

Modelo de Datos Difusos en UML: Un caso de Asesoramiento para Inserción de Publicidad en Programas de Televisión

Angélica Urrutia¹

y

Marcela Varas²

¹Dpto. de Computación e Informática, Universidad Católica del Maule, Chile, aurrutia@spock.ucm.cl

²Depto. Ingeniería Informática y Ciencias de la Computación, Universidad de Concepción, Chile, mvaras@udec.cl

Resumen

Un modelo de datos hoy en día es fundamental cuando se requiere de una base de datos que reúna todas las especificaciones de requerimiento de un sistema dado por un usuario. Aquí se presenta un modelado de datos para *un caso de asesoramiento para inserción de publicidad en programas de televisión* con la notación de UML, permitiendo incorporar atributos difusos (con referencial ordenado o no ordenado) con el uso de patrones y estereotipos en la definición de un patrón de diseño denominado “Patrón Atributos Difusos”. Además, para restricciones de multiplicidad se propone el uso de cuantificadores difusos (relativos o absolutos).

Palabras claves: Atributos Difusos en UML, Datos Imprecisos, Patrón Atributos Difusos, Modelo de Datos Difusos.

1 Introducción

En Rumbaugh et. al. [7] se afirma que un propósito fundamental de los modelos es que permiten “captar y enumerar exhaustivamente los requisitos y el dominio de conocimiento, de forma que todos los implicados puedan entenderlos y estar de acuerdo con ellos”. Es por ello, que este trabajo se centra en la captura de aquellos datos que posean un dominio impreciso que actualmente no son representados por los modelos, debido a la falta de expresividad de los lenguajes de modelación. En esta propuesta utilizamos la herramienta de modelado de datos en UML.

Originalmente, el nivel conceptual permite la utilización de tipos de datos elementales a los que se les llama clásicos (crisp). Estos tipos de datos son numéricos, alfanuméricos y datos binarios. Existe otro tipo de dato, que no son tratados en los modelos de datos conceptuales clásicos o tradicionales, que contienen incertidumbre o imprecisión en su información. Estos últimos se consideran datos “difusos” (fuzzy) que los humanos manejamos de forma cotidiana y natural, los cuales se pueden asociar a la teoría de conjuntos difusos [10]. Por concepto o información difusa entendemos información que encierra alguna imprecisión o incertidumbre.

Una investigación que profundiza en el modelado conceptual de datos es Ma et al. [5], estos autores proponen un tipo de atributos que indica el grado de importancia que éste tiene para cada entidad u objeto. Para el caso de nuestra propuesta usamos tipos de atributos que poseen dominios imprecisos o difusos, los cuales se encuentran definidos en una clase. Por otro lado, Geneste y Ruet [1][2], proponen una clase que incorpora un trapecioide de una etiqueta lingüística tratado como grado de pertenencia en una aplicación específica de un caso tratado en UML, esta propuesta es simplista, pues normalmente los atributos difusos tipo 2 obedecen a más de una etiqueta lingüística.

Este artículo presenta en la sección 2 una serie de conceptos básicos de la teoría de conjuntos difusos y atributos difusos, así como también, definiciones de notación en UML, la sección 3 presenta la definición de Patrón Atributos Difusos, en la sección 4 se desarrolla el caso de Asesoramiento de inserción de publicidad en programas de televisión utilizando el patrón definido en la sección 3, y finalmente cerramos el artículo con las conclusiones y trabajo futuro.

2 Conceptos básicos utilizados

El UML (Lenguaje Unificado de Modelado) es un lenguaje de modelado visual que se usa para especificar, visualizar, construir y documentar artefactos de un sistema de software. Captura decisiones y conocimiento sobre los sistemas que se deben construir, su objetivo es lograr que, además de describir con cierto grado de formalismo tales sistemas, puedan ser entendidos por los usuarios de aquello que se modela. Nuestra investigación hace uso de las herramientas de UML para el modelado de datos, especialmente se basará en estereotipos y patrones, como una forma especial de representar atributos con imprecisión utilizando la teoría de conjuntos difusos (lógica difusa).

Estereotipo en UML: Es un nuevo elemento del lenguaje definido sobre la base de algún elemento pre existente de UML. Extiende la semántica pero no la estructura de las clases del metamodelo. Permite representar una variación de un elemento existente que posee otra intención, o distinción de uso. La definición de un estereotipo se hace en forma explícita en la vista estática, mediante una relación de generalización con el elemento de UML que es base para su definición. El nombre del estereotipo debe ser distinto de los elementos de UML, y se denota entre comillas francesas («nombre estereotipo»). También puede considerar una notación gráfica distintiva.

Patrones: Solución ya probada y eficaz para algún problema de diseño que puede expresarse como un conjunto de principios y heurísticas. Los patrones se definen con un nombre, la solución propuesta y el problema que resuelve [4].

Conjuntos difusos: Zadeh en 1965 [11] definió el concepto de conjunto difuso, basándose en la idea de que existen conjuntos en los que no está claramente determinado si un elemento pertenece o no al conjunto. A veces, un elemento pertenece al conjunto con cierto grado, llamado grado de pertenencia. Por ejemplo, el conjunto de las personas que son altas es un conjunto difuso, pues no está claro el límite de altura que establece a partir de que medida una persona es alta o no lo es. Ese límite es difuso y, por tanto, el conjunto que delimita también lo será. Un conjunto difuso A sobre un universo de discurso U (ordenado) es un conjunto de pares dado por: $A = \{\mu_A(u) / u : u \in U, \mu_A(u) \in [0,1]\}$, Donde, μ es la llamada función de pertenencia y $\mu_A(u)$ es el *grado de pertenencia* del elemento u al conjunto difuso A . Este grado oscila entre los extremos 0 y 1, $\mu_A(u) = 0$, indica que u no pertenece en absoluto al conjunto difuso A , $\mu_A(u) = 1$, indica que u pertenece totalmente al conjunto difuso A .

Atributos difusos: Los atributos difusos se clasifican en tres tipos: Tipo 1: estos atributos son “*datos precisos*” (clásicos, sin imprecisión) que pueden tener etiquetas lingüísticas definidas sobre ellos. Los atributos de Tipo 1 reciben una representación igual que los datos precisos, pero pueden ser manejados en condiciones difusas. Por ejemplo, una persona es *alta*. Tipo 2: son atributos que pueden almacenar o tomar “*datos imprecisos sobre referencial ordenado*”. Estos atributos admiten tanto datos clásicos como difusos, en forma de distribuciones de posibilidad. Por ejemplo, la edad puede tener las etiquetas *niño, joven, adulto*, referenciadas sobre un conjunto entre 0 y 100. Tipo 3: son atributos sobre “*datos imprecisos sobre referencial no ordenado normalizado*”. Estos atributos son definidos sobre un dominio subyacente no ordenado, por ejemplo, el atributo “color del pelo” puede tener las etiquetas *rubio, pelirrojo y castaño*.

Nuestra propuesta sugiere usar los atributos difusos Tipo 1, 2 y 3 para caracterizar los *conceptos* del esquema del dominio, ya que los *conceptos* pueden ser tratados como clases de un modelo orientado objeto.

Cuantificadores difusos: Los cuantificadores difusos [9, 23, 24] permiten expresar cantidades o proporciones difusas para dar una idea aproximada del número de elementos de un subconjunto (o que cumplen cierta condición) o de la proporción de ese número en relación con el total de elementos posibles. Los cuantificadores pueden ser absolutos o relativos: Cuantificadores absolutos: expresan cantidades sobre el número total de elementos de un determinado conjunto, diciendo si este número es “grande”, “muchísimos”, “aproximadamente entre 5 y 10”, etc. Cuantificadores relativos: expresan mediciones sobre el número total de elementos que cumplen cierta característica dependiendo del total de elementos posibles, por lo que la verdad del cuantificador depende de dos cantidades. Este tipo de cuantificadores se usa en expresiones como “la mayoría”, “la minoría”, “aproximadamente 40 años”.

Nuestra propuesta sugiere usar los cuantificadores difusos (absolutos y relativos) para caracterizar algunos atributos difusos, así como también, la restricción de multiplicidad en UML.

3 Patrón Atributos Difusos en UML

Rumbaugh et. al. [7] definen un atributo en UML como la descripción de una ranura con el nombre de un tipo especificado en una clase; cada objeto de la clase tiene un valor independiente para el atributo. Un atributo se representa mediante una cadena de texto que puede dividirse en varias propiedades, su notación es:

«estereotipo» visibilidad nombre multiplicidad : tipo = valor-inicial {cadena de propiedades}.

El enfoque propuesto se basa en la definición de los estereotipos «FuzzyT1», «FuzzyT2» y «FuzzyT3» para cada uno de los atributos difusos del Tipo 1, 2 y 3 respectivamente. Los estereotipos mencionados corresponden a clases que incorporan los métodos y atributos necesarios para manejar las propiedades difusas de estos datos, definidos como:

- La clase estereotipo «FuzzyT1» tiene asociado una lista de etiquetas lingüísticas, de modo de posibilitar el manejo de los atributos difusos Tipo 1.
- La clase estereotipo «FuzzyT2» se compone de una lista de trapecios que tienen asociada una etiqueta lingüística. El tipo de dato (valores) de los atributos a, b, c y d, deben asignarse según el dominio del atributo difuso Tipo 2 que se utilizará.
- La clase estereotipo «FuzzyT3» tiene asociada una matriz que almacena los grados de similitud (gs) que se le atribuyen a los pares de etiquetas lingüísticas en consideración. No siempre la relación de similitud es simétrica, por lo que se diferencié el orden de las etiquetas en la relación (posición 1 y posición 2).

Para aplicar el estereotipo en un caso concreto, se debe hacer uso de la relación de especialización, y generar una clase del tipo de atributo requerido, y asignarle el tipo de dato correspondiente. Este proceso se realiza en la “vista de definición difusa”. En la vista estática del problema en cuestión se usará la clase como dominio del atributo difuso correspondiente, el cuál además será identificado con la etiqueta del estereotipo correspondiente.

Vista Estática Patrón Atributos Difusos.

Para poder modelar datos imprecisos o difusos en UML, nuestra propuesta está basada en una generalización, de tal forma que, el atributo difuso es definido en la superclase y el dominio del tipo de atributo difuso (Tipo 1, Tipo 2 o Tipo3) es definido en una subclase, obteniendo así, en cada subclase un tipo de representación correspondiente a un atributo impreciso. Esta subclase puede ser tratada como una función de distribución o una función de similitud según corresponda a cada uno de los casos. Además, se establece una multiplicidad 1..*. La Figura 1 muestra la representación en notación definida anteriormente.

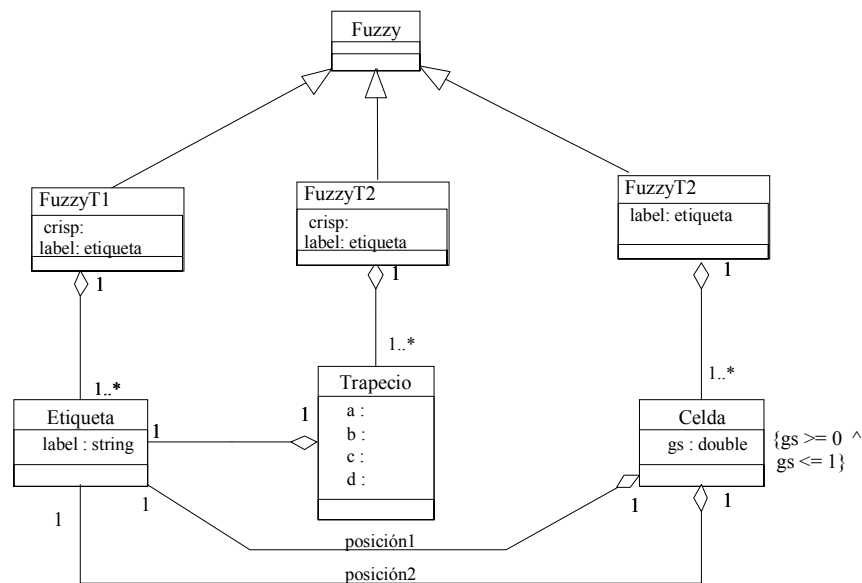


Figura 1: Patrón Atributos Difusos

El atributo clásico o crisp presente en FuzzyT1 y FuzzyT2 puede admitir valores precisos del tipo de dato especificado por el diseñador, además de los valores Unknown, Undefined y Null. El atributo etiqueta puede tomar como valor cualquiera de las etiquetas que están asociadas a la clase (que determina de ese modo el dominio de etiquetas posibles). Puede tomar el valor nulo cuando crisp tiene valor.

Consideramos que este tipo de representación de atributos difusos (Figura 1), en su notación muestra la semántica asociada a cada tipo de atributo, siendo un aporte a los diseñadores de bases de datos que contemplen en su modelado este tipo de expresiones. Además, el modelado en UML permite representar la estática correspondiente a los atributo, así como también, la dinámica que involucra métodos y operaciones.

4 Caso: Asesoramiento para inserción de publicidad en programas de televisión

A través de un estudio de marketing, se ha comprobado que los mayores de 40 años e inactivos profesionalmente, prefieren el horario de mañana, mientras que los activos de esa edad prefieren el horario de noche. Por otro lado, las personas menores de 40 años y activos prefieren el horario de tarde, pero si son inactivos prefieren el horario de madrugada y, contrariamente a todos los restantes, en una estación via satélite.

Se ha comprobado también que la televisión por cable más vista en horario de mañana es RTVE-1, por la tarde es Antena 3 y de noche Tele5. Entre las de satélite, la más vista por la madrugada es SuperSport.[10]

Utilizando las técnicas clásicas para identificar entidades a partir de un texto (búsqueda de sustantivos, etc.) se ha identificado el concepto, que hemos llamado: “televidente”. La Figura 2 muestra la modelización de éstos, en lenguaje UML.

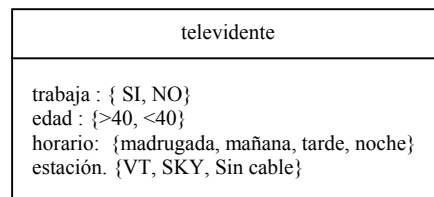


Figura 2: Representación gráfica de la clase “Televidente”.

4.1 Atributos Difusos

En la figura 2 el atributo “trabaja” indica si el televidente realiza o no una actividad profesional (sale de casa por un período determinado de tiempo). A simple vista se puede pensar que este atributo puede recibir, solamente, valores del tipo booleano, donde el SI indica que la persona es activa y el NO indica lo contrario. Sin embargo, si la persona es activa, el uso de un atributo difuso de Tipo T2 permite ampliar el conocimiento de las preferencias televisivas de los televidentes. En efecto, si utilizamos, por ejemplo, un cuantificador difuso tales como “*todos los días*”, “*aproximadamente 2 días*”, “*algunos días*”, es posible construir una función de distribución de posibilidad para representar algunas de estas expresiones. La Figura 3 muestra el cuantificador “algunos días” y su notación en UML.

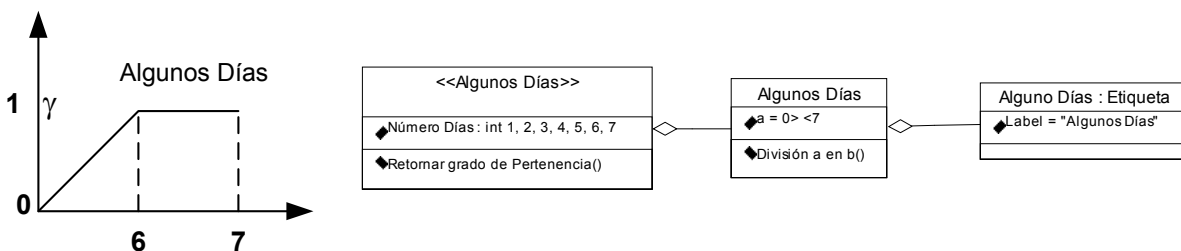


Figura 3: Uso del Patrón Atributos Difusos Tipo 2 para cuantificador difuso relativo “Algunos días”.

El atributo “*edad*” indica si la persona es menor o mayor que 40 años. En el caso, no se mencionan las preferencias de los televidentes de 40 años, ni tampoco la de los televidentes que tienen aproximadamente 40 años. En este sentido, el uso del atributo difuso de Tipo T2 con la etiqueta “aproximadamente 40” puede tratar la imprecisión de decir con que grado nos acercamos al valor 40. Por otra parte, la utilización de un atributo difuso de Tipo T2 asociado a las etiquetas lingüísticas, por ejemplo: “joven”, “maduro” y “mayor” pueden ser de gran ayuda en el momento de analizar los datos de la encuesta, ya que es posible obtener una función de distribución de posibilidad para cada una de las etiquetas. En que, por ejemplo, el conjunto difuso “joven” es un trapecio dado por sus 4 valores característicos {0/15, 1/20, 1/25, 0/30}; el conjunto difuso “maduro” es un trapecio dado por sus 4 valores característicos {0/25, 1/30, 1/40, 0/50}; el conjunto difuso “mayor” es un trapecio dado por sus 4 valores característicos {0/40, 1/50, 1/65, 0/90}. Otra ventaja de esta representación es que, por ejemplo, podemos decir que una persona de 26 años pertenece al conjunto difuso joven, con un grado de cumplimiento o pertenencia de 0.8. La Figura 4 muestra la etiqueta lingüística “aproximadamente 40” y su notación en UML.

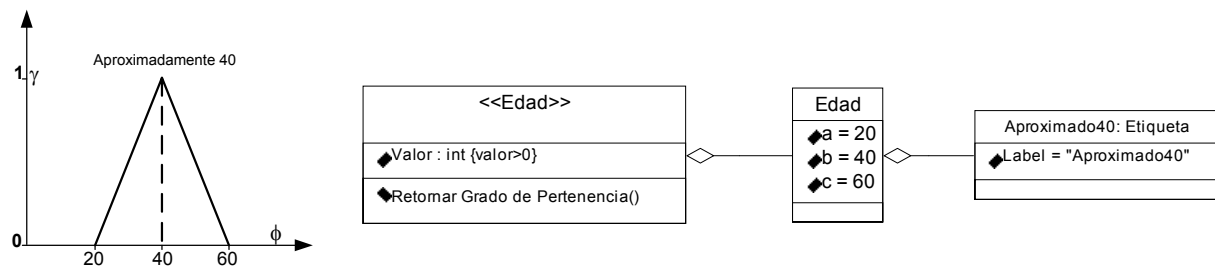


Figura 4: Uso del Patrón Atributos Difusos Tipo 2 para etiqueta lingüística “Aproximadamente 40”.

El atributo “*horario*” indica la jornada que prefiere el televidente. El estudio de marketing indica que esta puede ser en la mañana, en la tarde, en la noche o en la madrugada. Dado que estos valores pueden ser definidos en un referencial ordenado, este atributo puede ser tratado como de Tipo T2 del conjunto difuso de la jornadas {“madrugada”, “mañana”, “tarde”, “noche”} asociadas a una distribución de posibilidad, en que, por ejemplo, el conjunto difuso “madrugada” es un trapecio dado por 4 valores característicos {0/0, 1/2, 1/6, 0/8}; el conjunto difuso “mañana” es un trapecio dado por 4 valores característicos {0/6, 1/8, 1/12, 0/14}; el conjunto difuso “tarde” es un trapecio dado por 4 valores característicos {0/12, 1/14, 1/18, 0/20}; el conjunto difuso “noche” es un trapecio dado por 4 valores característicos {0/18, 1/20, 1/23, 0/24}, tal como lo muestra la Figura 5 para la clase horario con su respectivo conjunto difuso asociado, lo que significa una flexibilidad mayor en el momento de analizar la encuesta.

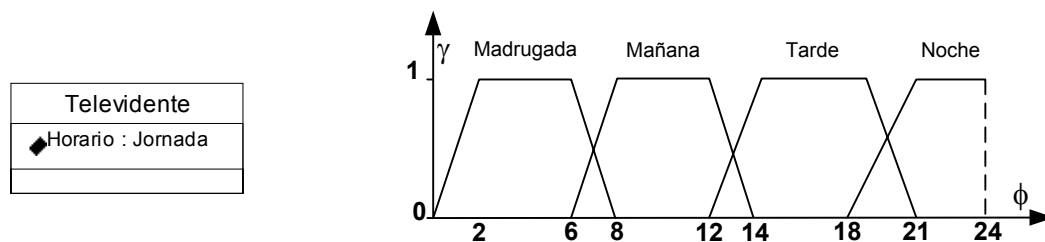


Figura 5: Clase televidente con atributos difusos Tipo 2 para etiquetas lingüísticas “jornadas”.

En clase televidente ha sido definido el atributo Horario como un atributo difuso Tipo 2, para lo cual en la Figura 1, el *Patrón Atributos Difusos* está representado por FuzzyT2 asociado al conjunto de etiquetas de Jornada. El modelo generado para este caso es mostrado en la Figura 6.

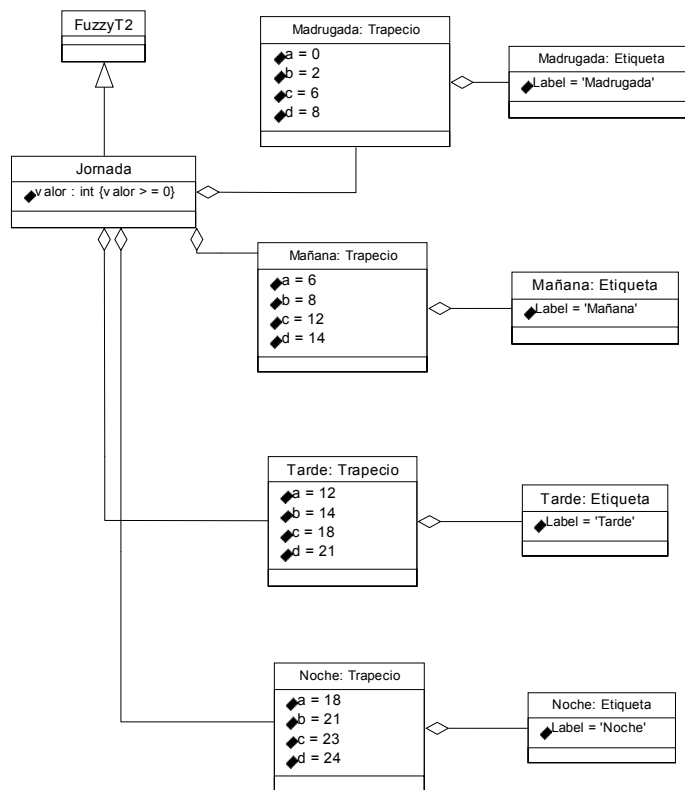


Figura 6: Clase televidente con Patrón Atributos Difusos Tipo 2 para etiquetas lingüísticas de “Jornadas”.

El atributo “*estación*” indica si el televidente prefiere ver la televisión por cable, por ejemplo VTR o prefiere ver los programas transmitidos a través de una antena parabólica, por ejemplo Sky Channel. Un tratamiento más interesante que se le puede dar a este tipo de datos es considerar ciertos grados de pertenencia que el televidente asocia a una estación, por ejemplo, un televidente puede preferir Sky Channel o VTR con un grado de 0.8, VTR o no ver cable con un grado de 0.2 y la televisión sin cable o SKY con un grado de 0.5, en consecuencia, el televidente puede preferir cualquiera de las estaciones pero con un cierto umbral de preferencia. En ese caso, se trata de un dato difuso del Tipo 3 con una función de similitud asociada o bien solamente con los grados asociados. La Figura 7 muestra la clase televidente con el tipo de estación, asociada a una tabla de similitud.

Televidente	VTR	SKY	Sin Cable
◆ Tipo Estación : Estación	1	0.8	0.2
	0.8	1	0.5
	0.2	0.5	1

Figura 7: Clase televidente con atributos difusos Tipo 3 para relación de similitud “Estación”.

En clase televidente ha sido definido el atributo Tipo estación como un atributo difuso Tipo 3, para lo cual en la Figura 1, el *Patrón Atributos Difusos* está representado por FuzzyT3 asociado a la relación de similitud de Estación. El modelo generado para este caso es mostrado en la Figura 8.

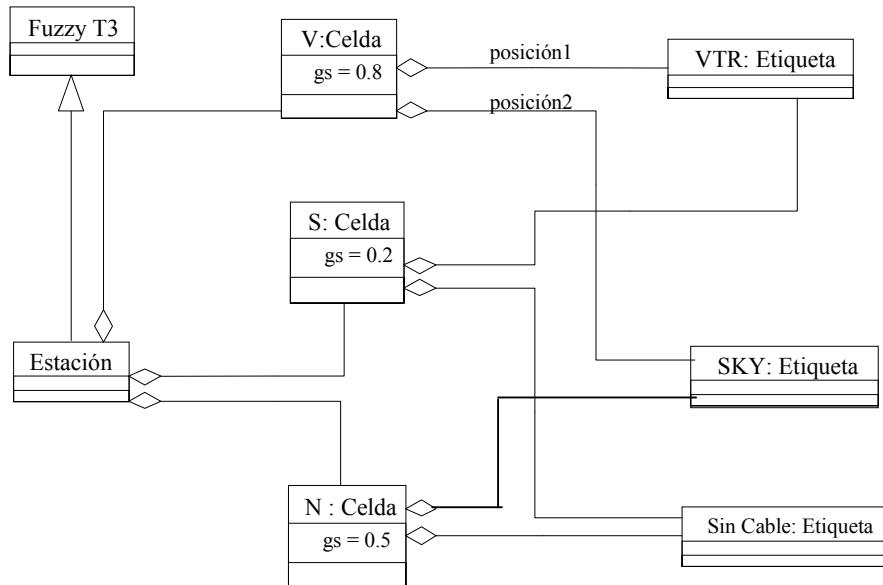


Figura 8: Clase televidente con Patrón Atributos Difusos Tipo 2 para etiquetas lingüísticas “Jornadas”.

La celda V corresponde a VTR, S a SKY y N a Sin Cable.

4.2 Relación entre clases

De la descripción del caso de estudio, algunos atributos de los expuestos anteriormente pueden ser definidos como clases relacionadas con televidente, por ejemplo tipo de estación y tipo de jornada. Lo que emerge a simple vista son las restricciones de multiplicidad entre clases, como sabemos estas pueden ser “uno a uno” (1..1), “uno a muchos” (1..*), “muchos a muchos” (*..*).

Por consiguiente, el tipo de restricción de multiplicidad depende de la significación y del contexto de la situación a modelar. Supongamos, entonces, la restricción siguiente: “Un televidente puede seleccionar uno o más canales de un tipo de estación según su preferencia horaria”. Desde esta perspectiva, el uso de cuantificadores difusos sobre la multiplicidad, por ejemplo, el cuantificador difuso relativo “*aproximadamente*” o “*casi todos*” pueden ampliar el horizonte de modelización ya que nos entregaran un porcentaje de instancias que se relacionan entre una clase y otra.

En efecto, ahora es posible modelar una restricción que exprese la condición que un televidente seleccione *casi todos* los canales de un tipo de estación, en cuyo caso, el cuantificador difuso relativo “*casi todos*” es una multiplicidad entre las clases; en este caso televidente y canal (HBO, TNT, ...) de un tipo de estación (VTR, SKY...). La Figura 9 muestra la notación de multiplicidad “*casi todos*”. Un televidente tiene como preferencia casi todos los tipos de estación, así como un tipo de estación tienen casi todos los canales asociados. Cabe destacar que cada “*casi todo*” se obtienen a partir de la división entre el número máximo de instancias de cada clase sobre la restricción mínima de instancias definidas para cada caso.

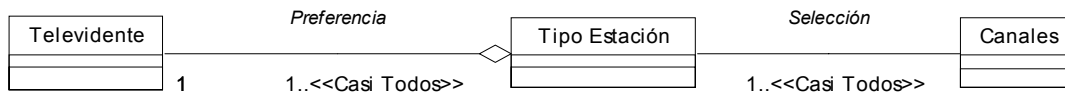


Figura 9: Clase televidente y Tipo de estación con multiplicidad de cuantificador difuso “Casi todos”.

Por otro lado, si consideramos Horario como una clase, se puede definir una restricción entre las clases Televidente y Horario, en esta restricción se encuentran las instancias que definen los horarios de preferencia que tienen los televidentes encuestados dentro del conjunto difuso de {mañana, tarde, noche, madrugada} como las definidas en la Figura 6. Un ejemplo del uso de un cuantificador difuso “casi todos” definido como multiplicidad entre Televidente y Horario se muestra en la Figura 10. En casi todo representa que; “casi todas las instancias de televidente prefieren un determinado tipo de jornada”. Al igual que la restricción anterior el casi todo se establece por la división del número total de televidentes y la restricción mínima asociada. Generando un valor entre 0 y 1.

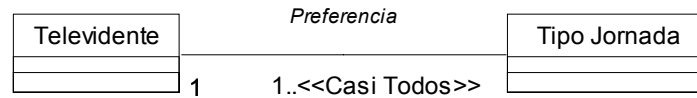


Figura 10: Clase Televidente y Tipo de Jornada con multiplicidad de cuantificador “Casi todos”.

Estas clases se pueden mezclar con un tipo de clases ternarias donde la multiplicidad puede estar dada por “casi todos los televidentes prefieren casi todos los canales a un cierto horario”.

5 Conclusiones y trabajos futuros

Las bases de datos difusas han sido ampliamente estudiadas con el objetivo de permitir el almacenamiento de datos imprecisos o difusos y la consulta de forma imprecisa de los datos existentes. Sin embargo, tradicionalmente la aplicación de la lógica difusa a las bases de datos ha prestado escasa atención al problema del modelado conceptual. La extensión del modelo UML para el tratamiento de datos difusos, ha sido estudiado en pocas publicaciones, pero en ninguna se referencia la posibilidad de modelar datos imprecisos en forma genérica utilizando etiquetas lingüísticas, referencias ordenado o de semejanza utilizando las herramientas que ofrece la teoría de conjuntos difusos.

En [9] los autores han propuesto un modelado de atributos imprecisos utilizando la notación EER, la gran diferencia con nuestra propuesta en UML, es que la semántica está presente en el modelado de estereotipos y representada en patrones de *Atributos Difusos*, no en el diccionario de datos como proponen los autores de [9], siendo UML más flexible y dando una gran expresividad al modelo de datos imprecisos con el tratamiento de la teoría de conjuntos difusos sobre todo para el caso en estudio de Asesoramiento para inserción de publicidad en programas de televisión.

Como trabajo futuro se pretende depurar los estereotipos y patrón propuesto además de seguir aplicando la teoría de conjuntos difusos al resto de las notaciones en UML, como son la multiplicidad y herencia, permitiendo así un modelado en UML aún más flexible.

6 Referencias

- [1] Geneste L., Ruet M., Fuzzy Case Based Configuration (2002): 15th European Conference on Artificial Intelligence, ECAI'2002, Workshop on Configuration, Lyon, France.
- [2] Geneste L., Ruet M. (2001): Experience based configuration, 17th International Conference on Artificial Intelligence, IJCAI'01, Workshop on Configuration, Seattle, Washington, USA, 4-10 august 2001.
- [3] Galindo J. (1999): Tratamiento de la Imprecisión en Bases de Datos Relacionales: Extensión del Modelo y Adaptación de los SGBD Actuales. Ph. Doctoral Thesis, University of Granada (Spain). (www.lcc.uma.es).
- [4] Larman C. (1999), “UML y Patrones”, Prentice Hall.
- [5] Ma Z. M., Zhang W. J., Ma W. Y., Chen Q. (2001) Conceptual Design of Fuzzy Object-Oriented Databases Using Extended Entity-Relationship Model. International Journal of Intelligent System. Vol 16. pág 697-711, 2001
- [6] Marín N., Pons O., Vila M.A. (2000): Fuzzy Types: A New Concept of Type for Managing Vague Structures. International Journal of Intelligent Systems, 15, pp. 1061-1085.

- [7] Rumbaugh J., Jacobson J., Booch G., (1999): "The Unified Modeling Language Reference Manual", Addison Wesley.
- [8] Urrutia A., Piattini M. (2001): Transformation of imprecise data to linguistic labels for model E/R. Conferencia SCI/2001. 7th International Conference on Information System Analysis and Synthesis (ISAS2001), Pág. 351, 355. Orlando, USA.
- [9] Urrutia A., Galindo J. (2001): "Notación para datos con imprecisión en un modelo conceptual EER difuso", UC-Maule Revista Académica de la Universidad Católica del Maule, diciembre N° 27, pág. 39-48.
- [10] Jiménez L., Urrutia A.: (2002): "Extensión del conocimiento del dominio de commonKads con lógica difusa", 5° Workshop Iberoamericano de Ingeniería de Requisitos y Ambiente de Software, pág. 389-393, IDEAS'02, Cuba.
- [11] Zadeh L. A. (1965): Fuzzy Sets. Information and Control, 8, pp. 338-353.