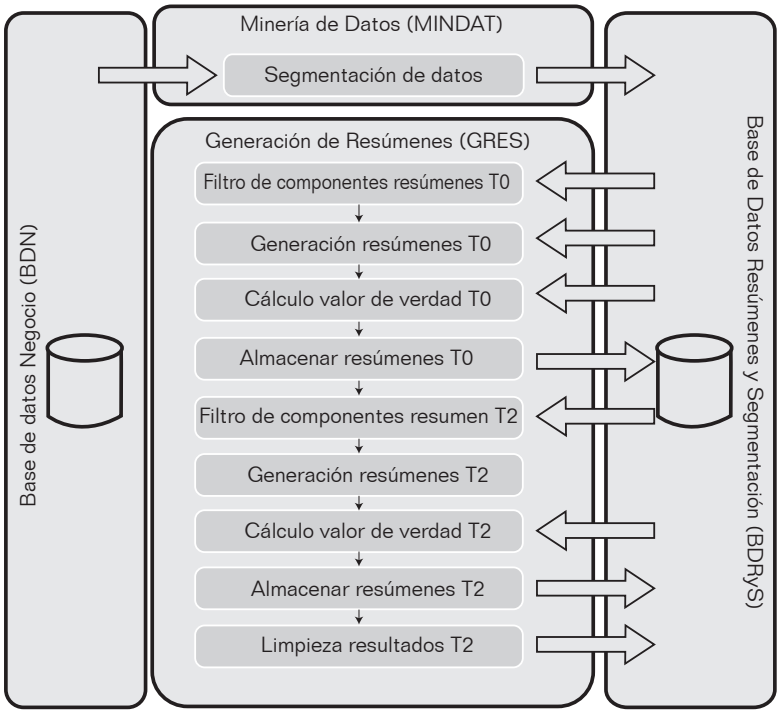

Fig. 1. Componentes del Proceso PROGREL, definido para la generación de resúmenes lingüísticos.

También es interesante poner atención a la ejecución secuencial de los subprocesos dentro de GRES. Esto se debe a que cada subproceso requiere como datos de entrada los de salida de su antecesor. Esto obliga a mantener un orden riguroso al momento de ejecutar los subprocesos.

PROGREL considera la utilización de dos modelos de datos, uno para los datos del negocio (BDN) y otro para los datos de la segmentación y los resúmenes lingüísticos (BDRyS). El modelo de datos del negocio considera los datos que se analizarán con una estructura multidimensional. El modelo de datos asociado a los resúmenes lingüísticos y la segmentación está destinado a almacenar los resultados de la ejecución del algoritmo de segmentación, como también almacenar las definiciones necesarias para construir los resúmenes y obtener su valor de verdad (para este último, ver Figura 2).

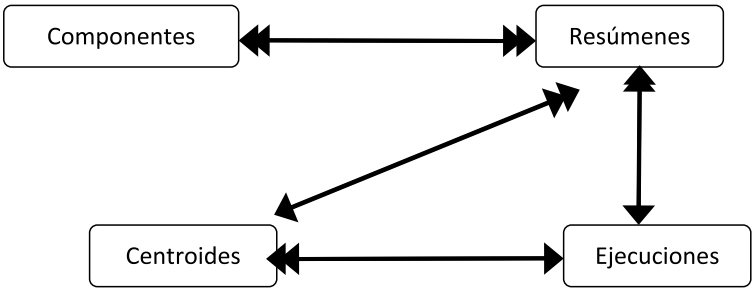

Fig 2. Modelo de Datos del módulo BDRyS, de PROGREL.

En cuanto al módulo MINDAT, en éste se encuentra un subproceso llamado "Segmentación de datos". Este subproceso corresponde a la ejecución del algoritmo de segmentación, que para este trabajo fue K-Means, uno de los más conocidos en su área. Éste es el punto del proceso donde se obtiene la información base para generar los resúmenes lingüísticos; el resultado de su ejecución corresponde a la información de entrada para generar los resúmenes lingüísticos. Como los datos de entrada corresponden a un cubo de datos, cada una de las dimensiones de este cubo corresponderá a una dimensión dentro del vector de posición en el algoritmo de segmentación. En otras palabras, si se está trabajando con un cubo de datos, de N dimensiones, el vector que describe a un punto dentro del espacio de trabajo del algoritmo de segmentación, constará de N componentes una por cada dimensión del cubo. El resultado de la ejecución del algoritmo de segmentación es almacenado en el módulo BDRyS.

El módulo GRES corresponde al "cerebro" de PROGREL. En este módulo es donde se realiza la mayor cantidad de procesamientos y donde las definiciones consideradas para la generación de los resúmenes son cruciales para obtener buenos resultados. Está compuesto por nueve subprocesos, de los cuales cinco realizan lecturas sobre BDRyS, y tres realizan escrituras sobre BDRyS. La ejecución de los subprocesos comienza con la generación de los resúmenes Tipo 0 Segmentos para luego seguir a los de Tipo 2 Multidimensionales.

A. PROGREL: Resúmenes T0

En esta sección se explican las partes principales de PROGREL, comenzando por las tareas que desarrollan los cuatro subprocesos iniciales, enfocados al trabajo de obtener resúmenes de Tipo 0 Segmentos.

El subproceso "Filtro de componentes resúmenes T0" realiza un análisis sobre el resultado obtenido desde el modulo MINDAT. Por cada uno de los centroides generados por el algoritmo de segmentación, se revisa cuál es la posición de éste y qué componentes difusos están descritos cercanos a esta posición. Por ejemplo, considerar que dentro de un espacio bidimensional, donde las dimensiones son "nivel socioeconómico" y "lugar", un centroide está posicionado entre los niveles B y C, para el "lugar" entre el centro de la ciudad y la zona norte de la misma. Como se observa en la Figura 3, no es correcto decir que este centroide pertenece sólo a un "nivel socioeconómico" y a un único "lugar". Por lo tanto, es necesario considerar todos los valores posibles de las dimensiones a las cuales el centroide pertenece, aun cuando a algunas pertenece menos que a otras. Esto entrega el valor de pertenencia del centroide a ese valor de la dimensión, lo que es usualmente llamado "grado de pertenencia".

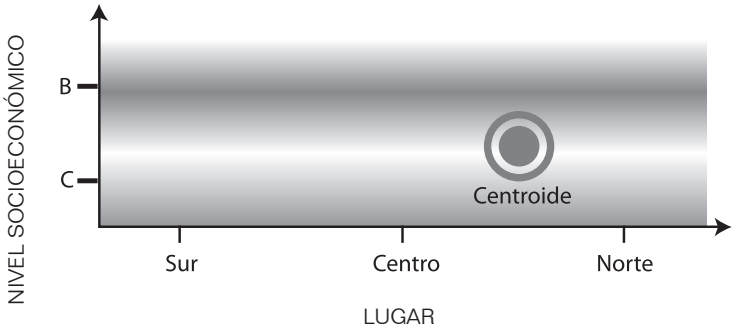

Fig 3. Gráfico de Pertenencia de un Centroide.

Es importante tener en cuenta, durante todo este proceso, que se están trabajando con conceptos difusos, usualmente palabras del lenguaje natural, que se encuentran descritas difusamente. Por lo tanto, el grado de pertenencia es un factor vital a tomar en cuenta al momento de estimar cuál de los resúmenes es el más verídico, es decir, cuál presenta mayor valor de verdad.

Teniendo en cuenta los componentes obtenidos desde el subproceso anterior, a continuación se ejecuta un subproceso en el cual se generan los resúmenes lingüísticos. Estos resúmenes se basan en los componentes seleccionados por cada centroide. Este subproceso recibe el nombre de "Generación de Resúmenes T0".

Por cada uno de los centroides se realizan combinaciones de los componentes obtenidos en el paso anterior, es decir, por cada dimensión se escoge un componente y se concatena con un componente de otra dimensión. De esta manera, se generan todas las combinaciones posibles de componentes por cada dimensión. Es importante tener en cuenta que cada una de estas combinaciones entrega un valor de verdad probablemente diferente, por lo que cada resumen es distinto. Permutaciones de las distintas combinaciones de los componentes son descartadas, dado que una permutación nos entrega la misma información. Luego, por cada uno de los resúmenes T0 generados se calcula el valor del porcentaje de los datos que representa. Éste es un factor muy relevante al momento de comparar los resúmenes con los de otros centroides, dado que un resumen que represente el 10% de los datos entrega menor información global de los datos, respecto del resumen de un centroide que represente al 40%.

A modo de ejemplo, se considera el mismo caso descrito anteriormente, asociado a dos dimensiones: "lugar" y "nivel socioeconómico". Para el centroide de ejemplo, se tienen los siguientes componentes para la dimensión lugar: Centro y Norte, y para la dimensión nivel socioeconómico se tienen los siguientes componentes: B y C. Considerando estos aspectos se obtendrían las combinaciones de resúmenes T0 expuestas en la Tabla II.

TABLA II

EJEMPLOS DE RESÚMENES LINGÜÍSTICOS

Resumen T0	Nivel SocioEconómico	Lugar
#1	B	Centro
#2	B	Norte
#3	C	Centro
#4	C	Norte

A continuación, se realiza el cálculo del valor de verdad para cada uno de dichos resúmenes. Este subproceso obtiene el nombre de "Cálculo valor de verdad T0", dado que corresponde a los resúmenes de Tipo 0. El cálculo del valor de verdad es realizado utilizando la fórmula (5) descrita anteriormente, correspondiente a los resúmenes Tipo 0.

Tras obtener los valores de verdad para cada uno de los resúmenes Tipo 0, se continúa al subproceso llamado "Almacenar resúmenes T0". En este subproceso se almacenan los datos obtenidos desde los subprocesos ante-

riores en BDRyS. Cada resumen Tipo 0 es almacenado junto a su valor de verdad y componentes asociados. Estos datos son almacenados para luego ser desplegados junto a los resúmenes Tipo 2 Multidimensional.

B. PROGREL: Resúmenes T2

El subproceso llamado "Filtro componentes resumen T2" es muy similar al subproceso "Filtro componentes resumen T0". Se diferencia en que los componentes son obtenidos desde los que conforman los resúmenes Tipo 0, es decir que los componentes que son utilizados en los resúmenes Tipo 2, sólo pueden ser algunos de los que están presentes en los resúmenes Tipo 0. Esto debido a que en los resúmenes Tipo 0 cada uno de los componentes describe una dimensión del centroide en particular, y con los resúmenes Tipo 2 se intenta alcanzar un nivel de generalidad mayor, por lo que no es posible agregar componentes que no están presentes al nivel de detalle.

Los resúmenes Tipo 2 generados no corresponden a descripciones promedios de lo entregado con los resúmenes Tipo 0. Los resúmenes Tipo 2 entregan una visión general, que intentan relacionar los segmentos entre sí, dando información que sea común para varios segmentos.

Suponer que se tiene un plano con tres segmentos. Cada uno de estos segmentos estaría compuesto por un centroide y sería descrito por uno o más resúmenes Tipo 0. Éstos sólo entregan información local para cada uno de los segmentos, no sobre algún tipo de relación de éstos entre sí. Gráficamente, como se puede observar en la Figura 4, cada segmento (círculo de color entero) contiene resúmenes Tipo 0. Si se relacionan dos o más segmentos, se obtiene un resumen Tipo 2 Multidimensional (Línea punteada), por ejemplo al relacionar el segmento celeste con el verde obtener un resumen Tipo 2 color rojo.

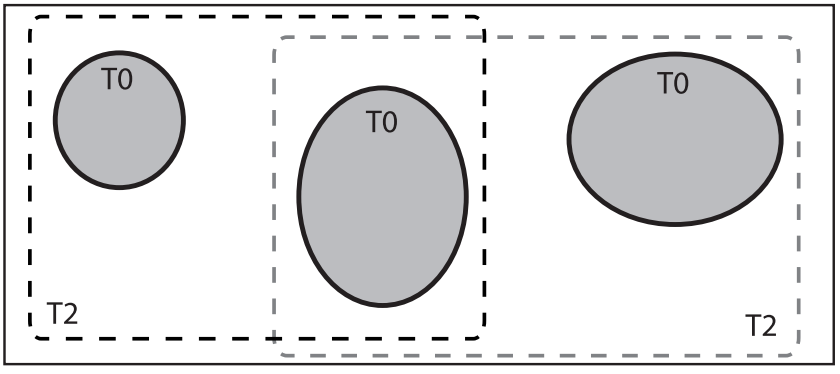

Fig. 4. Nivel de información para los tipos de resúmenes.

Por lo tanto, el subproceso "Filtro componentes resúmenes T2" realiza una selección de los componentes que se encuentran presentes en los resúmenes Tipo 0 obtenidos anteriormente. Esta selección se basa en el valor de verdad que tiene cada uno de los resúmenes en que el componente está presente.

Por cada componente presente en algún resumen Tipo 0, se escoge el valor de verdad promedio de todos los resúmenes en que participa y se

multiplica por la cantidad promedio de elementos de los segmentos de los cuales es parte. De esta manera, cada componente obtiene una calificación ponderada por el tamaño de los segmentos en que participa, y por el valor de verdad de los resúmenes en que está presente. Esta calificación favorece a los componentes que pertenecen a segmentos de mayor tamaño, por lo que los resúmenes que se generen entregarán una mayor cantidad de información relacionada con los segmentos más grandes. Como cada componente tiene asociada una calificación ponderada, explicada anteriormente, se considera que un componente es mejor que otro al tener una calificación de mayor tamaño. Luego se procede a escoger una cantidad arbitraria de componentes. Por ejemplo, en el caso práctico explicado más adelante, se escogieron los mejores seis.

El siguiente subproceso, llamado "Generación resúmenes T2", es el encargado de generar las combinaciones de componentes que corresponderán a los resúmenes. A diferencia de los resúmenes Tipo 0, estos resúmenes están conformados por cuatro tipos de componentes diferentes: cuantificador, descriptor, conector y resumidor. En el subproceso anterior, se realizó el filtrado para escoger los componentes candidatos del tipo descriptor y resumidor. Los componentes del tipo cuantificador corresponden a palabras que describen una cantidad absoluta o relativa, como por ejemplo: "Muchos", "Algunos", "La mitad", "Casi todos", entre otras similares. Los conectores, descritos anteriormente, pueden ser palabras de conjunción, como son "y", "o" o alguna otra similar.

Teniendo en cuenta la protoforma Tipo 2 multidimensional establecida anteriormente, se decidió generar todas las combinaciones posibles entre los distintos descriptores y resumidores permitidos, para luego mezclarlos con todos los posibles cuantificadores y conectores. La cantidad de descriptores y resumidores que se utilizarán como máximo en los resúmenes se determinó de manera arbitraria, aunque por lo general se recomienda no usar más de tres con fin de construir resúmenes simples y fáciles de comprender por el usuario. Para el caso de los descriptores, la situación es análoga.

Se establecieron ciertas reglas para asignar los descriptores y resumidores. Si un componente asociado a la dimensión X está siendo utilizado como un descriptor, no puede haber un resumidor asociado a la misma dimensión X. Esto, debido a que no tiene lógica presentar un resumen en donde el argumento y la conclusión no tengan sentido lógico. Por ejemplo, decir "Todas las compras en el año 2005 ocurrieron durante el año 2003", es un resumen válido desde el punto de vista de la protoforma, pero desde el punto de vista lógico, no tiene sentido. Las permutaciones entre distintas posiciones de los componentes del mismo tipo, en el caso de los resumidores y descriptores, fueron descartadas dado que no alteran el resultado. No importa si todos los descriptores o todos los resumidores, se refieren a la misma dimensión, dado que no afecta el sentido lógico del resumen.

Es posible tener un resumen en donde un componente participa como resumidor, y otro resumen similar en donde el mismo componente participa como descriptor. El encargado de definir los descriptores y resumidores puede establecer si un componente tiene un comportamiento dual o solamente es de un tipo de componente, entendiendo como comportamiento dual la presencia en distintos resúmenes, no de manera simultánea en el

mismo resumen, dado que si estuviera dentro del mismo resumen se estaría violando la primera regla explicada anteriormente.

A continuación, por cada segmento generado por el módulo MINDAT, se revisa cual es su centroide y los valores para las distintas dimensiones asociadas a los componentes que se estén revisando, tanto descriptores como resumidores. Luego se revisa cuales, son los conectores que participan, dado que éstos determinan qué operación se realizará entre los valores obtenidos anteriormente. También se calcula la desviación porcentual para cada segmento, donde se consideran todas las dimensiones en que los componentes participan. Luego, estos cálculos son ponderados por la cantidad de elementos asociados a los segmentos, dado que la cantidad de elementos del segmento representa el peso que éste tendrá en determinar la veracidad del resumen. Por último, se obtiene el resultado de los cálculos explicados anteriormente, valor que es analizado en la definición difusa del cuantificador asociado al resumen. El valor resultante de esta operación corresponde al valor de verdad asociado al resumen Tipo 2 multidimensional.

Luego se ejecuta el subproceso encargado de almacenar los resúmenes que tengan un valor de verdad distinto de cero. Este subproceso es llamado "Almacenar resúmenes T2". Cada resumen es almacenado con su valor de verdad, cuantificador, descriptores, resumidores y conectores, en el módulo BDRyS.

Para finalizar con PROGREL, se ejecuta el subproceso llamado "Limpieza resultados T2". Este subproceso revisa los resúmenes almacenados anteriormente y descartan algunos de ellos con la finalidad de entregar como resultado final un conjunto reducido de resúmenes que entreguen la mejor información. El concepto de mejor información está basado en dos propiedades: la primera se refiere a que los resúmenes con conectores "y" son más simples de leer y más fáciles de entender. Esto se debe a que los resúmenes que utilizan en su mayoría conectores "y" son más restrictivos, logrando entregar una visión menos global al lector. Por lo que éste es más fácil de entender las restricciones que las ambigüedades entregadas por el conector "o". Es por esto que a cada resumen se le asigna un peso dependiendo de la cantidad de conectores "y" que contenga. Entre los resúmenes con los mismos componentes se escoge el que tenga mayor peso. En caso de tener dos o más resúmenes con el mismo peso, se escoge aleatoriamente alguno de ellos.

Cabe mencionar que en BDRyS se encuentran almacenados los resúmenes generados anteriormente, tanto los Tipo 0 como los Tipo 2. Todos los datos generados durante alguno subproceso de PROGREL, que no son almacenados en BDRyS, se guardan en instancias temporales para ser comunicadas a los subprocesos que le siguen, debido a que estos datos aún no están listos como para almacenarlos.

V. Caso de prueba

Como caso de prueba se utilizaron datos relacionados a estadísticas sobre los nacimientos en México [10], los que consistían en aquéllos ocurridos

entre los años 1990 y 2007 especificando en qué estado federado mexicano habían ocurrido, como también en qué lugar (hospital público, hospital privado, domicilio particular). Así, desde un comienzo se identificaron las tres dimensiones asociadas al problema: Fecha, Lugar y Estado Federado. Se desarrolló un proceso para obtener los resúmenes lingüísticos que incorpora el algoritmo de segmentación K-Means implementado en su forma más básica [11]. Se realizaron tres pruebas, con diferentes cantidades de centroides (3, 5 y 6). Parte de los resultados obtenidos se muestran en la Tabla III.

TABLA III

RESULTADOS PARA EL CASO DE PRUEBA

Número de Centroides, Tipo de Resumen Lingüístico	Resumen Lingüístico (valor de verdad asociado)
3, T0 Segmentos	El segmento N°1 representa el 34% de los datos, y se encuentra centrado en los nacimientos ocurridos en Querétaro de Arteaga, durante el año 1993 y en un hospital público (0,989)
3, T2 Multidimensional	Una gran cantidad de los nacimientos en Querétaro de Arteaga y durante el año 2000 fueron en un domicilio particular y en un hospital público (1)
5, T0 Segmentos	El segmento N°1 representa el 19,5% de los datos, y se encuentra centrado en los nacimientos ocurridos en el Distrito de México, durante el año 2001 y en un hospital público (0,999)
5, T0 Segmentos	El segmento N°2 representa el 6.5% de los datos, y se encuentra centrado en los nacimientos ocurridos en San Luis Potosí, durante el año 2003 y en un hospital público (0,986)
5, T2 Multidimensional	La mayoría de los nacimientos en un hospital público, o en un domicilio particular fueron en el Distrito Federal y en Tlaxcala (1)
5, T2 Multidimensional	La mayoría de los nacimientos en un hospital público, o en un domicilio particular fueron en el Distrito Federal y en el Distrito de México (0,804)

VI. Conclusiones

Las protoformas definidas, Tipo 0 Segmentos y Tipo 2 Multidimensional, cumplieron el objetivo de entregar información resumida en base a un conjunto de datos multidimensional. Las fórmulas para el cálculo de valor de verdad entregaron resultados consistentes con los datos, pero su construcción debe ser reformulada y analizada en mayor profundidad. Aun cuando fueron construidas basadas en las fórmulas establecidas por los autores referenciados, la incorporación de nuevas variables no demostró ser la correcta, ni la incorrecta tampoco, estudio que amerita un análisis más exhaustivo a futuro.

Los resúmenes lingüísticos demostraron ser una herramienta capaz de entregar información descriptiva de varias dimensiones de los datos. Aun cuando su construcción está basada en conceptos difusos, sus resultados entregan información acertada y correcta para el contexto sobre el cual se está trabajando.

Referencias

[1] R. Yager. "A New Approach to the Summarization of Data,". *Information Sciences*, 28, pp. 69-86.

[2] L. Zadeh, "A Computational Theory of Dispositions," *Proceedings of the 10th international conference on Computational Linguistics*, 1984, pp. 312-318.

[3] J. Kacprzyk, S. Zadrozny. "Protoforms of Linguistic Database Summaries as a Tool for Human Language for Data Mining," *IJSSCI* (1), 2005, pp 100-111.

[4] J. Kacprzyk. "Fuzzy Logic for Linguistic Summarization of Databases," *Systems Research Institute*, Polish Academy of Sciences, Warsaw, Poland. 1999.

[5] J. Kacprzyk, S. Zadrozny. "Linguistic database Summaries y their Protoforms: towards Natural Language based Knowledge Discovery Tool," *Information Sciences*, 173(4), 2005, pp. 281-304.

[6] J. Kacprzyk, A. Wilbik, S. Zadrozny. "A Linguistic Approach to a Human-Consistent Summarization of Time Series Using a SOM Learned with a LVQ-Type Algorithm," *ICANN* (2), 2006, pp.171-180.

[7] J. Kacprzyk, A. Wilbik, S. Zadrozny. "On some types on Linguistic Summaries of Time Series," *Proceedings of the 3rd. International IEEE Conference on Intelligent Systems*, 2006, pp. 373-378.

[8] J. Kacprzyk, A. Wilbik. "Linguistic Summarization of Time Series using Linguistic Quantifiers: Augmenting the Analysis by a Degree of Fuzziness," Fuzzy Systems, 1(6), 2008, pp. 1146-1153.

[9] F. M. Montesino, A. Barriga, D. R. Lopez, S. Sanchez-Solano. "Linguistic Summarization of Network Traffic Flows," *Fuzzy Systems*, 1(6), 2008, pp. 619–624.

[10] Gobierno de México. SINAIS, Sistema Nacional de Información en Salud. <www.sinais.salud.gob.mx/basesdedatos/index.html>

[11] S. P. Lloyd."Least Squares Quantification in PCM," *IEEE Trans-actions on Information* Theory28 (2), 1982, pp. 129–137.