Capítulo 3

Estado del Arte

"No leas para contradecir o refutar, ni para creer o dar por bueno,
ni para buscar materia de conversación o de discurso,
sino para considerar y ponderar lo que lees".

Francis Bacon, escritor inglés

El diseño de bases de datos es un proceso complejo que abarca varias decisiones a muy distintos niveles. La complejidad se controla mejor si se descompone el problema en subproblemas y se resuelve cada uno de éstos independientemente, usando métodos y técnicas específicas. Así, el diseño de bases de datos se descompone en: diseño conceptual, diseño lógico y diseño físico. El diseño de bases de datos, representa un enfoque orientado a los *datos* para el desarrollo de los sistemas de información, donde, la atención completa del proceso de diseño se centra en los datos y sus propiedades. Con un enfoque orientado a los datos, por lo general clásicos, primero se diseña la base de datos, luego las aplicaciones que la usan.

Por otro lado, el auge de los modelos de datos con capacidad para manejar *imprecisión* e *incertidumbre* genera la necesidad de implementar nuevos tipos de herramientas para manejar esa imprecisión. En esta investigación nos centraremos en esta temática, en especial aplicando la *teoría de conjuntos difusos*.

En los últimos años, la teoría de conjuntos difusos se ha asentado como una herramienta eficaz para extender los modelos y metodologías existentes para la representación y manejo del conocimiento, con la imprecisión e incertidumbre asociada tradicionalmente al lenguaje humano. Paralelamente al desarrollo de dichas extensiones difusas, surge la necesidad de ampliar las herramientas y aplicaciones existentes, para abordar con esta nueva herramienta los problemas clásicos. Es así como encontramos algunos modelos conceptuales en este ámbito producto de investigaciones en: Yazici y Merdan (1996), Ma et al. (2002), Chen (1998) entre otros, y algunas implementaciones de SQL difusos en: Galindo et al. (1998) y Galindo (1999).

En concreto, uno de los modelos que más se han visto beneficiados por la aplicación de la teoría de conjuntos difusos ha sido el modelo relacional de bases de datos, en cuanto a aplicaciones se refiera, siendo varias las extensiones propuestas entre las que destacamos las que son presentadas en al Apéndice III y apartado 3.3.

En este apartado se recopilan conceptos de modelos conceptuales de datos, conceptos de conjuntos difusos y se resumen algunas de las investigaciones que hacen referencia a los modelos de datos conceptuales y otras que han sido consideradas de aporte a esta tesis.

3.1 Modelado Conceptual de Datos

Batini et al. (1994) dicen que el diseño conceptual parte de la especificación de requerimientos y su resultado es el esquema conceptual de la base de datos. Un esquema conceptual es una descripción de alto nivel de la estructura de la base de datos, *independientemente* del software del SGBD que se use para manipularla. Un modelo conceptual es un lenguaje que se usa para describir esquemas conceptuales. El propósito del diseño conceptual es describir el *contenido de información* de la base de datos, más que las *estructuras de almacenamiento* que se necesitarán para manejar esa información.

Un modelo de datos es una serie de conceptos que pueden utilizarse para describir un conjunto de datos y operaciones para manipular los mismos. Cuando un modelo de datos describe un conjunto de conceptos de una realidad determinada, se llama modelo conceptual de datos. Los conceptos de un modelo de datos se construyen por lo regular usando mecanismos de abstracción y se describen mediante representaciones lingüísticas y gráficas; es decir, puede definirse una sintaxis y puede desarrollarse una notación gráfica como partes de un modelo de datos.

En especial esta tesis extiende el modelo conceptual ER/EER, se utiliza para ellos la nomenclatura de: De Miguel et al. (1999), Elmasri y Navathe (2002), Atzeni et al. (1999) y Batini et al. (1994). Estos autores en su metodología definen en primera instancia las entidades e interrelaciones luego incorporan sus atributos y restricciones.

En Batini et al. (1998) se expone que el bloque común de construcción a todos los modelos de datos es una pequeña colección de mecanismos de abstracción primitivos:

Clasificación, agregación y generalización. Entonces tenemos tres tipos de abstracción en un diseño conceptual:

- **Abstracción de clasificación**: Se usa para definir un concepto como una clase o entidad de un objeto de la realidad, caracterizado por propiedades comunes. Por ejemplo, tenemos que el concepto Vehículo es la clase cuyos miembros de clasificación son: auto, bicicleta, camión, carreta, etc.

- **Abstracción de agregación**: Define una nueva clase a partir de un conjunto de (otras clases) que representan sus partes componentes. La abstracción por agregación se representa por un árbol de un nivel en el cual todos los nodos son clases; la raíz representa la clase creada por agregación de las clases representadas por las hojas. Cada rama del árbol indica que una clase hoja es una *parte de* la clase representada por la raíz. Este concepto lo enfoca a dos tipos agregación: de clases y de atributos, por ejemplo, Vehículo es una agregación de clases de (chasis, ruedas, motor...), y Persona es una agregación de atributo de (código, nombre, edad, etc.).

- **Abstracción de generalización**: Define una relación de subconjunto entre los elementos de dos (o más) clases. Por ejemplo, la clase *vehículos* es una generalización de *bicicleta*, *auto* y *camión*. El proceso inverso a este es llamado especialización.

Además, identifican correspondencias que se establecen entre dos clases, *cardinalidad mínima* (es el número mínimo de correspondencias en las que cada elemento de una clase puede participar), *cardinalidad máxima* (es el número máximo de correspondencia en las que cada elemento de la clase puede participar).

También una abstracción de generalización establece una correspondencia entre las clases genéricas y las clases subconjuntos, que son llamadas propiedades de cobertura: *cobertura total o parcial*, *cobertura exclusiva* o *superpuesta* (disjunta o solapada).

Cada una de estas componentes y otras se expondrán, con mayor detalle, en el capítulo 4, para mayor facilidad del lector en el seguimiento de cada una de las extensiones del modelo ER/EER hacia el modelo conceptual de datos difusos.

3.2 Representaciones de Difusas

Antes de comenzar describiendo brevemente la lógica difusa y su origen, preguntémonos ¿qué significa *fuzzy*?; término sobre el cual se sustenta una forma de expresar las leyes, modos y formas del conocimiento científico (lógica).

Originalmente el término *fuzzy* procede de *fuzz*, que sirve para denominar la pelusa que recubre el cuerpo de los polluelos al poco de salir del huevo. Este término en inglés significa "confuso, borroso, no definido o desenfocado". Este término se traduce por "*flou"* en francés y se pronuncia "*aimai"* en japonés. La traducción de esta palabra al castellano es difuso o borroso, aunque *fuzzy*, en los ámbitos académico y tecnológico, está aceptado tal cual, de forma similar a como lo es *"bit"*. *Fuzzy* significa ambiguo o vago, en el sentido del razonamiento humano, más que en la acepción de probabilidad de algo.

La lógica difusa nació cuando Lotfi A. Zadeh publicó un artículo titulado *"Fuzzy Sets"* (Conjuntos Difusos) (Zadeh, 1995). En este artículo el Dr. Zadeh presentó unos conjuntos sin límites precisos los cuales, según él, juegan un importante papel en el reconocimiento de formas, interpretación de significados, y especialmente abstracción, la esencia del proceso de razonamiento del ser humano.

En la lógica clásica sólo es posible tratar información que sea totalmente cierta o totalmente falsa; no le es posible manipular aquella información imprecisa o incompleta inherente a un problema que contiene datos que permitirían una mejor resolución del mismo. Con ello se podría decir que la lógica difusa es una extensión de los sistemas clásicos, como el propio Zadeh indica (Zadeh, 1992). La lógica difusa es la lógica que soporta modos de razonamiento aproximados en lugar de exactos. Su importancia radica en que muchos modos de razonamiento humano, en especial el razonamiento según el sentido común, son aproximados por naturaleza.

Esta lógica es una lógica multievaluada, sus características principales, presentadas por Zadeh en la referencia antes mencionada son:

- En la lógica difusa, el razonamiento exacto es considerado como un caso particular del razonamiento aproximado.
- Cualquier sistema lógico puede ser trasladado a términos de lógica difusa.

- En lógica difusa, el conocimiento es interpretado como un conjunto de restricciones flexibles, es decir, difusas, sobre un conjunto de variables.
- En lógica difusa, todo problema es un problema de grados.

Se podría decir que la lógica difusa permite a los ordenadores trabajar no sólo con métodos cuantitativos sino también cualitativos, se trata pues de un intento de aplicar una forma más humana de pensar en la programación de computadoras.

A continuación se resumen algunos conceptos de la teoría de conjuntos difusos que forman parte de nuestro estudio, tanto para la representación de datos imprecisos, así como también, en la flexibilización de restricciones. Aquí se tratarán temas, tales como: conjuntos difusos, modelos difusos, representación de los datos difusos, cuantificadores difusos.

3.2.1 Conjuntos Difusos

Como se ha dicho anteriormente Zadeh (1965) expone 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. Por ejemplo, el conjunto de las personas que son *altas* es un conjunto difuso (o conjunto borroso), pues no está claro el límite de altura que establece a partir de qué medida una persona es alta o no lo es. Ese límite es difuso y, por tanto, el conjunto que delimita también lo será.

La interpretación original de conjunto difuso proviene de una generalización del concepto clásico de subconjunto ampliado a la descripción de nociones "*vagas*" e "*imprecisas*". Esta generalización se realiza como sigue:

- La pertenencia de un elemento a un conjunto pasa a ser un concepto "*difuso*" o "borroso" (en este trabajo utilizará como difuso). Para algunos elementos puede no estar clara su pertenencia o no al conjunto.

- Dicha pertenencia puede ser cuantificada por un grado. Dicho grado se denomina habitualmente como "*grado de pertenencia*" (otros grados se muestran en el apartado 3.2.2) de dicho elemento al conjunto y toma un valor habitual en el intervalo [0,1].

Otras veces, la pertenencia puede ser representada por etiquetas como "mucho", "poco", etc., que a su vez, son conjuntos difusos definidos en dicho intervalo [0,1].

Mediante esta herramienta se pueden representar de forma adecuada conceptos con "imprecisión". Es necesario hacer notar que muchos de estos conceptos naturales dependen, en mayor o menor medida, de la persona que lo expresa y en todo caso es muy importante que sean definidos correctamente, utilizando algún método (Dubois y Prade, 1986).

Un conjunto difuso *A* sobre un universo de discurso *U* (ordenado) se define como un conjunto de pares tal que:

$$A = \{\mu_A(u) /u : \ u \in U, \mu_A(u) \in [0,1]\} \tag{3.1}$$

Donde, $\mu_A(u)$ se denomina *grado de pertenencia* del elemento *u* al conjunto difuso *A*. Este grado oscila entre los extremos 0 y 1 para conjuntos discretos:

- $\mu_A(u) = 0$, indica que *u* no pertenece en absoluto al conjunto difuso *A*.
- $\mu_A(u) = 1$, indica que *u* pertenece totalmente al conjunto difuso *A*.

A veces, en lugar de dar una lista exhaustiva de todos los pares que forman el conjunto, se da una definición para la función $\mu_A(u)$, llamada función característica o función de pertenencia, también en ocasiones función de posibilidad, o simplemente *distribución de posibilidad* que es la que utiliza este trabajo.

Si la distribución de posibilidad sólo produce valores del conjunto {0,1}, entonces, el conjunto que genera no es difuso, sino clásico o a veces llamado "*crisp*".

Por ejemplo, si la edad (en años enteros) es el universo de discurso "joven", el conjunto difuso que representa dicho concepto podría expresarse en la forma: Joven = {0/14, 1/20, 1/25, 0.9/26, 0.8/27, 0.1/29, 0/30}, o bien, como una función de distribución de posibilidad, que en ese caso, se supone que el dominio subyacente es continuo.

Para un dominio con referencial ordenado, llamaremos *etiqueta lingüística* a aquella palabra (o identificador), en lenguaje natural, que exprese un conjunto difuso, que puede estar formalmente definido o no. Con este concepto, se puede asegurar que en nuestra vida cotidiana utilizamos multitud de etiquetas lingüísticas: "joven", "viejo", "frío", "caliente", etc.

Además la definición intuitiva de esas etiquetas no sólo puede variar de un individuo a otro y del momento particular, sino que también varía dentro del contexto en que se aplique. Por ejemplo, no medirán la misma altura un individuo "*alto*" y un edificio "*alto*". El problema de definir bien las etiquetas o los conjuntos difusos ha sido estudiado por muchos autores (Pedrycz, 1995; Zadeh, 1975), y aunque aquí no tratamos ese problema sí queremos dejar clara su importancia para conseguir sistemas útiles y prácticos.

Tanto las *etiquetas lingüísticas* como los datos con *imprecisión* pueden ser representados por conjuntos difusos. Además, dependen fundamentalmente de la naturaleza del universo de discurso sobre el que se define el conjunto difuso, o dominio subyacente.

Por consiguiente, un conjunto difuso A se define como una distribución de posibilidad μ_A que hace corresponder los elementos de un dominio o Universo de discurso (U) con elementos del intervalo [0,1]:

$$\mu_A: U \rightarrow [0,1] \tag{3.2}$$

Un número difuso es un conjunto difuso, donde U es un dominio numérico (normalmente los números reales R). En la Figura 3.1 se representa la función de posibilidad del número difuso "Aproximadamente n". El valor margen indica los límites del conjunto difuso. Es fácil observar que cuanto más cerca esté un número del valor n, su grado de pertenencia a "aproximadamente n" será mayor. En la Figura 3.1, Φ es el domino subyacente (R en este caso) y γ es el grado de pertenencia de cada valor Φ.

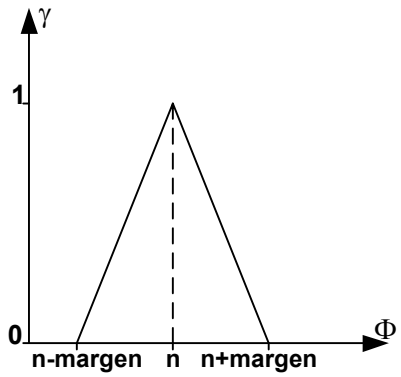

Figura 3.1: Función "Aproximadamente n" (n +/- margen).

Por otro lado, para dominios de referencial no ordenados se pueden definir etiquetas o escalares. Entre esas etiquetas puede existir una *función de similitud* que se define para cada par de valores del dominio D y establece una relación de similitud o proximidad para medir la similitud o parecido entre cada dos elementos del dominio (véase ejemplo en la Tabla 4.1). Normalmente, los valores de similitud están normalizados en un intervalo [0,1], correspondiendo el 0 al significado "*totalmente diferente*" y el 1 al significado "*totalmente parecido*" (o iguales). Por tanto, una relación de similitud puede ser vista como una función s_r, tal que:

$$s_r : D \times D \rightarrow [0,1]$$
$$s_r(d_i, d_j) \rightarrow [0,1] \text{ con } d_i, d_j \in U \text{ con } i,j \in R \qquad (3.3)$$

Para un determinado umbral γ, los valores que sean similares con un grado mayor a γ, serán indistinguibles y podrán ser considerados idénticos. Así, se pueden construir clases de equivalencia de forma que los elementos de una misma clase son indistinguibles para el grado de similitud γ, para dominio no discretos. La función similitud puede no cumplir la propiedad numérica (función de proximidad). Además, un concepto determinado puede ser expresado como una distribución de posibilidad sobre el dominio de esas etiquetas, como veremos más adelante.

A partir de estos sencillos conceptos se ha desarrollado toda una teoría matemática e informática que facilita la resolución de problemas (Bezdek, 1981; Petry, 1996; Pedrycz y Gomide 1998), tales como: control de sistemas, simulación, reconocimiento de patrones, sistemas de información o conocimiento, visión por ordenador, vida artificial...

3.2.2 Modelos de Bases de Datos Relacionales Difusas

Se han encontrado algunos modelos para el tratamiento de la incertidumbre como por ejemplo: La aproximación de Codd (valores nulos), Modelos estadísticos y probabilísticos, Modelos básicos de bases de datos difusas (añaden simplemente un grado difuso), Modelo de Buckles-Petry (Petry, 1996; Buckles, Petry (1982)), Modelo de Prade-Testemale (Dubois y Prade, 1998), Modelo de Umano-Fukami (Umano, 1982), Modelo de Zemankova-Kaendel, Modelo GEFRED (Medina et al., 1994). Estos modelos permiten almacenar y/o tratar información imprecisa o incierta y han sido utilizado en aplicaciones de bases de datos difusas. En el Apéndice III se expone una explicación de cada uno de estos modelos.

Estos modelos consideran el concepto de relaciones difusas debido a su implementación en bases de datos relacionales. Sin embargo, hay varias formas de permitir la información imprecisa o incierta en entidades o interrelaciones difusas. Además, esas formas pueden ser mezcladas para permitir una mayor flexibilidad. Algunas de las más importantes formas de modelar información difusa (Galindo, 1999) y (Galindo et al., 2001a) en las entidades o interrelación son:

- **Valores difusos en los atributos**: Esto se refiere a que en los atributos de una entidad se pueden contener valores difusos y también operar con ellos. Los tipos de estos valores pueden ser muy variados, tales como: valor nulo, valor desconocido, valor indefinido, distribución de posibilidad, función de similitud, etc., algunos ejemplos se encuentran en la Tabla 3.1 (Galindo, 1999).

- **Grado en cada valor de un atributo**: Esto implica que cada valor de un atributo puede tener asociado un grado, generalmente en un intervalo [0,1], que mida el nivel de difuminado de dicho valor. El significado de estos grados puede ser variado (Bosc et al., 1987; Medina et al., 1994).

- **Grado en toda la instancia de la entidad**: Esto es similar al caso anterior, pero aquí el grado está asociado a toda la instancia de la entidad y no exclusivamente a un valor particular de una instancia. Puede medir el grado con que esa instancia (o tupla) pertenece a esa relación (o tabla) de la base de datos (Tahani, 1977; Umano y Fukami, 1994).

- **Grado en un conjunto de valores de diversos atributos**: Este es un caso intermedio entre los casos anteriores. Aquí el grado está asociado a algunos atributos. Este puede ser un caso poco usual pero puede ser a veces muy útil (Galindo, 1999).

El dominio de estos grados se puede encontrar en un intervalo [0,1], pero también permiten otros valores, como por ejemplo una distribución de posibilidad. Además el significado de esos grados es variado. Dependiendo de este significado el tratamiento de los datos podrá ser diferente. Los posibles significados más importantes de los grados son los siguientes:

- **Grado de cumplimiento** (satisfacción): Una propiedad puede cumplirse con cierto grado entre dos extremos. La propiedad se cumple totalmente (usualmente grado 1), y la propiedad no se cumple en absoluto (usualmente grado 0). Esto suele emplearse tras establecer alguna condición sobre los valores de una entidad o interrelación (i.e. una selección con una condición difusa) y los grados expresarán en qué medida esa condición ha sido satisfecha (Tahani, 1977; Medina et al., 1994).

- **Grado de incertidumbre**: El grado de incertidumbre expresa la seguridad con que se conoce un dato determinado. Si se está seguro de la veracidad de dicho grado, éste será 1 y si se está seguro de su falsedad dicho grado será 0. Los valores entre 0 y 1 expresan distintos niveles de incertidumbre, indicando que no se está completamente seguro. Este significado es, en cierta forma, bastante parecido al anterior, pues puede extenderse esa incertidumbre como expresada sobre la pertenencia de una instancia (tupla) a la relación (tabla) de la base de datos. (Umano y Fukami, 1994).

- **Grado de posibilidad**: Mide la posibilidad de la información utilizada. Este significado es similar al anterior, pero se ve este grado como más débil, ya que contiene información que es más o menos posible y no más o menos cierta (Mouabdid, 1994).

- **Grado de importancia**: Distintos objetos (instancias, atributos, etc.) pueden tener diferentes importancias, de forma que existan objetos más importantes que otros (Mouabdid, 1994; Bosc et al., 1997).

En el apartado 3.2.1 se ha definido además el **grado de pertenencia** a un conjunto difuso, el cual se añade a los presentados aquí. Tanto el grado de cumplimiento, incertidumbre, posibilidad, importancia y pertenencia han sido estudiado por diversos autores (Dubois y Prade (1998); Petry (1996); Medina (1994), etc.), cada uno de ellos establece una fórmula de cálculo, que permite distinguir un tipo de grado de otro. En Galindo (1999) y Galindo et al. (2001b) se discute la diferencia de utilizar el grado de importancia y de posibilidad en la división. Para que estos grados sean tratados en un modelo conceptual de datos, en el capítulo 4 se mostrará la notación asociada a estos conceptos para un modelo de datos conceptual difuso.

3.2.3 Representación de los Datos Difusos

Los diferentes tipos de datos que constituyen la definición de dominio difuso generalizado son definidos en Medina et al. (1994) y utilizados en Galindo (1999), algunos ejemplos se encuentran representados en la Tabla 3.1.

A. Un escalar simple (ej. Tamaño = Grande, representado mediante la distribución de posibilidad 1/Grande).
B. Un número simple (Ej. Edad = 28, representado mediante la distribución de posibilidad 1/28).
C. Un conjunto de posibles asignaciones excluyentes de escalares (Ej. Aptitud = {Mala, Buena}, representado mediante {1/Mala, 1/Buena}.
D. Un conjunto de posibles asignaciones excluyentes de números (Ej. Edad ={20,21}, representados mediante {1/20, 1/21}.
E. Una distribución de posibilidad en el dominio de los escalares (Ej. Aptitud = {0.6/Mala, 1.0/Regular}.
F. Una distribución de posibilidad en el dominio de los números (Ej. Alto = {0/185, 1.0/195, 1.0/205, 0/215}, números difusos, etiquetas lingüísticas, etc.).
G. Un número real perteneciente a [0, 1] representando el grado de cumplimiento (Ej. Calidad = 0.9).
H. Un valor desconocido UNKNOWN dado por la distribución de posibilidad UNKNOWN = $\{1/u : u \in U\}$ sobre el dominio, *U,* considerado.
I. Un valor indefinido UNDEFINED dado por la distribución de posibilidad UNDEFINED = $\{0/u : u \in U\}$ sobre el dominio U, considerado.
J. Un valor nulo dado por la expresión Null = {1/UNKNOWN, 1/ UNDEFINED}.

Tabla 3.1: Ejemplo de datos imprecisos propuestos en el modelo GEFRED.

Considerando la propuesta de estos autores, en esta tesis, se utilizarán los siguientes criterios de representación de datos difusos:

Datos precisos (*crisp*, clásicos): Se empleará la representación que proporcione el modelo de datos EER (numéricos, alfanuméricos, fecha, etc.).

Datos imprecisos (*fuzzy*, difusos): Los datos de naturaleza imprecisa soportados por los modelos difusos actuales (Tabla 3.1) pueden ser clasificados en dos grupos, con distinta representación para cada uno de ellos: sobre referencial ordenado y sobre referencial no ordenado. Por *referencial*, se entiende el dominio subyacente del atributo en cuestión. Así, por ejemplo, un atributo *altura* puede ser considerado impreciso (difuso) para almacenar valores como "alto" (véase Figura 3.3), "bajo", siendo el dominio subyacente sobre el que se construyen las distribuciones de posibilidad es ordenado y corresponde a los centímetros de altura posible. En cambio, el *color del pelo* está sobre un referencial no ordenado, ya que no existe un orden entre los colores: "rubio", "castaño" y "negro".

En los apartados siguientes se detallan las distintas representaciones para dominios de referencial ordenado y no ordenado que se relacionan con la Tabla 3.1, siendo estos necesarios para modelar datos imprecisos (difusos) que trata este trabajo.

3.2.3.1 Datos Imprecisos sobre Referencial Ordenado

Este grupo de datos contiene distribuciones de posibilidad sobre dominios continuos o discretos para los cuales existe una relación de orden. Cada dato de este tipo tiene asociado una función de pertenencia o distribución de posibilidad. Por cuestión de simplicidad de representación y de eficiencia en el cálculo, se adoptan las siguientes representaciones para este tipo de datos (Medina, 1994; Galindo, 1999):

Distribución de posibilidad trapezoidal: Esta representación determina la función de pertenencia asociada al dato mediante el uso de cuatro parámetros [α, β, γ, δ], tal y como se muestra en la Figura 3.2. Utiliza funciones de pertenencia normalizadas, que son aquellas que poseen un núcleo no vacío, y tienen posibilidad máxima (1) para, al menos, un valor del dominio subyacente.

La función de distribución para este caso puede estar dada por:

$$\mu(x) \begin{cases} 0 & \text{si } x \le \alpha \\ (x-\alpha)/(\beta-\alpha) & \text{para } \alpha < x < \beta \\ 1 & \text{para } \beta \le x \le \gamma \\ (\delta-x)/(\gamma-x) & \text{para } \gamma < x < \delta \\ 0 & \text{si } x \ge \delta \end{cases}$$

La Figura 3.2 muestra su representación de distribución de posibilidad gráfica.

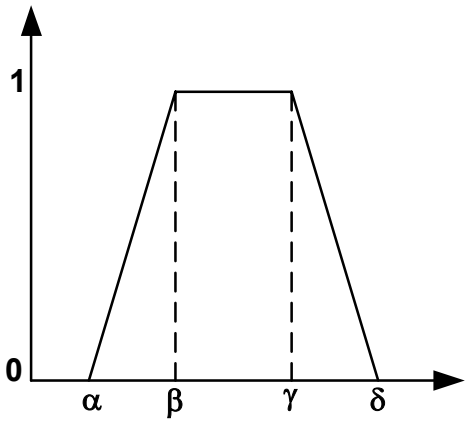

Figura 3.2: Formato de una distribución de posibilidad trapezoidal.

Etiquetas lingüísticas: Los datos expresados mediante una etiqueta lingüística hacen referencia a un concepto impreciso, a veces subjetivo, que lleva asociado una distribución de posibilidad. Por ejemplo, la etiqueta lingüística "*alto*", puede llevar asociada la distribución de posibilidad en representación trapezoidal que muestra la Figura 3.3. En la Tabla 3.1 corresponde al ejemplo F.

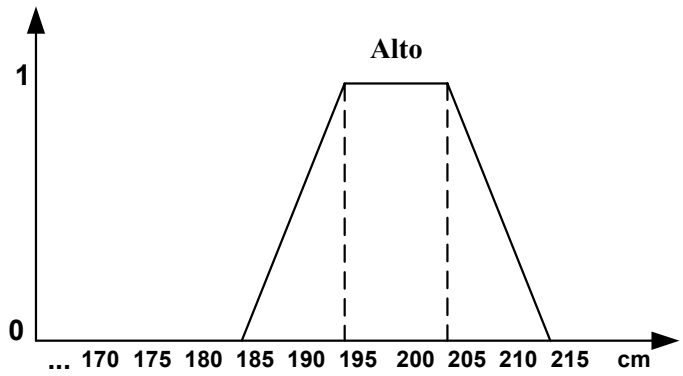

Figura 3.3: Ejemplo de una etiqueta lingüística para el concepto "alto".

Valores aproximados: Dado un valor "*n"* perteneciente a un dominio subyacente, se puede dar una representación del concepto impreciso "aproximadamente n" mediante un valor, llamado "margen", a partir del cual se construye su función de pertenencia como una distribución de posibilidad triangular, como lo muestra la Figura 3.4. Nuevamente se emplea la función de pertenencia normalizada.

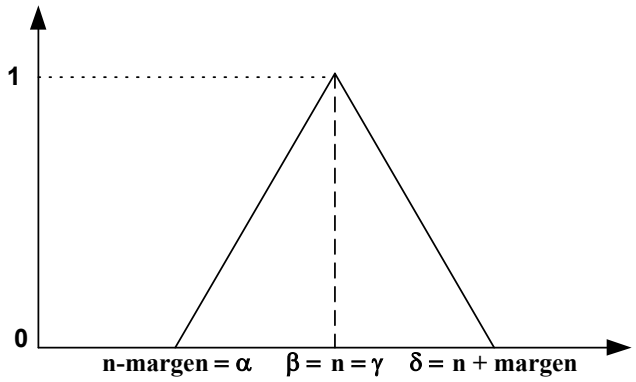

Figura 3.4: Distribución de posibilidad para "aproximadamente n".

Intervalos de posibilidad: Son un caso de distribución de posibilidad trapezoidal en el que las pendientes de ambos lados del trapecio son infinitas y por tanto los valores entre los dos extremos son los únicos que son totalmente posibles (posibilidad 1), como se muestra en la Figura 3.5. Se opera con ellos de forma similar a como se hace con la distribución de posibilidad trapezoidal. En la Tabla 3.1 corresponde al ejemplo D.

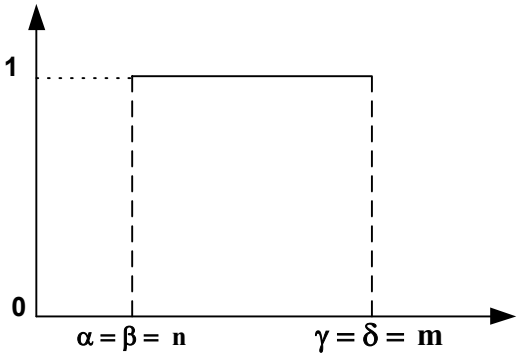

Figura 3.5: Distribución de posibilidad para intervalo [n,m].

3.2.3.2 Datos sobre Referencial no Ordenado

Este grupo de datos está construido sobre dominios subyacentes discretos no ordenados en los que se encuentran definidas "relaciones" de similitud o proximidad entre valores que lo construyen. Para este tipo de datos se tiene que proporcionar una forma para la representación de los mismos, así como para las "relaciones" definidas sobre los valores del dominio. Los diferentes tipos de datos que puede representar dentro de este grupo son (Medina, 1994; Galindo, 1999):

Escalares Simples: Se consideran como una distribución de posibilidad con una única pareja de datos en la que el valor de posibilidad 1:{(1/d)}, o sea, se considera que el valor del escalar d es el único posible y su posibilidad es 1 (para que éste sea normalizado). Ejemplo: {1/grande} mostrado en Tabla 3.1, corresponde al ejemplo A.

Distribución de posibilidad sobre escalares: A un dato impreciso de este tipo se le asocia una representación en la que se describen los valores del dominio de discurso que la componen los respectivos valores de posibilidad para cada uno de ellos, por ejemplo, $\{(p_1/d_1),...,(p_n/d_n)\}$. Uno de los p_i debe ser 1 para que la distribución esté normalizada. Un caso en que se observa en la Tabla 3.1, corresponde al C. Otro ejemplo, puede ser la cercanía entre dos a más barrios asociados al conjunto difuso {0.5/norte, 1/oriente, 0.2/plaza_españa}, o bien el color del pelo asociado al conjunto {rubio, pelirrojo, negro}.

3.2.3.3 Valores Especiales: UNKNOWN, UNDEFINED y NULL

Por otra parte, existen otros tres valores especiales que pueden encontrarse en cualquiera de los tipos de datos "imprecisos" que acabamos de describir. Estos tres valores son tomados en el sentido de los modelos de Umano-Fukami (véase apéndice III):

UNKNOWN (desconocido, pero aplicable): Un dato de este tipo refleja el desconocimiento con respecto al valor que toma un atributo. Sin embargo, se sabe que el atributo puede tomar algún valor del dominio de discurso. Esto implica que es posible que tome cualquiera de ellos, por tanto se representa el tipo UNKNOWN mediante la distribución de posibilidad $\{1/u, \forall\ u \in U\}$ donde U es el dominio subyacente. La Figura 3.6 muestra gráficamente esta distribución de posibilidad, la cual toma el valor 1 para todo el dominio subyacente o sea, todos los valores posibles y totalmente posibles. En la Tabla 3.1 corresponde al ejemplo H.

UNDEFINED (no aplicable): Cuando un atributo toma el valor UNDEFINED refleja el hecho de que ninguno de los valores del dominio sobre el que está definido es aplicable. Esto se puede entender como que ninguno de los valores del dominio es posible, por lo que se representa mediante la distribución de posibilidad, $\{0/u, \forall\ u \in U\}$ donde U es el dominio subyacente. La distribución de posibilidad se muestra en la Figura 3.6, la cual toma el valor 0 para todo el dominio subyacente. En la Tabla 3.1 corresponde al ejemplo I.

NULL (ignorancia absoluta): Sobre un atributo tenemos un valor Null cuando no aportamos información, bien por qué no la conocemos (UNKNOWN) o porqué no es aplicable (UNDEFINED). Mediante el conjunto {1/UNKNOWN, 1/UNDEFINED} podemos modelar este tipo de datos. En la Tabla 3.1 corresponde al ejemplo J.

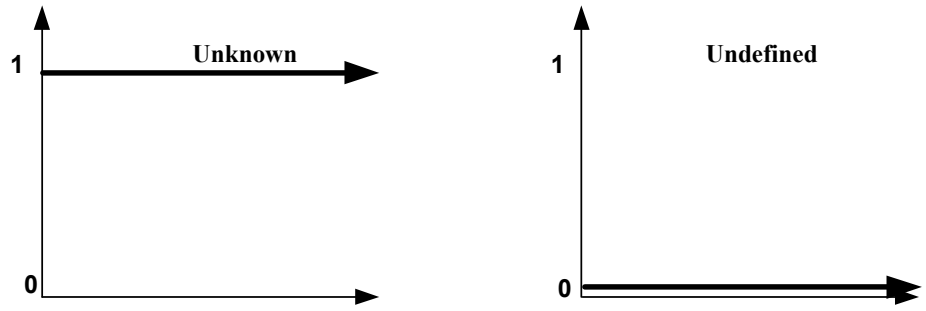

Figura 3.6: Distribución de posibilidad para los tipos UNKNOWN y UNDEFINED.

Obsérvese que estos tres valores especiales son aplicables tanto a datos con referencial ordenados, como no ordenados.

3.2.4 Atributos con Tratamiento Impreciso

Utilizando los dominios de referencial ordenado y no ordenado, como los examinados, en este trabajo, se definen distintos tipos de atributos para el tratamiento de datos imprecisos. A partir de la definición de estos dominios, este apartado define los atributos que pueden modelarse con dominios difusos.

Los atributos toman valores de un dominio, por lo tanto se puede decir que el atributo le da una determinada interpretación al dominio (o, a los dominios) en un contexto determinado. Es así, como se tiene el caso de los atributos clásicos o precisos, y atributos que son imprecisos. Estos últimos pueden ser tratados en un SGBD, pero con un formato especial, para representar atributos difusos. La clasificación adoptada en esta tesis, se basa en criterios de representación y de tratamiento de datos "imprecisos" que clasifican los atributos *según el tipo del dominio subyacente*. Lo que permite el tratamiento de información imprecisa, o el tratamiento impreciso de datos sin imprecisión. Los *atributos difusos*, han sido clasificados en cuatro tipos, los tres

primeros Tipo 1, Tipo 2 y Tipo 3 han sido tratados en Medina (1994), Vila et al. (1999), Galindo (1999), Marín et al. (2000), siendo estos atributos los más usuales, su implementación es discutida en Galindo et al. (1998) y Galindo (1999). Sin embargo, para el caso de los atributos difusos Tipo 4, son propios de esta tesis, y han sido propuestos para diferenciar los atributos difusos Tipo 3 que tienen relación de similitud y los que no. A continuación se presentan cada una de ellos.

- **Tipo 1**: Estos atributos son *datos precisos*, clásicos o también llamados "*crisp*" (tradicionales, sin imprecisión), que pueden tener etiquetas lingüísticas definidas sobre su dominio. Este tipo de atributo reciben una representación igual que los datos precisos, pero admiten que se puedan utilizar en condiciones difusas (con comparadores difusos, constantes difusas, umbrales de cumplimiento, etc.). Por tanto, son atributos clásicos que admiten el tratamiento impreciso. Las etiquetas lingüísticas definidas sólo se usarán en las condiciones difusas en las consultas (Galindo, 1999).

- **Tipo 2**: Estos atributos admiten tanto datos clásicos (crisp) como difusos (imprecisos), en forma de distribuciones de posibilidad sobre un dominio subyacente ordenado. Este tipo de atributos es el que se usará para extender una base de datos clásica con las ventajas que pueden recoger "*datos imprecisos sobre referencial ordenado"*. Permiten también, la representación de información incompleta en forma de datos de tipo UNKNOWN, UNDEFINED y NULL (Medina, 1994).

- **Tipo 3**: Son atributos sobre "*datos de dominio discreto no ordenado con analogía*". Estos atributos son definidos sobre un dominio subyacente no ordenado, por ejemplo, color del pelo. En estos atributos se definen algunas etiquetas ("rubio", "castaño", etc.), que son escalares con una *relación de similitud* (o proximidad) definida sobre ellas, de forma que esta relación indique en qué medida se parecen entre sí cada par de etiquetas. Este tipo de datos acepta escalares simples (como "rubio") para los que se supone grado de posibilidad 1 (1/rubio). Además, de ese tipo de dato elemental también se admiten distribuciones de posibilidad sobre ese dominio (por ejemplo: "1/rubio, 0.6/pelirrojo, 0.4/castaño") siempre normalizados. También aceptan valores del tipo UNKNOWN, UNDEFINED, y NULL.

Una variación de este tipo consiste en utilizar funciones de similitud que no cumplan la propiedad simétrica, o que la similitud sólo se cumpla en un sentido. Esto se puede representar con un grafo dirigido, donde cada nodo es una etiqueta.

- **Tipo 4**: Estos atributos son propuestos en esta tesis, se definen de la misma forma que los atributos Tipo 3, sin la necesidad de que exista una relación de similitud entre las etiquetas (o valores) del dominio. En efecto lo que importa es grado asociado a cada etiqueta de forma individual, sin evaluar la similitud entre las etiquetas. Por ello que, para este tipo de atributo, será en principio, más habitual encontrar valores no normalizados. Un ejemplo podría ser el tipo del rol que cumple un cliente en una inmobiliaria, por ejemplo con qué grado (de importancia) un cliente es "demandante" u "ofertante" de un inmueble.

También, existen algunos dominios difusos que son mostrados en la Tabla 3.1, como por ejemplo los tipos UNKNOWN, UNDEFINED y Null pueden ser definidos en el sentido de Umano y Fukami (ver apéndice III).

3.2.5 Cuantificadores Difusos

Dentro de las temáticas de la teoría de conjuntos difusos que se requiere en esta tesis, además de las mostradas anteriormente, consideramos a los **cuantificadores difusos** tanto relativos como absolutos**,** los cuales son utilizados para definir restricciones difusas en un modelo de datos. Pudiendo ser posible de utilizarlos en otras instancias.

Los cuantificadores difusos o lingüísticos (Yager, 1983; Zadeh, 1983), 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. Al igual que en el modelo clásico, los cuantificadores pueden ser absolutos o relativos:

- Los **cuantificadores absolutos** expresan cantidades sobre el número total de elementos de un determinado conjunto, diciendo si este número es "*grand*e", "*pequeñ*o", "*muchos*", "*pocos*", "*muchísimos*", "*aproximadamente entre 5 y 10*", etc. En estos casos se observa que la verdad del cuantificador depende de una única cantidad. Por eso, la

definición de los cuantificadores difusos absolutos es, como veremos, muy similar a los números difusos.

- Los **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 cuantificador se usa en expresiones como "*la mayoría*", "*la minoría*", "*aproximadamente la mitad*", etc. En estos casos, para evaluar la verdad del cuantificador se requiere hallar la cantidad total de elementos que cumplen la condición y ponderarla respecto a la cantidad total de elementos que podrían cumplirla (incluyendo los que la cumplen y los que no la cumplen). Véase ejemplo de la Figura 3.7.

Los cuantificadores difusos absolutos son definidos como conjuntos difusos (Zadeh, 1983) en el intervalo [0,+∞) y los relativos como conjuntos difusos en el intervalo [0,1]. O sea, un cuantificador es representado como una función Q, cuyo dominio puede ser absoluto o relativo:

$$Q_{abs} : R^{+} \rightarrow [0,1]$$
$$Q_{rel} : [0,1] \rightarrow [0,1]$$

En el primer caso Q_{abs} se define en R^{+}, en cambio, en el segundo caso el dominio de Q_{rel} es [0,1] porque en ese intervalo toma valor la división del número de elementos que cumplen cierta condición entre el número total de elementos existentes.

Para saber en qué grado se cumple el cuantificador sobre los elementos que cumplen cierta condición, se aplica la función Q del cuantificador al valor de cuantificación Φ (véase Figura 3.7):

- Si Q es absoluto, el valor Φ es el número de elementos que cumplen cierta condición.
- Si Q es relativo, Φ es la división del número de elementos que cumplen cierta condición entre el número total de elementos existentes.

Dicho de otro forma (a elementos que cumplen la condición y b elementos existentes):

$$\Phi \begin{cases} a \text{ si } Q = Q_{abs} \\ a/b \text{ si } Q = Q_{rel} \end{cases} \tag{3.4}$$

Si la función del cuantificador (absoluto o relativo), $Q(\Phi)$, toma el valor 1, indica que dicho cuantificador se satisface completamente, el valor 0 indica, por el contrario, que el cuantificador no se cumple en absoluto. Cualquier valor intermedio indica un grado de cumplimiento intermedio del cuantificador. Cada uno de estos grados de cumplimiento son tomados en el intervalo [0,1], y son representados por γ en la Figura 3.7.

Ejemplo 3.1: Un *cuantificador difuso relativo* es por ejemplo, "casi todo", definido tal como muestra la Figura 3.7. Por otro lado, *un cuantificador difuso absoluto*, es por ejemplo, "aproximadamente 8" definido tal como lo muestra la Figura 3.4, con n=8. En cambio, un cuantificador difuso absoluto es "aproximadamente entre 50 y 60", definido tal como se muestra en la Figura 3.2, con β=50 y α=60.

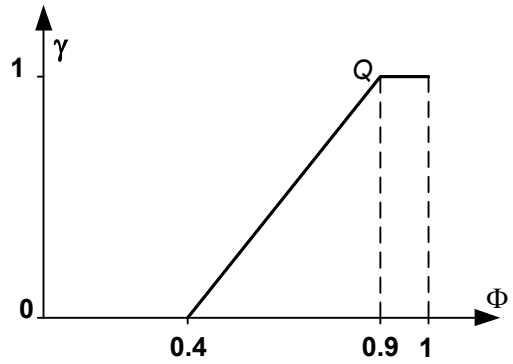

Figura 3.7: Cuantificador difuso relativo "casi todo": $x \in [0.4, 0.9] \leftrightarrow y = 2(x - 0.4)$.

Aplicados en el contexto de las bases de datos, los cuantificadores difusos permiten expresiones sobre el número de instancias que satisfacen una determinada condición o la proporción respecto del total. Su utilidad está demostrada por la flexibilidad que ofrecen para efectuar consultas que involucren estos cuantificadores, como por ejemplo, en la utilización de la operación de división del álgebra relacional en bases de datos difusas o clásicas (Galindo et al., 2001c).

Ejemplo 3.2: A continuación mostramos cinco ejemplos de cuantificadores difusos:

a) Cuantificador absoluto existencial ($\exists$), definido como:

$$Q_{\exists}(x) = \begin{cases} 0 & \text{si } x=0 \\ 1 & \text{si } x>0 \end{cases}$$

b) Cuantificador absoluto "*aproximadamente 2*" definido como una función triangular, de la siguiente forma (con n=2 y margen =1 en la Figura 3.1):

$$Q(x) = \begin{cases} 0 & \text{si } x \leq 0 \text{ o si } x \geq 3 \\ x-1 & \text{si } 1<x\leq 2 \\ 3-x & \text{si } 2<x<3 \end{cases}$$

c) Cuantificador relativo "*algunos*", definido como: $Q(x)=x$.

d) Cuantificador relativo "casi todo": Véase Figura 3.7.

e) Cuantificador relación universal ($\forall$) definido como:

$$Q_{\forall}(x) = \begin{cases} 1 & \text{si } x=1 \\ 0 & \text{si } x<1 \end{cases}$$

Finalmente, se debe diferenciar entre un cuantificador que se refiere a un conjunto de instancias y una etiqueta lingüística que hace referencia a un conjunto difuso del dominio de un atributo. Es por ello, que se propone usar cuantificadores difusos para definir algunas restricciones difusas y las etiquetas lingüísticas para definir valores de los atributos y entidades difusas.

3.2.6 Otros Conceptos sobre Conjuntos Difusos

Sobre conjuntos difusos se definen una serie de conceptos que nos permiten tratar y comparar conjuntos difusos. A continuación se considera Ω al tipo de referencial de un conjunto, obteniendo las siguientes definiciones:

Igualdad *(Equality)*: Dos conjuntos difusos A y B sobre Ω se dicen iguales si cumplen:

$$A = B \Leftrightarrow \forall x \in \Omega,\ \mu_A(x) = \mu_B(x)$$

Inclusión *(Inclusion)*: Dados dos conjuntos difusos A y B sobre Ω, se dice que A está incluido en B si cumplen:

$$A \subseteq B \Leftrightarrow \forall x \in \Omega,\ \mu_A(x) \leq \mu_B(x)$$

Núcleo (*Core*): El núcleo de un conjunto difuso A, definido sobre Ω es un subconjunto de dicho universo que satisface:

$$\text{Kern}(A) = \{x \in \Omega,\ \mu_A(x) = 1\}$$

Altura *(Height)*: La altura de un conjunto difuso A, definido sobre Ω se define como:

$$\mathrm{Hgt}(A) = \sup_{x \in \Omega} \mu_A(x)$$

Conjunto difuso Normalizado: Un Conjunto Difuso es normalizado, sí y sólo sí:

$$\exists x \in \Omega, \mu_A(x) = \mathrm{Hgt}(A) = 1$$

Teoría de la Posibilidad: Esta teoría se basa en la idea de variables lingüísticas y cómo estas están relacionadas con los conjuntos difusos (Zadeh, 1965). Así, se puede evaluar la *posibilidad* de que una determinada variable X sea o pertenezca a un determinado conjunto *A*, como el grado de pertenencia de los elementos de X en *A* definida como: Sea un conjunto difuso *A* definido sobre Ω con su función de pertenencia $\mu_A(x)$ y una variable X sobre Ω (que desconocemos su valor). Entonces, la proposición "X es *A*" define una *Distribución de Posibilidad*, de forma que se dice que la "posibilidad" de que "X = u", vale $\mu_A(u)$, para todo valor u perteneciente al conjunto *A*.

3.3 Investigaciones en Modelos Difusos

En la actualidad, casi toda la información que se maneja acerca del mundo real es incompleta, imprecisa, incierta o vaga (Parson, 1996).

Motro (1995) sostiene que la incertidumbre depende del contexto y la clasifica en:

- Incertidumbre, no es posible determinar si la información es verdadera o falsa, por ejemplo, "Juan puede tener 38 años".

- Imprecisión, la información disponible no es lo suficientemente específica, por ejemplo, "la edad de Juan está entre 37 y 43 años", —disyunción— "la edad de Juan es 34 ó 43 años", —negativa— "la edad de Juan no es 37, o incluso desconocido.
- Vaguedad, el modelo incluye elementos (predicados o cuantificadores) que son inherentemente vagos, por ejemplo, "Juan está al principio de su edad", "Juan está al final de su juventud". No obstante, una vez definido esos conceptos este caso coincidiría con el anterior (imprecisión).
- Inconsistencia, contiene dos o más afirmaciones que no pueden ser verdaderas al mismo tiempo, por ejemplo, "Juan tiene entre 37 y 43 o Juan tiene 35", este es un caso especial de disyunción.

- Ambigüedad, algunos elementos del modelo carecen de una semántica completa (o de un significado completo), por ejemplo, no queda claro si los salarios son anuales o mensuales.

Este autor sugiere que estos conceptos pueden ser tratados mediante la Inteligencia Artificial.

Zadeh (1965) introduce la lógica difusa, la cual se explicó en el apartado 3.2, para satisfacer este tipo de datos. Este autor observó, que la lógica clásica no representaba datos tales como "*persona atractiva*", "*bastante azul*" o "*mediana edad*", términos típicos del razonamiento humano. La lógica tradicional al ser bivaluada sólo podía trabajar con los conceptos: sí o no, blanco o negro, verdadero o falso, 0 ó 1, lo que permitía una representación muy limitada del conocimiento. Aunque existen otras lógicas que admiten más valores de verdad, llamadas lógicas multivaluadas, la lógica difusa es una extensión de ellas, admitiendo infinitos niveles o grados de verdad. Este autor define el concepto grado de pertenencia o grado de verdad en un intervalo [0,1] dentro de la teoría de conjuntos difusos.

A partir de este contexto, para analizar y discutir algunas de las investigaciones existentes actualmente de ésta temática y para facilitar nuestro estudio, hemos clasificado éstas en: Modelado conceptual de incertidumbre, Dependencias difusas, Implementación de incertidumbre. Estos temas son detallados a continuación.

3.3.1 Modelado Conceptual de Incertidumbre

Las principales metodologías de diseño de bases de datos (Batini et al., 1994; Connolly et al., 1998; Elmasri y Navathe, 2000; De Miguel et al., 1999; Atzeni et al., 1999; Gardarin, 1999) no han prestado atención al modelado de datos con incertidumbre, a pesar de que en el intento de modelar el mundo real la imprecisión está rara vez ausente.

Excepciones a la escasa atención al modelado conceptual impreciso o difuso y un buen acercamiento para comprender la problemática que plantea el modelado de incertidumbre en EER son:

3.3.1.1 Propuesta de Yazici y Merdan (1996)

Estos autores proponen una extensión del modelo IFO, en este trabajo se extiende un motor de inferencia de un sistema experto clásico (usando el modelo EER) al tratamiento de datos imprecisos, de forma especial, para aquellos datos que presenten similitud en una etiqueta, por ejemplo "color". En su propuesta formalizan una nomenclatura utilizando la notación EER, la extienden a dominios difusos, e implementa en lenguaje Prolog.

A esta extensión la llaman ExIFO, y exponen mediante ejemplos la implementación y validación de la representación de un esquema conceptual difuso considerando una representación de atributos inciertos. En el modelo agregan tres nuevos constructores, y usando estos constructores es posible representar explícitamente atributos (tipo modelo IFO) que posean valores inciertos. Las tres clases de incertidumbre distinguidas son:

1. El valor verdadero de un dato puede pertenecer a un conjunto específico de valores.
2. El valor verdadero de un dato puede ser incompleto.
3. El valor del dato verdadero es no conocido, nulo.

En general el valor del dato verdadero es disponible, pero en términos descriptivos la ausencia del dato preciso es impreciso (*fuzzy*).

Junto con las otras primitivas del modelo IFO los nuevos constructores F-set para el caso 1, I-set para el caso 2 y Null-set para el caso 3, son usados para describir el significado asumido con la información incierta en el diseño conceptual. El constructor F-set permite definir un conjunto de valores de un dominio especificado llamado F-set, y este es un subconjunto de valores del constructor F-set de tipo estructura que genera las instancias verdaderas en F-set.

El F-set es usado para construir una instancia en forma de un conjunto, cuyos elementos son relacionados entre ellos con la semántica OR, este tipo de constructor es usado para capturar las instancias que tienen valor difuso. La representación de valores de atributos que pueden ser construidos es representado por el *constructor F-set*, por ejemplo, un atributo `COLOR` con un dominio D={rojo, naranjo, amarillo azul}, donde se establece que el atributo difuso especifica instancias de un subconjunto, por ejemplo, $O_{F\text{-}set}$ ={naranjo, amarillo}. Véase Figura 3.8 a).

El constructor I-set también representa un dominio especificado de un tipo de instancias dadas que el constructor I-set define, pero este envuelve semánticas distintas que el constructor F-set. Para el constructor I-set se especifica solamente uno de los valores dados del conjunto y el rango correspondiente es una instancia en el tipo de estructura I-set, además I-set es un constructor de semántica XOR, desde uno y sólo uno de los valores. Por ejemplo: el atributo `PUB-TIME` (tiempo de publicidad) es un atributo con valor incompleto definido por un *constructor I-set* con un dominio de un conjunto de años con dominio D={1000 – 2000} y las instancias específicas están dadas por $O_{I\text{-}set}$ ={1990 – 1992}. Véase Figura 3.8 b).

Por último, para el caso del constructor Null-set la técnica es exactamente la misma, se tiene un dominio del constructor Null-set que define el dominio de un atributo que puede tomar valores nulos. Por ejemplo, el constructor *Null-set* de atributo `TEL-NO` con dominio D={tel#, unk, dne, ni} y $O_{Nul\text{-}set}$ ={dne}, la representación gráfica la muestra la Figura 3.8 c).

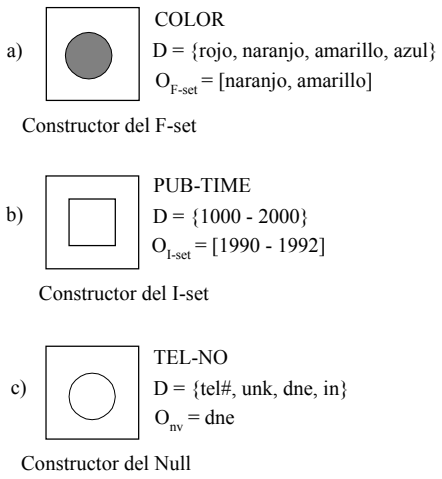

Figura 3.8: Ejemplo para representación de incertidumbres en el Modelo ExIFO. a) Atributo Valor- Fuzzy, b) Atributo Valor-Incompleto, c) Atributo Valor-Nulo.

Estos autores proponen además, relaciones (interrelaciones) con incertidumbre, para su representación gráfica, usan un rombo con una "F" en su interior. Con ello especifican que ciertas instancias son difusas ya que corresponden a un conjunto difuso, indicando un grado de pertenencia a la relación y este grado toma valores entre 0 y 1. Por ejemplo, un auto puede considerarse con un grado 0.6 de pertenencia a vehículo, un ómnibus puede considerarse con un grado 0.9. La Figura 3.9 a) muestra el modelo ExIFO con toda sus componentes, en cambio, la Figura 3.9 b) muestra un ejemplo de la aplicación del modelo ExIFO.

Otras publicaciones posteriores a la comentada se encuentran en Yazici et al. (1999) donde utilizan la extensión del modelo ExIFO para exponer, que su modelo, genera diseños de tablas en segunda forma normal (2FN), y con algunas sentencias SQL, que le permiten la implementación de las especificaciones de requerimientos recopiladas con la aplicación de esquemas utilizando el modelo ExIFO. Recuérdese que el modelo ExIFO va desde los atributos a las clases o entidades en forma de agregación y generalización de Batini et al. (1994), como se explicó en el apartado 3.1, lo que facilita llegar a establecer la forma normal que genera dicho diseño.

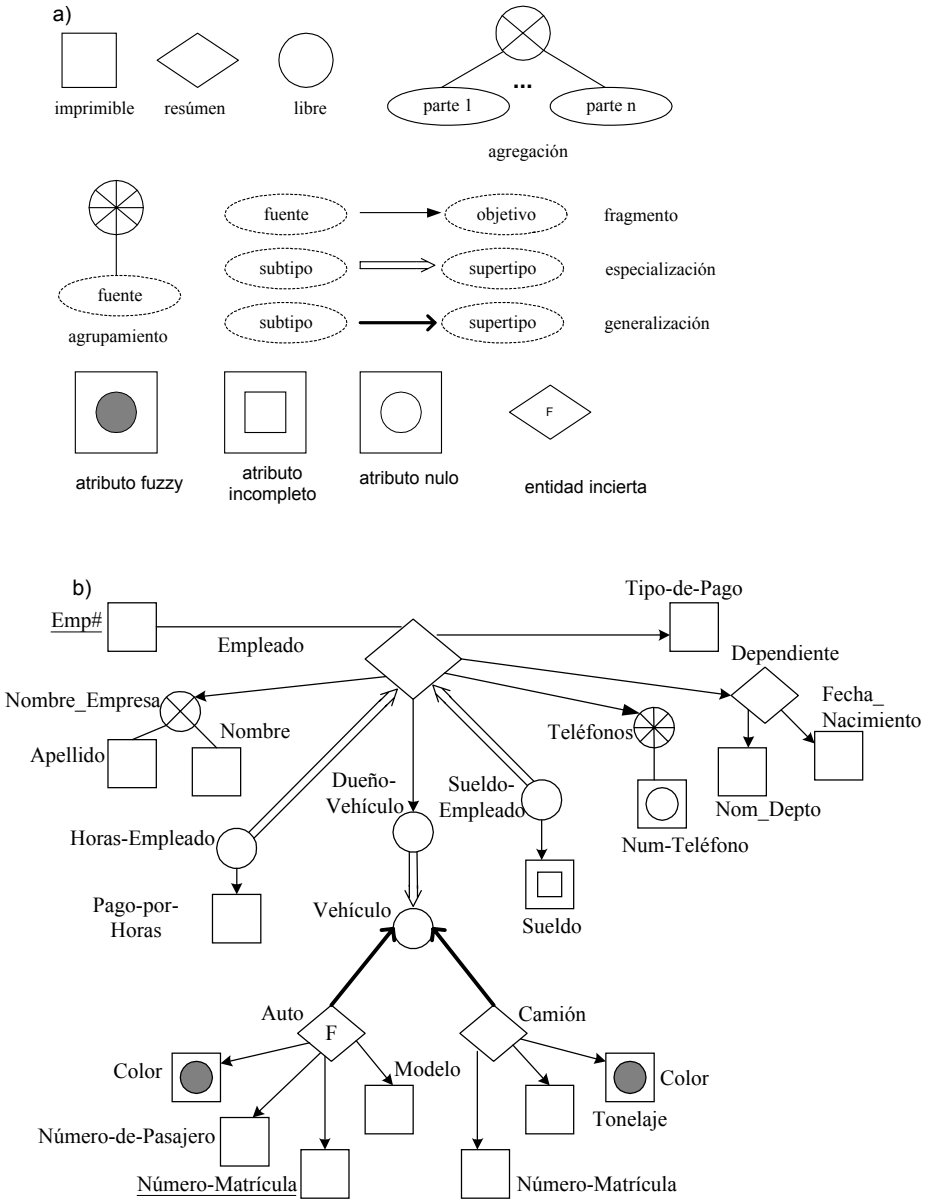

Figura 3.9: Modelo ExIFO difuso propuesto por Yazici y Merdan, (1996). a) Notación. b) Ejemplo Empleado-vehículo.

Por último, en Yazici y Cinar (2000) se expone el modelo ExIFO con una formalización matemática para los casos de atributos difusos, atributos inciertos y atributos nulos, así como también, una mayor gama de aplicaciones de este tipo de modelos de datos.

Obsérvese que la propuesta de extensión de modelo IFO a difuso sólo abarca una pequeña parte de lo que a modelado de incertidumbre se refiera. Los autores se preocupan de tres tipos de atributos específicos, dejando de lado otros tipos de atributos que son propuestos en esta tesis, como aquellos que tengan una relación de semejanza asociada o bien algunas etiquetas lingüísticas, más aún, no profundizan en otros elementos del modelo IFO (entidades, interrelaciones, etc.) y nada en lo que a restricciones se refiera.

3.3.1.2 Propuesta de Chen (1998)

En este libro el autor desarrolla una metodología de representación de semántica para datos con imprecisión, aplicando una distribución de posibilidad en etiquetas lingüísticas asociadas a conjuntos difusos.

El autor define que una variable lingüística X está formada por la tupla (T,U,G,M), donde: T es el conjunto de términos lingüísticos de X, U es el universo del discurso, G el conjunto de reglas sintácticas que genera el elemento T, y M es el conjunto de reglas semánticas traducidas desde T que corresponden al subconjunto difuso de U. Todo esto para definir un modelo conceptual y su respectiva representación matemática. Por ejemplo, sea X = `Edad`, T es generado vía G por el conjunto {muy viejo, viejo, edad media, joven, muy joven}, cada término de T es particularmente manejado vía M por conjuntos difusos entre 0 y 1. Un esquema de la variable lingüísticas se muestra en la Figura 3.10.

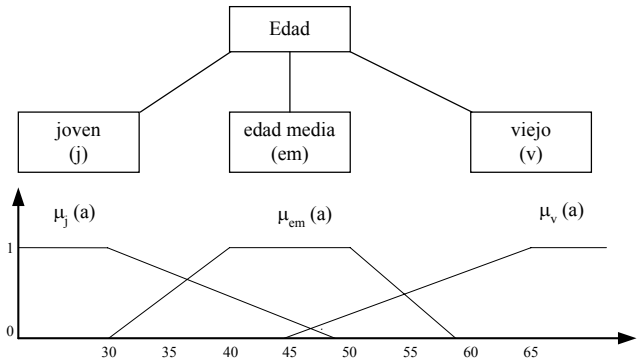

Figura 3.10: Variable lingüística "`Edad`" con sus respectivos valores y modelo conceptual.

En otro apartado son definidas las entidades interrelaciones y atributos difusos, también establece un tipo de correspondencia entre una entidad y entidad difusa, además del conjunto de valores que obtiene un grado de pertenencia a un conjunto difuso: 1:1, 1:N, N:M, incorporando la difusidad al modelo ER, y para cada una de ellas definen la representación gráfica extendiendo así el ER al que llaman Fuzzy ER. La Figura 3.11 muestra la notación propuesta por este autor.

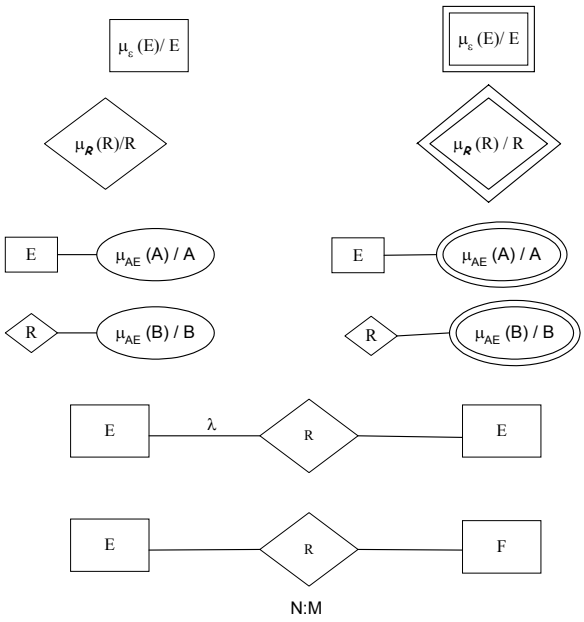

Figura 3.11: Notación Fuzzy ER propuesta por Chen (1998).

El modelo Fuzzy ER propone un modelo generado por $M = (E, R, A)$ expresado por E tipo de entidad, R tipo de interrelación, y A atributos, extendiéndose a tipos de etiquetas que genera a $LI(M) = (E, R, A_E, A_R)$ proponiendo cuatro tipos de conjuntos, donde μ es el grado de pertenencia al conjunto (la notación se muestra en la Figura 3.11):

- $E = \{\mu_E (E)/E / E \in D_E \text{ y } \mu_E (E) \in [0,1]\}$.
- $R = \{\mu_R (R)/R / R$ es el tipo de interrelación envuelta en el tipo de entidad en D_E y $\mu_r (E) \in [0,1]\}$.
- $A_E = \{\mu_{AE} (A)/A / A$ es el tipo de atributo del tipo de entidad E, y $\mu_{AE} (A) \in [0,1]\}$.
- $A_R = \{\mu_{AR} (B)/B / B$ es el tipo de atributo del tipo de la tipo interrelación R, y $\mu_{AR} (B) \in [0,1]\}$.

Posteriormente, en otro apartado se define un atributo difuso para la generalización y especialización/generalización en la disyunción, solapamiento (véase Figura 3.12 a)), así como también para la categoría y subclase compartida unión e intersección (véase Figura 3.12 b)), extendiendo así las restricciones difusas en un modelo EER. En la propuesta siempre se hace referencia a etiquetas lingüísticas y a la función de trapezoidal sobre un atributo o entidad específica, no a un conjunto de atributos distintos o entidades distintas. Este autor al igual que Yazici y Merdan (1998) establecen su modelo de datos a partir de los atributos y con uso de las herramientas de generalización y especialización forman la clase o entidad de los objetos.

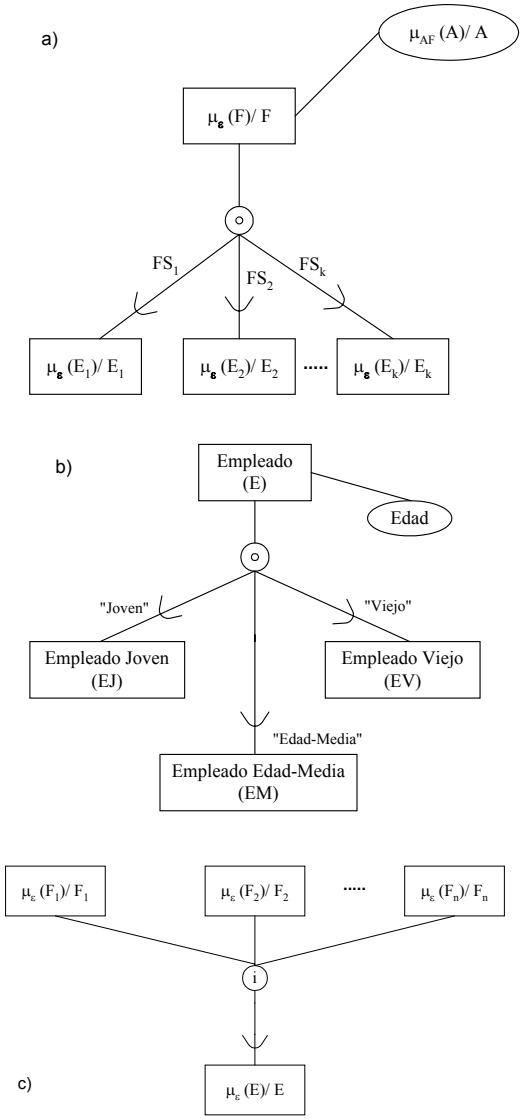

Figura 3.12: Notación propuestas por Chen (1998). a) Especialización solapadas, b) Empleado en una especialización solapada por atributo difuso edad, c) Subclase compartida intersección.

Con relación al diseño de bases de datos, el autor plantea el problema de anomalías de actualización con el uso de una primera forma normal difusa utilizando dependencias funcionales difusas, además, también propone la segunda y tercera forma normal. Cada una de ellas con una formalización correspondiente, y propone diferentes propuestas de dependencias funcionales difusas que deriva a cada forma normal, definiendo así, los algoritmos de descomposición de cada una. La nomenclatura de las formas normales difusas son especificadas agregando una F al final de la notación tradicional de la dependencias funcionales, generando así, la 1FNF, la 2FNF y la 3FNF.

Otras propuestas parecidas encuentran en Zvieli y Chen (1986), que proponen un modelo que soporta atributos difusos en las entidades y las interrelaciones, estos autores introducen tres niveles de *fuzziness* en el modelo ER:

1. El primer nivel plantea que los conjuntos de entidades, relaciones y atributos pueden ser difusos respecto al modelo. O sea, ellos tienen un grado de pertenencia al modelo conceptual. Por ejemplo, la entidad *Radio* puede ser difusa y tener un grado de pertenencia 0.7 como parte integrante de otra entidad *Coche*. La Figura 3.14 caso 1) muestra un ejemplo de este nivel y otro ejemplo está en la Figura 3.13.
2. El segundo nivel está relacionado con las ocurrencias difusas de entidades y relaciones. Por ejemplo, una entidad *Empleados jóvenes*, tiene un grado de pertenencia asociado a sus instancias.
3. El tercer nivel permite valores difusos en los atributos de las entidades e interrelaciones. Por ejemplo, el atributo `Calidad` de una entidad que representa a un *jugador de baloncesto*, puede considerarse como el atributo difuso.

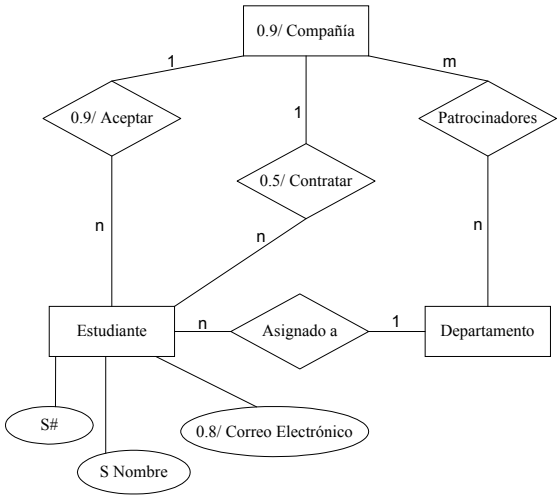

Figura 3.13: Ejemplo del tipo de entidad compañía con grado parcial 0.9.

El primer nivel puede ser útil, pero al final debemos decidir si tal entidad, relación o atributo aparecerá o no en la implementación. Por eso en nuestro estudio no hemos dado importancia al primer nivel. El segundo también es útil, pero es importante considerar distintos significados de los grados. El tercer nivel es útil, y es similar a escribir el tipo de dato de algún atributo, ya sea que los valores difusos pertenecen a tipos de datos difusos. No obstante, este último nivel es importante para reflejar la potencia de un modelo conceptual, pero también aquí hay que considerar los distintos tipos de atributos difusos que pueden existir (véase Tabla 4.5).

3.3.1.3 Propuesta de Ma et al. (2001)

Según estos autores un modelo interrelación extendido ER juega un papel crucial en el rol de los conceptos de bases de datos relacionales, así como también, en las bases de datos orientadas a objeto. Estos autores en su propuesta trabajan con los tres niveles de Zvieli y Chen, (1986) e introducen el modelo interrelación extendido difuso (FEER. Fuzzy Extended Entity-Relationship) para manejar objetos complejos en el mundo real a un nivel conceptual. Sin embargo, sus definiciones de generalización, especialización, categoría y agregación imponen condiciones muy restrictivas.

El objetivo de esta investigación es extender las bases de datos orientadas a objetos utilizando la teoría de conjuntos difusos, una propuesta inicial es extender el modelo entidad interrelación a conceptos difusos, para satisfacer las imperfecciones de no poder representar el mundo real a partir de los datos que sean imprecisos o inciertos. Para ello, utilizan etiquetas conceptuales y aprovechan el mapeo de la teoría de conjuntos difusos para representar atributos, entidades e interrelaciones difusas. La investigación aborda conceptos de información imprecisa o incertidumbre, conjuntos difusos y distribución de posibilidad, para representar incertidumbre. Proponen los autores un modelo FEER, definiendo tanto el concepto como la representación gráfica de atributos entidades e interrelaciones, así como también, la agregación pero sólo para etiquetas lingüísticas.

Además presentan una transformación del esquema FEER (modelo Entidad Interrelación Extendido Fuzzy) a un esquema FOODB (Base de Datos Orientada a Objeto Fuzzy).

La notación propuesta en el FEER es la que muestra la Figura 3.14, donde hemos resumido la propuesta de los autores en tres Figuras, dos de notación y una de ejemplo.

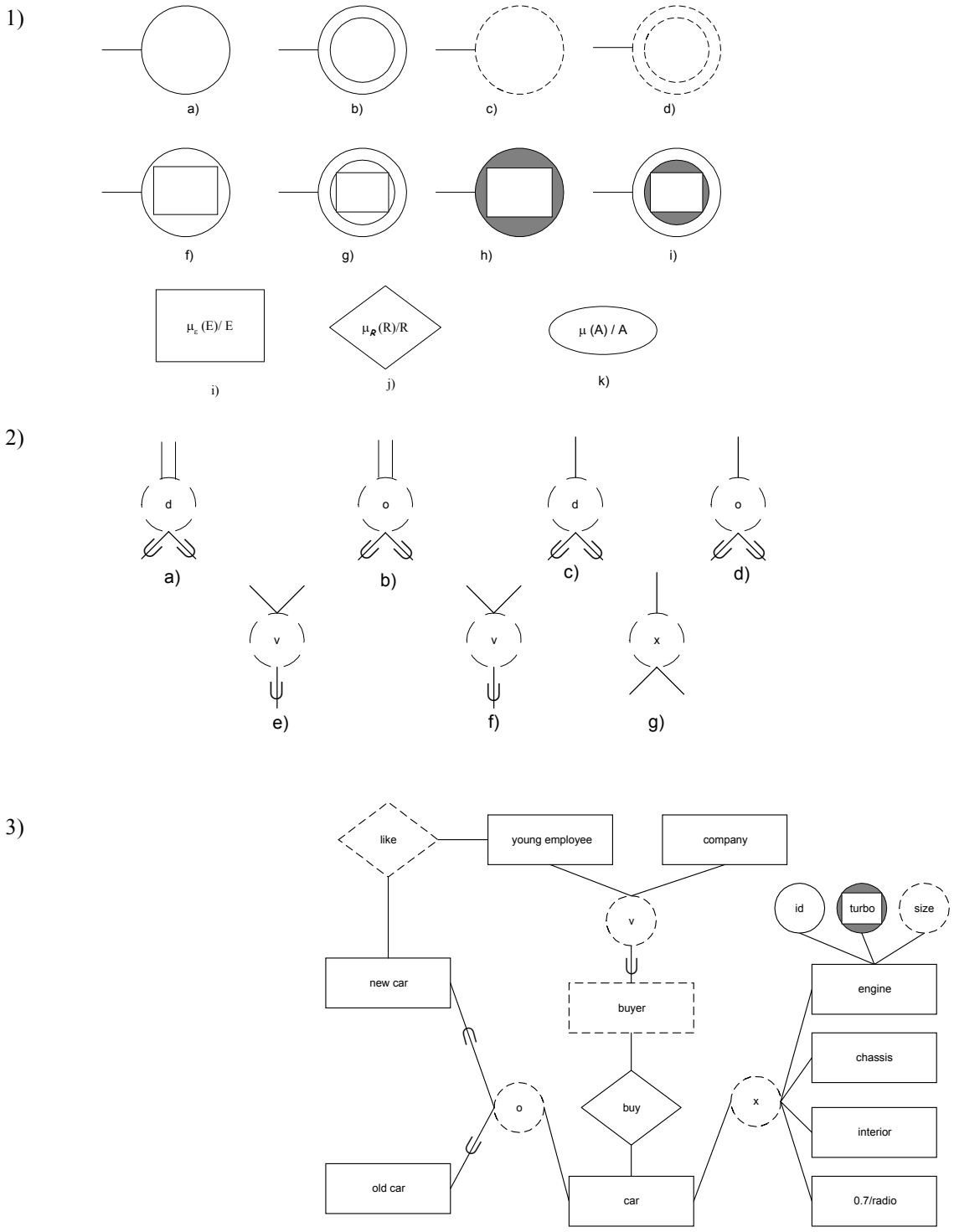

Figura 3.14: Notación de FEER de Ma et al. (2001). 1) Atributos, entidades e interrelaciones difusas, 2) Especialización, agregación y categorías difusas, 3) Ejemplo de uso de la notación.

La Figura 3.14 en 1) muestra: a) Tipo de atributo con valor simple, b) tipo de atributo multivaluado, c) tipo de atributo disjunto, d) tipo de atributo conjuntivo, e) tipo de atributo nulo, f) tipo de atributo abierto a nulo, g) tipo de atributo disjunto impreciso, h) tipo de atributo

impreciso conjuntivo, i) entidad con grado relacionado, j) interrelación con grado relacionado, k) atributo con grado relacionado.

La Figura 3.14 en 2) muestra: a) especialización disjunta total difusa, b) especialización solapada total difusa, c) especialización disjunta parcial difusa, d) especialización solapada parcial difusa, e) subclases difusas con múltiples superclases difusas, f) categorías difusas, y g) agregación difusa.

La Figura 3.14 en 3) muestra un ejemplo de Fuzzy EER utilizando algunas de las notaciones propuestas por Ma et al. (2001). Es así como la entidad "car" es una superclase de dos subclases difusas "new car" y "old" en una especialización solapada. Además, la entidad difusa "young employee" con instancias difusas de la entidad "company" compuesta por la categoría *unión* de la entidad difusa "buyer". También "young employee" tiene una interrelación difusa "like". Por último la entidad "car" es una *agregación* de las entidades "radio" que asociada con grado 0.7, "interior", "chassis", y "engine" que muestra algunos atributos difusos como *size* y *turbo*.

Obsérvese que la propuesta de estos autores, si bien usan conjuntos difusos en casi todas las componentes del modelo ER/EER, lo hacen aplicando los tres niveles de Zvieli y Chen (1986), y siempre referenciado a la distribución de posibilidad, no hay mención a la distribución de similitud, por otro lado, la notación propuesta para el caso de los diferentes tipos de atributos difusos es un poco engorrosa y difícil de recordar (Figura 3.14, parte 1)). Además, las restricciones propuestas a la especialización son, sólo considerando, grados de pertenencia al conjunto difuso, no hacen referencia a otro tipo de grados como los expuestos en el apartado 3.2.2. Una diferencia entre la propuesta de estos autores y la de esta tesis, es el uso de cuantificadores difusos para flexibilizar este tipo de restricciones.

3.3.1.4 Propuesta de Chaudhry et al. (1994)

Estos autores proponen un método para diseñar bases de datos difusas siguiendo el modelo EER. A su propuesta la llaman FRDBS (Fuzzy Relational Data Base System). En dicha metodología, se encuentra un conjunto de pasos para representar imprecisión siguiendo el modelo EER como una extensión, prestando especial interés en convertir bases de datos clásicas (*crisp*) en difusas. La forma de hacerlo es definiendo *n* etiquetas lingüísticas para *n* conjuntos

difusos, después cada instancia de una entidad *crisp* es transformada en un conjunto de instancias difusas, creando así una nueva entidad con estas instancias que contienen los grados correspondientes a cada etiqueta asociada al valor *crisp*, Véase Figura 3.15.

La metodología es utilizada para un caso aplicado al *diseño y control de bases de datos de semiconductores*. En este trabajo se propone que a partir de una "entidad regular" se debe crear una "entidad débil difusa", esto se da, por la diversidad de valores que puede tomar un atributo que se defina impreciso. A esta entidad la definen como DBFdifusa(DBFdifusa(B_0)) (B_0 dato impreciso), incorporando el grado de pertenencia como un atributo más a esta entidad, además de todos los atributos de la entidad original.

Los autores exponen en su trabajo algunos conceptos de conjuntos difusos para la definición de una base de datos difusa, prestando importancia a los modelos probabilísticos y dependencias funcionales entre los datos de tipo difuso, en ello utilizan la función comparación difusa. Proponen la aplicación de una metodología que consta de tres pasos: primero se modela los requerimientos de la aplicación con la notación ER, segundo transforman el modelo ER, de la etapa primera, a un modelo relacional (tablas), y el tercero se normalizan las relaciones obtenidas en la etapa segunda. El modelo propuesto y sus etapas son aplicadas a un ejemplo práctico, en sus trabajos futuros, pretenden extender dicha metodología a un modelo orientado a objetos para obtener una mayor complementación de su propuesta.

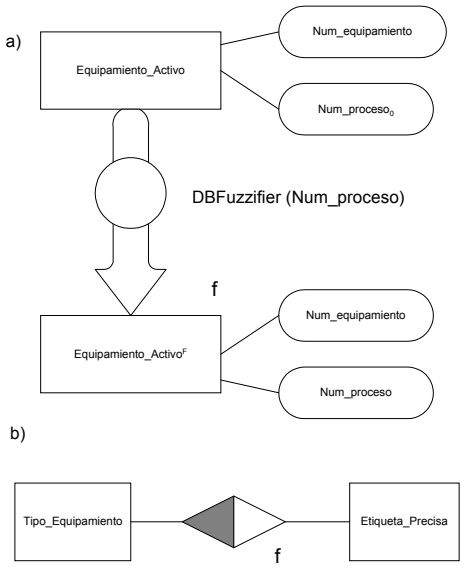

Figura 3.15: Modelo propuesto por Chaudhry et al. (1994). a) Ejemplo de transformación DBFuzzifier, b) Interrelación difusa para etiquetas.

En Chaudhry et al. (1999) los autores extienden el modelo presentado en Chaudhry et al. (1994), haciendo hincapié en su propuesta al ejemplo de *procesos de control*. En cada proceso se observan valores imprecisos que están asociados a etiquetas lingüísticas, y cada valor pasa por un proceso que denominan "DBFuzzifier construct". Aquí formalmente llaman a su metodología FERM, la cual extiende el modelo de datos ER al tratamiento de datos imprecisos considerando en su transformación los tres pasos mencionados anteriormente.

Obsérvese que estos autores sólo han trabajado un caso específico y sobre éste la extensión de su modelo propuesto. En este sentido, sólo enfocaron aquellos atributos asociados a más de un valor de un atributo difuso en una entidad (atributo multivaluado) y han dejado de lado otros tipos de datos inciertos o imprecisos, las entidades difusas y las interrelaciones. Nunca proponen ni en sus trabajos futuros, algún tipo de restricción.

3.3.1.5 Propuesta de Vert et al. (2000)

En este trabajo, los autores utilizan la notación de modelo de datos introducida por Oracle y la aplicación de la teoría de conjuntos difusos en el tratamiento de colecciones de conjuntos y subconjuntos (clases) difusos, para un caso específico de Sistemas de Información Geográfico (GIS). En su propuesta definen una notación ERD (Diagrama Entidad Relación), como una extensión del modelo ER para representar, con la teoría de conjuntos difusos, problemas de manejo de información geográfica.

Esta investigación enfoca *funciones que son discretizadas*, denotadas como: D(), para problemas difusos definidos por campos continuos de datos. D() es un miembro especializado de *clases de funciones* denotadas como: M(), donde los valores están seleccionados sobre definición de temporalidad continua de campos que son espaciales. La propuesta de los autores se centra en el manejo de datos en un GIS y modelo ER, ejemplificado con un modelo. Llaman a su modelo "Modelo ERD difuso" (ERD es una propuesta de extensión al modelo de datos usado por Oracle, de relacional a relacional difuso). En esta investigación se utiliza una notación convencional y extensión (para lo cual consideran; clases de funciones M() como relaciones difusas, función difusa discretizada D(), verdad o indefiniciones, efectos de cardinalidad de D() y M()). En la extensión del modelo definen notación difusa para manejar conjunto de datos imprecisos. Como trabajo futuro los autores proponen la implementación del modelo ERD propuesto en un gestor de bases de datos Oracle.

Obsérvese que tanto Chaudhry et al. (1994) como Vert et al. (2000) son casos específicos y de ahí la extensión del modelo, en cambio las propuestas de Yazici y Merdan (1996), Chen (1998) y Ma et al. (2001) son más generales.

3.3.1.6 Propuestas en Modelos Orientados a Objetos

Trabajos en cuanto a modelos orientados a objetos se refieran, hemos encontrado: Marín et al. (2000a), Geneste y Ruet (2001) y (2002), Goswami y Kumar (2001), los cuales resumimos a continuación:

Marín et al. (2000a): Estos autores, en su investigación incorporan la vaguedad al modelo orientado a objeto. A partir de esto manifiestan que se puede considerar la presencia de la vaguedad en el tipo asociado a una clase en sí, paralela e independientemente de una *visión borrosa* del conjunto de objetos que forma una clase. Ofrecen una perspectiva nueva, para lo cual definen el concepto de *tipo difuso*, así como también una visión de la *instanciación* y la *herencia* aplicada a dicho concepto.

En una de las secciones los autores proponen la estructura de una clase, explican que el caso más claro lo constituyen aquellos problemas en los que se trata el concepto con diferentes niveles de especificidad o precisión y para lo cual procuran encontrar la estructura para definir la imprecisión. Utilizan un ejemplo de "objeto imagen", dando énfasis a las estructuras de rasgos mínimos (fichero, formato, versión), para ello crean dos niveles de especificidad. A partir de este ejemplo, definen una notación y formalización del concepto tipo difuso acorde.

En otra sección tratan la instanciación y herencia de tipos difusos con una formalización acorde. Este mecanismo debe permitir elegir un "α-corte" de atributos del tipo que se requiere y para un nuevo objeto a crear. En sus líneas futuras plantean la posibilidad de implementación adecuada y resolver problemas de herencia múltiple.

Otra investigación similar es Marín et al. (2000b), estos autores, definen distintos tipos de estructuras de vaguedad, ordenados en diferentes etiquetas de precisión y amplitud. Proponen una forma de incorporarlo a un sistema de bases de datos orientada a objetos.

Geneste y Ruet (2001) y (2002): Estos autores tratan la búsqueda de información a partir de una metodología de *razonamiento a partir casos*. Proponen un modelado en UML de etiquetas lingüísticas, todo esto es sólo en un plano teórico. La metodología propuesta tiene tres pasos, que permite encontrar los casos relevantes en la base de casos de experimento, luego determina el dominio relevante para adoptar la solución seleccionando el nuevo problema. Finalmente, utilizando *propagación de restricciones* permite guiar la adaptación generada. Usan la representación de UML en Rational Rose para definir los datos difusos como una clase de instancias de los valores de un trapecio.

Goswami y Kumar (2001): Los autores en este trabajo presentan una extensión al modelado de datos orientado a objeto difuso. La propuesta permite almacenar, manipular y utilizar información imprecisa, que es introducida en una clase de objeto. El modelo propuesto es un formalismo que consiste en tres partes: en la primera se especifican los requerimientos del usuario en términos de un lenguaje natural al que le llaman FRSL (Fuzzy Requirement Specification Language). En la segunda parte, el FRSL es traducido, por una parte en *diagramas*, que permiten construir un modelo conceptual orientado a objeto difuso llamado, FDNER (Fuzzy Diagrams Nested Entity Relationship). Se establecen expresiones para la agregación, generalización y los atributos, por otra parte, en las *expresiones* permiten construir un protocolo llamado FUPME (Fuzzy Update Protocol Model Expressions). Finalmente, estos dos modelos (diagramas y expresiones) son implementados en un FOOD (Fuzzy Object Oriented Diagram), introduciendo una extensión estática y dinámica para modelo orientado a objetos difusos.

Otros trabajos relacionados se pueden encontrar en Yazici y Koyuncu (1997) y Vila et al. (1998).

3.3.2 Dependencias Difusas

Otros trabajos interesantes de analizar son aquellos que contemplan dependencias difusas planteadas sobre entidades y atributos que utilizan la teoría de conjuntos difusos, si bien no son parte de un modelo conceptual de datos, estos trabajos aportan (con la imprecisión, conjuntos difusos y requerimientos) que se puede modelar a través de distribución de posibilidad, sobre los atributos en forma individual o atributos de toda una entidad. En este tema existen diversos trabajos de investigación, algunos de los cuales se presentan a continuación:

3.3.2.1 Propuesta de Raju y Majumdar (1988)

Estos autores, tratan la aplicación de la lógica difusa en un ambiente de base de datos relacional con el objetivo de capturar más significado de los datos. Para ello, presentan operadores del álgebra para relaciones difusas y la aplicabilidad de la lógica difusa en capturar restricciones de integridad con una medida difusa, semejante a "EQUAL", para comparar valores de dominio. La definición de dependencia funcional clásica es generalizada a dependencia funcional difusa (ffd) con la función EQUAL. Además, se ejemplifica la problemática de implicación de las dependencias ffds y se propone un conjunto de estados y completos axiomas de inferencia.

En uno de sus apartados extienden el modelo relacional clásico para tratar información difusa, para lo cual consideran restricciones de integridad que pueden involucrar conceptos difusos. Por ejemplo, restricciones de integridad difusas, tal como "el sueldo de los empleados casi igualmente calificados será *más o menos igual*". Este tipo de integridad se presenta naturalmente en bases de datos difusas, pero no en una clásica. Con este caso trabajan diferentes tipos de restricciones de integridad, tal como dependencia funcional, dependencia multivaluada, dependencia de unión, etc., identificando conjuntos de estados y completas reglas de inferencia para tales dependencias, las cuales son discutidas y propuestas por los autores. Por otra parte, se sugieren varios algoritmos para diseñar esquemas de bases de datos normalizadas desde un conjunto asociado a las dependencias de los datos.

En su trabajo los autores definen algunos conceptos de la teoría de bases de datos relacionales clásicas, y los extienden con algunas definiciones básicas de la teoría de conjuntos difusos para formar bases de datos relacionales difusas. Discuten relaciones difusas y restricciones de integridad difusas. A su vez, son definidas algunas dependencias funcionales difusas y reglas de inferencia asociadas. Por último, trata la extensión de la unión difusa en relaciones difusas con presencia de dependencias funcionales difusas.

Queremos añadir que una buena definición de los operadores del álgebra relacional difusa puede encontrarse en Medina (1994) y Galindo et al. (2001b) y Galindo (1999).

3.3.2.2 Propuesta de Cubero et al. (1994 y 1998)

En este trabajo se definen operadores de proyección y combinación (join) para bases de datos relacionales difusas. Se plantean operadores de semejanza para desarrollar el concepto de dependencia funcional difusa para un conjunto de atributos difusos, y así, conservar la integridad y las restricciones de una base de datos. Los autores manifiestan que si la relación *r* satisface una dependencia entre un conjunto de atributos, X e Y, ellos deben preservar esta dependencia dentro de las proyecciones existentes sobre XY de *r*, para lo cual analizan cada una de estas proyecciones y dependencias difusas. En su trabajo encuentran las siguientes definiciones: relaciones difusas en bases de datos, dependencia funcional difusa, proyecciones difusas y reunión natural difusa con un formalismo matemático.

3.3.2.3 Propuesta de Carrasco et al. (2000a)

Estos autores definen las dependencias generalizadas, globales y graduales. La idea de estas dependencias difusas es sencilla, consiste en cambiar en la definición de las dependencias funcionales clásicas el operador "=", por uno difuso, como por ejemplo, "*aproximadamente igual*", así como también "*aproximadamente mayor*", "*aproximadamente menor*", etc. todo ello para descubrir otro tipo de relaciones entre los atributos. Según los autores muchas veces este tipo de dependencias se dan para un porcentaje alto de instancias de la tabla, pero no para todas. Para ello, definen el concepto de "*confianza*" (confidence) que no es nada más que el porcentaje de instancias que cumplen la dependencia en la base de datos, definiendo así las dependencia de confianza. Este concepto (bastante sencillo) se podría usar al igual que las dependencias funcionales clásicas y ser de gran utilidad, por ejemplo, para aplicaciones de minería de datos.

Las *dependencias graduales* son algo parecido a la idea de *regla difusa*. Este término gradual es utilizado para definir dependencias entre atributos que más bien se asemejan a reglas difusas. En este artículo mencionan que un ejemplo para la dependencia gradual podría ser "*cuanto más altos son los jugadores de baloncesto, mayor calidad tienen*". También se definen las *dependencias globales* que abarcan tanto las funcionales como las graduales, además del hecho de que no todas las instancias tengan que cumplir la dependencia confianza.

Otros trabajos similares se encuentran en: Carrasco et al. (2000a), Chen et al. (1994), Cubero y Vila (1994), Chen (1998).

3.3.3 Implementación de la Incertidumbre

La incertidumbre o imprecisión en bases de datos al nivel de implementación ha permitido crear bases de datos difusas, que aplican la lógica difusa a la tecnología de las bases de datos. Algunas de estas bases de datos son una extensión de las bases de datos relacionales, existiendo algunas propuestas de tipo teórico en este ámbito y otros. En este sentido resumimos en forma muy breve algunas investigaciones.

3.3.3.1 Propuesta de Medina et al. (1994)

Los autores muestran los principales modelos (Prade-Testemale, Umano y Fukami, Buckles_Petry entre otros) para dar solución al tratamiento de la información "imprecisa" en bases de datos relacionales, especialmente aquellos que utilizan la teoría de conjuntos difusos. Aquí, proponen una extensión a los modelos existentes tratando de englobar, de alguna forma, a todos estos modelos, al que denominan *modelo generalizado* GEFRED (GEFRED. A Generalized Model of Fuzzy Relational Databases, que es propuesto en la tesis doctoral de Medina (1994)). Este modelo constituye una síntesis ecléctica de los diferentes modelos publicados (Apéndice III) para tratar el problema de representación y tratamiento de la información difusa mediante bases de datos relacionales. Una de las principales ventajas de este modelo es que consiste en una abstracción general que permite tratar diferentes enfoques, incluso aunque estos puedan parecer muy dispares. Este modelo se define sobre un dominio difuso generalizado, para relaciones difusas generalizadas, mostrando ejemplos y definiciones matemáticas de cada una de sus propuestas. Además, se propone una metabase de datos difusa que es el diseño lógico, en tablas, de la BDRD llamado FIRST (FIRST. A Fuzzy Interface for Relational SysTems, explicado en el Apéndice IV), para extender el modelo teórico en un SGBD tradicional.

3.3.3.2 Propuesta de Galindo (1999)

Este autor propone un tratamiento de la imprecisión en bases de datos relacionales. Incorpora un servidor FSQL basándose en GREFED y FIRST en una extensión al SQL (Galindo et al. 1998), un lenguaje de consulta difuso o Fuzzy SQL (SQL difuso) que es un prototipo para bases de datos difusas construido sobre el SGBD Oracle. El servidor FSQL permite el almacenamiento

mediante FIRST de información difusa y su tratamiento a través de un lenguaje especial para el manejo y definición de información difusa. El lenguaje FSQL es una extensión del lenguaje SQL normal, que permite recuperar los grados de cumplimiento asociados a los datos, flexibilizando consultas y haciéndolas muy expresivas, a la vez establece umbrales de cumplimiento para condiciones difusas con comparadores difusos Un ejemplo que se plantea en FSQL es: *"seleccione las personas jóvenes con grado mínimo 0.75*". En esta investigación se utiliza el modelo generalizado GEFRED, muestra ejemplos y definiciones para cada caso, y enfatiza el operador relacional de división por medio de la aplicación de cuantificadores difusos en las consultas (Galindo et al. 2001b).

También en uno de sus apartados se explican los objetivos y características básicas de una arquitectura cliente-servidor para FSQL, que son utilizados como interfaz entre el usuario y el servidor FSQL propuesto. Además, se explica el funcionamiento de la comunicación entre el cliente y el servidor FSQL, y cómo programar un cliente FSQL, así como una serie de operaciones útiles en estos programas. Se incluye una breve referencia sobre cómo esta implementado el programa FQ (un cliente de FSQL). Como trabajo futuro, expone la creación de un cliente Visual en Java, que pueda ser ejecutado a través de internet. Finalmente, contiene ejemplos prácticos de esta propuesta.

Este trabajo ha sido un aporte importante para formular el modelo FuzzyEER propuesto en esta tesis, ya que, los conceptos de diseño lógico del FSQL y algunos otros utilizados en su creación, han sido los que se han tratado de modelar.

Otras publicaciones relacionadas con esta propuesta se encuentran en Galindo et al. (1998), (2001b) y (2001c).

3.4 Discusión del Capítulo

En cuanto a modelos de datos, podemos decir hay una escasa investigación de modelado conceptual difuso, las presentadas aquí son las más cercanas al objetivo de la tesis. Unas propuestas interesantes son, por ejemplo, la investigación de Chen (1998) y Ma et al. (2001), dando importancia a modelar con qué grado de pertenencia cada atributo se asocia con la entidad, o también con qué grado un atributo pertenece a una etiqueta lingüística. Sin embargo, aún cuando estas investigaciones están referidas a modelar la imprecisión con etiquetas

lingüísticas, no tratan las representaciones asociadas al tipo de dato que representa según su dominio y significado de diferentes tipos de grados, como las que proponemos en el capítulo 4. Yazici y Merdan (1996) muestran tres tipos distintos de datos a modelar en un modelo de datos difuso, pero tampoco hacen referencia a los distintos grados que se pueden modelar. En cuanto a la generalización es propuesta como herramienta de abstracción.

Por otro lado, en cuanto a las restricciones (participación y tipo de correspondencia entre interrelaciones), donde se utilizan uno o dos umbrales de cuantificadores difusos relativos o absolutos, los autores comentados anteriormente, no hacen mención en ninguna de sus propuestas ni tampoco al uso de la teoría de conjuntos difusos a restricción de jerarquías o categorías, como trataremos en esta tesis.

Resumiendo, ninguna de las investigaciones discutidas anteriormente, hacen intentos de modelar restricciones, aplicando cuantificadores difusos ni relativos, ni absolutos. Por otro lado, los autores siempre se refieren a distribución de distribución, rara vez a la distribución de similitud (véase Tablas 4.5 y 4.5). Además, la mayoría de las investigaciones son referidas a modelados de datos para casos específicos, como sistemas geográficos, manufacturación, etc., escasamente a la extensión del modelado conceptual en su conjunto.

Hay otros autores que ni siquiera proponen notaciones gráficas, solo conceptos como Kerre y Chen (2000) y Ruspini (1986).

En cuanto a dependencias funcionales, aún cuando en ninguno de los artículos tratados se ve reflejado el uso de un modelo conceptual difuso, han servido como etapa de requerimiento para conocer qué es lo que se debe modelar. Algunas de las componentes o conceptos usados en las dependencias, tanto funcionales difusas como graduales u otras, se han tratado de representar en el modelo conceptual que se propone en esta tesis.

De estas investigaciones se ha tomado el uso de cuantificadores difusos, tanto relativos como absolutos y comparadores difusos aplicados a consultas en bases de datos difusas. En nuestro caso, los cuantificadores difusos son utilizados para definir el modelado de restricciones de interrelaciones en un modelo FuzzyEER. Para el caso de las dependencias entre los atributos difusos, tratadas en algunas de estas investigaciones, han servido de aporte para la definición del concepto de atributos que modelan *grado de un conjunto de valores de diversos atributos*, concepto que se detalla en el apartado 4.1.3.

En cuanto a implementaciones difusas, las investigaciones presentadas en este apartado, se han centrado preferentemente en el desarrollo de aplicaciones y la construcción de bases de datos difusas (fase dos y tres de la metodología de diseño de bases de datos), descuidando la fase primera de modelado conceptual. El interés de estos autores ha estado en el tratamiento de la imprecisión de datos y la creación de una extensión del SQL a Fuzzy SQL con su respectiva implementación. Lo mismo ocurre con el uso de lenguajes como es el caso de Prolog en el ExIFO. En su gran mayoría existe una formalización de sus propuestas usando la lógica difusa. Cada uno de estos artículos, han servido de algún modo, como especificación de requerimientos de lo que se desea modelar, además han proporcionado distintos conceptos difusos que son utilizados.

Para el caso de la FIRST expuesta en Medina et al. (1994) y Medina (1994) y el FSQL expuesto en Galindo (1999), sirve como etapas posteriores, pudiendo extenderla con las definiciones propuestas en el capítulo 4, con un diseño lógico adecuado, en la aplicación de la metodología de diseño de bases de datos.