

# Razonamiento Aproximado y Adaptable en el Procesamiento de Consultas Vagas

TESIS DOCTORAL

Autoría de:

CLAUDIA JIMÉNEZ RAMÍREZ M.Sc.



Director:

HERNÁN DARÍO ÁLVAREZ ZAPATA Ph.D.

DOCTORADO EN INGENIERÍA, SISTEMAS E INFORMÁTICA  
FACULTAD DE MINAS  
UNIVERSIDAD NACIONAL DE COLOMBIA  
SEDE MEDELLIN  
2008

# TABLA DE CONTENIDO

<b>1</b>	<b><u>INTRODUCCIÓN</u></b>	<b>2</b>
1.1	PLANTEAMIENTO DEL PROBLEMA	3
1.2	OBJETIVOS	5
1.2.1	<i>OBJETIVO GENERAL</i>	5
1.2.2	<i>OBJETIVOS ESPECÍFICOS</i>	5
1.3	METODOLOGÍA Y ALCANCE	6
1.4	MÉTODO DE EVALUACIÓN	7
1.5	APORTES ORIGINALES	7
1.6	ESTRUCTURA DEL DOCUMENTO	8
<b>2</b>	<b><u>EL LENGUAJE Y SU VAGUEDAD</u></b>	<b>10</b>
2.1	EL LENGUAJE NATURAL.	10
2.2	LENGUAJES FORMALES ARTIFICIALES.	13
2.3	TIPOS DE GESTIÓN DEL CONOCIMIENTO EN LA INGENIERÍA LINGÜÍSTICA	14
2.3.1	RECUPERACIÓN DE INFORMACIÓN.	14
2.3.2	SISTEMAS FLEXIBLES DE CONSULTA-RESPUESTA	14
2.3.3	DESCUBRIMIENTO DE NUEVO CONOCIMIENTO.	15
2.4	EL MODELO OBJETO-RELACIONAL Y LOS LENGUAJES DE CONSULTA	16
<b>3</b>	<b><u>REPRESENTACIÓN Y MANEJO DE LA VAGUEDAD</u></b>	<b>20</b>
3.1	LA INCERTIDUMBRE ORIGINADA POR LA VAGUEDAD	20
3.1.1	<i>VARIABLES LINGÜÍSTICAS</i>	21
3.1.2	<i>TEORÍAS DE CONJUNTOS</i>	22
3.1.3	<i>OPERACIONES DE CONJUNTOS DIFUSOS</i>	29
3.1.4	<i>REGLAS DE INTERPRETACIÓN DEL LENGUAJE TEÓRICO PRUF</i>	33
3.2	TÉCNICAS DE RAZONAMIENTO APROXIMADO	35
3.3	SISTEMAS EXPERTOS O SISTEMAS BASADOS EN CONOCIMIENTO	37
3.4	LA DISCRIMINACIÓN Y LA CLASIFICACIÓN DE OBJETOS	39
3.4.1	<i>TÉCNICAS DE REPRESENTACIÓN DE UNA COLECCIÓN DE OBJETOS</i>	40
3.4.2	<i>REGLAS DE DECISIÓN</i>	49
3.4.3	<i>GRANULACIÓN DIFUSA O BORROSA</i>	51
3.4.4	<i>TÉCNICAS DE DISCRIMINACIÓN DIFUSA BASADAS EN REGLAS HEURÍSTICAS</i>	53

## **4 PROPIEDADES DESEABLES EN EL MODELO CONCEPTUAL DE RAZONAMIENTO** **56**

4.1	LOS MODELOS CONCEPTUALES EN LA INGENIERÍA	56
4.2	RESTRICCIONES IMPUESTAS AL NUEVO MODELO DE RAZONAMIENTO	58
4.3	RESTRICCIONES GENERALES SOBRE EL NUEVO MODELO CONCEPTUAL	58
4.3.1	<i>RESTRICCIÓN 1. CONFIABILIDAD</i>	58
4.3.2	<i>RESTRICCIÓN 2. EXTENSIBILIDAD</i>	59
4.3.3	<i>RESTRICCIÓN 3. GENERALIDAD</i>	59
4.3.4	<i>RESTRICCIÓN 4. CORRECCIÓN</i>	59
4.3.5	<i>RESTRICCIÓN 5. RIQUEZA SEMÁNTICA O POTENCIA EXPRESIVA</i>	60
4.3.6	<i>RESTRICCIÓN 6. FORMALIDAD O RIGOR</i>	60
4.3.7	<i>RESTRICCIÓN 7. ROBUSTEZ.</i>	60
4.4	RESTRICCIONES SOBRE EL SISTEMA DE INFERENCIA	61
4.5	RESTRICCIONES SOBRE LOS MARCOS DE COGNICIÓN	62
4.5.1	<i>RESTRICCIÓN 8. ORDENAMIENTO APROPIADO</i>	62
4.5.2	<i>RESTRICCIÓN 9. NÚMERO DE ELEMENTOS JUSTIFICABLE</i>	62
4.5.3	<i>RESTRICCIÓN 10. GRÁNULOS DISTINGUIBLES</i>	62
4.5.4	<i>RESTRICCIÓN 11. NO A TRES, AL TIEMPO</i>	63
4.5.5	<i>RESTRICCIÓN 12. COBERTURA (COMPLETITUD)</i>	63
4.5.6	<i>RESTRICCIÓN 13. COMPLEMENTARIEDAD (CRITERIO <math>\Sigma</math>)</i>	63
4.5.7	<i>RESTRICCIÓN 14. ACEPTACIÓN COMPLETA</i>	63
4.5.8	<i>RESTRICCIÓN 15. NORMALIDAD</i>	63
4.5.9	<i>RESTRICCIÓN 16. CONVEXIDAD</i>	64
4.5.10	<i>RESTRICCIÓN 17. UNIMODALIDAD</i>	64

## **5 EL MODELO CONCEPTUAL DEL RAZONAMIENTO** **66**

5.1	TRABAJOS RELACIONADOS CON LA FLEXIBILIZACIÓN DEL LENGUAJE SQL	66
5.2	MODELO CONCEPTUAL DEL RAZONAMIENTO PROPUESTO	74
5.3	MODELO ESTRUCTURAL DEL DOMINIO DEL RAZONAMIENTO	75
5.3.1	<i>SUPUESTOS DEL MODELO</i>	75
5.3.2	<i>CONCEPTOS BÁSICOS</i>	76
5.3.3	<i>RESTRICCIONES DEL MODELO</i>	77
5.4	MODELO DE COMPORTAMIENTO DEL SISTEMA	79
5.4.1	<i>ANÁLISIS PRELIMINAR DE LA CONSULTA</i>	81
5.4.2	<i>DELIMITACIÓN DEL CONTEXTO LINGÜÍSTICO DE LA PREGUNTA</i>	85
5.4.3	<i>CONCRECIÓN DE LA VAGUEDAD</i>	86
5.4.4	<i>OBTENCIÓN DE LA RESPUESTA</i>	88

## **6 LA CONCRECIÓN DE LA VAGUEDAD** **90**

6.1	PREGUNTAS DE TIPO I (CON TÉRMINOS VAGOS SIMPLES)	90
6.1.1	<i>REGLAS SINTÁCTICAS PARA CONDICIONES VAGAS SIMPLES</i>	91
6.1.2	<i>REGLAS SEMÁNTICAS PARA CONDICIONES VAGAS SIMPLES</i>	92

6.1.3	TÉCNICA DE DISCRIMINACIÓN PARA LA DEFINICIÓN DE REGLAS SEMÁNTICAS	93
6.2	PREGUNTAS DE TIPO II (CON TÉRMINOS VAGOS SIMPLES ACOMPAÑADOS DE UN MODIFICADOR)	106
6.2.1	REGLAS SINTÁCTICAS PARA LOS MODIFICADORES LINGÜÍSTICOS	106
6.2.2	REGLAS SEMÁNTICAS PARA LOS MODIFICADORES LINGÜÍSTICOS	106
6.2.3	REGLAS DE INTERPRETACIÓN DE COMPOSICIONES	114
6.3	PREGUNTAS DE TIPO III (CON CUANTIFICADORES LINGÜÍSTICOS)	124
6.3.1	REGLAS SINTÁCTICAS PARA LOS CUANTIFICADORES RELATIVOS	125
6.3.2	REGLAS SEMÁNTICAS PARA LOS CUANTIFICADORES RELATIVOS	126
6.3.3	REGLAS SINTÁCTICAS PARA COMPARADORES VAGOS	128
6.3.4	REGLAS SEMÁNTICAS PARA COMPARADORES VAGOS	129
6.4	PREGUNTAS DE TIPO IV (CON CALIFICADORES LINGÜÍSTICOS)	132
6.4.1	REGLAS SINTÁCTICAS PARA VALORES LINGÜÍSTICOS DE LA VERDAD	133
6.4.2	REGLAS SEMÁNTICAS PARA VALORES DE VERDAD LINGÜÍSTICOS	134
6.5	ESPECIFICACIÓN FORMAL DE LA CONCRECIÓN DE LA VAGUEDAD	138
6.5.1	REGLAS SINTÁCTICAS DE LA ORDEN DE CONSULTA	139
6.5.2	REGLAS SEMÁNTICAS DE LA ORDEN DE CONSULTA	140
6.6	EXTENSIONES ADICIONALES DEL LENGUAJE SQL, CONSIDERANDO EL ESTÁNDAR SQL:1999	143
6.6.1	FUNCIONES Y PROCEDIMIENTOS ALMACENADOS	143
6.6.2	CURSORES	144
6.6.3	DISPARADORES	144
6.6.4	TIPOS DE DATOS DEFINIDOS POR EL USUARIO	145
6.6.5	VISTAS	145
6.6.6	LA CLÁUSULA FROM DE UNA CONSULTA	146
6.6.7	LA DECLARACIÓN CASE EN UNA CONSULTA	146
6.7	NUEVAS SENTENCIAS DE DEFINICIÓN DE DATOS	147

## **7 EVALUACIÓN DEL MODELO PROPUESTO** **148**

7.1	FACTIBILIDAD TÉCNICA DE SISTEMAS BASADOS EN EL MODELO PROPUESTO	148
7.1.1	ARQUITECTURA GENERAL DEL SISTEMA FLEXIBLE DE CONSULTA-RESPUESTA	148
7.1.2	DISEÑO DETALLADO DEL SISTEMA DE CONSULTA-RESPUESTA	150
7.2	CONFIRMACIÓN TEÓRICA DEL MODELO PROPUESTO	166
7.2.1	CONFIABILIDAD	167
7.2.2	EXTENSIBILIDAD	167
7.2.3	GENERALIDAD	168
7.2.4	CORRECCIÓN	168
7.2.5	RIQUEZA SEMÁNTICA O POTENCIA EXPRESIVA	176
7.2.6	FORMALIDAD O RIGOR	177
7.2.7	ROBUSTEZ	177
7.2.8	CONSIDERACIONES SOBRE LA EFICIENCIA DE UN SISTEMA BASADO EN EL MODELO PROPUESTO	178
7.3	CONCLUSIONES	181
7.4	TRABAJO FUTURO	183

# Resumen

La imprecisión o la vaguedad es un tipo de incertidumbre o de imperfección inherente al lenguaje natural, modelo habitual de comunicación humana. Puesto que en la interacción humano-máquina, los términos vagos deben precisarse para obtener respuestas de esta última, en esta investigación se aborda el problema de la representación y manejo de la vaguedad como una estrategia clave para aproximar el lenguaje estándar de consulta a bases de datos, el SQL, al lenguaje natural.

Para la interpretación de las consultas vagas se concibió una máquina de inferencia que puede adaptarse a los distintos contextos delimitados por cada consulta. Ello se logró dotando al sistema de inferencia subyacente en el lenguaje de capacidades de razonamiento no deductivo para que, por su propia cuenta y dinámicamente, descubra los modelos particulares que representan adjetivos calificativos y otros términos vagos dependientes del contexto, expresados en las consultas. La máquina de inferencia identifica la semántica de etiquetas lingüísticas, examinando los modelos teóricos definidos para los diferentes patrones sintácticos con los que puede encajar el texto de la consulta y estima el valor de los parámetros usando los datos disponibles en la base de datos referentes al contexto. Por esto, desde una perspectiva general, el modelo propuesto constituye un aporte cuya principal novedad consiste en la delegación de la obtención de los modelos que representan los conjuntos rotulados con alguna etiqueta lingüística, a la máquina de inferencia, con el propósito de obtener respuestas confiables de los sistemas flexibles de consulta a bases de datos objeto-relacionales.

El modelo conceptual propuesto incluye una extensión del lenguaje estándar SQL de consulta a bases de datos, generalizando los operadores IS y LIKE existentes e incorporando otros nuevos para representar adverbios de cantidad, cuantificadores y valores lingüísticos de la verdad. Estos operadores permiten representar y operar con términos vagos simples o complejos, en los cuales se incluye aquellos que están formados por una combinación lineal de condiciones simples, vagas o concretas. También se aporta un mecanismo útil y sencillo de extracción de nuevo conocimiento para caracterizar, de manera aproximada, las asociaciones existentes entre los objetos de una base de datos.

**Palabras Claves:** *Lenguaje Flexible de Consulta, Bases de Datos Difusas, Representación de la Vaguedad*

# Abstract

Vagueness is a kind of uncertainty or imperfection inherent in natural language, usual pattern of human communication. Since in the human-machine interaction, vague terms need to be transformed into crisp values in order to obtain answers from the query systems, this research addresses the problem of representation and management of vagueness as a key strategy to bring standard database query language SQL closer to natural language.

An inference machine was conceived for the interpretation of vague queries that can be adaptable to multiple contexts. This was achieved by giving non deductive reasoning capabilities to the inference system so that, by its own and dynamically, discover the models that represent qualifying adjectives and other vague terms that depends on linguistic context. The inference machine identifies the meaning of the labels by examining the theoretical fuzzy models defined for the different syntactic patterns which can match the query text and obtains parameter values using contextual data available in the database. For this reason, from a general perspective, the proposed model represents a contribution of which the main novelty is the delegation for obtaining the models that represent linguistic labels to the inference machine, in order to get reliable answers from flexible querying systems.

The proposed conceptual model includes an extension of the standard query language SQL by overriding existing operators IS and LIKE and incorporating new ones to represent linguistic labels used as qualifying adjectives, quantifying adverbs, quantifiers and truth values. The proposal covers operators to represent and manipulate simple or complex vague terms, including those that are derived from a linear combination of conditions, vague or specific. It also provides a useful and easy way to extract new knowledge to characterize, approximately, associations between objects in a database.

**Keywords:** *Flexible Query Answering Language, Fuzzy Databases, Vagueness Representation*

# Agradecimientos

En primer lugar, quiero ofrecer mis agradecimientos a la Universidad Nacional de Colombia, la institución donde he recibido toda mi formación académica de postgrado y donde me he desempeñado como profesora por casi 15 años.

Agradezco sinceramente todo el apoyo constante, la lectura cuidadosa de mis escritos y los consejos del Dr. Hernán Darío Álvarez, director de esta Tesis Doctoral. Su trabajo de dirección es digno de tomarlo como ejemplo.

También quiero dar las gracias al Dr. Guillermo Morales Luna, del Departamento de Computación del Centro de Investigaciones y de Estudios Avanzados (CINVESTAV) del I.P.N, Ciudad de Méjico, por su invitación y acogida para realizar una pasantía de investigación, bajo su tutoría, y por todos sus aportes. En esa misma institución, quiero extender mis agradecimientos al Dr. Feliú Sagols, del Departamento de Matemáticas.

Adicionalmente, quiero agradecer a la Dra. Teresa de Pedro Lucio por su invitación para la realización de otra pasantía investigativa en el Instituto de Automática Industrial (IAI) del Consejo Superior de Investigaciones Científicas, en España. También al Dr. Miguel Ángel Sicilia, director de la Escuela de Ingeniería Informática de la Universidad de Alcalá de Henares, por el tiempo que me dedicó y la documentación que me facilitó para apoyar mi trabajo de investigación.

Por último, quiero manifestar mis agradecimientos a los estudiantes, Carlos Mario Soto, de la Maestría en Ingeniería de Sistemas, y a los estudiantes de pregrado Juliana Lopera, Lina Marcela Velásquez y Martín Rico, por su colaboración. Y también a mi familia, por su ánimo y comprensión.

## Lista de Símbolos y Abreviaturas

### LISTA DE SÍMBOLOS

$A^c$	Complemento del conjunto $A$ .
$ A $	Cardinalidad del conjunto $A$ .
$E$	Etiqueta lingüística.
$\mathcal{F}$	Marco de Cognición.
$\cap$	Operador de la Intersección de Conjuntos.
$\cup$	Operador de la Unión de Conjuntos.
$\alpha$	Umbral o grado mínimo de pertenencia a un conjunto difuso.
$\wedge$	Conjunción.
$\vee$	Disyunción.
$\prec$	Orden parcial en un marco de cognición.
$\pi$	Operador de la proyección en Álgebra Relacional.
$\Pi$	Modelo de conjunto difuso.
	Medida de posibilidad de ocurrencia de un conjunto difuso.
$\neg$	Negación.
$\oplus$	Operador difuso para la disyunción.
$\otimes$	Operador difuso para la conjunción.
$P(A)$	Probabilidad de ocurrencia del evento $A$ .
$P_q$	Percentil $q$ -ésimo, con $q = 0, 100$ . Valor en el dominio que deja a su izquierda, un porcentaje $q$ de los elementos cuando un conjunto de datos es ordenado.
$\mathcal{P}(U)$	Conjunto potencia sobre el conjunto universo $U$
$Q_i$	Cuartil $i$ -ésimo, con $i = 1,3$
$\sigma$	Operador de la selección en álgebra Relacional
$R$	Relación de una base de datos.
$R^*$	Relación derivada.
$\mu_A$	Función de pertenencia del conjunto difuso $A$ .
$\mu_A(x)$	Grado de pertenencia del elemento $x$ al conjunto difuso $A$ .

### LISTA DE ABREVIATURAS

BNF	Backus-Naur Form. Notación o Forma Normal de Backus-Naur.
CWA	Closed World Assumption. Supuesto de Mundo Cerrado.
DCB	Diagrama de Cajas y Bigotes.
DDL	Data Definition Language. Lenguaje de Definición de Datos.
DML	Data Manipulation Language. Lenguaje de Manipulación de Datos.
FATI	First Aggregate Then Infer. Primero Agregar y luego Inferir.
FITA	First Infer Then Aggregate. Primero Inferir y luego Agregar.

## LISTA DE ABREVIATURAS

FQAS	Flexible Query Answering Systems. Sistemas Flexibles de Consulta-Respuesta.
IQ	Intelligence Quotient. Cociente de Inteligencia.
ITL	Information Technology Laboratory. Laboratorio de Tecnología de la Información.
NIST	National Institute of Standards and Technologies. Instituto Nacional de Estándares y Tecnologías
IR	Information Retrieval. Recuperación de Información.
ISO	The International Organization for Standardization.
IEC	The International Electrotechnical Commission.
KBS	Knowledge Based Systems. Sistemas Basados en Conocimiento.
KDD	Knowledge Discovery in Databases. Descubrimiento de Conocimiento en Bases de Datos.
PLN	Procesamiento de Lenguaje Natural.
PRUF	Possibilistic Relational Universal Fuzzy.
LOWA	Linguistic Ordered Weighted Averaging. Promedio Ponderado Ordenado Lingüístico
LWA	Linguistic Weighted Averaging. Promedio Ponderado Lingüístico
OMG	Object Management Group. Grupo de Manejo de Objetos.
OWA	Ordered Weighted Averaging. Promedio Ponderado Ordenado.
SQL	Structured Query Language. Lenguaje de Consulta Estructurado.
UML	Unified Modeling Language. Lenguaje de Modelado Unificado.

## Lista de Tablas

Tabla 1. Conceptos de la Teoría de Conjuntos .....	19
Tabla 2. Propiedades del Algebra Booleana .....	22
Tabla 3. Clasificación de los individuos adultos, según su inteligencia .....	42
Tabla 4. Estadísticas para la estimación de la tendencia central .....	42
Tabla 5. Conocimiento obtenido de expertos .....	55
Tabla 6. Comparadores difusos de FSQL .....	68
Tabla 7. Distribución relativa de las notas .....	72
Tabla 8. Distribución de probalilidades de las notas .....	73
Tabla 9. Posibilidades para las agrupaciones de las notas .....	73
Tabla 10. Modelos teóricos para etiquetas simples, considerando tres clases.....	95
Tabla 11. Modelos teóricos para etiquetas simples, considerando dos clases.....	96
Tabla 12. Modelos para la potencia (HP) de los autos, considerando diferentes contextos .....	100
Tabla 13. Modelo teórico para las clases borrosas, basado en funciones no lineales .....	104
Tabla 14. Modelo teórico para el adverbio “muy”, basado en funciones lineales ...	112
Tabla 15. Modelo teórico de las clases borrosas “extremadamente” .....	113
Tabla 16. Algunos operadores propuestos para las conectivas lógicas.....	116
Tabla 17. Modelos propuestos para la representación de etiquetas lingüísticas .....	141
Tabla 18. Modelos para agregaciones, cuantificadores y calificadores de verdad ...	142
Tabla 19. Resultados de la ejecución de una consulta vaga .....	157
Tabla 20. Relación “Tipo_Termino” .....	160
Tabla 21. Relación “Termino” .....	160
Tabla 22. Relación “Modelo” .....	160
Tabla 23. Relación “Patrón” .....	161
Tabla 24. Relación “Consulta” .....	161
Tabla 25. Relación “Cond_vaga” .....	162
Tabla 26. Relación “Var_ling” .....	163
Tabla 27. Relación “Conj_Terminos” .....	163
Tabla 28. Estadísticas de resumen de los autos, considerando diferentes variables	175

## Lista de Figuras

Figura 1. Conceptos Básicos de la Teoría de Conjuntos Difusos.....	24
Figura 2. Mapeos en un sistema difuso o borroso.....	38
Figura 3. Distribución normal del IQ según dos pruebas diferentes.....	41
Figura 4. Histogramas de frecuencias con diferente resolución.....	44
Figura 5. Polígono de frecuencias acumuladas para las edades de los encuestados	45
Figura 6. Ejemplo de la función suavizada de la densidad del núcleo .....	46
Figura 7. Ejemplo de un Diagrama de Cajas y Bigotes.....	46
Figura 8. Distribución de los autos, considerando diferentes variables.....	48
Figura 9. Histograma de frecuencias para las notas definitivas.....	72
Figura 10. Modelo Estructural en la Interpretación de la Vaguedad.....	79
Figura 11. Proceso General de Razonamiento en la Resolución de Preguntas Vagas	80
Figura 12. Diagrama de actividades en el análisis preliminar de la consulta .....	85
Figura 13. Diagrama de actividades de la delimitación del contexto.....	86
Figura 14. Diagrama de Actividades de la Concreción de la Vaguedad.....	87
Figura 15. Diagrama de actividades de la obtención de la respuesta.....	89
Figura 16. Partición nítida de la distribución de los datos.....	93
Figura 17. Explicación del proceso de difuminado .....	94
Figura 18. Representación gráfica de la partición difusa en tres clases.....	96
Figura 19. Representación gráfica de la partición difusa en dos clases.....	97
Figura 20. Partición difusa en cuatro categorías.....	98
Figura 21. Partición difusa en cinco categorías.....	98
Figura 22. Partición difusa en seis categorías.....	99
Figura 23. Modelos de potencia “baja”, según diferentes contextos .....	100
Figura 24. Conjuntos borrosos para el rendimiento de los autos.....	101
Figura 25. Dos marcos de cognición para la potencia de los autos.....	101
Figura 26. Formas trapezoidales y su acentuación con el adverbio “muy” .....	102
Figura 27. Función Z y su representación gráfica.....	103
Figura 28. Función S y su representación gráfica .....	103
Figura 29. Modelos no lineales para la representación de conjuntos borrosos .....	105
Figura 30. Modelos lineales y no lineales, considerando tres clases.....	105
Figura 31. La concentración y dilatación, basada en la función de potencia.....	108
Figura 32. Efecto de los modificadores sobre la potencia “alta” de los autos .....	110
Figura 33. Partición borrosa para representar el adverbio “muy” .....	113
Figura 34. El valor máximo como operador de la disyunción.....	117
Figura 35. La suma como operador de la disyunción.....	118
Figura 36. Dos operadores para representar la conjunción.....	118
Figura 37. El cuantificador “Most” en SummarySQL.....	127
Figura 38. Cuantificadores relativos.....	128
Figura 39. Distribución de un conjunto difuso triangular.....	129
Figura 40. Representación del término “parecido a...” .....	131
Figura 41. Funciones propuestas por Baldwin para calificadores de la verdad .....	135
Figura 42. Partición para los calificadores lingüísticos de la verdad.....	136

Figura 43. Arquitectura para el sistema de consulta-respuesta.....	149
Figura 44. Interfaz de consultas con condiciones vagas simples o compuestas .....	151
Figura 45. Ventana emergente “Establecer relaciones” .....	152
Figura 46. Sección “Mostrar” de Interfaz de Consulta .....	153
Figura 47. Ventana emergente para formar expresiones.....	153
Figura 48. Sección “Condiciones” de la Interfaz de Consulta de Características ....	154
Figura 49. Especificación de una Combinación Lineal de Condiciones Simples .....	155
Figura 50. Interfaz para validación de reglas de asociación .....	158
Figura 51. Ejemplo de uso de la interfaz para validación de reglas .....	159
Figura 52. Diagrama de secuencias para la resolución de preguntas vagas.....	164
Figura 53. Formas lineales versus no lineales .....	170
Figura 54. Distribución de la tasa de fallas de un aparato electrónico .....	173
Figura 55. Distribución hipotética de las notas de un curso .....	174
Figura 56. Discriminación difusa de las notas hipotéticas .....	174
Figura 57. Diferencias entre dos notas consecutivas.....	175

# Capítulo 1

## 1 Introducción

En este Capítulo introductorio se presenta una síntesis del trabajo de investigación realizado. Se plantea el problema que se pretende resolver y se presentan los objetivos, el alcance y el método de trabajo definidos en la propuesta de investigación aprobada por un comité doctoral nombrado por la Universidad Nacional de Colombia. En la parte final del Capítulo se describen, de manera resumida, los logros obtenidos.

El objetivo central de esta investigación es hacer más flexibles los lenguajes de consulta a bases de datos objeto-relacionales, con el propósito de permitirle al usuario de estos sistemas expresar vagamente o de manera imprecisa las condiciones o las restricciones que deben cumplir los objetos de los cuales quiere conocer ciertas propiedades.

Los términos vagos que aquí se tratan, se caracterizan por ser dependientes del contexto lingüístico que delimita cada consulta y no de la percepción o del juicio subjetivo las personas. También se excluyen aquellos cuya vaguedad sea dependiente de contextos extralingüísticos como los originados por las diferentes culturas o creencias religiosas. Buena parte de los términos vagos que usamos en el lenguaje cotidiano son del tipo del objeto de estudio de este trabajo investigativo como son los términos “costoso” o “grande”, usados como adjetivos para calificar o clasificar los objetos de interés, de acuerdo con alguno de sus atributos. Son términos que se consideran imprecisos porque su significado, en términos cuantitativos, no tiene bordes claramente definidos. Y su semántica puede variar si se consideran diferentes contextos físicos.

Por la dependencia del contexto de los términos vagos, que sólo es conocido en el momento de la consulta, fue necesario proveer de capacidades de razonamiento no deductivo, a la máquina de inferencia encargada de la interpretación la vaguedad, con el objeto de encontrar, dinámica y autónomamente, el significado de las palabras vagas de cada consulta.

A la máquina de inferencia le corresponde determinar el contexto o el conjunto de elementos que sirven de marco de referencia para precisar la semántica de un término vago, mediante un proceso de discriminación difusa y con base en los datos que se tienen disponibles de ese contexto, que constituyen los hechos de la base de conocimientos. Con la concepción de un sistema de inferencia con

capacidades de razonamiento no deductivo, se logra la flexibilidad deseada en el lenguaje de interacción humano-máquina.

En este trabajo de investigación se consideraron modelos matemáticos provenientes de Teorías de Conjuntos Difusos propuestas para representar y operar con términos imprecisos o vagos y de la Estadística. Los primeros modelos enunciados, han surgido a partir del trabajo investigativo inicial del profesor Lofti A. Zadeh, basado en una lógica multivaluada (Zadeh, 1965 y 1975). Dicha teoría ha contribuido significativamente en otros trabajos de investigación, para la descripción y el análisis de grupos de una gran variedad de objetos, como también para el control automático de dispositivos que utilizan la experiencia o el conocimiento de expertos. Se puede citar la obra (Dubois y Prade, 2000) como uno de los compendios más completos de los muchos que existen sobre esta temática. El resultado de ese esfuerzo de investigación ha producido avances significativos en el terreno de la representación de la imprecisión lingüística.

## **1.1 Planteamiento del Problema**

Gracias a los avances tecnológicos en la Informática y las Telecomunicaciones, se ha incrementado en forma exponencial la cantidad de información digital disponible y de forma parecida, se ha incrementado la cantidad y variedad de usuarios finales que tienen acceso a ella. Hoy, ya es difícil concebir el trabajo o cualquier actividad académica e investigativa, sin el apoyo de los sistemas interactivos de consulta-respuesta a bases de datos, por el valioso conocimiento que nos pueden brindar.

A pesar de la gran ayuda que hoy ofrecen los sistemas consulta-respuesta al permitirnos plantear preguntas de manera espontánea o ad hoc, aún falta superar algunos problemas de rigidez en el lenguaje de interacción humano-máquina. Uno de ellos es la limitación para representar y manejar los términos vagos o imprecisos que un usuario pudiera especificar en los criterios de las consultas.

Habitualmente, un usuario consultor de una base de datos no puede emplear adjetivos vagos, que usa cotidianamente en su lenguaje natural como “bajo” o “costoso”, para calificar los objetos contenidos en una base de datos y así restringir lo que le interesa saber de ellos. Mucho menos, puede formular preguntas más complejas, pero también corrientes, como “¿La mayoría de los trabajadores actuales con “alto” desempeño, en esta empresa, son casados?” o “¿Se puede considerar “apto” el aspirante B para el cargo Y, considerando su escolaridad, sus años de experiencia y los resultados de su entrevista?”.

Lo más corriente es que un consultor no sepa si un objeto cualquiera se puede considerar “alto” o “reciente” en un contexto específico. Por eso, tiene que recurrir a la imaginación para adivinar valores concretos que expresará en los criterios de la consulta o realizar una serie de procedimientos o de cálculos previos, antes de formularla de una manera que sea aceptada por el sistema interactivo de consulta-respuesta utilizado.

El empleo de términos vagos en la formulación de una consulta resulta una manera más sencilla y cómoda de interactuar con un sistema informático. Incluso, a veces, puede ser suficiente o más conveniente que el sistema nos responda también vagamente, por ser una manera más simple y comprensible de ofrecernos conocimiento o información, tal como respondería un ser humano. Es el caso de una pregunta sobre el nivel de colesterol de una persona, puede ser preferible que la respuesta sea proporcionada de manera calificada, como “muy alto”, que una respuesta numérica que al usuario, no conocedor del tema, no le informe nada.

El significado de un calificativo vago, como los recientemente expuestos, es relativo al contexto o población que sirve de referencia o marco de comparación de los objetos de interés. Un estudiante con rendimiento académico “alto” en un curso, por ejemplo, puede no considerarse así, si lo comparamos con todos los estudiantes de su facultad o con los estudiantes de otro curso en el que esté inscrito. En otras palabras, para comprender una consulta vaga es necesario contextualizarla.

En las propuestas estudiadas para flexibilizar el lenguaje de consulta a sistemas de bases de datos, no se permite la contextualización de las consultas puesto que los modelos difusos para representar dichos términos deben definirse, con la ayuda de expertos, o proporcionárselos al sistema de consulta-respuesta, antes de poder interactuar con él (Galindo y Piattini, 2003; Goncalves y Tineo, 2001; Kacprzyk and Zadrozny, 2001; Bosc y Pivert, 1995; Medina et. al, 1994). Por esto, son propuestas que no permiten que un sistema de consulta-respuesta se adapte a los distintos contextos que pueden surgir, aún dentro de un mismo entorno o dominio de aplicación.

Los sistemas de consulta-respuesta, contruidos siguiendo las propuestas recién mencionadas requieren de expertos humanos, a lo largo de todo su ciclo de vida, para que los modelos usados sigan teniendo algún grado de validez en la interpretación de la vaguedad. Por otro lado, el grado de subjetividad inmerso en las propuestas antes citadas, incide negativamente en la confiabilidad de los sistemas de consulta-respuesta ya que se pueden producir diferentes respuestas a una misma consulta, a pesar de tener los mismos datos. Por estas razones, posiblemente, dichas propuestas aún no han sido acogidas por las grandes casas desarrolladoras de software comerciales como una estrategia clave para aproximar los lenguajes artificiales de consulta-respuesta, al lenguaje natural.

A diferencia de las propuestas anteriores, en (Zapata, 2002) se propone que el sistema de consulta-respuesta encuentre la semántica de los términos vagos, basándose en las distribuciones de frecuencias, a partir de los datos disponibles en la base de datos. Sin embargo, es una propuesta que no tiene la cobertura de las propuestas anteriores debido al reducido número de términos vagos que permite representar. También presenta la limitación de subdividir el Universo de Discurso en sólo dos clases o categorías y otras limitaciones que se detallarán más adelante, pues han servido de base para las mejoras sugeridas en este trabajo de investigación.

Por las limitaciones detectadas en las propuestas estudiadas, en esta investigación se abordó el problema de concebir una máquina de inferencia que

permita la flexibilización del lenguaje de los sistemas interactivos de consulta-respuesta, dotándola con la inteligencia suficiente para que interprete la vaguedad en las consultas, de manera dinámica y autónoma, considerando el contexto lingüístico enmarcado por la misma consulta. Todo ello con el propósito de permitir el uso de un lenguaje más parecido al que los humanos usamos naturalmente, en las consultas a bases de datos, logrando que el usuario piense que está interactuando con un experto colaborador, en lugar de una máquina. Con ello, no sólo se facilitaría la interacción humano-máquina, sino que se aumentaría la satisfacción de los usuarios en la resolución de sus cuestionamientos, gracias a la confiabilidad de las respuestas que le ofrezca el sistema.

## **1.2 Objetivos**

### **1.2.1 Objetivo General**

El objetivo general del presente trabajo de investigación consiste en flexibilizar el lenguaje artificial de los sistemas de consulta-respuesta, con la modificación o la adición de nuevos operadores que capturen la semántica de los términos vagos, simples o complejos, según el contexto que se establece dinámicamente en el momento de una consulta y que puedan ser implementados sobre cualquier lenguaje de interfaz de diálogo hombre-máquina con sistemas de bases de datos que soporten el modelo objeto-relacional.

### **1.2.2 Objetivos Específicos**

A continuación se presentan los objetivos específicos que se desean alcanzar para cumplir con el objetivo general, divididos en los niveles corrientes de evolución de un trabajo de investigación.

#### **1.2.2.1 Nivel Perceptivo**

- Describir los modelos convencionalmente utilizados para representar y dar tratamiento a la incertidumbre originada por la vaguedad, con el fin de potenciar su utilización como parte sustantiva de una interfaz de humano-máquina inteligente<sup>1</sup>.
- Caracterizar distintos métodos de producir inferencias, que intentan reproducir esquemas mentales del cerebro humano y la estructura de la lógica subyacente.
- Establecer las propiedades deseables de un modelo genérico de razonamiento aproximado y adaptable<sup>2</sup> en la comprensión del lenguaje vago de los sistemas consulta-respuesta y sus indicadores de satisfacción.

---

<sup>1</sup> La inteligencia de una interfaz (o lenguaje artificial), aquí se entiende como su capacidad de encontrar en forma aproximada, pero de forma más autónoma y objetiva, el significado de los conceptos vagos dependientes del contexto.

### **1.2.2.2 Nivel Aprehensivo**

- Identificar los alcances y limitaciones de las técnicas aproximadas para la comprensión de los conceptos vagos inmersos en el lenguaje de usuario final a sistemas de consulta-respuesta.
- Contrastar los distintos métodos de inferencia estudiados, analizando su complejidad y sus propiedades estructurales.

### **1.2.2.3 Nivel Comprensivo**

- Proponer al menos una modificación en los elementos propios y metodológicos para una mejor aproximación al verdadero significado de los conceptos vagos en el lenguaje de consulta o interfaz humano- máquina, según su contexto.
- Elaborar un modelo conceptual del razonamiento aproximado que represente las reglas semánticas necesarias para que un lenguaje artificial sea más parecido al natural al admitir preguntas vagas y se adapte de manera dinámica al contexto definido por la consulta.

### **1.2.2.4 Nivel Integrador**

- Comprobar, mediante un ejemplo, la traducibilidad del modelo conceptual propuesto implementándolo en el lenguaje de consulta de una interfaz Web.
- Evaluar la efectividad del modelo propuesto de razonamiento aproximado para lograr la flexibilización deseada del lenguaje de consulta, con base en los indicadores de satisfacción planteados como un objetivo específico en el nivel perceptivo.

## **1.3 Metodología y Alcance**

En este trabajo se intentó seguir un proceso de investigación riguroso, con el fin de exponer y confirmar la hipótesis planteada sobre la posibilidad de mejorar los sistemas interactivos de consulta en la comprensión de los términos vagos, considerando el contexto lingüístico que enmarca cada consulta.

La definición del modelo conceptual de razonamiento y su validación constituyeron un proceso iterativo pero incremental, entre las fases reconocidas en un trabajo de investigación o de descubrimiento de nuevo conocimiento: la fase perceptiva, la aprehensiva, la comprensiva y la integradora, según se indica en (Bloom, 1990). El método de trabajo aplicado se puede catalogar como un método principalmente “arriba-abajo”, pues se partió de la definición general de las funciones y las restricciones que debía cumplir el sistema de inferencia, inmerso en un gestor de bases de datos objeto-relacional, en la interpretación de la vaguedad de las consultas, sin entrar en detalles en los procesos, hasta llegar a la especificación completa del modelo de la solución propuesto.

Sobre el alcance de la investigación, el sistema de inferencia debe poder capturar el significado de términos vagos, simples o compuestos, que califican a los objetos contenidos en una base de datos objeto-relacional, que sean dependientes del contexto especificado en la consulta. Debido a esto, no se incluye la interpretación de términos vagos subjetivos, como “bonito” o de conceptos como “bueno” que dependen de factores extralingüísticos como la religión o el entorno socio-cultural.

En este trabajo se aborda el problema de la vaguedad en la calificación de los atributos cuyo dominio es cuantitativo, excluyendo atributos con dominios nominales representados con cadenas de caracteres o mediante imágenes. El método de razonamiento para comprender la vaguedad especificada en las consultas, se circunscribe a los sistemas de consulta-respuesta a bases de datos cuyo soporte lógico sea el modelo objeto-relacional. La razón de la elección de este modelo de bases de datos, consiste en que éste extiende al modelo relacional, su predecesor. Como es un modelo más general, se espera que lo que se proponga también sea válido para su modelo particular: el modelo relacional puro.

Los datos de ejemplo utilizados para la explicación de la propuesta o para presentar los resultados que arroja un sistema basado en ella, fueron extractados de una base de datos de referencia disponible en la Web y utilizada por otros investigadores para acreditar sus teorías o proposiciones y garantizar la reproducibilidad del trabajo de investigación aquí realizado.

#### **1.4 Método de Evaluación**

Para la evaluación del modelo de razonamiento aproximado en la interpretación de las consultas vagas propuesto en la presente investigación, se estableció un conjunto de propiedades deseables como modelo conceptual, de manera general, y específicamente como modelo de un sistema de inferencia difusa. Sobre este último aspecto, se consideraron algunas restricciones que la literatura sugiere que deben cumplir estos modelos para que la representación de los términos vagos en las consultas flexibles a bases de datos, se caracterice por consistencia lógica e interpretabilidad.

Adicionalmente, el presente trabajo de investigación incluyó la creación y experimentación de una interfaz humano-máquina para certificar la traducibilidad del modelo de razonamiento propuesto. Mediante el prototipo diseñado con las características del modelo conceptual propuesto, se corroboró que se pueden diseñar y construir sistemas mejorados de consulta-respuesta a bases de datos objeto-relacionales, a partir de esta propuesta.

#### **1.5 Aportes Originales**

Después de concluir este trabajo de investigación se ha confirmado la hipótesis planteada en la propuesta de tesis sobre la posibilidad de ampliar y fortalecer el modelo de razonamiento del sistema de inferencia inmerso en un sistema flexible de consulta-respuesta, en la interpretación de los términos vagos expresados en las solicitudes de los usuarios.

- Se ha aportado un modelo de razonamiento aproximado para la resolución de la vaguedad en las condiciones de las consultas a bases de datos objeto-relacionales, con las características que se describen a continuación.
- Permite que la interpretación de las consultas vagas se ajuste de forma dinámica a su contexto lingüístico, aumentando la validez de las respuestas y, por ende, se generen sistemas de consulta-respuesta más confiables.
- Permite que la interpretación de las etiquetas vagas usadas como adjetivos calificativos para restringir los objetos de una base de datos, se ajuste a distintos niveles de granularidad, en el marco de cognición.
- Define una técnica heurística alternativa para la representación de modificadores lingüísticos que representan los adverbios de cantidad como “muy” o “extremadamente”, que emula mejor el modo de proceder de los humanos.
- Extiende el lenguaje de consulta para representar y manejar términos vagos que pueden deducirse de otros, ampliando las capacidades de derivación de términos complejos, con la representación de condiciones vagas que se pueden expresar como una combinación lineal de otras simples, sean éstas vagas o concretas.
- Extiende propuestas previas para la flexibilización de los lenguajes de consulta a bases de datos, por permitir que un sistema interactivo de consulta-respuesta responda con valores lingüísticos de la verdad, cuando se quiera saber si existen asociaciones entre las propiedades de los objetos.
- Aporta un mecanismo útil y sencillo de extracción nuevo de conocimiento para caracterizar, de manera aproximada, las asociaciones existentes las propiedades de los objetos de la base de datos y usarlo en la toma de decisiones. Esto se logró, mediante la representación del concepto de cuantificador difuso y de los valores lingüísticos de la verdad, dentro del lenguaje de consulta.

## **1.6 Estructura del Documento**

La memoria de este trabajo de investigación se estructura como se describe a continuación. En el presente Capítulo, se introduce el trabajo realizado, se presentan los objetivos y la metodología seguida. También se presentan los aportes obtenidos como resultado del trabajo de investigación.

En el Capítulo 2 se aborda el objeto de estudio, la vaguedad del lenguaje, y se presenta sintéticamente el área de conocimiento o disciplinar donde se orientan los aportes de este trabajo de investigación. En el Capítulo 3 se describen las técnicas o modelos empleados para la representación de la vaguedad en los lenguajes artificiales y las técnicas para sistemas de inferencia que permiten representar y operar con términos vagos. Posteriormente, en el Capítulo 4, se presentan todos los requerimientos o restricciones que deben ser considerados en la definición del modelo de razonamiento aproximado y adaptable, en el procesamiento de las

consultas vagas. En el Capítulo 5 se presenta el modelo conceptual propuesto de una manera general y en el Capítulo 6, se profundiza en las técnicas para la generación dinámica de los modelos que representan los distintos tipos de vaguedad admisibles en una consulta, analizando las adiciones o cambios sintácticos y semánticos en el lenguaje. Para cada tipo de vaguedad contemplado se presentan pruebas experimentales o comparativas con propuestas previas y se consideran las restricciones impuestas al modelo.

En ese Capítulo 6 también queda reflejado el proceso de síntesis o integrador de los resultados de los distintos análisis, pues se formalizan las reglas sintácticas y semánticas requeridas para la interpretación de tipo de consulta vaga admisible en el modelo de comunicación humano-máquina propuesto. En el Capítulo 7 se presentan los resultados del estudio de su factibilidad técnica y de la evaluación del nuevo modelo de razonamiento para el procesamiento de las preguntas vagas, considerando el grado de ajuste del mismo a las restricciones preestablecidas. En este mismo capítulo se finaliza la memoria con las conclusiones y el trabajo futuro.

# Capítulo 2

## 2 El lenguaje y su Vaguedad

El propósito de este Capítulo, es describir el objeto de estudio del presente trabajo de investigación: el lenguaje de consulta de bases de datos objeto-relacionales. También se presentan las áreas de la Inteligencia Artificial comprometidas con la gestión del conocimiento y el procesamiento del lenguaje natural, con el fin de diferenciar un sistema flexible de consulta-respuesta de los sistemas de recuperación de información, que son otros tipos de sistemas informáticos interactivos para la realización de consultas a bases de datos, pero con capacidades más limitadas.

Después de esa descripción, se presenta el modelo lógico objeto-relacional que subyace dentro de los sistemas gestores de bases de datos más recientes, en el cual se enfocan los aportes de este trabajo investigativo.

### **2.1 El Lenguaje Natural.**

El lenguaje se define como el sistema de comunicación con el que el ser humano expresa sus sentimientos, sus conocimientos y sus razonamientos. Es indudable que un paso decisivo en la evolución del género humano fue la adquisición de un vínculo entre el pensamiento y los símbolos materiales; ya que por primera vez establecía una relación entre representaciones mentales y símbolos externos construidos deliberadamente. La comunicación llevada a cabo inicialmente por medio de señales gestuales y vocales, culmina con la escritura de signos, para conformar lo que es conocido como lenguaje natural. Aún existen lenguajes humanos que tienen una representación sonora pero carecen de un modelo escrito con el que se pueda expresar todo lo que se puede manifestar y decir oralmente (Llisterri, 2003). No obstante, el lenguaje ha permitido designar las cosas y razonar acerca de ellas, así como también crear nuevos significados, independientemente de si poseen modelo escrito o no.

La riqueza semántica del lenguaje como resultado de su uso, en múltiples situaciones, ha permitido el gran poder expresivo que hoy lo caracteriza constituyéndose en una herramienta fundamental para los razonamientos profundos y para la transmisión del conocimiento. Los símbolos de un lenguaje son transportadores de conocimiento que pueden apelar a cualquiera de nuestros sentidos para hacernos llegar el contenido semántico y el sentido perceptivo que encierran.

El lenguaje natural, es la forma más potente de representación del conocimiento y, por tanto, de gestionar la información (Codina, 2001). Más aún, la

generación, representación y transferencia del conocimiento a través del lenguaje fueron algunos de los factores determinantes para la supervivencia de nuestra especie sobre otros homínidos (Betranpetit y Junyent, 2000). Sin embargo, la flexibilidad que otorga el lenguaje para representar múltiples conceptos con las mismas palabras o frases puede dar origen a interpretaciones erradas o confusas por parte del receptor, poniendo obstáculos a una buena comunicación.

La ambigüedad del lenguaje hace que el receptor de un mensaje dude sobre la verdadera semántica de las palabras o frases contenidas en él. La ambigüedad, ocurre cuando una palabra o frase posee múltiples significados o interpretaciones, como sucede con la palabra “banco” pues puede hacer referencia a un establecimiento financiero, a un gran cúmulo de arena, de datos o a un asiento, entre muchas otras cosas, y de ahí que demande más de una representación semántica, aún en el mismo contexto. Una palabra vaga también es ambigua pues presenta variaciones en su significado pero los cambios no suelen ser tan abruptos, que conduzcan a dominios totalmente diferentes como en el ejemplo anterior. Se dice que la vaguedad presenta variaciones pero dentro de un mismo significado (Bosc, 1979). Así pues, una palabra vaga es ambigua, pero una ambigua no necesariamente es vaga, dado que la palabra ambigua suele representar objetos concretos y no abstractos, que pueden determinarse exactamente en el proceso de desambiguación.

Los conceptos vagos no tienen un significado universal cuando son subjetivos o porque son dependientes del contexto físico o situacional que se establece cuando se emite un mensaje. Los conceptos vagos subjetivos varían en su significado según un juicio individual como el concepto de “belleza” o “amistad” siendo, por lo tanto, más opiniones o juicios que conceptos. Otros términos cambian de significado según el contexto y no por la percepción particular que un individuo tiene de ellos. Ejemplos de este último tipo de vaguedad, son los conceptos de juventud, extensión, lejanía y riqueza ya que su significado es relativo a la región o a cualquier otro contexto considerado. Por ejemplo, una persona de 40 años puede considerarse “joven” para desempeñar alguna función, como la gerencia de una empresa, pero se puede considerar “vieja” en otro contexto diferente.

Para los conceptos vagos, no hay valores fijos y bien delimitados para establecer lo que se puede considerar “alto” o “importante” o diferenciar entre términos cercanos como “apto” y “muy apto”, aún dentro de un mismo contexto. No existen unas fronteras claramente definidas para calificar un objeto porque el paso de una clase a la otra es gradual, haciendo difícil la clasificación de los mismos dentro de un grupo u otro.

Los conceptos vagos, además de ser variables en el espacio o contexto físico, pueden ser variables en el tiempo: una persona de cierta edad se consideraba “vieja” años atrás, mientras que hoy podría no ser considerada así, ya que los avances en la medicina han permitido aumentar la expectativa de vida de las personas. Debido a esto, la semántica de las palabras vagas puede depender también del contexto histórico.

Un contexto histórico es un tipo de contexto situacional, de igual forma que un contexto sociocultural o religioso, que pueden implicar diferencias semánticas significativas para algunos conceptos vagos como la bondad de un individuo. Ese tipo de contexto es llamado extralingüístico porque no es posible deducirlo del conjunto de unidades lingüísticas que preceden o siguen a la unidad en cuestión porque son externas a los actos lingüísticos (Cerezo, 1994). En algunos casos, es necesario considerar factores extralingüísticos que proporcionen datos para saber cuáles acepciones son las que hay que activar para entender adecuadamente un mensaje. Sin embargo, de acuerdo con el alcance definido para este trabajo de investigación, los factores extralingüísticos, no van a ser considerados.

Los términos vagos se pueden clasificar en dos categorías: simples y complejos. Para la interpretación de los términos vagos simples se requiere conocer únicamente el dominio de una única variable cuantitativa. Así ocurre con el concepto de documento "reciente" ya que sólo se requiere de la fecha de su elaboración. Por otra parte, los términos complejos requieren de más de una variable para aproximarnos a su verdadero significado. Por ejemplo, la complejidad física de una persona está determinada no sólo por el peso, sino también por la altura de la persona y es posible que las variables no aporten, en la explicación de un término vago complejo, en la misma proporción. Puede ser que importe más el peso que la estatura para explicar la complejidad física de un individuo, por ejemplo. Es por esto, que se ha considerado relevante representar los términos vagos complejos definidos de esta manera, con el propósito de aumentar la potencia expresiva del modelo de razonamiento requerido para responder preguntas formuladas vagamente en un sistema de consulta-respuestas a bases de datos objeto-relacionales.

La determinación de la semántica de las palabras o frases en un contexto dado es un problema fundamental de la Lingüística. El lingüista Charles Bally dice en su obra "El lenguaje y la vida humana" (Bally, 1977): "En el lenguaje diario, no hay palabra que no tenga varios sentidos y que no se preste a confusión". Esta propiedad única de los lenguajes naturales se conoce en Lingüística con el nombre polisemia. Este término se define formalmente como la pluralidad de significados de una palabra o de cualquier signo lingüístico, con independencia de la naturaleza de los signos que lo constituyen.

La Semántica, como rama de la Lingüística, trata aspectos relacionados con el sentido o percepción del mensaje, con los referentes, las condiciones de verdad de las proposiciones y el análisis del discurso. Por su lado, la Pragmática es el estudio de la comprensión del lenguaje natural y del modo en que el contexto influye en la interpretación del significado (Jurafsky, 2005). Por esto, la Pragmática es considerada como la parte de la Semántica que se ocupa del significado dependiente del contexto. La Pragmática intenta entender las relaciones entre los signos y las interpretaciones, o las relaciones entre la oración y el mundo externo, mientras la Semántica se enfoca en los objetos reales o en las ideas a las que las palabras se refieren y la Sintaxis, en las relaciones entre los signos (Bach, 2005). De acuerdo con esto, el problema de la interpretación de una expresión polisémica, por la vaguedad de las palabras que la conforman, se puede considerar un asunto de la Pragmática.

## **2.2 Lenguajes Formales Artificiales.**

Aunque el español o cualquier otro lenguaje son producto del artificio humano, su creación a lo largo de siglos no fue plenamente consciente y racional como ha sido la creación de lenguajes con fines más restringidos, como los lenguajes de programación, concebidos para la comunicación con las computadoras o como un lenguaje matemático requerido para la comunicación científica. Debido a esta característica, los lenguajes formales no sirven para expresar emociones ni percepciones, sino tan sólo una estrecha gama de fenómenos o problemas definidos lógicamente. Tampoco son lenguajes que posean una representación oral, ya que su fin no es que se entiendan de esta manera (pocos estarían dispuestos a dialogar con una máquina con un lenguaje como el Java), lo cual constituye otra diferencia importante con lo que se denomina lenguaje natural.

Un lenguaje artificial de programación actual tiene como objetivo fundamental lograr una comunicación humano-máquina similar a la comunicación humano-humano. El Procesamiento de Lenguaje Natural (PLN), que intenta simular el comportamiento lingüístico humano, se define como "una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación humano-máquina y permitan una comunicación mucho más fluida y menos rígida que con los lenguajes formales" (Moreno, 1999).

Otros autores, suelen usar términos parecidos al PLN como la Lingüística Computacional o Ingeniería Lingüística. Ésta última es definida como "la aplicación de los conocimientos sobre la lengua al desarrollo de sistemas informáticos que puedan reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas" (Comisión Europea, 2002). Sin importar cual nombre sea el más apropiado, estas tecnologías comprenden una serie de procedimientos relacionados con el tratamiento informático del lenguaje. En general, suelen distinguirse las técnicas que se aplican al tratamiento de la lengua hablada, de las propias del procesamiento del texto escrito. El presente trabajo de investigación aborda el último caso.

Los lenguajes naturales permiten mayor flexibilidad y espontaneidad en la interacción humano-humano o humano-máquina pero ponen obstáculos en la comunicación, bien sea por la ambigüedad o por la vaguedad que los caracteriza. Por el contrario, los lenguajes formales evitan la mala interpretación de lo que se quiere decir, por su rigor. Pero se requiere de un nivel alto de conocimiento y mayores esfuerzos por parte del emisor de una consulta, para poder lograr una comunicación fluida.

A diferencia del lenguaje formal, donde el significado de una expresión o frase sólo está influido por aspectos gramaticales como la notación y la sintaxis, en los lenguajes naturales el significado específico y contextual de sus componentes intervienen en la validez de un mensaje y por eso, añaden tanta complejidad a su estudio.

De acuerdo con esto, el presente trabajo de investigación, cubre en forma simultánea dos áreas de trabajo de la Ingeniería Lingüística: la Comprensión de Lenguaje y la Gestión del Conocimiento. Dado que esta última área incluye la Recuperación de Información, los Sistemas Flexibles de Consulta-Respuesta y el Descubrimiento de Nuevo Conocimiento, se procederá a describir cada una de estas tareas analíticas.

## **2.3 Tipos de Gestión del Conocimiento en la Ingeniería Lingüística**

### **2.3.1 Recuperación de Información.**

La recuperación de información (Information Retrieval o IR, en inglés) consiste en la búsqueda de información en documentos o de información sobre documentos (los metadatos), generalmente, a través de la Web. También incluye la búsqueda de cualquier tipo de información almacenada en bases de datos, a través de una interfaz de usuario. La función principal de este tipo de sistemas es la de recoger la solicitud de un usuario, procesarla y generar una respuesta adecuada. Se ha establecido que los Sistemas de Recuperación de Información se diferencian de los Recuperadores de Datos en que éstos últimos son más limitados pues sólo traen los ítems que cumplen plenamente los criterios de la consulta (Large, 1999). En los sistemas de recuperación de información, en cambio, se admite el cumplimiento parcial de los requerimientos del usuario. Esto se hace seleccionando los ítems que cumplen, en alguna medida, las condiciones de la consulta.

Para que un sistema de recuperación de información sea efectivo en el intento de satisfacer las necesidades de los usuarios debe, de alguna manera, interpretar la solicitud del usuario y traer la colección de elementos requeridos y ordenarlos según el grado de relevancia. Esta interpretación involucra la extracción de información semántica de los ítems de datos. La dificultad no es sólo saber cómo extraerla, sino cómo usarla para decidir su relevancia o grado de cumplimiento. Por esto, la noción de relevancia es el centro de este tipo de sistemas (Baeza y Ribeiro, 1999). El grado de relevancia se determina, usualmente, con la coincidencia de las palabras declaradas por el usuario y las existentes en los documentos, utilizando técnicas estadísticas convencionales, basadas en una lógica bivaluada.

### **2.3.2 Sistemas Flexibles de Consulta-Respuesta**

Estos tipos de sistemas pertenecen a los sistemas de recuperación de información pero son de mayor complejidad puesto que, dada una colección de elementos, el sistema debe responder a las preguntas realizadas, aunque la consulta pueda formularse de manera imprecisa o incompleta (Bosc, Motro y Pasi, 2001). Por esto, la palabra clave para describir los sistemas flexibles de consulta-respuesta (abreviadamente FQAS por sus siglas en inglés), es la tolerancia a la imprecisión en la formulación de las consultas y también en la representación y manejo de la información incierta. Se dice que los sistemas flexibles de consulta-respuesta pueden considerarse los sucesores de los sistemas IR, pero requieren una mayor comprensión del texto (Llopis, Fernández y Vicedo, 2001).

En (Zadeh, 2006) se señala que los sistemas flexibles de consulta-respuesta se distinguen de los sistemas de recuperación de información, por sus capacidades deductivas para la concreción (presitiation) de la vaguedad. Por lo tanto, la definición de un sistema flexible consulta-respuesta, encaja plenamente con las características del sistema bajo estudio en el presente trabajo de investigación.

### 2.3.3 Descubrimiento de Nuevo Conocimiento.

Como una extensión a la gestión del conocimiento, en los últimos años el Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases, KDD), ha recibido especial atención debido a la disponibilidad actual de grandes cantidades de datos y a la necesidad de convertirlos en información útil y en nuevo conocimiento. Este proceso también es conocido como Minería de Datos: un proceso que, a través del análisis y la cuantificación de relaciones en los datos, permite extraer patrones comunes, asociaciones o modelar el comportamiento dinámico de distintos fenómenos de la naturaleza o la sociedad (Han y Kamber, 2001). Pero siendo rigurosos, la Minería de Datos es el proceso central del KDD pues se necesitan otros procesos antes y después de aplicar técnicas para extraer el nuevo conocimiento. La totalidad de los pasos en el proceso de descubrimiento de nuevo conocimiento que han sido planteados en (Mitra y Acharya, 2003) se describen a continuación.

- **Comprensión del dominio de aplicación.** Incluye la recolección de la información *a priori* relevante sobre los objetivos del trabajo que se aborda, del dominio del problema y de los supuestos que se cumplen.
- **Selección de datos.** Se eligen los datos que se consideran relevantes para el análisis y se hace un proceso de integración de datos, si están almacenados en distintas fuentes o con distintos formatos.
- **Preprocesamiento de datos.** Este proceso se necesita para depurar la información, chequear inconsistencias o preparar los datos para la minería. El preprocesamiento puede incluir tareas de:
  - **Limpieza de datos.** Consiste en la detección y corrección de datos errados o inconsistentes.
  - **Transformación.** En ocasiones, los datos deben ser transformados o consolidados en una forma apropiada para la minería. Puede ser necesario resumir la información recolectada, realizar cambios de escala o reducir la dimensionalidad del problema, antes de aplicar una técnica de minería de datos, en particular.
- **Minería de Datos.** Es el proceso central, en el cual se aplican los métodos o técnicas que permitan el análisis de los datos para encontrar relaciones implícitas o patrones previamente desconocidos. Son diversas las técnicas que se pueden aplicar para el análisis discriminante de los datos, para la clasificación de los objetos de interés o para el hallazgo de asociaciones entre variables, entre otras tareas de la minería, que se eligen según las características de los datos y las

restricciones que se deben considerar en un caso determinado. Por esto, la minería de datos se puede realizar con un amplio espectro de técnicas de las Bases de Datos, de la Estadística y de la Inteligencia Artificial, con el fin de obtener la caracterización de un dominio particular.

- **Evaluación de los hallazgos.** Este proceso es necesario para validar los modelos construidos o decidir si aceptar o no, las hipótesis planteadas como tesis o nuevo conocimiento. En la validación de los modelos, es preciso determinar el grado de bondad del ajuste de los datos reales a los modelos, realizando pruebas experimentales o validando el cumplimiento de las restricciones impuestas de antemano, a esos nuevos modelos.
- **Interpretación y presentación del conocimiento.** Es el paso final del proceso de descubrimiento de nuevo conocimiento, donde se aplican técnicas para la visualización y la representación del conocimiento “minado”.

Ya descritos los tres sistemas para la Gestión del Conocimiento (la Recuperación de la Información, los Sistemas Flexibles de Consulta-Respuesta y el KDD) de los que se ocupa la Ingeniería Lingüística, se puede afirmar que los tres demandan, generalmente, un razonamiento de tipo no deductivo para el cálculo del grado de relevancia, el manejo de la imprecisión y para las tareas de minería de datos, respectivamente.

En el presente trabajo de investigación, para la interpretación y manejo de la vaguedad o imprecisión en el lenguaje de consulta, se requiere que el propio sistema realice, de manera autónoma y adaptable a cada contexto, tareas de minería de datos. Más concretamente, se requiere que pueda realizar tareas de discriminación y clasificación para determinar el grado de cumplimiento de los objetos de interés, a las condiciones establecidas vagamente en las consultas.

## 2.4 El Modelo Objeto-Relacional y los Lenguajes de Consulta

El modelo relacional de datos, que sirvió de base al modelo objeto-relacional, fue propuesto por E.F. Codd en 1972 para superar la complejidad y las limitaciones de los modelos que en ese entonces eran los utilizados en los sistemas gestores de bases de datos: el modelo jerárquico y en red, según el recuento histórico reportado en (Date, 2001). El concepto matemático que subyace bajo el modelo relacional es el concepto de relación, tal como se define en la teoría clásica de conjuntos. Una relación  $R$  es un subconjunto del producto cartesiano de varios dominios:  $R: D_1 \times D_2 \times \dots \times D_k$  que puede ser descrita por extensión, mediante una lista de los elementos que la componen, o por comprensión mediante una fórmula de derivación que genera una relación derivada, denominada en Bases de Datos como “vista”. El concepto de relación, así descrita, le proporciona el nombre al modelo relacional. A los elementos de una relación se les conoce como tuplas. Una relación  $R$  del producto cartesiano  $D_1 \times D_2 \times \dots \times D_k$ , se dice que tiene orden (o grado)  $k$  de acuerdo con el número de atributos o propiedades consideradas.  $R$  tiene una

cardinalidad (o extensión) de  $n$ , si éste es el número de tuplas que hacen parte de dicha relación.

El esquema de una relación  $R$  es definido mediante un conjunto finito de atributos  $A_1, A_2, \dots, A_k$ . Cada atributo  $A_i$  tiene asociado un dominio  $D_i$ , con  $1 \leq i \leq k$  de donde toma sus posibles valores. El esquema relacional de una base de datos se representa como un conjunto de relaciones  $R_i (A_{i1}:D_1, A_{i2}:D_2, \dots, A_{ik}:D_k)$  donde el orden de los atributos o de las tuplas no interesa, pero sí que cada tupla pueda diferenciarse de las demás, por al menos el valor de uno de sus atributos. Una relación definida de esta manera, está representando el concepto básico de la Teoría de Conjuntos basada en la lógica booleana.

Para el modelo relacional, Codd diseñó dos lenguajes teóricos para expresar las operaciones entre relaciones, que se diferencian entre sí por el nivel de abstracción. El primero de ellos, el Álgebra Relacional, es un lenguaje de especificación formal en el cual las consultas se resuelven aplicando operadores a las relaciones de una base de datos, los operandos. Con este lenguaje pueden usar operadores tradicionales de conjuntos, como la unión o la intersección, y otros especializados como la restricción (denotada usualmente por el símbolo  $\sigma$ ) que sirve para seleccionar las tuplas de una relación que satisfagan las condiciones especificadas, la proyección (denotada por  $\pi$ ) para extraer los atributos o características que se desean visualizar de una relación y la reunión o "join" que consiste en una operación más compleja por ser una restricción del producto cartesiano de dos o más relaciones. El Cálculo Relacional es el otro lenguaje teórico propuesto por Codd, por medio del cual las consultas se expresan formulando restricciones lógicas que las tuplas de la relación derivada, deben satisfacer. El Cálculo Relacional se basa en la lógica de primer orden y tiene dos variantes: el Cálculo Relacional de Dominios, donde las variables esperan componentes (atributos) de las tuplas y el Cálculo Relacional de Tuplas donde las variables esperan tuplas.

El Álgebra Relacional y el Cálculo Relacional son lenguajes teóricos equivalentes, es decir, todas las consultas que se pueden formular utilizando Álgebra Relacional pueden también formularse utilizando el Cálculo Relacional y viceversa. Sin embargo, los lenguajes basados en el Cálculo Relacional tienen mayor poder expresivo, pues una consulta planteada en este lenguaje se traduce, regularmente, en varias órdenes del álgebra para poder resolverla. Esto significa que el Cálculo Relacional es menos procedimental o más declarativo. El lenguaje estándar de bases de datos SQL (acrónimo de Structured Query Language) se parece más al cálculo relacional de tuplas que al álgebra.

Una de las características del lenguaje de consulta a bases de datos es que se cumple la ley de la clausura: el resultado de cualquier operación aplicada sobre relaciones siempre es otra relación. La relación resultante de una operación está constituida por la tuplas que evalúan como cierto un predicado y quedan excluidas las demás, incluyendo en este último grupo a tuplas cuyo valor de verdad sea indeterminado, ya que en el modelo relacional se amplía la lógica bivaluada para

admitir y operar con valores nulos, no aplicables o desconocidos y representados con la palabra NULL.

En los últimos años se han propuesto y materializado algunas extensiones al modelo relacional como el llamado Modelo Objeto-Relacional, para incorporarle, a los sistemas gestores de bases de datos, las mejores características de la Tecnología Orientada por Objetos. Este último modelo admite la declaración y el manejo de atributos objetovaluados o compuestos de otros atributos y permite también múltiples valores para un atributo. Es decir, el Modelo Objeto-Relacional permite que los datos no sean atómicos, sino estructurados de diferentes maneras. Esta era una exigencia impuesta más por problemas de implementación que por violación del concepto de relación matemática. Otra nueva e importante característica del Modelo Objeto-Relacional es la posibilidad de definir nuevos tipos de datos, a partir de los nativos. Por eso, en (Date, 2001) se señala que un modelo objeto-relacional es ni más ni menos que el modelo relacional verdadero. El Modelo Objeto-Relacional subsume al Modelo Relacional “puro”, evitando tener que modificar datos existentes en bases de datos relacionales. De ahí, que lo que se proponga para la flexibilización del lenguaje de consulta, es válido también para las bases de datos relacionales.

Un lenguaje de consulta de bases de datos se ha caracterizado por ser declarativo. Gracias a esto, y a la admisión de operaciones de conjunto, el lenguaje de consulta permite una alta productividad y simplicidad en la programación. Así, una sola sentencia en SQL puede equivaler a muchas más líneas de código en un lenguaje procedimental, orientado a operaciones sobre elementos individuales. Sin embargo, este tipo de lenguaje aún demanda un alto dominio de la persona que quiera consultar una base de datos, pues se requiere que conozca todas las palabras y las reglas sintácticas o semánticas del lenguaje para formular las consultas. Tratando de evitar esto, existe una proliferación de interfaces gráficas de usuario que sustituyen a las órdenes y a la sintaxis tradicional, aportando facilidades para restringir o delimitar los elementos que se desean traer de una base de datos o de cualquier otro tipo de repositorio, pero que al final se deben traducir a un lenguaje de bases de datos como el SQL (Zloof, 2001).

En un sistema interactivo de consulta-respuesta, una consulta es una expresión o fórmula bien formada del lenguaje, empleada para solicitar la derivación de una nueva relación con las tuplas que cumplen totalmente (con grado 1) las condiciones establecidas. De estas tuplas se especifican las propiedades que quieren visualizarse, sean éstas derivadas o no. Por esto, una consulta tiene tres componentes principales: la proyección, la lista de las clases de objetos o relaciones involucrados que especifican el producto cartesiano que se debe considerar y, opcionalmente, una serie de condiciones que restringen las tuplas en la relación resultante.

La proyección, restringe las propiedades de la nueva relación que se desean visualizar, extraídas o calculadas de todas las propiedades definidas para los objetos mencionados en la consulta. Por esto, es una lista de expresiones que pueden construirse con los atributos, las funciones, las variables o las constantes predefinidas en una base de datos, en la cláusula SELECT de la consulta. La

restricción del Algebra Relacional, que es la operación que permite filtrar tuplas indeseadas en la relación resultante, se especifica en la cláusula WHERE mediante un conjunto de expresiones lógicas que se concatenan con las conectivas lógicas AND y OR. También se puede emplear el operador de la negación, NOT, dentro de las expresiones de la restricción. Además de estas componentes, una consulta puede incluir una cláusula GROUP BY para especificar agrupamientos o que se muestren las tuplas resultantes en un orden determinado, con la cláusula ORDER BY.

La sintaxis general de una consulta, en SQL, considerado el lenguaje estándar de consultas a bases de datos relacionales, tiene la forma:

```
SELECT proyección
FROM lista de tablas o vistas del producto cartesiano
[WHERE lista de condiciones]
[ORDER BY lista de propiedades en la relación resultante]
[GROUP BY lista de propiedades en la relación resultante
  [HAVING lista de condiciones de grupo]]
```

En los criterios de la consulta se utilizan operadores de comparación, como la igualdad (representada con el símbolo "=") o la diferencia (<>), aplicables a los distintos tipos de atributos. También se puede utilizar el operador IS para verificar si un valor es nulo (desconocido o faltante). Por ejemplo, la sentencia SQL para mostrar el nombre de los empleados del departamento de sistemas, con una estatura entre 1.70 y 2 metros, inclusive, es:

```
SELECT nombre
  FROM empleado, departamento
  WHERE empleado.id_depto = departamento.id AND
        departamento.nombre = "Sistemas" AND
        empleado.estatura BETWEEN 1.70 AND 2
```

Como se expresó recientemente, en los sistemas flexibles de consulta-respuesta, el usuario no tiene por qué conocer detalles del lenguaje SQL puesto que la interacción humano-máquina se realiza a través de un interfaz gráfica, donde el usuario va llenando o seleccionando los datos para formar sus consultas. Pero el acceso a un sistema de bases de datos no puede restringirse a este tipo de interfaz, ya que es posible que algunos programas o procedimientos que se deban correr en lote (o fuera de línea) también requieran de la interpretación y la operación con términos vagos por lo que debe extenderse el lenguaje SQL y no sólo el lenguaje de interfaz. Tomando el ejemplo de la consulta anteriormente dado, con esta Tesis Doctoral se espera obtener una respuesta confiable cuando se plantea así:

```
SELECT nombre
  FROM empleado, departamento
  WHERE empleado.id_depto = departamento.id AND
        departamento.nombre= "Sistemas" AND
        empleado.estatura IS "alta"
```

# Capítulo 3

## 3 Representación y Manejo de la Vaguedad

Tanto en la vida cotidiana como en el trabajo investigativo, hacemos afirmaciones que no son verdades absolutas. Tal como se señala en (Berkran y Trubatch, 1997): “La construcción de la ciencia y de los sistemas operativos ha evolucionado del idealismo de las matemáticas exactas, ya que a veces se quedan cortas manejando y representando la realidad de la vida”.

Convencionalmente, el diseño de las técnicas aplicables en la Ingeniería se ha basado en teorías ideales de la Matemática, la Física, la Biología o la Química. De igual modo, la Computación, que está basada en una lógica bivaluada, sólo es concebida para reconocer ceros y unos, lo que es compatible únicamente con la dualidad cierto/falso, sin admitir otros valores de verdad. Debido a esto, más recientemente se han propuesto técnicas alternativas para admitir lógicas multivaluadas que permitan representar conceptos abstractos vagos y operar con ellos.

Para encontrar la semántica de cada término vago en una consulta que más se ajuste su contexto lingüístico, se requiere de un proceso de discriminación o clasificación borrosa de los objetos, en tiempo de ejecución.

En el presente Capítulo se inicia con la descripción de la incertidumbre originada por la vaguedad, para luego proseguir con la descripción de las técnicas de discriminación consideradas para su aplicación directa, o su extensión, en la resolución del problema de la interpretación de la vaguedad de las consultas a bases de datos objeto-relacionales, motivo de este trabajo de investigación.

### 3.1 *La Incertidumbre Originada por la Vaguedad*

Cuando el ser humano trata de describir los fenómenos naturales o hacer inferencias a partir de su estudio, se ve enfrentado a diversos tipos de incertidumbre originados por la información imperfecta o por los mecanismos usados para hacer inferencias.

Las técnicas e instrumentos utilizados en la recolección de información no son perfectas, de ahí que se tengan datos imprecisos, inexactos o errados. Unas veces, los datos son imprecisos porque son tomados con equipos que tienen una limitada gradación en las unidades de medición o por las lecturas deficientes hechas por los

encargados de las mediciones. Otras veces, es la naturaleza misma del fenómeno estudiado la que genera lecturas fluctuantes o variables, convirtiéndose en otra fuente de imprecisión en los datos.

Tradicionalmente, en la investigación científica experimental, se han utilizado modelos probabilistas para la representación y manejo de la incertidumbre. Un modelo probabilista incorpora el concepto de aleatoriedad como una componente inexplicable de la variación en el comportamiento de un fenómeno dado. Por eso, una variable aleatoria representa un fenómeno que no es predecible exactamente, sino con algún grado de confianza. La producción de oro de un país, la humedad, la temperatura de una región o el rendimiento académico de un estudiante son ejemplos de variables aleatorias, pues a pesar de tenerse un buen registro de su comportamiento histórico que permite conocer parcialmente la característica estudiada, no podemos determinar otros nuevos valores sin tener alguna probabilidad de equivocarnos en nuestras afirmaciones o predicciones.

La incertidumbre, también es originada por los modelos usados y esto ocurre aún en el caso ideal de tener datos libres de imprecisiones. Generalmente los modelos son meras aproximaciones o simplificaciones de porciones de la realidad y por eso no podemos estar seguros, de manera absoluta, de las inferencias que realicemos con ellos. Además, si dichos modelos son expresados mediante un lenguaje simbólico con algunas componentes en lenguaje natural, la vaguedad puede estar presente, lo que puede conducir a múltiples interpretaciones de los términos empleados, generando de esta manera otro tipo de incertidumbre.

Hasta mediados de los años sesenta del siglo pasado, no se disponían de técnicas para la representación de la incertidumbre originada por la vaguedad de la información o del lenguaje. Con el trabajo investigativo de Zadeh, surge una corriente para el tratamiento de la vaguedad aplicando los conceptos y técnicas basadas en Lógica Difusa (Zadeh, 1975). Luego aparece otra técnica alternativa: la Teoría de los Conjuntos Rugosos (Pawlack, 1982). Técnicas conocidas como de "Razonamiento Aproximado" (Approximate Reasoning, en inglés) que son usadas para representar y operar con datos cuantitativos que pueden ser transformados en etiquetas que representan categorías ordinales que denominó variables lingüísticas.

### **3.1.1 Variables Lingüísticas**

Puesto que las palabras son menos precisas que los números, el concepto de variable lingüística sirve para ofrecer una caracterización aproximada de los fenómenos complejos o mal definidos, en concordancia con su descripción en términos cuantitativos convencionales. La utilidad de las variables lingüísticas radica en que permiten resumir la información y por lo tanto reducir la complejidad (Zadeh, 1975). Para el resumen de la información, es usual emplear una técnica de partición o granulación borrosa (fuzzy granulation, en inglés) para determinar la colección de términos o etiquetas posibles  $\{E_1, E_2, \dots, E_m\}$ , asignables a una variable lingüística.

La definición formal de Zadeh establece que una variable lingüística se caracteriza por la quintupla  $(X, T(X), U, G, S)$ , en la cual  $X$  es el nombre de la variable en cuestión,  $T(X)$  es el conjunto de términos o valores lingüísticos que toma la variable  $X$  en el *Universo de Discurso*  $U$ ,  $G$  son las reglas sintácticas que generan nuevos términos de  $T(X)$  y  $S$  son las reglas semánticas que asocian cada valor o etiqueta lingüística  $E$  con su significado  $S(E)$ , por medio con un conjunto difuso en  $U$ , entre todos los posibles (Zadeh, 1965). Usualmente, una regla sintáctica toma la forma de una gramática para generar los nombres de los valores que toma la variable  $X$  y la regla semántica  $S$  define un procedimiento algorítmico para la cuantificación del significado de cada valor.

La semántica de las etiquetas lingüísticas se puede representar usando la Teoría de los Conjuntos Difusos, como lo propuso Zadeh desde sus trabajos iniciales. Pero también se podrían representar usando la Teoría de los Conjuntos Rugosos propuesta en (Pawlak, 1982) que también ha sido base de otras propuestas en Computación Granular (Zhang, Liang y Liu, 2004; Yao, 1999). Por esto, a continuación se describirán las propiedades de la teoría clásica de conjuntos, para luego describir las dos teorías de conjuntos recién mencionadas. Todo ello con el objeto de justificar la elección tomada para la representación de la vaguedad en este trabajo de investigación.

### 3.1.2 Teorías de Conjuntos

**Tabla 2. Propiedades del Álgebra Booleana**

Idempotencia	$A \cap A = A$ $A \cup A = A$
Identidad	$A \cap U = A$ $A \cap \phi = \phi$ $A \cup \phi = A$ $A \cup U = U$
Commutatividad	$A \cap B = B \cap A$ $A \cup B = B \cup A$
Asociatividad	$A \cap (B \cap C) = (A \cap B) \cap C$ $A \cup (B \cup C) = (A \cup B) \cup C$
Ley Distributiva	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
Ley de Absorción	$A \cap (A \cup B) = A$ $A \cup (A \cap B) = A$
Ley de no contradicción	$A \cap A^c = \phi$
Medio Excluido	$A \cup A^c = U$
Involución	$(A^c)^c = A$
Leyes de De Morgan	$(A \cap B)^c = A^c \cup B^c$ $(A \cup B)^c = A^c \cap B^c$

En la teoría clásica de conjuntos, dado un conjunto  $U$  no vacío, el conjunto de todos sus subconjuntos  $\mathcal{P}(U)$ , junto con las operaciones para la unión ( $\cup$ ), la intersección  $\cap$  y el complemento ( $^c$ ) forman la estructura  $(\mathcal{P}(U), \cup, \cap, ^c)$  que cumple con las propiedades básicas de un álgebra de Boole que se presentan en la Tabla 2. Dichas propiedades son válidas para conjuntos cualesquiera  $A, B$  y  $C \in \mathcal{P}(U)$ .

En la teoría de conjuntos clásica, la pertenencia o no de un elemento  $x$  de  $U$  al conjunto  $A$ , puede expresarse numéricamente mediante la función característica:

$$A: x \rightarrow \{0,1\}, \text{ donde } A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$$

Las nuevas teorías de conjuntos generalizan la teoría clásica, mediante la transformación de la función característica  $A: x \rightarrow \{0,1\}$  de cada conjunto en el conjunto potencia. La función de pertenencia se extiende a un mapeo  $A: x \rightarrow [0,1]$  o  $A: x \rightarrow \{0,1, \text{Indeterminado}\}$  en el caso de la Teoría de Conjuntos Difusos y en la Teoría de los Conjuntos Rugosos, respectivamente. Ambas teorías de conjuntos se pueden considerar generalizaciones de la clásica porque al aplicarse una operación definida en ellos, sobre conjuntos nítidos o concretos, proporcionan los mismos resultados que cuando se aplican las operaciones del álgebra booleana. Esta restricción es llamada, por algunos autores, como el “Principio de Preservación de lo Clásico” (Pradera et al., 2006).

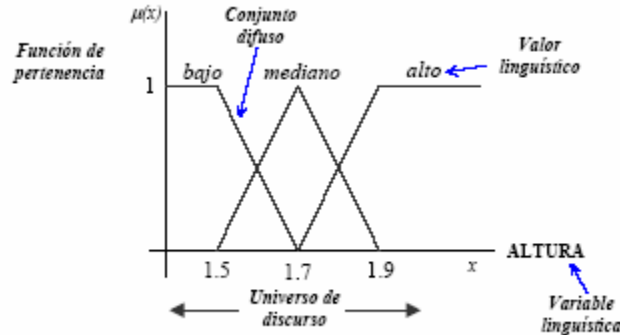
Debido al relajamiento que se pretende con las lógicas multivaluadas, varias propiedades de la teoría clásica como las leyes de no contradicción, la ley del medio excluido, la involución, la ley distributiva, la idempotencia o las leyes de De Morgan, pueden dejar de ser válidas en las nuevas teorías de conjuntos (Galindo, 2001).

### 3.1.2.1 Teoría de Conjuntos Difusos

Esta teoría se desarrolla a partir del concepto de *conjunto*, que se define como una colección de objetos de cualquier tipo denominados elementos o miembros del conjunto, del mismo modo que en la clásica. En ambas teorías se denomina *Universo de Discurso*, al dominio o al tipo de los objetos contenidos en el conjunto. De igual modo, un conjunto difuso se puede denotar por extensión, declarando explícitamente cuales son los miembros del conjunto o por comprensión. Sin embargo, lo más corriente será declarar un conjunto difuso por comprensión, definiendo las propiedades o las restricciones impuestas a los elementos del dominio para poderlos considerar miembros del conjunto.

Según (Parsons, 1996) la Teoría de los Conjuntos Difusos es una generalización de la Teoría de Conjuntos clásica realizada con el fin de representar aquellas clases de objetos que no permiten definir criterios precisos para determinar la membresía

de sus elementos. Una forma de resolver este problema consiste en asignar un grado de pertenencia a cada objeto de un dominio para indicar cuánto pertenece a un conjunto dado. Los grados pueden ser combinados de modo que es posible calcular el grado de pertenencia de un elemento a un conjunto derivado de una composición lógica de otros conjuntos. Con fines ilustrativos, en la Figura 1 se muestran los conceptos básicos de dicha teoría.



**Figura 1. Conceptos Básicos de la Teoría de Conjuntos Difusos**

En la literatura se observan algunas variaciones sobre lo que se puede considerar una Teoría de Conjuntos Difusos y todas sus componentes. Por lo tanto, ahora se presentan todas las definiciones que se adoptan en este trabajo de investigación.

**Definición 1.** (Teoría de Conjuntos Difusos). Sea  $\mathcal{P}(U) = \{A; A: U \rightarrow M\}$  el conjunto de todos los conjuntos difusos sobre  $U$ , con grados de pertenencia en el conjunto ordenado  $(M, \leq)$ , equipado con un orden parcial  $\prec$  y con tres operaciones  $\cap, \cup: \mathcal{P}(U) \times \mathcal{P}(U) \rightarrow \mathcal{P}(U), ^c: \mathcal{P}(U) \rightarrow \mathcal{P}(U)$  que representan la intersección, la unión y el complemento de un conjunto, respectivamente. Entonces, la estructura algebraica  $(\mathcal{P}(U), \prec, \cap, \cup, ^c)$  se llamará una Teoría de Conjuntos Difusos cuando se satisfagan las propiedades siguientes, que buscan expresar el “Principio Básico de Preservación de lo Clásico” (Pradera et al., 2006):

- i.  $(M, \leq)$  es acotado. Es decir, al menos contiene un elemento mínimo  $0_M$  y un elemento máximo  $1_M$ .
- ii. Para los conjuntos difusos  $A, B \in \mathcal{P}(U)$ ,  $A \prec B \leftrightarrow A \subseteq B$ .
- iii. Para los conjuntos difusos  $A, B \in \mathcal{P}(U)$ ,

$A \cap B = A \hat{\cap} B, A \cup B = A \hat{\cup} B, A^c = A \hat{c}$ , donde  $\hat{\cap}, \hat{\cup}, \hat{c}$  representan los operadores clásicos de la intersección, la unión y de la negación.

**Definición 2.** (Conjunto difuso). Un conjunto es una generalización de un conjunto nítido o concreto. Sea  $U$  un dominio y sea  $x$  un valor particular cualquiera en ese

dominio. Un conjunto borroso  $A$  está definido como un conjunto de pares ordenados:

$$A = \{(x, \mu_A(x)) \mid x \in U\}$$

Donde la *medida borrosa*  $\mu_A(x)$  es llamada el grado de pertenencia de  $x$  al conjunto  $A$ . Según (Jang, Sun y Mizutani, 1997), el grado de pertenencia de un elemento a un conjunto borroso no tiene un significado absoluto sino relativo, pues toma su significado de la comparación con los grados de pertenencia de otros elementos  $u \in U$ . Por ejemplo, si  $\mu_A(x) > \mu_A(y)$  se está afirmando que  $x$  pertenece con mayor grado al conjunto borroso  $A$  que  $y$ . Si  $A$  representa un concepto vago, entonces se dice que  $x$  se parece más a ese concepto que  $y$ .

El conjunto vacío y el universal en la Teoría de Conjuntos Difusos, se definen mediante los axiomas que aparecen, a continuación.

$$\mu_\emptyset(x) \equiv 0 \quad \forall x \in U \quad (\text{Conjunto vacío})$$

$$\mu_U(x) \equiv 1 \quad \forall x \in U \quad (\text{Conjunto universal})$$

**Definición 3.** (Teoría Estándar de Conjuntos Difusos). Una Teoría  $(\mathcal{P}(U), \prec, \cap, \cup, ^c)$  es estándar si cumple las condiciones (Martín del Brio y Sanz Molina, 2002):

i.  $M = [0, 1]$ .

ii. Para los conjuntos difusos  $A, B \in \mathcal{P}(U)$ ,

$$A \prec B \leftrightarrow \mu_A(x) \leq \mu_B(x) \quad \forall x \in U .$$

iii. Para los conjuntos difusos  $A, B \in \mathcal{P}(U)$ , los operadores  $(\cap, \cup, ^c)$  se definen como:

$$(A \cap B)(x) = \mu_{A \cap B}(x) = T(\mu_A(x), \mu_B(x))$$

$$(A \cup B)(x) = \mu_{A \cup B}(x) = S(\mu_A(x), \mu_B(x))$$

$$A^c(x) = \mu_{A^c}(x) = No(\mu_A(x))$$

Para cualquier  $x \in U$ , donde  $T : [0,1] \times [0,1] \rightarrow [0,1]$  es una *norma triangular continua*,  $S : [0,1] \times [0,1] \rightarrow [0,1]$  es una *conorma triangular continua* y la función  $No : [0,1] \rightarrow [0,1]$  es una *negación fuerte*<sup>3</sup>.

**Definición 4.** (Medida borrosa). En términos formales una medida borrosa  $f$ , se define como una función evaluada sobre conjuntos, tal que:

$$f: \mathcal{P}(U) \rightarrow [0,1]$$

---

<sup>3</sup> Estos nuevos términos se definen un poco más adelante.

Donde,  $\mathcal{P}(U)$  es el conjunto cuyos elementos son todos los subconjuntos de  $U$ , también conocido como el conjunto potencia de  $U$ .

Para considerar una función  $f$  como una medida borrosa, debe cumplir los axiomas siguientes:

**Axioma 1.** (condiciones de borde):  $f(\emptyset) = 0$  y  $f(U) = 1$

**Axioma 2.** (monotonía):  $\forall A, B$  subconjuntos de  $U$ , si  $A \subseteq B$ , entonces  $f(A) \leq f(B)$

**Axioma 3.** (continuidad) : Para cada secuencia  $A_i$  de subconjuntos de  $U$ ,

$$\text{Si } A_1 \subseteq A_2 \subseteq \dots \text{ o } A_1 \supseteq A_2 \supseteq \dots \text{ entonces } \lim_{i \rightarrow \infty} (f(A_i)) = f(\lim_{i \rightarrow \infty} A_i)$$

De acuerdo con la definición dada a una medida borrosa y a los axiomas que deben cumplir, se deben resaltar dos tipos especiales de medidas borrosas como son las medidas de la *probabilidad* y de la *posibilidad*. Los modelos usados en Estadística se han basado en el concepto de probabilidad para representar las variables aleatorias, mientras que los modelos difusos, se basan en el concepto de posibilidad para representar el grado de pertenencia a un conjunto difuso o compatibilidad de un objeto con el conjunto difuso.

La definición clásica de probabilidad de un evento consiste en la razón dada por el número de casos favorables al evento y el número de casos posibles en un dominio dado. Se considera como evento a un subconjunto factible del espacio muestral o *Universo de Discurso*. Para determinar un valor numérico que represente adecuadamente la probabilidad, es corriente considerar el comportamiento pasado del fenómeno bajo estudio, por medio de las frecuencias de ocurrencias observadas de los eventos. Esta es una forma natural de razonar de los seres humanos: si un fenómeno se ha comportado de cierta manera en el pasado, entonces si no ocurre algo anómalo o atípico, se esperaría que se comportara en forma parecida en el futuro cercano (Kallenberg, 2002). Esta técnica se basa en una definición de la probabilidad de ocurrencia del evento  $A$  como:

$$P(A) = \lim_{n \rightarrow \infty} \frac{na}{n}$$

Donde  $na$  es el número de veces que ocurrió el evento  $A$  en  $n$  observaciones o experimentos realizados. La medida  $p$  (de probabilidad) es una función que asigna a cada evento  $A$  un valor real no negativo  $p(A)$  en el intervalo  $[0,1]$ . La función de densidad de probabilidad de una variable aleatoria, es una función a partir de la cual se obtiene la probabilidad de ocurrencia de cada evento posible en el dominio considerado. Su integral, en el caso de variables aleatorias continuas, es la distribución de probabilidad. En el caso de variables aleatorias discretas la distribución de probabilidad se obtiene a través de la sumatoria de la función de densidad.

En Lógica Difusa, se propone la medida borrosa de *posibilidad* para representar nuestra percepción del grado de pertenencia o de encajamiento, de cada

elemento, a un conjunto borroso en  $U$ . Para cualquier conjunto  $A$  en el *Universo de Discurso*,  $\pi(A)$  denota el grado en que  $A$  se considera "posible". Un evento tiene un grado de posibilidad cero cuando es imposible y tiene un grado de uno cuando es totalmente posible.

Formalmente, para un *Universo de Discurso*  $U$ , una medida de posibilidad puede definirse a partir de la función de distribución de posibilidad  $\pi : U \rightarrow [0,1]$  de forma tal que para cualquier conjunto  $A$  definido sobre  $U$  y para cualquier  $x$  elemento de  $U$ , se cumple lo siguiente:

$$\Pi(A) = \max(\pi(x) \exists x \in A)$$

$$\Pi(\emptyset) = 0$$

$$\Pi(U) = 1$$

$$\Pi\left(\bigcup_{i=1}^n A_i\right) = \sup_{i=1, n} \Pi(A_i)$$

$$\Pi(A \cap B) \leq \min(\Pi(A), \Pi(B))$$

Con la última propiedad, se está admitiendo que puede ser imposible la ocurrencia simultánea de dos eventos aunque ellos sean posibles en forma separada.

**Definición 5.** (Soporte). El soporte del conjunto difuso  $A$  es el conjunto concreto o nítido que contiene todos los elementos del universo  $U$  con un grado de pertenencia diferente de cero:

$$\text{Soporte}(A) = \{x \in U / \mu_A(x) > 0\}$$

Si el soporte de  $A$  consiste de un punto solamente, es llamado conjunto unitario (*singleton*). Si el grado de pertenencia de este punto es uno,  $A$  es denominado un *conjunto unitario concreto* (Zimmermann, 1993).

**Definición 6.** (Alfa -corte). El alfa-corte de un conjunto difuso  $A$  sobre  $X$  se define como:

$$\alpha\text{-corte}(A) = \{x \in U / \mu_A(x) \geq \alpha\}$$

Un alfa-corte es, por definición, una generalización del soporte. Un alfa-corte es *estricto*, si se cambia la desigualdad  $\geq$  por  $>$ .

**Definición 7.** (Núcleo) El núcleo o corazón de un conjunto difuso  $A$  es definido por los elementos cuyo grado de pertenencia es total:

$$\text{Núcleo}(A) = \{x \in U / \mu_A(x) = 1\}$$

**Definición 8.** (Altura). La altura de un conjunto difuso  $A$  sobre  $X$  se define como:

$$\text{Altura}(A) = \sup_{x \in X} \mu_A(x)$$

**Definición 9.** (Normalidad). El conjunto  $A$  es llamado *normal* si su altura es 1, y *subnormal* si es menor que este valor. Un conjunto difuso subnormal se puede normalizar, dividiendo su función de pertenencia por  $\sup_{x \in X} \mu_A(x)$ .

$$\text{Normal}(A) \leftrightarrow \text{Altura}(A) = 1$$

**Definición 10.** (Conjunto convexo). La convexidad de un conjunto difuso significa que los grados de pertenencia de los elementos en un intervalo cualquiera, no son menores que los dos valores de pertenencia de los extremos del intervalo considerado. Más formalmente,

$$\forall a, b, x \in U \text{ y } A \in P(U) : a \leq x \leq b \rightarrow \mu_A(x) \geq \min\{\mu_A(a), \mu_A(b)\}$$

La convexidad es *estricta*, si se usa la desigualdad  $>$  en lugar de  $\geq$ .

**Definición 11.** (Número difuso). Es un conjunto difuso que representa una cantidad o un valor escalar. Es un conjunto convexo, normal donde la función de pertenencia es continua y el núcleo consiste de un sólo valor que es precisamente el escalar que representa.

**Definición 12.** (Cardinalidad de un conjunto difuso). El *tamaño*, o *cardinalidad*, de un conjunto difuso  $A$  en un universo  $U$  es la suma, sobre los elementos del universo, de los grados de pertenencia a  $A$  (Morales, 2000):

$$\text{Cardinalidad}(A) = |A| = \sum_{x \in U} \mu_A(x)$$

Por lo tanto, el peso relativo de cada objeto  $x$  del universo  $U$ , en el conjunto difuso  $A$ , es:

$$p_A(x) = \frac{\mu_A(x)}{|A|}$$

**Definición 13.** (Marco de cognición). Un marco de cognición es definido como una colección finita de conjuntos difusos sobre el mismo *Universo de Discurso*, en la cual es conveniente dar un orden (Ios et al., 1998). Por lo tanto, un marco de cognición puede ser formalizado con la tripleta:

$$\text{Marco} = \langle U, \mathcal{F}, \prec \rangle$$

Donde,  $\mathcal{F} = \{A_i; i \in I\}$  es una familia de conjuntos difusos con un orden parcial sobre  $\mathcal{F}$ , inducido por los números reales. Esto es,

$$A_i \prec A_j \leftrightarrow \mu_{A_i}(x) \leq \mu_{A_j}(x), \forall x \in U$$

Por simplicidad se puede asumir que si  $i \leq j \leftrightarrow A_i \prec A_j$ .

**Definición 14.** (Punto de cruce). El elemento  $x$  de  $U$ , se llama punto de cruce si el grado de pertenencia a dos conjuntos adyacentes en el marco, es 0.5:

*Punto de cruce*  $(x) \rightarrow \mu_A(x) = 0.5 \cap \mu_B(x) = 0.5 \exists A, B \in \mathcal{F} / A \cap B \neq \emptyset$

**Definición 15.** (Partición difusa). Una familia de conjuntos difusos  $\mathcal{F}: \{A_i : i \in I\}$  se llama una partición difusa si y sólo si, se cumplen las dos propiedades siguientes.

1.  $A_i \neq \emptyset \forall i \in I$

2.  $\bigcup_{i \in I} A_i = U$

La única diferencia de la partición difusa con la denominada partición matemática es que esta última exige que se cumpla que la intersección entre cualquier par de conjuntos de  $\mathcal{F}$ , sea vacía. Es decir, que los conjuntos sean mutuamente excluyentes. Por esto, se puede decir que las teorías de conjuntos difusos, también generalizan el concepto clásico de partición.

**Definición 16.** (Partición consistente). Una familia de conjuntos difusos  $\mathcal{F}: \{A_i : i \in I\}$  forma una partición consistente si  $\exists x_0 \in U$  tal que  $\mu_{A_i}(x_0) = 1 \rightarrow \mu_{A_j}(x_0) = 0 \forall A_j \in \mathcal{F}$ .

### 3.1.3 Operaciones de Conjuntos Difusos

Del mismo modo que en la teoría clásica de conjuntos, se han propuesto una serie de operadores para realizar operaciones algebraicas sobre conjuntos difusos que permitan la inferencia deductiva de otros o la comparación de dos o más conjuntos difusos que representan conceptos vagos.

En la teoría clásica de conjuntos, los elementos de  $\mathcal{P}(U)$  pueden ser diferenciados por sus funciones características  $A : x \rightarrow \{0,1\}$ . Así mismo, ocurre los conjuntos difusos, pero considerando la función característica  $\mu : x \rightarrow [0,1]$ . Por lo tanto, la estructura  $(\mathcal{P}(U), \cap, \cup, ^c)$  es isomórfica con la estructura algebraica  $([0,1], \wedge, \vee, \neg)$  (Pradera et al., 2006). Esto significa que las operaciones de conjuntos pueden definirse por medio de operaciones sobre los grados de pertenencia. De estos grados se infiere un nuevo concepto vago representado con un conjunto difuso.

**Definición 18.** (Igualdad). Dos conjuntos difusos son iguales, si solo si, el grado de pertenencia a los dos conjuntos es el mismo, para todo elemento del universo  $U$ :

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

**Definición 19.** (Subconjunto). Un conjunto difuso  $A$  es subconjunto de  $B$ , si para cualquier elemento de  $U$  su grado de pertenencia es menor para el conjunto  $A$  que para el  $B$ :

$$A \subset B \Leftrightarrow \mu_A(x) \leq \mu_B(x) \forall x \in U$$

**Definición 20.** (Complemento). Dado un conjunto difuso  $A$ , su complemento se define como el conjunto  $A^c$  cuya función de pertenencia viene dada por:

$$A^c(x) = \mu_{A^c}(x) = C(\mu_A(x)) \forall x \in U$$

Donde la función  $C: [0,1] \rightarrow [0,1]$  debe cumplir, al menos, las dos primeras propiedades:

**Condiciones de borde.**  $C(0) = 1 \wedge C(1) = 0$

**No creciente.**  $\mu_A(x) \leq \mu_B(x) \Rightarrow C(\mu_A(x)) \geq C(\mu_B(x))$

**Involución.**  $C(C(\mu_A(x))) = \mu_A(x)$

**Definición 21.** (Intersección). En la teoría estándar de conjuntos difusos, la intersección se define como:

$$(A \cap B)(x) = \mu_{A \cap B}(x) = T(\mu_A(x), \mu_B(x))$$

Donde la función de transformación  $T: [0,1] \times [0,1] \rightarrow [0,1]$  es una norma triangular. Esto es  $(T, [0,1])$  tiene una estructura que debe cumplir las siguientes propiedades o leyes (Klement, Mesiar y Pap, 2000):

**Conmutativa.**  $T(\mu_A(x), \mu_B(x)) = T(\mu_B(x), \mu_A(x))$

**Asociativa.**  $T(\mu_A(x), T(\mu_B(x), \mu_C(x))) = T(T(\mu_A(x), \mu_B(x)), \mu_C(x))$

**Monotonía.** Si  $\mu_A(x) \leq \mu_C(x) \wedge \mu_B(x) \leq \mu_D(x) \Rightarrow T(\mu_A(x), \mu_B(x)) \leq T(\mu_C(x), \mu_D(x))$

**Elemento Identidad.**  $T(\mu_A(x), 1) = \mu_A(x)$

En ocasiones las normas triangulares se suelen restringir con restricciones adicionales como la *continuidad* y la *subidempotencia*. La restricción de continuidad, exigida por las teorías difusas estándares, previene que un pequeño cambio en el grado de pertenencia individual a uno de los conjuntos difusos  $A$  o  $B$ , involucrados en la operación de intersección, produzca un cambio grande (discontinuo) en el grado de pertenencia a la intersección  $A \cap B$ . La subidempotencia expresa que el grado de pertenencia a la intersección de un conjunto consigo mismo, no puede ser mayor que el grado de pertenencia al conjunto. Esto es,  $T(\mu_A(x), \mu_A(x)) \leq \mu_A(x)$  (Fodor, 2004).

**Definición 22.** (Unión). En la teoría estándar de conjuntos difusos, la unión de conjuntos difusos se define como:

$$(A \cup B)(x) = \mu_{A \cup B}(x) = S(\mu_A(x), \mu_B(x))$$

Donde el operador  $S$ , es una "s-conorma". La s-conorma (o unión generalizada) es una función  $S: [0,1] \times [0,1] \rightarrow [0,1]$  que debe cumplir las siguientes propiedades o leyes (Klement, Mesiar, y Pap, 2000):

**Conmutativa.**  $S(\mu_A(x), \mu_B(x)) = S(\mu_B(x), \mu_A(x))$

**Asociativa.**  $S(\mu_A(x), S(\mu_B(x), \mu_C(x))) = S(S(\mu_A(x), \mu_B(x)), \mu_C(x))$

**Monotonía.** Si  $\mu_A(x) \leq \mu_C(x) \wedge \mu_B(x) \leq \mu_D(x) \Rightarrow S(\mu_A(x), \mu_B(x)) \leq S(\mu_C(x), \mu_D(x))$

**Elemento Neutro.**  $S(\mu_A(x), 0) = \mu_A(x)$

Las normas triangulares también se suelen restringir con las restricciones adicionales de continuidad y superidempotencia:  $S(\mu_A(x), \mu_A(x)) \geq \mu_A(x)$ .

**Definición 23.** (Relación dual entre operaciones borrosas). Si los operadores para la intersección, la unión y el complemento cumplen las leyes de De Morgan generalizadas,  $C(T((x, y))) = S(C(x), C(y))$  y  $C(S((x, y))) = T(C(x), C(y))$ , entonces se dice que la t-norma y la s-conorma son *duales* con respecto al complemento difuso.

**Definición 24.** (La Agregación). A la hora formar expresiones complejas, también debe mencionarse otra de las conectivas lógicas debe proporcionar un lenguaje de consulta: la agregación. Este operador se considera un operador más general que permite definir un conjunto borroso a partir de otros. Según Klir y Yuan (1995), para que una función  $+: [0, 1]^p \rightarrow [0, 1]$  sea considerada una operación de agregación sobre  $p$  conjuntos borrosos, debe satisfacer, por lo menos, las tres primeras propiedades siguientes:

i) **Condiciones de borde:**

$$+(0, 0, \dots, 0) = 0 + 0 + \dots + 0 = 0$$

$$+(1, 1, \dots, 1) = 1 + 1 + \dots + 1 = 1$$

ii) **Monotonía:**

$$\mu(x_i) \leq \mu(y_i) \forall i = 1, p \Rightarrow \mu(x_1) + \mu(x_2) \dots + \mu(x_p) \leq \mu(y_1) + \mu(y_2) \dots + \mu(y_p)$$

iii)  $+(\mu(x_1), \mu(x_2), \dots, \mu(x_p))$  es continua en  $[0, 1]^p$

iv)  $+(\mu(x_1), \mu(x_2), \dots, \mu(x_p))$  es simétrica, en todos sus argumentos.

v) **Idempotencia:**

$$+(\mu(x_i), \mu(x_i), \dots, \mu(x_i)) = \mu(x_i)$$

Cuando un operador de la agregación cumple con las tres primeras condiciones de las recién enunciadas, Klir y Yuan, lo consideran un operador *esencial* de la agregación.

En (Grabish, Orlovski y Yager, 1998) se definen otras propiedades básicas que debe cumplir un operador de la agregación. Éstas son:

vi) **Neutralidad:**

$$+(\mu(x_1), \mu(x_2), \dots, \mu(x_p)) = +(\mu(\varphi(x_1)), \mu(\varphi(x_2)), \dots, \mu(\varphi(x_p)))$$

Donde  $\varphi$  es una transformación isomórfica. Esto es, una función  $\varphi: [0, 1] \rightarrow [0, 1]$  monótona creciente que cumple las condiciones de borde. Algunas

transformaciones isomórficas son  $\varphi(a) = \frac{2a}{a+1}$  o  $a^\lambda$ ,  $\lambda > 0$ .

- vii) **Compensación:**  
 $\otimes(\mu(x_1), \mu(x_2), \dots, \mu(x_p)) \leq +(\mu(x_1), \mu(x_2), \dots, \mu(x_p)) \leq \oplus(\mu(x_1), \mu(x_2), \dots, \mu(x_p))$
- viii) **Asociatividad:**  
 $+(\mu(x_1), \mu(x_2), \dots, \mu(x_p)) = +(+(\mu(x_i), \mu(x_j)), \dots, \mu(x_p))$
- ix) **Desagregación:**  
 $Si +(\mu(x_1), \mu(x_2)) = \mu'(x_s) \Rightarrow +(\mu(x_1), \mu(x_2), \dots, \mu(x_p)) = +(\mu'(x_s), \mu'(x_s), \dots, \mu(x_p))$

### 3.1.3.1 Teoría de Conjuntos Rugosos

La Teoría de Conjuntos Rugosos fue concebida por Zdzislaw Pawlak en los años 70 (Pawlak, 1970). Las nociones básicas de esta teoría son las nociones de aproximación inferior y superior de un conjunto. La aproximación inferior determina los objetos del dominio que se sabe, con certeza, pertenecen al subconjunto de interés, mientras que la aproximación superior permite una descripción de los objetos que, posiblemente, pertenecen al subconjunto.

En el marco de esta teoría, se define una base de conocimientos como un par  $(U, R)$ , siendo  $U$  un conjunto finito no vacío y  $R$  una familia de relaciones de equivalencia sobre  $U$ . Una relación binaria es una relación de equivalencia  $\approx$ , si cumple con las propiedades siguientes.

1. Es reflexiva:  $\forall a \in U, a \approx a$
2. Es simétrica:  $\forall a, b \in U, a \approx b \Leftrightarrow b \approx a$
3. Es transitiva:  $\forall a, b, c \in U, a \approx b, b \approx c \Rightarrow a \approx c$

El término conjunto aproximado o rugoso, se refiere a que si bien un concepto no puede definirse exactamente, éste puede aproximarse mediante dos conjuntos exactos, denominados  $R$ -aproximación inferior y  $R$ -aproximación superior. El conjunto  $R$ -aproximación inferior de  $X$ , denotado por  $\underline{R}X$ , es el conjunto de objetos de  $U$  que con total certeza pueden ser clasificados como elementos de la clase o categoría  $X$  utilizando el conocimiento  $R$ . Formalmente,

$$\underline{R}X = \{x \in U : [x]_R \subseteq X\}, \text{ donde } [x]_R \text{ es una clase de equivalencia } R \text{ sobre } U.$$

La  $R$ -aproximación superior de  $X$ , denotada por  $\overline{R}X$ , es el conjunto de objetos del universo que posiblemente pueden ser clasificados como elementos de la categoría  $X$  utilizando el conocimiento  $R$ . Formalmente,

$$\overline{R}X = \{x \in U : [x]_R \cap X \neq \emptyset\}$$

De acuerdo con la aproximación inferior y la superior de un subconjunto  $X \subseteq U$ , una región de  $U$  puede subdividirse en tres regiones disjuntas denominadas: región

positiva o  $POS_R(X)$ , región negativa o  $NEG_R(X)$  y región frontera o  $BND_R(X)$ . La región positiva corresponde a la aproximación inferior del conjunto rugoso que representa los objetos de  $U$  que pueden clasificarse con plena confianza como miembros del conjunto  $X$ , usando el conocimiento  $R$ . La región negativa es el conjunto de objetos que no pertenecen al conjunto  $X$ , de acuerdo con el conocimiento derivado de la relación de equivalencia definida. Luego, la región negativa está formada por aquellos elementos de  $U$  que pertenecen al complemento de  $X$ . La región frontera de un conjunto rugoso corresponde al área indiscernible (o de no especificidad) de  $U$ , significando que ninguno de los objetos de dicha región pueden clasificarse, exactamente, en el conjunto  $X$  o en su complemento; al menos utilizando el conocimiento  $R$ . Por lo tanto, dichas regiones se definen así:

$$\begin{aligned} POSR(X) &= \underline{RX}, \\ NEGR(X) &= U - \overline{RX} \\ BND_R(X) &= \overline{RX} - \underline{RX} \end{aligned}$$

De acuerdo con lo anterior, los grados de pertenencia de un objeto a un conjunto definido en  $U$  serían tres: 1 si un objeto está en la región positiva, 0 si está en la negativa y NULL (indiscernible) cuando se encuentre en la región frontera. Se trata entonces de una lógica trivaluada, en lugar de la lógica difusa donde el grado de pertenencia de los objetos a un conjunto definido en  $U$  puede tomar infinitos valores en el intervalo  $[0, 1]$ . Por lo tanto, se puede afirmar que un conjunto rugoso es un conjunto difuso que presenta discontinuidades.

Habiendo ya presentado las Teorías de los Conjuntos Difusos y Rugosos que pueden emplearse para representar los términos vagos, se continúa con la especificación declarativa, de las reglas semánticas definidas en el lenguaje PRUF para poder operar con ellos.

### **3.1.4 Reglas de Interpretación del Lenguaje Teórico PRUF**

El lenguaje PRUF, como acrónimo de Possibilistic Relational Universal Fuzzy (Zadeh, 1978), fue propuesto para la representación y manejo de etiquetas o valores lingüísticos en un sistema de inferencia difusa cualquiera. Es un lenguaje de representación teórico y declarativo, pues no especifica la manera de implementarlo en un lenguaje de programación particular. En este lenguaje, los operandos (los términos vagos o las etiquetas lingüísticas) se representan mediante conjuntos no convencionales y los operadores se definen mediante reglas de inferencia.

Por lo anterior, Zadeh propuso una representación lingüística de los gránulos de información por medio de restricciones suaves (soft constraints, en inglés), sin importar si las técnicas para derivar los gránulos estuvieran basadas en la Teoría de Probabilidades, los Conjuntos Rugosos o en los Conjuntos Difusos (Zadeh, 1997).

El lenguaje PRUF ha sido la base teórica de la mayoría de las propuestas para la extensión del lenguaje de consulta a bases de datos relacionales (ver, por ejemplo, a Kacprzyk and Zadrozny, 2001; Gonçálves y Tineo, 2001; Galindo et al., 1998; Bosc y Pivert, 1995; Medina, Pons y Vila, 1994). Los operadores que admite el lenguaje PRUF caen dentro de las categorías siguientes.

#### **3.1.4.1 Adverbios de cantidad o intensidad.**

El papel de este tipo de adverbios, en un lenguaje, consiste en ampliar o acentuar el sentido de un adjetivo como cuando nos referimos a los productos “extremadamente pesados” o “muy costosos”. Para este tipo de adverbio, en PRUF, se define la regla de interpretación de tipo I, como se describe en seguida.

Si la consulta  $C$  tiene la forma  $C$  es  $mE_j$ , siendo  $E_j$  un término lingüístico primario y  $m$  un modificador lingüístico, entonces la compatibilidad de la  $i$ -ésima tupla con el término, estará determinada por medio de su función de pertenencia y de una función de acentuación.

#### **3.1.4.2 Composición de proposiciones.**

Se pueden componer proposiciones vagas usando las conectivas lógicas de la conjunción, de la disyunción y también la negación para formar proposiciones más complejas. En PRUF, se conoce como la regla de interpretación de tipo II. Dicha regla señala que si la consulta  $C$  tiene la forma  $C$  es  $mE_j$  y  $mE_k$ , entonces la similitud de la  $i$ -ésima tupla con ese término, estará dada por medio de  $\mu_{ij} \wedge \mu_{ik}$  seleccionando un operador apropiado para representar la conjunción. De esta manera, un sistema podrá dar respuestas a solicitudes como “Mostrar el nombre y el formato de los documentos que sean cortos y con bajo grado de dificultad sobre el tema  $X$ ”. Si la consulta  $C$  tiene la forma “ $C$  es  $mE_j$  o  $mE_k$ ”, entonces la similitud de la  $i$ -ésima tupla con ese término, estará dada por medio de  $\mu_{ij} \vee \mu_{ik}$  y seleccionando el operador apropiado para la disyunción. De esta forma un sistema puede dar respuestas a solicitudes como “Mostrar el nombre de los alumnos de alto rendimiento académico o que sean buenos deportistas”.

Si una consulta contiene un criterio que incluye la negación y cualquier combinación de los operadores antes descritos, la respuesta dependerá de las funciones de pertenencia a cada uno de los componentes del término agregado y al significado que se le dé a la negación.

#### **3.1.4.3 Cuantificadores Difusos.**

Los cuantificadores de este tipo, extienden los cuantificadores universales usuales para admitir términos como “casi todos”, “muchos”, “bastantes”, “pocos”. Se busca con ellos describir las clases de objetos de interés. Con este tipo de cuantificadores, los usuarios podrían formular preguntas corrientes como: “Mostrar

el nombre de los cursos que no fueron aprobados por la mayoría de los estudiantes, en este semestre". Los cuantificadores borrosos son interpretados como lo especifica la regla de tipo III en el lenguaje PRUF.

Dicha regla establece que la consulta  $C$  tiene la forma "QN son  $F$ ", donde  $Q$  es un cuantificador borroso,  $F$  es un subconjunto borroso de  $U$ , el Universo de Discurso, y  $N$  es un descriptor. Un descriptor puede ser el nombre de una clase concreta o de una etiqueta lingüística. Una forma de expresar un cuantificador difuso es definiendo una función de  $X$ , donde  $X$  es un atributo con dominio en el intervalo  $[0,1]$ .

#### **3.1.4.4 Calificadores difusos.**

La regla de traducción de tipo IV, en PRUF, es definida para una consulta  $C$  con la forma " $N$  es  $t$ ", donde  $t$  es un valor de verdad, una probabilidad o una posibilidad. De ahí que, las maneras principales de calificación de una proposición son (Ribeiro y Moreira, 2003):

- Calificación mediante valores de verdad. Por ejemplo: "Es muy cierto que Juan es glotón".
- Calificación mediante probabilidades. Por ejemplo: "Es poco probable que él padezca la enfermedad de Hodgkin".
- Calificación mediante posibilidades. "Si Juana come mucho es posible que sea gorda".

Takahashi señala que con las cuatro reglas de interpretación definidas en PRUF, se incluyen la mayoría de proposiciones encontradas en los lenguajes de consulta (Takahashi, 1995). Esto se evidencia en la literatura sobre lenguajes flexibles de consulta que se mencionó recientemente. Sin embargo, en esta investigación se pretende extender el lenguaje PRUF para cubrir la interpretación de los términos vagos que se derivan de una combinación lineal de otros términos, caso que también se considera bastante útil para los usuarios.

## **3.2 Técnicas de Razonamiento Aproximado**

El modelo ideal del razonamiento, sea éste humano o mecánico, es el razonamiento exacto, pero la imperfección de la información, las características del mundo real, las deficiencias de los modelos y la poca disponibilidad de información, o de tiempo, para la toma de decisiones hacen que el razonamiento sólo sea aproximado, en la mayoría de los casos.

La razón es un término para referirse a la facultad de la mente para representar conceptos abstractos y operar con ellos. El razonamiento es un procedimiento intelectual mediante el cual, partiendo de unos datos conocidos (las premisas o los

hechos) se llega por inferencia a otros datos desconocidos, pero que se derivan de aquellos y a los que llamamos la conclusión.

En un proceso de razonamiento para el descubrimiento de nuevo conocimiento, el hombre necesita del lenguaje para hilar sus ideas y de la lógica para validarlas. La lógica es esencial pues permite analizar la validez de los argumentos o las producciones intelectuales y llegar a determinar (o al menos estimar) el grado de verdad o falsedad de los razonamientos.

Existen tres maneras básicas de producir inferencias: la inducción, la deducción y la presunción (o abducción), pero los razonamientos mixtos también son muy plausibles (Peirce, 1992). El razonamiento inductivo no es un examen exhaustivo de todos los objetos de interés en un dominio determinado, debido a las características de los fenómenos, las restricciones físicas, tecnológicas o económicas que se tengan en un proceso de descubrimiento de nuevo conocimiento determinado. En el razonamiento inductivo, a partir de un subconjunto de instancias particulares se contrastan hipótesis, que permiten concluir en una tesis o teoría general. De la observación repetida de objetos o de acontecimientos de la misma índole se establece una conclusión para todos los objetos o fenómenos de la misma naturaleza que los estudiados. Otro tipo de razonamiento no deductivo consiste en obtener una conclusión a partir de premisas en las que se establece una comparación o analogía entre elementos o conjuntos de elementos distintos, llamado razonamiento analógico (Pearl, 2000).

Por otro lado, en el razonamiento deductivo se parte de premisas generales para particularizar. En este caso, la validez de las premisas implica lógicamente la validez de la conclusión. En el razonamiento deductivo, si la generalidad es cierta, las conclusiones siempre serán ciertas, a no ser que haya información errada sobre la generalidad o que esté vagamente definida. En cambio, el razonamiento no deductivo suele generar incertidumbre en las conclusiones. La única posibilidad de conclusiones completamente ciertas en un proceso inductivo demanda que no exista variabilidad en el comportamiento del fenómeno bajo estudio. Mientras menor sea esta variabilidad, más confiables podrán ser las conclusiones y se requerirá, por tanto, menor cantidad de casos específicos (o ejemplares) para poder hacer inferencias o generalizaciones válidas. Así, por ejemplo, sólo es necesaria una pequeña muestra de sangre para estimar la cantidad de leucocitos que posee una persona pues se supone que éstos se distribuyen uniformemente en el torrente sanguíneo.

En el razonamiento no deductivo, la verdad de las premisas no convierte en verdadera la conclusión, ya que en cualquier momento podría aparecer una excepción que haga falsa una hipótesis. De ahí que la conclusión de un razonamiento no deductivo sólo pueda considerarse probable y, por eso, el nuevo conocimiento que se obtiene por medio de esta modalidad de razonamiento es siempre incierto o

discutible. Este tipo de razonamiento sólo es una síntesis incompleta o aproximada de todas las premisas.

Cuando se hace un proceso de reconocimiento o de discriminación de objetos, donde se trata de definir o caracterizar algo nuevo, se realiza un proceso de razonamiento no deductivo puesto que se deben realizar comparaciones con otros objetos existentes, sean éstos concretos o abstractos, para identificar las características que distinguen a un grupo de objetos, de los demás. Posteriormente, se continúa con una síntesis, la conclusión, para describir de manera abstracta el objeto o el grupo de objetos identificados, descartando detalles irrelevantes, mediante algún modelo.

Resumiendo, el razonamiento aproximado consiste en realizar inferencias aunque los hechos no satisfagan plenamente las reglas, por medio de la extensión del razonamiento clásico que usa los esquemas conocidos como el *modus ponens* (MP) y el *modus tollens* (MT). El primero de ellos utiliza un mecanismo de encadenamiento hacia delante para obtener las conclusiones y el segundo, un encadenamiento hacia atrás. Esto se observa en las dos definiciones siguientes.

$$\text{MP} = \frac{\begin{array}{l} \text{Regla : si } x \text{ es } A \rightarrow y \text{ es } B \\ \text{Hecho : } x \text{ es } A \end{array}}{\text{Conclusión : } y \text{ es } B}$$

$$\text{MT} = \frac{\begin{array}{l} \text{Regla : si } x \text{ es } A \rightarrow y \text{ es } B \\ \text{Hecho : } y \text{ no es } B \end{array}}{\text{Conclusión : } x \text{ no es } A}$$

Puesto que los computadores pueden ser dotados de mecanismos para realizar inferencias, la capacidad de razonamiento deja de ser exclusiva de los humanos, sino que se extiende a las máquinas. Esta capacidad ha sido altamente aprovechada para la construcción de sistemas expertos o basados en conocimiento (KBS, su sigla en inglés) en la Inteligencia Artificial.

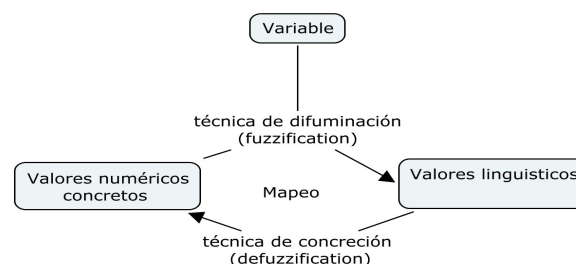
### **3.3 Sistemas Expertos o Sistemas Basados en Conocimiento**

En la Inteligencia Artificial, se considera la Ingeniería del Conocimiento como la disciplina que orienta en la concepción y desarrollo de sistemas basados en conocimiento (abreviadamente, KBS por sus siglas en inglés). Existe el consenso de considerar dicha construcción como el desarrollo de programas de computador con capacidades para la resolución de problemas, comparables a las realizadas por expertos en el dominio (Studer, Benjamins y Fensel, 1998). Por esto, el concepto de KBS subsume al concepto de sistema experto que antes prevalecía.

La estructura de un KBS consta de una base de conocimientos y una máquina de inferencia. La base de conocimientos de un sistema KBS contiene los datos específicos del dominio del problema, en un momento determinado, y las reglas necesarias para realizar las inferencias que permitan resolver problemas o tomar decisiones (Negnevitsky, 2004). La máquina de inferencia define cómo hacer uso de la base de conocimientos para producir las respuestas deseadas. Dicha máquina es el “cerebro” de los sistemas KBS que se encarga de derivar nueva información u obtener conclusiones a partir de los datos de entrada y las reglas definidas en la base de conocimientos.

En los sistemas de inferencia difusa más comunes, las reglas de toman la forma simple  $Si P \rightarrow Q$ . De acuerdo con el valor de verdad de  $P$ , se hace válido  $Q$  en un *Universo de Discurso* determinado. Las reglas de inferencia deductiva incorporan la vaguedad en el antecedente o consecuente, empleando valores vagos o etiquetas lingüísticas para representar el conocimiento. La relación condicional  $Si P \rightarrow Q$  en lógica bivaluada tiene una única tabla de verdad, pero en los sistemas de inferencia borrosa presenta múltiples interpretaciones, todas con algún grado de validez dentro de un cuerpo formal (el contexto). En el antecedente puede existir una combinación de etiquetas relacionadas mediante las conectivas lógicas  $\wedge$ ,  $\vee$  y la negación.

Un sistema de inferencia difusa para el control automático de aparatos recibe valores de entrada concretos, como la temperatura  $x$  del calentador de agua, y las reglas de inferencia son definidas con términos vagos en el consecuente, como en la regla: *Si la temperatura es 40.5  $\rightarrow$  “caliente”*. En este tipo de sistemas, los valores recibidos como entrada deben ser sometidos a un proceso de difuminado para asociarle la etiqueta lingüística que más se ajuste, antes aplicar otras reglas de inferencia o tomar las decisiones del caso. En un sistema de inferencia difusa para la comprensión de términos vagos en el lenguaje de consultas, por el contrario, los valores de entrada son vagos y por eso deben ser sometidos a un proceso inverso para concretar o precisar esos valores y, así, las solicitudes de los usuarios de un sistema flexible de consulta-respuesta puedan ser interpretadas por el sistema gestor de bases de datos. En la Figura 2 se muestran, entonces, los dos tipos de mapeo posibles entre términos vagos y concretos o nítidos que caracterizan a los Sistemas de Inferencia Borrosos: el difuminado o emborronamiento (fuzzification, en inglés) y la concreción o desambiguación (defuzzification, en inglés).



**Figura 2. Mapeos en un sistema difuso o borroso.**

Una máquina de inferencia puede dar su respuesta en forma concreta con valores numéricos (enfoque Takagi-Sugeno) u ofrecer una respuesta vaga como salida (enfoque Mamdani) (Álvarez y Peña, 2001). Para ofrecer respuestas cualificadas con alguna etiqueta lingüística, cuando los valores sean numéricos, la máquina de inferencia deberá incluir un módulo de difuminado donde se realiza un proceso deductivo.

En el sistema de inferencia para la valoración de expresiones compuestas o agregadas en el antecedente de una regla, se pueden aplicar dos mecanismos básicos de razonamiento: la inferencia basada en la activación de reglas individuales (modelo FITA, como acrónimo de First Infer Then Aggregate) o inferencia basada en la activación de una composición de las reglas (modelo FATI, como acrónimo de First Aggregate Then Infer). En el primer mecanismo, cada regla que se activa se valora individualmente y después se concatenan los aportes individuales en una conclusión final. En el segundo método, se obtiene un compuesto de las reglas del sistema como una relación borrosa y sobre ésta se valora el antecedente. Los modelos FITA y FATI son equivalentes si las entradas al sistema de inferencia son valores escalares pero en los sistemas de inferencia difusos sólo se esperan resultados parecidos pues, generalmente, en el álgebra de las medidas difusas no siempre es válida la ley distributiva (Czogala y Leski, 2000).

Tradicionalmente, en los KBS propuestos para interpretar la vaguedad de las consultas, los modelos para representar los términos vagos deben predefinirse de antemano, con la ayuda del conocimiento experto. Pero como el contexto, que es el que le da sentido a una palabra vaga, es variable y sólo es conocido en el momento de la consulta, la máquina de inferencia deberá llevar a cabo un proceso de minería de datos para la derivación de los modelos o las reglas de inferencia, válidas en ese contexto. Significa que la máquina de inferencia también debe realizar inferencias o razonamientos de tipo no deductivo, realizando un análisis comparativo o analógico de los elementos u objetos de interés que se circunscriben al contexto delimitado por la consulta. Este tipo de análisis, es un proceso conocido como discriminación y clasificación de objetos.

### **3.4 La Discriminación y la Clasificación de Objetos**

El objetivo básico de la discriminación es reconocer las diferencias entre grupos de objetos y poder describirlas en forma gráfica o algebraica para lograr un mejor entendimiento de un determinado entorno. En la clasificación, como proceso predictivo, el objetivo consiste en encontrar modelos o formular reglas que permitan asignar nuevos objetos a las clases rotuladas con alguna etiqueta lingüística (Johnson y Wichern, 2001). Puesto que para realizar una clasificación es necesario haber realizado previamente una discriminación, estos dos conceptos suelen ir muy ligados.

A principios del siglo pasado Fisher propuso las primeras técnicas estadísticas para encontrar reglas discriminantes en estudios taxonómicos (Fisher, 1936). Más recientemente, en Inteligencia Artificial, surge el área denominada

“Aprendizaje de Máquinas” (Machine Learning, en inglés) que también se concentra en este tipo de técnicas para el reconocimiento de patrones y la clasificación de los objetos. Pero además se ocupa de la complejidad algorítmica de las implementaciones requeridas para que los computadores puedan “aprender” de la información disponible (Perlich y Provost, 2006).

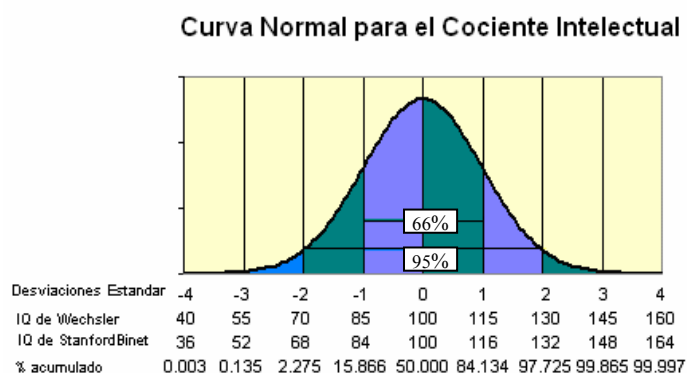
En el proceso de reconocimiento de patrones o de identificación de rasgos comunes en los objetos de un dominio, se requiere de un sensor que “lea” o capture las señales observadas en ellos, de un mecanismo de extracción de la información a partir de las señales y de un modelo descriptivo para la discriminación y la clasificación de nuevos objetos. Este proceso de reconocimiento se basa, generalmente, en la disponibilidad de un conjunto de ejemplares que hayan sido clasificados o descritos previamente. Este conjunto es llamado el conjunto de entrenamiento o la muestra de aprendizaje. Y la estrategia de reconocimiento es conocida como *aprendizaje supervisado* pues se requiere de una guía (que en este caso es la muestra de aprendizaje), para realizar la tarea de encontrar el modelo de interés. El aprendizaje se denomina *no supervisado*, cuando al sistema no se le ofrece una catalogación *a priori* de los patrones, ni se le informa sobre el número de grupos o clases en que se deben agrupar una colección de datos (Dyer, 2006). Por lo tanto, en el aprendizaje no supervisado la máquina debe establecer por su cuenta la cantidad de grupos y los modelos descriptivos de cada uno de ellos, basándose en las regularidades estadísticas de los datos.

Por lo general, el modelo “aprendido” o ajustado tras el proceso de discriminación se representa por medio de reglas de clasificación, de árboles de decisión, de una red neuronal o de una fórmula matemática (para mayor profundidad véase a O’Hagan (1994), Hastie, Tibshirani y Firedman (2001) y a Webbs (2002), entre otros. La mayoría de los modelos, se basan en la distancia de los elementos, con respecto al prototipo de cada grupo, para calcular las diferencias y similitudes entre objetos en un espacio multidimensional. Entre las medidas de disimilitud, basadas en distancias geométricas, se mencionan la distancia Euclidiana Cuadrática y la distancia de Manhattan. Cuando los datos para algunas de las dimensiones involucradas son categóricos, en lugar de numéricos, se usan otras medidas de disimilitud como el grado de desacuerdo.

### **3.4.1 Técnicas de Representación de una Colección de Objetos**

La Estadística Clásica ha supuesto que la mayoría de las distribuciones de datos se pueden ajustar al modelo probabilista normal o de Gauss. En especial, cuando el tamaño del conjunto de observaciones es grande. En tal sentido, existen múltiples técnicas para el tratamiento y análisis de los datos, entre ellas las usadas para la discriminación y clasificación de objetos. Es así como, para la clasificación de las personas adultas, según su cociente intelectual (en forma abreviada, *IQ*, por sus siglas en inglés) se realizó una partición matemática del dominio de esta variable, usando como modelo teórico la distribución normal. El *IQ* es una variable que se deduce de un conjunto de pruebas de razonamiento abstracto y de comprensión verbal, entre otras.

Debido a la simetría y a otras características de la distribución normal, se espera que: el 66% de los datos, aproximadamente, se concentren alrededor de la media poblacional a una distancia máxima de una desviación estándar, el 95% de los datos a dos desviaciones y el 99% a una distancia de tres desviaciones estándar (Johnson y Wichern, 2001). En la Figura 3, se muestra la curva normal definida de acuerdo con dos pruebas diferentes: el IQ de Wechsler y el IQ de Stanford Binet (Wechsler, 1944). Las líneas verticales delimitan una desviación estándar.



**Figura 3. Distribución normal del IQ según dos pruebas diferentes**

La clasificación de los individuos propuesta por Terman, basada en la prueba Stanford Binet, fue diseñada para que el valor medio o típico fuera 100. Considerando una desviación estándar de 16, las reglas de clasificación sugeridas por este autor se resumen en la Tabla 3. Allí se observa que la clase intermedia (llamada “promedio”) concentra el 50% de los datos.

A partir de esa clase intermedia se definen las demás, con menor porcentaje de observaciones mientras más alejadas se encuentren del valor medio. También se observa en la Tabla 3, que en la partición del *Universo de Discurso*, un individuo pertenece con grado 1, a una de las categorías establecidas y con grado 0, a las restantes. Así, un individuo que obtenga 128 se le debe considerar “genio”, mientras que el que obtenga 127 no puede catalogarse de esa manera y sólo por un punto de diferencia. Esta limitante, es la que se pretende resolver con una partición difusa, en lugar de la partición matemática o nítida.

**Tabla 3. Clasificación de los individuos adultos, según su inteligencia**

Categoría	Limites del IQ	Porcentaje esperado
Genio o Superdotado	128 o más	2.2%
Superior	120-127	6.7%
Brillante Normal	111-119	16.1%
Promedio	91-110	50%
Lento Normal	80-90	16.1%
Limitrofe	66-79	6.7%
Defectuoso	65 o menos	2.2%

Una partición del *Universo de Discurso* basada en la distribución normal tiene la ventaja de depender de sólo dos parámetros: la media y la desviación estándar. Pero es importante señalar que la distribución normal no siempre es el modelo teórico probabilista más apropiado para representar una colección de datos u observaciones, pues es corriente que existan sesgos o asimetrías en la forma de la distribución y tratar de representar la distribución de los datos sólo con dos parámetros (la media como medida de localización y la varianza como medida de dispersión) no sólo resulta insuficiente, sino inexacto. Esta inexactitud ocurre porque la media aritmética y a la varianza muestral, como estimadores de los dos parámetros de la distribución normal, no son medidas robustas a los valores extremos (Montgomery y Runger, 2003).

Como ejemplo, en la Tabla 4 se presenta el comportamiento de algunas funciones de los datos o estadísticas usadas corrientemente para la estimación de la localización de la tendencia central de la distribución, para una muestra  $M = (1, 1, 1, 1, 1, 100)$ . Allí se puede observar que la media aritmética está muy distante de las estimaciones realizadas por las otras funciones de los datos porque se ha dejado influenciar por el valor extremo. La mediana y la media recortada, por el contrario, ignoran cualquier valor atípico y por eso son mejores representaciones de la tendencia central cuando la distribución no se puede suponer normal.

**Tabla 4. Estadísticas para la estimación de la tendencia central**

Estadística	Estimación
Media geométrica	1.93
Media aritmética	15.14
Mediana	1.00
Media recortada al 25%	1.00

Por lo anterior, cuando la distribución de los datos no presenta un comportamiento que se pueda considerar normal, en la Estadística, se han propuesto los modelos denominados *no paramétricos* (non parametric models). Este término no quiere decir que tales modelos carecen de parámetros, sino que el número y la naturaleza de los mismos pueden ser flexibles y no preestablecidos de antemano (Huber, 2004). Significa que la distribución de los datos no tiene que ser definida a priori, pues los datos observados son los que la determinan y por eso, también se les denomina modelos libres o independientes de la distribución de los datos (distribution free models). Dentro de esta categoría de modelos, los más usados para describir la distribución de un conjunto de datos son las tablas o histogramas de frecuencia, el polígono de frecuencias acumuladas, los modelos basados en percentiles, la estimación de las densidades de probabilidad del núcleo (kernel density estimation) y las funciones empíricas de distribución acumulada (Wasserman, 2005). Cada una de ellas se describirá con cierto detalle, para justificar la opción elegida en la obtención de los valores particulares de los parámetros de los

conjuntos difusos que representarán los términos vagos relativos al contexto lingüístico, a partir de los datos disponibles.

### 3.4.1.1 Modelos Basados en las Frecuencias de los Datos

Las tablas de frecuencias han sido ampliamente utilizadas en aquellos casos donde la cantidad de valores de una variable hace necesario agruparlos en clases, rangos o intervalos para facilitar su descripción e interpretación. Una tabla de frecuencias también pueden presentarse de manera gráfica para facilitar aún más la comprensión de la distribución de los datos sobre su dominio o *Universo de Discurso*. Al diagrama de barras de las frecuencias absolutas o relativas con que ocurre cada intervalo o categoría disjunta, donde la amplitud de la barrar corresponde a un intervalo del dominio y la longitud a la frecuencia, se le conoce con el nombre de *histograma de frecuencias*. Cuando se unen los puntos medios de los intervalos de clase por medio de una línea, se le conoce como *polígono de frecuencias*.

El problema principal en la construcción de una tabla o histograma de frecuencias es determinar el número adecuado de intervalos para la agrupación de los datos. Existen diversas reglas empíricas acerca de cuál debería ser un número de intervalos o grupos adecuados. Entre las reglas más conocidas se tienen:

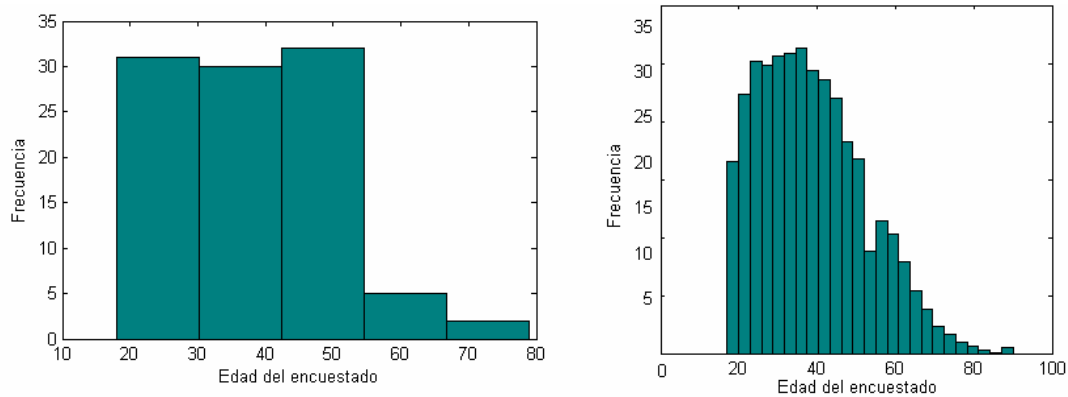
1.  $K = \sqrt{n}$
2.  $5 \leq K \leq 20$
3.  $K = 1 + 3.33 \log_{10}(n)$

Donde  $n$  es el número de observaciones y  $K$ , el número de intervalos o grupos. Si se usa la primera regla, se corre el riesgo de incrementar drásticamente el número de intervalos a medida que aumenta el número de datos. Por eso casi siempre se recomienda usar la segunda o la última regla que es conocida como la regla de Sturges. Particularmente, Soong recomienda dividir el rango o recorrido de la variable en 12 intervalos de clase, pues considera que más de estos valores no resumiría lo suficiente un conjunto de datos y usar menos intervalos, impediría describirlo adecuadamente (Soong, 2004).

Con propósitos ilustrativos en la

Figura 4 se muestran dos histogramas de frecuencias relativas de las edades de los encuestados en el Censo de Estados Unidos realizado en el año de 1994. Estos datos constituyen una muestra de 32561 del total de encuestados seleccionada por Barry Becker usando como condiciones de filtrado la edad (mayor de 16 años) y que trabajara al menos una hora a la semana, entre otras condiciones (US Census Bureau, 2002). En el histograma de la izquierda construyó el histograma usando cinco agrupaciones ( $K = 5$ ), mientras que en el histograma de la derecha se usó un  $K = 25$ . En la figura se observa el efecto de acercamiento o “zoom” al aumentar el número de agrupaciones. La granularidad más fina del histograma de la derecha permite apreciar mejor la distribución de los datos y por lo tanto la describe mejor, pero

tendría el inconveniente de la cantidad de parámetros del modelo: los intervalos de clase.

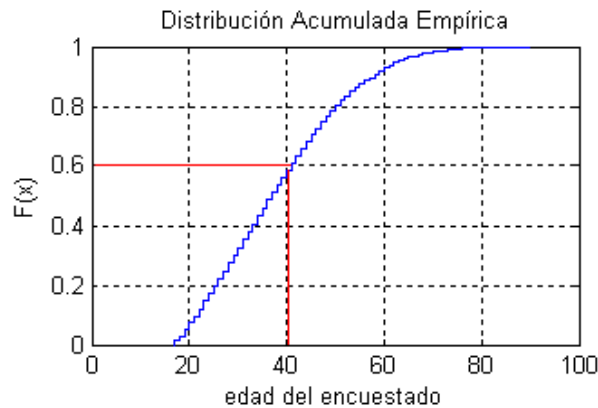


**Figura 4. Histogramas de frecuencias con diferente resolución**

Cuando los datos ya han sido agrupados en un histograma, se pierde la información proporcionada por los datos de manera individual, por lo que para poder ofrecer medidas descriptivas se salva el inconveniente determinando un representante de cada clase (el prototipo) que no es más que el punto medio del intervalo, conocido como la *marca* de clase. Por lo tanto, una alternativa, que reduciría sensiblemente el número de parámetros del histograma sería usar las marcas de clase en lugar de los límites de los intervalos, pero estas medidas dejan de ser representativas cuando los datos no tienen una distribución normal o uniforme dentro de cada uno de ellos, como ocurre con la media aritmética. Además, las marcas de clase tampoco son representativas de un intervalo vacío.

Por otro lado, a pesar de lo informativos que son los histogramas de frecuencia hacen difícil la comparación de dos o más distribuciones de datos, especialmente cuando se superponen en una sola gráfica. Por eso, otra forma frecuente de describir una distribución de datos se logra mediante el polígono de frecuencias acumuladas, también llamado *ojiva* por su forma. En este tipo de gráfico, el eje X corresponde a los límites superiores de cada intervalo definido en una tabla de frecuencias, no las marcas de clase y el eje Y corresponde a las frecuencias relativas acumuladas hasta esos límites. De acuerdo con esto, la gráfica resultante es una función escalonada, donde los escalones serán más pronunciados si existen pocos intervalos de clase. Cuando el número de intervalos se acerca al tamaño de la colección de datos se tendrá una gráfica casi suavizada, como se muestra en la

Figura 5, para la distribución acumulada de las frecuencias relativas de las edades de encuestados en el censo de Estados Unidos de 1994. Se dice que la distribución acumulada construida de esta forma es *empírica* pues los datos no se han ajustado a un modelo probabilista teórico en particular.



**Figura 5. Polígono de frecuencias acumuladas para las edades de los encuestados**

El polígono de frecuencias acumuladas ofrece una manera rápida y simple de determinar aproximadamente los percentiles de una distribución. Por ejemplo, en la

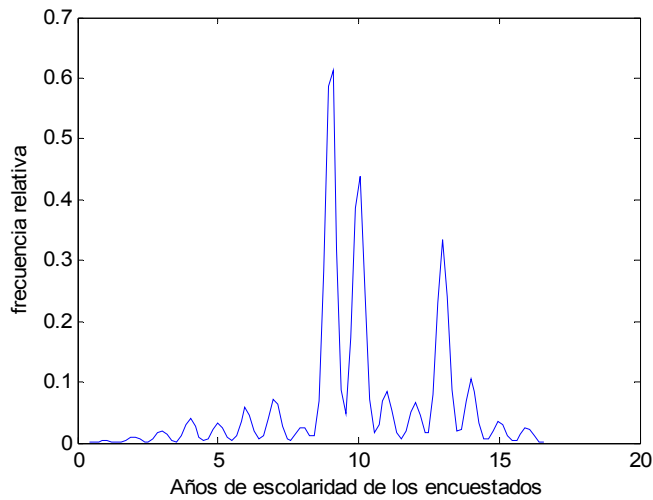
Figura 5, las líneas rojas están indicando la probabilidad aproximada de encontrar un sujeto en la muestra de los censados, con una edad inferior o igual a los 40 años es 0.6. Alternativamente, de manera inversa, que el percentil 60% es 40 años. Sin embargo, debe resaltarse que este valor es sólo una aproximación, pues el percentil 60% exacto es 41 años que es calculado con los datos no agrupados. En este ejemplo, estos dos valores son muy próximos por el número tan alto de agrupamientos de los datos, pero la estimación puede desmejorar cuando se consideran pocos intervalos de clase.

### 3.4.1.2 La Estimación de la Densidad del Núcleo

Una muestra de datos también puede describirse estimando su densidad en una forma no paramétrica, usando una función de suavización del núcleo de la misma (Webb, 2002). Mientras es difícil superponer dos histogramas para compararlos, como recién se dijo, también es fácil hacer la superposición de las estimaciones de las funciones de densidad suavizadas y de ahí, su utilidad.

La función de la densidad del núcleo produce una versión empírica de la función de densidad de probabilidad. Esto es, en lugar de seleccionar una densidad con una forma teórica particular y estimar los parámetros, la función produce una estimación suavizada de la densidad que se ajusta mejor a la distribución de datos. Sin embargo, el costo computacional para generar una curva como ésta puede ser prohibitivo cuando se tenga una gran cantidad de datos. Además, si se parte de esta función para la categorización difusa no se podría asegurar la convexidad de los subconjuntos derivados, dado que el conjunto que representa el Universo de Discurso no siempre es convexo, como se puede apreciar en el ejemplo de la

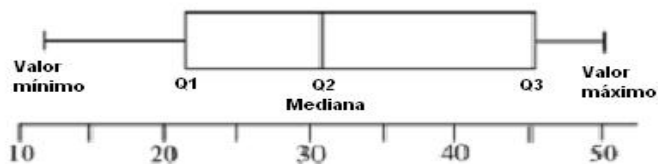
Figura 6.



**Figura 6. Ejemplo de la función suavizada de la densidad del núcleo**

### 3.4.1.3 Los Modelos Basados en Percentiles

Formalmente, un percentil  $P_q$  es un punto del dominio de una variable, bajo el cual se encuentra un porcentaje  $q$  de los valores de una colección de datos, cuando ésta es ordenada. Si se elige un número apropiado de percentiles, se obtiene un modelo que representará la forma y la localización de la distribución de los datos mejor que el modelo normal o gaussiano, pues con ellos se pueden también representar los sesgos o las asimetrías en las distribuciones. Particularmente, se ha comprobado que forma muy efectiva para la descripción de una colección de datos es la estadística de resumen de los cinco números (5-number summary, en inglés) compuesta por el valor mínimo, el valor máximo y los tres cuartiles de la distribución. Los cuartiles dividen la densidad de la distribución de los datos ordenados en cuatro áreas, cada una con el 25% de los datos. A partir de esta estadística de resumen se suele construir el Diagrama de Cajas y Bigotes (Box and Whisker Plot), uno de los modelos gráficos más informativos y usados para representar colecciones de datos o poblaciones (Johnson y Wichern, 2001).



**Figura 7. Ejemplo de un Diagrama de Cajas y Bigotes**

En un Diagrama de Cajas y Bigotes, o abreviadamente DCB, como el que se muestra en la Figura 7, se usa como aproximación o estimador de la tendencia central de la distribución de los datos a la mediana (equivalente al segundo cuartil  $Q_2$ ) y se dibuja como una línea que divide la caja, en dos secciones que serán iguales sólo cuando no existan sesgos en la distribución. El primer cuartil  $Q_1$  y el tercero  $Q_3$  determinan el rango intercuartil de la distribución y se representa gráficamente como la caja del diagrama. El valor mínimo y máximo de la variable de interés, junto con el primer y tercer cuartil, respectivamente, sirven para la construcción de los bigotes. Por lo tanto, se puede considerar que la caja de un DCB está representando los valores “medianos” y los bigotes están representando a los valores “pequeños” y “altos” de la distribución de los datos, si se realizara una partición de la manera convencional.

Con fines comparativos, en la Figura 8, se muestran dos maneras de representar una distribución de datos de las descritas anteriormente: los Histogramas de Frecuencia y los Diagramas de Cajas y Bigotes. En la comparación se consideraron algunas variables de la base de datos de los autos que será usada en esta investigación como referencia, la cual ha sido ampliamente usada en otros trabajos de investigación, en libros y en guías de usuario de paquetes de Estadística. Esta base de datos, junto con otras, es proporcionada por el Laboratorio de Tecnología de la Información (ITL) del Instituto Nacional de Estándares y Tecnología (NIST) del gobierno de Estados Unidos, con el propósito de permitir a los investigadores analizar o verificar el comportamiento de diversas técnicas estadísticas o de aprendizaje de máquinas (ITL, 2006). Se ha elegido esta base de datos porque contiene datos que se suponen ciertos y varias características de los carros de tipo cuantitativo que son las dos únicas condiciones que se imponen al modelo de razonamiento en la interpretación de la vaguedad (ver sección 5.1.1).

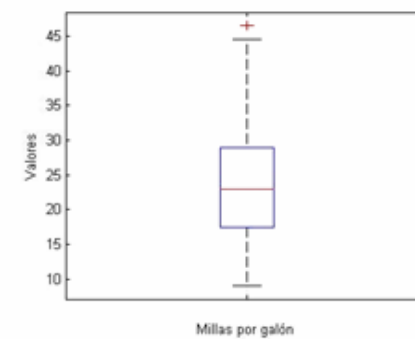
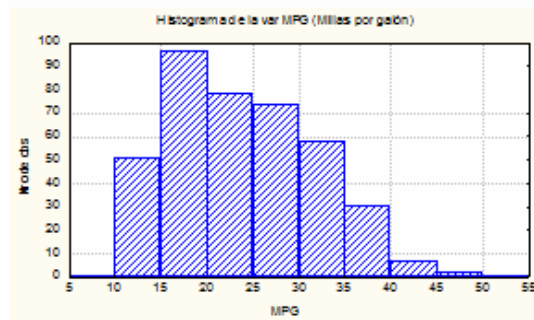
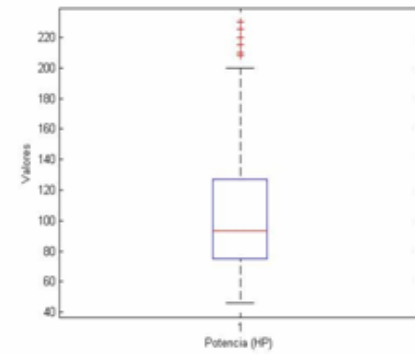
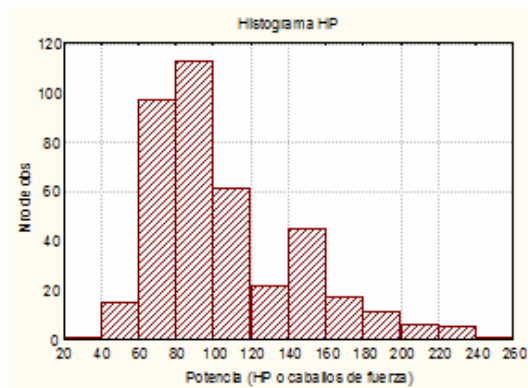
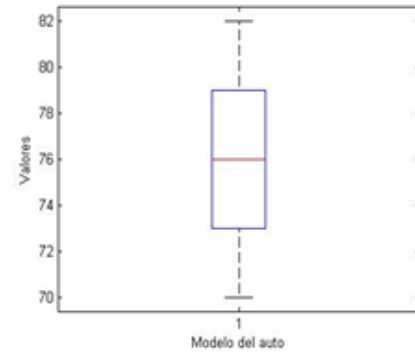
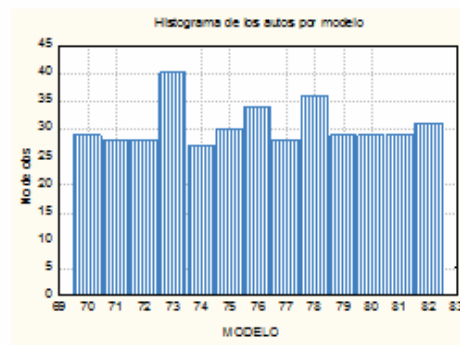
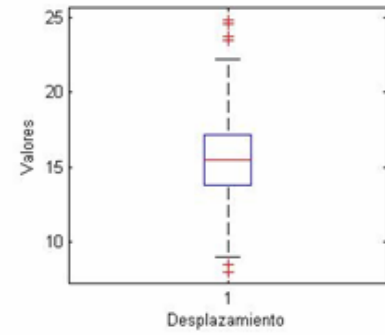
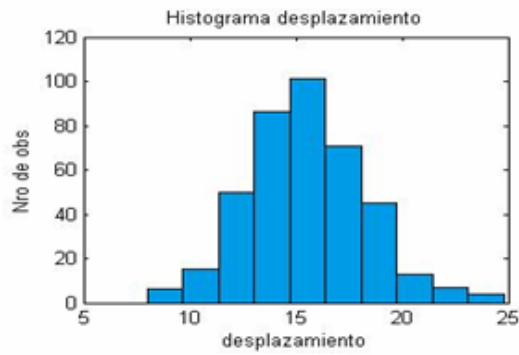


Figura 8. Distribución de los autos, considerando diferentes variables

En la Figura 8, se puede observar que la distribución de los autos, considerando su rendimiento (medido en millas por galón) o su potencia (medida en caballos de fuerza o HP) son distribuciones sesgadas positivamente, pues hay una mayor concentración de los datos a la izquierda o a los valores menores. Estos sesgos positivos también se detectan en el DCB de cada una de las variables mencionadas, pues el bigote superior es más largo que el inferior. Si la distribución de los datos fuera uniforme sobre el dominio de la variable, los bigotes y cada parte de la caja tendrían un tamaño igual, como ocurre para la distribución de los autos considerando su año de fabricación o modelo. Otro caso muy distinto se presenta cuando se considera el desplazamiento de los autos, pues en el histograma correspondiente se observa un comportamiento bastante ajustado a una distribución normal y por eso, el tamaño casi igual de los bigotes. En este histograma se aprecia una caja más pequeña del DCB, indicando que los datos están más concentrados en la parte central que a los extremos, como es lo esperado para una distribución normal.

Considerando el número de parámetros de un Diagrama de Cajas y Bigotes (la estadística de los 5 números) se puede decir es un modelo más resumido que un histograma o polígono de frecuencias relativas, pero con la ventaja de ser más parsimonioso. Un DBC informa dónde se encuentra la tendencia central, cómo se concentran los datos y si existen sesgos en la distribución. Además, a diferencia de un histograma, asegura que ninguno de los subconjuntos del *Universo de Discurso* en que se subdivide sea vacío, acorde con la definición de una partición, bien sea ésta una partición matemática o difusa.

Ya descritos los modelos libres o no paramétricos comúnmente empleados para describir una distribución de datos que sirven de base para la discriminación o categorización difusa de los objetos, se observa que los modelos más recomendables son los basados en percentiles, por su sencillez de cálculo y su facilidad de interpretación. Además, un modelo basado en percentiles aprovecha la información que proporcionan los datos de manera individual y garantizan la convexidad de los conjuntos que se infieran de él, lo que no puede asegurarse con un histograma de frecuencias o con la estimación del núcleo de la densidad de la distribución.

### 3.4.2 Reglas de Decisión

En la discriminación convencional (nítida) se consideran que existen  $k$  clases disjuntas denotadas  $C_1, \dots, C_k$ , en el que se puede subdividir el dominio de una variable categórica unidimensional o multidimensional. Un primer supuesto para definir la reglas de clasificación de los objetos, es considerar que la probabilidad *a priori* de ocurrencia de cada una de ellas  $P(C_1), \dots, P(C_k)$  es conocida. Cuando se desea clasificar un objeto en una clase de las posibles, en un dominio determinado, y del cual se conoce su distribución de probabilidad sobre las clases, la regla de decisión será asignar el objeto a la clase  $j$ , si (Webb, 2002):

$$P(C_j) > P(C_r) \quad r = 1..k; j \neq r$$

Con esta regla, se asigna cada objeto a una sola clase. Cuando un objeto tiene igual probabilidad de pertenecer a varias clases, éste es asignado a una de ellas, arbitrariamente.

Si se tiene una muestra de  $n$  observaciones de una variable  $X$ , con  $p$  dimensiones, relacionada con una variable  $y$  de salida, la regla de decisión puede ser corregida o mejorada para hacer la clasificación de los objetos. La nueva regla asigna al objeto a la clase  $\omega_j$  si la probabilidad de esa clase, dadas las observaciones, es la mayor entre todas las posibles:

$$P(C_j / X = x) > P(C_r / X = x) \quad r = 1..k; j \neq r \quad (3.1)$$

Esta regla decisión parte en  $k$  regiones disjuntas,  $\Omega_1, \Omega_2, \dots, \Omega_k$  al dominio de tal forma que si  $x \in \Omega_j$ , entonces pertenece a la clase  $C_j$ .

Por otro lado, la probabilidad de pertenencia *a posteriori* de un objeto a una clase dada, puede ser expresada en términos de la probabilidad *a priori*, y las funciones de densidad de probabilidad condicional de  $x$ , dada la ocurrencia de esa clase y aplicando el teorema de Bayes, resulta ser la siguiente

$$P(C_i / X = x) = \frac{P(X = x / C_i) P(C_i)}{P(X = x)}$$

Por lo tanto, la regla de interpretación de la ecuación 3.1 puede ser escrita de esta manera: asigne  $x$  a la clase  $C_j$  si:

$$P(X = x / C_j)P(C_j) > P(X = x / C_r)P(C_r) \quad r = 1..k; j \neq r$$

Esta regla es conocida como la regla de Bayes de *mínimo error*. Para el caso de considerar sólo dos clases en el dominio, esta regla se puede plantear como aparece, a continuación.

$$V(x) = \frac{P(x / C_j)}{P(x / C_r)} > \frac{P(C_j)}{P(C_r)} \Rightarrow x \in \Omega_j$$

La función  $V(x)$  es conocida como la *verosimilitud* cuyo umbral es  $P(C_j) / P(C_r)$ . Como la regla asigna un objeto a la clase que tenga mayor probabilidad *a posteriori*, se minimiza el error de una mala clasificación y de ahí, su nombre.

Por lo expuesto, se puede afirmar que la esencia de la técnica bayesiana es proporcionar una regla matemática que explique como podrían cambiar las creencias a la luz de nueva evidencia. Esto permite a los científicos combinar nueva información con la teoría previa. Pero infortunadamente, los supuestos del conocimiento de las distribuciones de probabilidades *a priori* no se cumplen, en muchos casos reales. Por esto, en Teoría de Decisiones surgió otra estrategia para definir las reglas de decisión, una técnica que considera supuestos pero sobre la forma de las reglas de inferencia o las *funciones discriminantes*.

Una función discriminante es una función aplicable sobre el objeto  $x$ , como la que a continuación se presenta, para el caso de dos clases y una constante  $c$ :

$$\begin{aligned}h(x) > c &\Rightarrow x \in C_1 \\ < c &\Rightarrow x \in C_2\end{aligned}$$

De acuerdo con lo anterior, el valor óptimo de la función discriminante  $h(x)$  se obtiene cuando  $c = P(C_j) / P(C_r)$  para el caso de dos clases. En el caso de  $k$  clases no se puede definir una sola función discriminante, sino varias:

$$h_i(x) > h_j(x) \Rightarrow x \in C_i \text{ para } j = 1, 2..k \text{ } j \neq i$$

Existen muchas formas de funciones discriminantes, que varían en complejidad desde las funciones lineales discriminantes, propuestas por Fisher (1936) en las cuales cada  $h_i(x)$  es una combinación lineal, hasta funciones no lineales como la *perceptron multicapa* de las Redes Neuronales y otras técnicas como los Árboles de Decisión o Clasificación (Hastie, Tibshirani y Friedman, 2001).

Puesto que en este trabajo investigativo, la distribución *a priori* de los datos contenidos en una base de datos es desconocida, la técnica de discriminación borrosa deberá basarse en supuestos sobre la forma de las funciones discriminantes, de modo parecido a lo que se realiza en la discriminación nítida o concreta.

### 3.4.3 Granulación Difusa o Borrosa

La granulación borrosa es considerada una técnica de descomposición del todo en sus partes (los gránulos) donde se admite algún grado de solapamiento entre esas partes, pues un gránulo derivado de este proceso no es una entidad física, sino abstracta o conceptual.

La diferencia principal de las técnicas difusas para la discriminación o descripción de un dominio particular tiene que ver con el dominio del grado de pertenencia de un elemento cualquiera a un conjunto o clase rotulada con alguna etiqueta lingüística. En las técnicas convencionales, este grado se considera una variable discreta dicotómica que admite sólo dos valores: "cierto" o "falso". Dichos valores son representados con los dígitos binarios 1 y 0, respectivamente, para significar que un elemento pertenece o no, a una clase determinada. En las técnicas de granulación o discriminación borrosa, los grados de pertenencia de un elemento del dominio a cualquier subconjunto o gránulo en un contexto dado, pueden tomar más de dos valores. También es admisible que un elemento pertenezca, simultáneamente, a más de un subconjunto del *Universo de Discurso*.

La Computación Granular es un paradigma que surge para la representación y el manejo del concepto "gránulo de información", definido como aquel que surge de un proceso de abstracción de los datos y de la derivación de conocimiento a partir de la información (Bargiela y Pedrycs, 2003). Es decir, el significado para un gránulo emerge de los datos como consecuencia de su transformación, su resumen o condensación. El aspecto clave en la Computación Granular es la comprensibilidad o interpretación de los gránulos.

Según Yao, la interpretación de un gránulo es alcanzada cuando se le asocia una etiqueta lingüística con significado (Yao, 2000). Esto quiere decir que un subconjunto del *Universo de Discurso* sólo es interpretable cuando se le puede asignar una etiqueta. En (Morales, 2000) también se hace énfasis en que una etiqueta lingüística no puede verse sólo como un nombre.

Una etiqueta lingüística es una tripleta (*Nombre, E, U*) donde *Nombre* puede ser asignado al conjunto difuso *E* en el universo *U*. El conjunto difuso *E* representa el significado de la etiqueta en *U*, en términos cuantitativos. Al ser una tripleta (*Nombre, E, U*), pueden ocurrir estas situaciones.

- i) Tanto *E* como *Nombre* son conocidas en *U*. Esto quiere decir que ya se tiene el significado de *Nombre* que es válido en el universo *U*.
- ii) *Nombre*, o el conjunto de términos  $T(X)$ , es conocido pero sus correspondientes significados son desconocidos. Por lo tanto, *A* debe aproximarse mediante propiedades o resúmenes estadísticos de los objetos de interés en *U*.
- iii) *Nombre* es desconocido pero *A* es conocido. Quiere decir que se tiene un conjunto de gránulos o agrupamientos, en el espacio multidimensional considerado pero no se conoce a  $T(X)$ .

En el problema que nos ocupa, se presenta la situación descrita en el literal ii) puesto que se trata de hallar la semántica de los términos vagos expresados en una consulta, lo desconocido, pero se conocen los nombres de las etiquetas lingüísticas.

Para realizar una granulación o discriminación borrosa, se han propuesto diversas técnicas que extienden las técnicas estadísticas o de la inteligencia artificial para admitir que un objeto pueda pertenecer a varias clases simultáneamente. Generalmente, dependen de la naturaleza de los datos y si los fines de la discriminación son descriptivos o predictivos. Entre las técnicas de granulación difusa no supervisadas que se han usadas con fines descriptivos, se puede mencionar al popular algoritmo C-medias difuso (Fuzzy C-Means, en inglés) que es visto como un difuminado de la técnica “*k*-Means” propuesta por Mac Queen (1967). Otra técnica más refinada es el algoritmo difuso C-medias ponderado (Weighted Fuzzy C-means) propuesto en (Tsekouras, 2005). En dichas técnicas se considera como estimador del prototipo de una clase, a la media aritmética y la medida de disimilitud se basa en las desviaciones cuadráticas medias con respecto al prototipo. También se han propuesto otras técnicas que extienden la técnica denominada K-medoides, que se basan en la mediana y no en la media aritmética o en la distribución normal, pero a expensas de una mayor complejidad computacional (Hastie, Tibshirani y Friedman, 2001). También se han propuesto otras técnicas de agrupamiento particionales alternativas basadas en redes neuronales o en computación evolutiva, usando algoritmos genéticos. Infortunadamente, casi todas esas técnicas están caracterizadas por su alta complejidad algorítmica y por eso, suelen ser catalogadas como “NP-Hard” (acrónimo de Non-deterministic Polynomial-time Hard). Estas técnicas no

supervisadas tampoco aseguran gránulos interpretables por no dar origen a conjuntos convexos.

Por lo anterior, una estrategia común para asegurar la interpretabilidad de los gránulos, consiste en la definición de un conjunto de restricciones que deben verificarse en ellos y en los modelos utilizados en la discriminación difusa. Son técnicas que se basan en métodos heurísticos, tratando de emular lo que haría el ser humano, en casos similares. De acuerdo con los objetivos establecidos, y como se había mencionado recientemente, esta será la estrategia a seguir en el presente trabajo de investigación.

#### 3.4.4 Técnicas de Discriminación Difusa Basadas en Reglas Heurísticas

Una técnica cualquiera, para dar solución a un problema de la Ingeniería, puede ser considerada como *heurística* si emula el modo de proceder humano que se guía por el sentido común. Por eso, al razonamiento aproximado basado en reglas heurísticas se le conoce también como razonamiento de sentido común (Guillaume, 2001; Casillas et al., 2003).

Para precisar un poco el término "heurístico", en (Harris, 2001) se afirma que casi todas las fórmulas o modelos estadísticos vienen en dos versiones, la heurística, que sugiere el racional bajo el procedimiento o modelo y una versión "computacional" que es algebraicamente equivalente a la versión heurística pero que puede ser útil por su rapidez o por su exactitud en los cálculos, aunque no exprese el significado del concepto. Por ejemplo, la fórmula para estimar la varianza de una población es:

$$\hat{\sigma} = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$$

Esta estadística o función de los datos es una regla heurística porque está indicando que el estimador de la varianza es aproximadamente la media aritmética de las desviaciones cuadráticas de cada elemento con respecto a la media muestral. Esto quiere decir que mediante la regla se está declarando su significado. Una versión computacional de esa misma regla es:

$$\hat{\sigma} = \frac{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}{n(n-1)}$$

La versión computacional, no hace explícito el significado del concepto, pero produce los mismos resultados de la regla heurística y es más apropiada ya que evita operar con números negativos y con fracciones decimales. Otra regla heurística será:

"sumar los valores de los grados de pertenencia" expresada como  $\sum_{i=1}^p \mu(x)$ . Este será

el estilo de reglas que se definirán en el presente trabajo de investigación, aunque también se deberán buscar sus reglas computacionales equivalentes, cuando sea el caso, por motivos de eficiencia o exactitud en los cálculos.

Generalmente, las reglas heurísticas usadas en la discriminación borrosa son basadas en el conocimiento *a priori* de las características del dominio en cuestión, a través de las definiciones de las funciones de pertenencia a los subconjuntos de una partición borrosa, sobre el dominio de un atributo o varios atributos de los objetos de interés (Branco, Evsukoff y Ebecken, 2006).

En la granulación difusa se representa cada clase etiquetada  $E_j \in Marco$  con un conjunto difuso de la forma:

$$E_j = \{(x, \mu_{E_j}(x)), x \in U\}$$

Cuando se definen las funciones de pertenencia de los objetos a las clases consideradas en un marco de cognición, con el juicio de un experto o grupos de expertos, la regla de decisión será asignar el objeto a la clase  $j$  si:

$$\mu_{E_j}(x) > \mu_{E_r}(x) \quad r = 1..k; \quad j \neq r$$

Con esta regla, se asignan todos los objetos a una sola clase. En caso de empates los objetos son asignados arbitrariamente. El valor de pertenencia  $\mu_{E_j}(x)$  corresponde al grado de encajamiento del ejemplar  $x$ , con el concepto que representa la clase etiquetada  $E_j$  (Nozaki, Ishibuchi y Tanaka, 1997; Rasmussen y Yager, 1999).

Cuando las funciones de pertenencia son desconocidas, el problema de decisión ya no es tan simple pues se debe recurrir a una técnica minería de datos, para calcular o estimar el grado de pertenencia de los objetos, de un dominio o *Universo de Discurso*  $X$ , a cada clase  $E_j \in Marco$ , con  $j = 1, k$ . Debido a esto, la discriminación o granulación borrosa debe considerar las clases asociadas a un conjunto de  $n$  observaciones, representadas por un conjunto de  $p$  atributos  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ , donde cada uno tiene como dominio a  $X_1, X_2 \dots X_p$ , respectivamente. Por lo tanto, para calcular o estimar los grados de pertenencia, de los objetos a las clases agregadas, se requiere de un procedimiento que realice la transformación o el mapeo del dominio  $X$  de la variable  $p$ -dimensional  $x_i$  al dominio o marco de cognición  $\Omega$  de las clases definidas para la variable  $Y$  de la mejor manera posible:

$$F: X \rightarrow \Omega$$

Las técnicas de discriminación borrosa basadas en la ponderación de las reglas difusas, que inicialmente fue propuesta en (Dubois, Prade y Testemale, 1988), consideran una matriz de pesos y una muestra de entrenamiento de la forma  $(x_i, y_i)$ ,

$i = 1..n$ , donde cada ejemplar  $x_i$  representa un conjunto de atributos observados o medidos y la variable  $y_i$  representa la clase asociada a ese ejemplar.

Para determinar los pesos o la influencia de cada valor de los atributos considerados en la ocurrencia de la clase, se suele recurrir al conocimiento empírico o recolectado a través de entrevistas con expertos. Dicho conocimiento se suele presentar en una tabla, donde las filas corresponden a los distintos valores de los atributos de la variable multidimensional  $x$  y las columnas corresponden a las clases  $E_j$ ; tal como se muestra en la Tabla 5.

**Tabla 5. Conocimiento obtenido de expertos**

Clase Valor	$E_1$	$E_2$		$E_p$
$a_{1i}$	$\varphi(a_{11}, E_1)$	$\varphi(a_{12}, E_2)$		$\varphi(a_{1p}, E_p)$
$\dots$	$\varphi(a_{i2}, E_1)$	$\varphi(a_{i2}, E_2)$		$\varphi(a_{i2}, E_p)$
$\dots$				
$a_{ni}$	$\varphi(a_{n1}, E_1)$	$\varphi(a_{n2}, E_2)$		$\varphi(a_{np}, E_p)$

La información presentada en la tabla anterior, se traduce en los pesos para las reglas de decisión: a mayor peso  $\varphi(a_{lm}, E_m)$ , mayor posibilidad de considerar que el valor  $a_{lm}$  de  $x_i$  pertenece a la clase  $E_m$ . Esto significa que las reglas son de la forma:

Si  $x_i$  es  $a_{ik}$ , entonces la clase es  $E_j$  con peso  $\varphi(a_{ik}, E_j)$

Los pesos para las reglas de decisión pueden asimilarse al concepto de probabilidad condicional (Branco, Evsukoff y Ebecken, 2006):

$$\varphi(a_{ik}, E_j) \approx P(E_j/x_i \text{ es } a_{ik})$$

Cuando no se tienen las probabilidades *a priori* definidas por expertos, la matriz de pesos suele construirse considerando cada peso como la frecuencia relativa de ocurrencia de cada clase  $E_j$ , en los objetos de la muestra. Después de esto, se acostumbra usar como operador de la agregación de las conclusiones parciales, para la variable  $x$  con  $p$  dimensiones, al producto u otro operador definido para la conjunción:

$$\mu_{\omega_j}(x) = \prod_{i=1}^p \mu_{\omega_j}(x_i) \approx P(\omega_j/x)$$

Esto significa que generalmente se aplica el modelo FITA, primero inferir para estimar las distribuciones marginales difusas *a priori* y luego aplicar un operador de agregación para obtener la respuesta.

# Capítulo 4

## 4 Propiedades Deseables en el Modelo Conceptual de Razonamiento

El objetivo específico fundamental en esta investigación, es elaborar un modelo conceptual del razonamiento aproximado que represente los elementos y las reglas necesarias para que un lenguaje artificial sea más parecido al natural, admitiendo vaguedad en las consultas.

El modelo conceptual que se proponga para la interpretación y operación con los términos vagos especificados en el lenguaje de consulta a sistemas de bases de datos objeto-relacionales debe cumplir un conjunto de características generales, deseables en todo modelo conceptual y otras restricciones particulares, como modelo de inferencia difusa.

Para especificar las restricciones generales que se le desean imponer al modelo, en este capítulo se procederá inicialmente con la descripción general de lo que se puede considerar como modelo conceptual en el ámbito de la Ingeniería.

### 4.1 *Los Modelos Conceptuales en la Ingeniería*

El proceso de modelado en la Ingeniería, en el nivel conceptual, consiste en aprehender el mundo exterior transformándolo en ideas y conceptos que supongan una imagen fiel del comportamiento del mundo real, con el fin de analizar su naturaleza y plantear soluciones o mejoras sobre aspectos tecnológicos.

Un modelo conceptual sirve para describir y comprender un fenómeno cualquiera expresando sus características, sus componentes y cómo se presenta o puede manifestarse, en un dominio o Universos del Discurso dado. Algunos modelos conceptuales permiten ir más allá de la descripción, pues tratan de ayudarnos en la explicación del comportamiento del fenómeno bajo estudio y en la predicción de su comportamiento futuro o bajo otras condiciones no estudiadas u observadas. Un modelo conceptual construido con fines predictivos puede ser descriptivo, pero para ello debe ser interpretable o comprensible para el ser humano. De ahí la importancia de la interpretabilidad de un modelo.

Un modelo conceptual se caracteriza por su generalidad o alto nivel de abstracción. Esto significa que un modelo conceptual de la ingeniería no incluye detalles relacionados con la tecnología o con otro tipo de recursos que se pueden emplear para suplir una necesidad o proponer una solución específica a un problema. Realizando una analogía con la música, la representación de una melodía

en un pentagrama es un modelo conceptual que puede ser aplicado múltiples veces, usando diferentes instrumentos musicales y un número variable de personas. En el pentagrama, cada símbolo presente y su posición, está expresando lo que el autor desea que se interprete como su melodía. Es un modelo conceptual comprensible para los músicos, por lo que puede ser compartido y replicado por ellos.

En la solución de problemas científicos, los modelos conceptuales usualmente se especifican mediante un lenguaje simbólico (usando números, letras y signos), en el que cada letra o signo representa un número u otra entidad matemática bien definida. Algunos de estos signos pueden ser desconocidos (las incógnitas) y deben ser determinados para resolver un problema particular. El proceso de obtener o estimar el valor de las incógnitas se conoce en algunas disciplinas como un proceso de identificación, en otras como instanciación o el ajuste de un modelo matemático teórico a un dominio particular.

El modelo teórico elegido para ajustarlo a un fenómeno dado, se elige considerando los supuestos, las restricciones y las teorías válidas en el dominio bajo estudio. El valor de las incógnitas puede derivarse de las teorías, las reglas o fórmulas, realizando un proceso de inferencia deductivo. Pero, infortunadamente, en pocas ocasiones se corre con la suerte de tener toda la información a priori requerida en un proceso de obtención de nuevo conocimiento. Esta situación conduce a hacer uso de nueva información: los hechos o las evidencias. Esto significa que para ajustar un modelo matemático es corriente realizar un proceso de inferencia no deductivo, para el descubrimiento o la aproximación al valor de las incógnitas de muchos modelos. Por lo tanto, los modelos conceptuales se construyen a partir de modelos teóricos y de la observación directa de un fenómeno en un experimento o a partir de fuentes de datos confiables, sobre los objetos de interés del dominio bajo estudio.

El ajuste o instanciación de un modelo teórico general, como el modelo de regresión, a un dominio particular da origen a un modelo conceptual y abstracto que es válido en ese contexto, en un momento determinado. Un modelo conceptual también puede instanciarse o refinarse para describir más detalles sobre las técnicas involucradas en el fenómeno que se modela, resolviendo así, otras incógnitas. El modelo generado puede considerarse un modelo lógico o de diseño porque ya tiene dependencias con la tecnología. Cuando este modelo lógico se instancia en un entorno particular, utilizando determinados recursos, se genera un modelo concreto totalmente libre de ambigüedades.

Por lo anterior, un proceso de instanciación o de ajuste a un modelo concreto, consiste en bajar de nivel de abstracción, de la generalidad a la particularidad, que es lo mismo que pasar de la incertidumbre y la vaguedad en la porción de la realidad bajo estudio, a la concreción.

Existen múltiples clasificaciones de los modelos conceptuales, dependiendo la característica en particular que se considere. Por lo tanto, para catalogar un modelo conceptual como “apropiado” se deben considerar varias características de manera simultánea, admitiendo que el aporte de cada propiedad a este concepto vago, sea diferente de acuerdo con el uso que se le vaya a dar al modelo y que el

grado de encajamiento o cumplimiento de un modelo particular a ciertas características puede ser parcial y no total.

Las características deseables en el modelo de razonamiento, en la interpretación de una consulta vaga de un sistema flexible de consulta-respuesta a bases de datos, conducen a definir las como restricciones impuestas al modelo para asegurar su calidad.

## ***4.2 Restricciones Impuestas al Nuevo Modelo de Razonamiento***

Se ha reconocido que una técnica basada en restricciones es un medio efectivo para la optimización o el mejoramiento de cualquier sistema, más cuando se dispone de información imperfecta (Apt, 2003). Es por ello, que en este trabajo se planteó como un objetivo específico el establecimiento a priori de una serie de restricciones o propiedades deseables en el modelo genérico de razonamiento para la interpretación de la vaguedad en las consultas.

El modelo de razonamiento aproximado basado en los datos que aquí se proponga para la interpretación y operación con términos vagos especificados en el lenguaje de consulta, debe cumplir un conjunto de características generales, deseables en todo modelo conceptual. Además, debe cumplir otras restricciones particulares, como modelo de inferencia difusa para que se pueda caracterizar por su consistencia lógica y capacidad de interpretación.

Infortunadamente, no existe un conjunto de restricciones universalmente aceptadas para caracterizar los modelos de inferencia difusos. Tal variabilidad es debida, en parte, al juicio subjetivo que motiva la inclusión de cada restricción en un proyecto particular.

No obstante lo anterior, las restricciones que más adelante se sobre el sistema de inferencia difuso han sido extractadas de otras propuestas de investigación en el marco de la granulación difusa en otros entornos diferentes a los sistemas de bases de datos (Frigg y Hartmann, 2006; Pradera et al., 2006; Hajek, 2005; Amo et al., 2004; Peña-Reyes y Sipper, 2003; Espinosa y Vandewalle, 2000; Jamei et al., 2001; Jiménez et al., 2001).

## ***4.3 Restricciones generales sobre el nuevo modelo conceptual***

### ***4.3.1 Restricción 1. Confiabilidad***

Los sistemas interactivos de consulta-respuesta requieren que el usuario pueda confiar con que cuenta con un colaborador experto que le va a ayudar a encontrar la información que sea más compatible con las restricciones declaradas vagamente en su petición. Y como la vaguedad aquí tratada puede ser variable en el espacio o en el tiempo, para lograr un modelo de razonamiento confiable en la interpretación de las solicitudes de los usuarios, se requiere que se adapte dinámicamente al contexto lingüístico de cada consulta.

Por otro lado, un modelo de inferencia o de razonamiento se puede considerar consistente, si en él no se especifican contradicciones y siempre produce las mismas respuestas cuando se tenga la misma información disponible sobre un contexto determinado. Por esto, los conjuntos difusos que representen los términos vagos de la consulta, no deben depender de los juicios de expertos.

De lo expuesto, el modelo de razonamiento propuesto se considerará confiable, no sólo si es adaptable a los distintos contextos lingüísticos, sino consistente.

#### **4.3.2 Restricción 2. Extensibilidad**

Esta característica se deduce de lo que el modelo propuesto extiende o generalice, en los modelos previos, así como también lo que permitirá extender en el futuro. El primer aspecto está relacionado con la incorporación de características adicionales, buscando completitud o casos no cubiertos por otras propuestas para la determinación de la semántica de las consultas vagas.

Como se expresó antes, en las propuestas estudiadas para la flexibilización del lenguaje de consulta, no se consideran términos vagos expresados como una combinación lineal de otros términos. Tampoco consideran las preguntas de tipo IV del lenguaje teórico PRUF de forma que un sistema interactivo de consulta-respuesta pueda responder con calificadores de la verdad (ver página 39). Por lo tanto, al resolver estos casos, el presente trabajo de investigación estará extendiendo modelos previos.

No obstante, se debe validar que las extensiones propuestas sean estrictamente aditivas. Es decir, que el modelo que se toma como el original, el lenguaje de consulta a bases de datos, debe poder ser considerado un caso restringido o más específico que el lenguaje que se derive de este trabajo investigativo.

Por otro lado, para determinar la capacidad de evolución o extensibilidad futura del modelo de razonamiento aproximado inmerso en el nuevo lenguaje, debe caracterizarse por su simplicidad para realizar los cambios o adiciones futuras.

#### **4.3.3 Restricción 3. Generalidad**

El modelo deberá ser válido para cualquier dominio de aplicación. Este dominio puede ser un centro de investigación molecular, una empresa productora de zapatos, una agencia de bienes raíces, entre muchos otros. También debe ser independiente del sistema gestor de bases de datos. Debe ser lo mismo implementarlo en Oracle o en PostgreSQL, por ejemplo.

#### **4.3.4 Restricción 4. Corrección**

Aunque se admite que el modelo de razonamiento no puede ser exacto, por la incertidumbre generada por la vaguedad y por requerir de resúmenes de datos para realizar las inferencias, debe ofrecer una buena aproximación al verdadero significado de la vaguedad especificada en la consulta. Los indicadores del

cumplimiento de esta característica serán su corrección interna y externa. La corrección interna del modelo de razonamiento estará determinada por la estructura de la lógica del razonamiento para hallar la semántica de cada término vago especificado en la consulta, según el contexto. Esto significa que el proceso de inferencia debe estar sujeto a las restricciones establecidas para los sistemas de inferencia que más adelante se detallarán.

La corrección externa, por otro lado, tiene que ver con los aspectos sintácticos o gramaticales. Para velar por la corrección externa, es indispensable que las extensiones propuestas para el lenguaje de consulta se ajusten a reglas gramaticales del lenguaje que trata de generalizar. Tampoco deben obligar al usuario a emplear una terminología desconocida para él, cuando formula sus consultas. Es decir, se requiere que el modelo sea “caja-negra” para el usuario final, aunque debe ser “caja-blanca” para el personal informático.

#### ***4.3.5 Restricción 5. Riqueza Semántica o Potencia Expresiva***

Un modelo es rico semánticamente, si ofrece una amplia gama de conceptos para cubrir la mayor cantidad de aspectos relevantes del dominio del problema. En un modelo de razonamiento, la riqueza semántica o potencia expresiva se ve reflejada en la cantidad y variedad de términos vagos a los que el propio sistema es capaz de encontrarle significado, de manera autónoma.

Para validar la potencia expresiva del lenguaje propuesto, se tomará como referencia al lenguaje teórico PRUF. Esto porque que ese lenguaje ha servido de base teórica para propuestas recientes, aplicables en sistemas difusos de bases de datos (Gongçalves y Tineo, 2006; Kacprzyk and Zadrozny, 2001; Galindo et. al., 1998; Bosc y Pivert, 1995; Medina, Pons y Vila, 1994). Por lo tanto, para evaluar si el modelo propuesto cumple esta restricción, se basará en el cumplimiento de las reglas de interpretación, agrupadas en las cuatro categorías que ya fueron descritas en el Capítulo 3.

#### ***4.3.6 Restricción 6. Formalidad o Rigor***

Un modelo es formal, si cada concepto tiene una interpretación única y bien definida. Esta cualidad determina que un modelo de comportamiento de un sistema pueda ser replicable por otros para llegar a resultados parecidos o pueda ser compartido con la comunidad científica para su validación. Además, determina la capacidad de traducción o transformación de un modelo a otro de más bajo nivel de abstracción y la determinación de la traza de ese proceso.

#### ***4.3.7 Restricción 7. Robustez.***

Se espera de un sistema de inferencia que sus respuestas o interpretaciones de la vaguedad no se vean afectadas por ligeras variaciones en los datos o por valores atípicos. La robustez de un modelo de inferencia difusa es determinada, generalmente, por la propiedad de continuidad de los operadores empleados para representar la semántica de las expresiones vagas. Pero adicionalmente, para tener

un modelo de inferencia robusto, se requiere que el modelo de razonamiento sea insensible a las múltiples formas de distribución de las variables u objetos de interés, en el contexto delimitado por una consulta.

#### ***4.4 Restricciones sobre el sistema de inferencia***

Hace ya buen tiempo Hobbs propuso una teoría de la granulación o discriminación difusa aplicable a los sistemas de inteligencia artificial, observando que el humano mira al mundo bajo diferentes niveles de granularidad y abstrae de allí las cosas que le sirven (Hobbs, 1985). Allí se afirma que nuestra inteligencia y flexibilidad depende de la capacidad de conceptuar sobre el mundo, con diferentes niveles de granularidad y de cambiar fácilmente de granulaciones o marcos de referencia.

Una capacidad parecida debe caracterizar al sistema de inferencia difusa que se proponga para encontrar los significados válidos de los términos vagos expresados en una consulta. Por esto, la restricción principal impuesta al sistema de inferencia borrosa es la flexibilidad para cambiar de marcos de cognición o de niveles de granularidad, dinámicamente. Para lograr esa flexibilidad, el proceso de concreción de la vaguedad demandará la identificación o el ajuste de un modelo de la distribución de los grados de pertenencia de los objetos de interés, por medio de resúmenes o estadísticas calculadas con los datos disponibles, en la base de datos.

Independientemente que un proceso de identificación o de ajuste de un modelo para la discriminación de los objetos, sea llevado a cabo por las personas o por las máquinas, se debe buscar que el modelo ajustado minimice la pérdida de información proporcionada por los datos individuales, considerando un número suficiente de parámetros para la descripción de la semántica de las etiquetas lingüísticas.

Un estimador es una estadística o función aplicable sobre un conjunto de datos u observaciones para resumir o describir de una manera condensada los datos o realizar inferencias (Lohninger, 1999). Para cada parámetro pueden existir varios estimadores diferentes, como ocurre con la tendencia central de una distribución de datos que puede estimarse con la media aritmética de los valores individuales, o con la mediana o la moda, entre otras estadísticas. En general, se elige el estimador que posea mejores propiedades que los restantes, como el insesgamiento, la eficiencia, la suficiencia y la robustez (Harris, 2001). Por lo tanto, se deberán considerar dichas propiedades cuando se definan las técnicas de discriminación difusa que deban llevarse a cabo por el sistema de inferencia difusa para la interpretación de los términos vagos en las condiciones de la consulta, según el contexto.

Por otro lado, cuando se definió la restricción número 3, sobre corrección de un modelo, se mencionó que el sistema de inferencia difusa debía caracterizarse por su corrección interna, la cual es determinada por la estructura lógica del razonamiento en la interpretación de los términos vagos. Por lo tanto, la lógica subyacente en las técnicas de discriminación difusa para hallar la semántica de los términos vagos, deben cumplir con las propiedades o restricciones que se mencionan

seguidamente. Pero antes de eso, debe señalarse que las restricciones planteadas en la literatura sobre la granulación difusa no todas tienen un soporte matemático, sino psicológico o guiado por el sentido común. De ahí, que no haya un acuerdo universal sobre las restricciones que deba cumplir una granulación o partición difusa de modo que se garantice su interpretabilidad.

Distintos autores han acogido técnicas basadas en restricciones con el fin de lograr la interpretabilidad de los gránulos de información (Mencar, 2004, Casillas et al., 2003; Guillaume, 2001). En el presente trabajo, de las restricciones estudiadas para el proceso para la obtención de los gránulos de información, se consideraron las propiedades o restricciones que se presentan a continuación.

#### ***4.5 Restricciones sobre los Marcos de Cognición***

Para construir gránulos de información o modelos interpretables de los términos vagos, un marco de cognición debe estar bien estructurado. Por esto, debe cumplir con las restricciones siguientes.

##### ***4.5.1 Restricción 8. Ordenamiento Apropiado***

Un marco de cognición es una tripleta  $\langle U, \mathcal{F}, \prec \rangle$ . De acuerdo con esta definición, el orden  $\prec$  o la ubicación espacial de los conjuntos difusos en el marco es relevante. Un conjunto borroso del universo  $U$  con etiqueta "baja" debe estar situado, en el marco de cognición  $\mathcal{F}$ , primero que el que representa la clase de los "medianos" y éste, antes que el conjunto con la etiqueta "alta" cuando los conjuntos considerados sean tres, por ejemplo.

##### ***4.5.2 Restricción 9. Número de Elementos Justificable***

El número de conjuntos difusos en el marco de cognición no debería ser muy alto, preferiblemente menor a  $7 \pm 2$ . Este número es sugerido por experimentos psicológicos (Miller, 1956) y considerado en un buen número de artículos relacionados con el modelado difuso interpretable (Peña-Reyes y Sipper, 2003; Jiménez et al., 2001; Jamei et al., 2001; Espinosa y Vandewalle, 2000). Esta restricción se puede decir que se cumple de manera natural puesto que la cantidad de conjuntos en un marco cognición corresponde al número de elementos del conjunto de términos de una variable lingüística y no se esperaría que se pudieran identificar, en cualquier caso, gran cantidad de etiquetas para una variable de este estilo.

##### ***4.5.3 Restricción 10. Gránulos Distinguibles***

Cualquier conjunto difuso en un marco de cognición debe ser distinguible de los restantes. Los conjuntos completamente disjuntos son fácilmente distinguibles pero las situaciones reales pueden hacer que los conjuntos no tengan esa característica, como cuando representan conceptos vagos. Intuitivamente se espera que los modelos de los conjuntos borrosos que representen las etiquetas lingüísticas

no se solapen demasiado para no perder poder diferenciador entre dos o más clases, en el marco de cognición.

#### 4.5.4 Restricción 11. No a tres, al tiempo

Cualquier elemento del *Universo de Discurso* no posee más de dos grados de pertenencia mayores que cero, a los distintos conjuntos en el marco cognitivo:

$$\forall x \in U : |\{A \in Marco : \mu_A(x) > 0\}| \leq 2$$

#### 4.5.5 Restricción 12. Cobertura (completitud)

Un marco de cognición debe ser completo. Es decir, cada elemento del *Universo de Discurso* debe pertenecer, como mínimo, a un conjunto difuso del marco:

$$\forall x \in U \exists A \in Marco : \mu_A(x) > 0$$

#### 4.5.6 Restricción 13. Complementariedad (criterio $\Sigma$ )

Para cada elemento en el *Universo de Discurso*, todos los valores de membresía del elemento a los conjuntos marco de cognición deben sumar 1:

$$\forall x \in U : \sum_{A \in Marco} \mu_A(x) = 1$$

El criterio de complementariedad asegura que la división del *Universo de Discurso* en clases o categorías, corresponde a una partición de Ruspini, conocida con este nombre gracias a su autor (Ruspini, 1969). En este tipo de partición se cumple la propiedad de cobertura del marco de cognición y se admite que un objeto  $x$  puede pertenecer a clases distintas, pero el grado total de membresía está distribuido entre las clases. En (Hermann, 1997) se motiva al cumplimiento de la complementariedad para lograr el cumplimiento de las dos restricciones siguientes, que se generan como su consecuencia directa.

#### 4.5.7 Restricción 14. Aceptación Completa

Cuando un valor de pertenencia disminuye de uno, otro valor se incrementa de cero.

$$\forall x \in U \forall A \in Marco \exists B \in Marco | \{A\} : \mu_A(x) < 1 \rightarrow \mu_B(x) > 0$$

Desde un punto de vista semántico, la aceptación completa implica que cuando un concepto no representa totalmente a un elemento, habrá otro que lo pueda representar mejor, pero se garantiza que el grado de pertenencia de un elemento siempre pertenezca con grado uno, al dominio considerado.

#### 4.5.8 Restricción 15. Normalidad

Un conjunto difuso  $A$  debe ser normal. Es decir, en él existe al menos un elemento (denominado prototipo) cuya pertenencia a cualquier conjunto  $A$  en el marco de cognición sea completa.

$$\exists x \in U \forall A \in \text{Marco} : \mu_A(x) = 1$$

La restricción de normalidad implica que por lo menos un elemento de  $U$  debería exhibir un encajamiento total con el concepto representado por el conjunto difuso (Peña-Reyes y Sipper, 2003; Espinosa y Vandewalle, 2000).

Según Joslyn la normalización de un conjunto difuso  $A$  puede considerarse un mapeo de conjuntos difusos a distribuciones de posibilidades (Joslyn, 1994):

$$\text{Normalizar}(A) \mapsto \Pi(A)$$

Ese mismo autor también establece una relación de equivalencia entre una función de pertenencia  $\mu$  y la distribución de probabilidades  $p$ , si se cumple la propiedad de complementariedad. De acuerdo con esto, se puede concluir que un conjunto difuso podría estar definido mediante una distribución de posibilidades o de probabilidades, dependiendo de las propiedades de  $\mu$ . Lo que además confirma que las distribuciones de probabilidades y posibilidades son casos especiales de la función de pertenencia (Joslyn, 1994). Estas relaciones se presentan para hacer uso de ellas si son requeridas en la discriminación de los grupos de objetos, de acuerdo con el contexto y el tipo de vaguedad involucrada.

#### 4.5.9 Restricción 16. Convexidad

Un conjunto borroso debe ser convexo. La convexidad de un conjunto difuso se puede considerar como un complemento a la normalidad, pues asegura que el concepto representado por ese conjunto va perdiendo evidencia, a medida que los elementos se distancian de los prototipos (Mencar, 2004):

$$\forall a, b, x \in U : a \leq x \leq b \rightarrow \mu_A(x) \geq \min\{\mu_A(a), \mu_A(b)\}$$

La convexidad de un conjunto es estricta si en la fórmula anterior se cambia el símbolo “ $\geq$ ” por la desigualdad estricta “ $>$ ”. La convexidad estricta se define como una restricción necesaria para los números y vectores de números difusos que representan cantidades etiquetadas como “Aproximadamente  $x$ ”, siendo  $x$  un número escalar, puesto que la evidencia que una cantidad tome un valor específico debe disminuir monótonamente, a medida que la distancia a ese valor aumenta.

Técnicas tan conocidas como el algoritmo Fuzzy C-Means para el proceso de granulación difusa, no siempre dan origen a conjuntos difusos convexos (Bezdek, 1981). Por lo que se requeriría de ciertas transformaciones complejas para obtener gránulos convexos y, por lo tanto, interpretables como se propone en (Peña, 2001; Mencar, 2004).

#### 4.5.10 Restricción 17. Unimodalidad

Un conjunto difuso que represente a un escalar o a un vector de escalares debería ser unimodal (Mencar, 2004). Es decir, debe existir uno y sólo un elemento cuyo grado de pertenencia sea máximo:

$$\exists p \in U : \mu_A(p) = \max \mu_A(x) \forall x \in U \wedge \forall q \neq p \in U : \mu_A(q) < \mu_A(p)$$

Esta última restricción no es obligatoria para todos los conjuntos borrosos pues se espera que éstos contengan más de un prototipo y no uno solamente. Por esto, la restricción de una sola moda se necesita como requerimiento exclusivo de los números difusos.

Con esta última restricción quedan especificadas todas las restricciones a las cuales debe ajustarse el sistema de inferencia difusa en este trabajo de investigación, para representar y operar con etiquetas lingüísticas o términos vagos dependientes del contexto, expresados en las condiciones de una consulta. Todas ellas tiene como objetivo asegurar que la estructura lógica del la máquina de inferencia sea sólida y consistente con una teoría de conjuntos difusos.

# Capítulo 5

## 5 El Modelo Conceptual del Razonamiento

En este capítulo, se describen los trabajos relacionados para la extensión del lenguaje de bases de datos con el objeto de admitir consultas vagas, donde se destacarán las bondades y las limitaciones detectadas en ellos.

Luego de la presentación de los trabajos relacionados, se presenta el modelo conceptual del razonamiento aproximado y adaptable que aquí se propone, ajustado a las restricciones descritas en el Capítulo anterior o a las impuestas por la definición misma de los conceptos utilizados.

### 5.1 Trabajos Relacionados con la Flexibilización del Lenguaje SQL

Los trabajos de investigación relacionados con el manejo de la vaguedad en bases de datos se suelen clasificar, de acuerdo con los problemas que abordan, en dos categorías principales: los que se ocupan de la aspectos estructurales para la representación y manejo de tipos de datos expresados en forma vaga y, por otro lado, los trabajos que abordan la problemática del procesamiento de las consultas con términos vagos, pero con datos concretos o no difusos en la base de datos. Como en esta última categoría se clasifica el presente trabajo, se describirán de algunas de las propuestas más reconocidas en el ámbito mundial, entre ellas las dos más consolidadas para la flexibilizar el lenguaje de consultas para la interacción con sistemas de bases de datos como son FSQL y SQLf (Urrutia, Tineo y González, 2008).

Desde hace ya buen tiempo existen varias propuestas para abordar el problema de la vaguedad en las consultas. En (Tahani, 1977) ya se propone un método basado en la Lógica Difusa para determinar el grado de cumplimiento de las condiciones de filtrado de las consultas. Esta propuesta estaba orientada al lenguaje de manipulación de datos más reconocido en ese entonces: el SEQUEL. Para ilustrar un poco la propuesta de Tahani, supóngase que la relación de empleados tiene el siguiente esquema:

Empleado(identificación, nombre, fecha\_nacimiento, salario, fecha\_ingreso)

De acuerdo con esta relación, una pregunta vaga como ¿cuáles de los empleados “jóvenes” o de los “recientemente vinculados” devengan un salario “alto”? se plantearía, según Tahani, como aparece a continuación.

```
SELECT nombre
FROM empleado
WITH (fecha_nacimiento = "joven" OR fecha_ingreso = "recién
vinculado") AND salario= "alto"
```

Para resolver esa pregunta, se propone hacer uso de los grados de pertenencia de cada tupla  $t$  a cada uno de los atributos borrosos. Y para encontrar el grado de pertenencia global a las condiciones especificadas en la consulta se propone como operador para la conjunción al mínimo valor de pertenencia de los operandos y para la disyunción al máximo valor de pertenencia de los operandos, acorde con la teoría estándar. Por eso, para el ejemplo, la fórmula de cálculo sería así:

$$\mu_{condición\ global}(t) = \min(\max(\mu_{joven}(t), \mu_{recien\ vinculado}(t)), \mu_{alto}(t))$$

Debe observarse que para poder calcular el valor de pertenencia global es necesario definir de antemano, los valores de pertenencia marginales de cada tupla a cada categoría difusa establecida por los términos vagos presentes en la fórmula de la consulta. Tahani sugiere hacerlo explícitamente dentro de la base datos. Esto hace que sea necesario extender el modelo de datos (las relaciones existentes en la base de datos) creando nuevas columnas para poder especificar el grado de pertenencia de cada tupla, a cada uno de los términos vagos posibles.

Posteriormente, el lenguaje teórico PRUF (Zadeh, 1978) da origen a trabajos que se orientan particularmente a extender el lenguaje de bases de datos SQL (acrónimo de Structured Query Language). Uno de estos, es el lenguaje FSQL propuesto inicialmente por Medina en 1994 y, luego, redefinido e implementado por Galindo, donde se incorporan además de las etiquetas lingüísticas, algunos operadores o comparadores vagos que permiten extender el lenguaje estándar SQL, admitiendo vaguedad en las consultas (Galindo, Urrutia y Piattini, 2005).

La sintaxis de una consulta vaga en FSQL, para una pregunta con una condición vaga simple es:

```
SELECT proyección
FROM relaciones
WHERE atributo comparador etiqueta [THOLD umbral]
```

La palabra THOLD es la abreviatura usada por FSQL para establecer el umbral de cumplimiento mínimo (threshold, en inglés) que el usuario especifica de manera opcional. Dicho umbral es un alfa corte, por lo que debe ser un valor en el intervalo (0,1]. Un ejemplo de una consulta con la sintaxis de FSQL es:

```
SELECT * FROM Personas
WHERE estatura FEQ $alto THOLD 0.5
```

Los comparadores lógicos que propone FSQL están basados en los conceptos de necesidad que es una medida de creencia y la posibilidad que es una medida de incertidumbre, definidos por Zadeh y se muestran en la Tabla 6.

**Tabla 6. Comparadores difusos de FSQL**

Comparador de Posibilidad	Comparador de Necesidad	Significado
FEQ o F=	NFEQ o NF=	Fuzzy EQual (Posiblemente/Necesariamente Igual)
FDIF, F!= o F<>	NFDIF, NF!= o NF<>	Fuzzy DIFferent (Posiblemente/Necesariamente Diferente)
FGT o F>	NFGT o NF>	Fuzzy Greater Than (Posiblemente/Necesariamente mayor que)
FGEQ o F>=	NFGEQ o NF>=	Fuzzy Greater or Equal (Posiblemente/Necesariamente mayor o igual a)
FLT o F<	NFLT o NF<	Fuzzy Less Than (Posiblemente/Necesariamente menor que)
FLEQ o F<=	NFLEQ o NF<=	Fuzzy Less or Equal (Posiblemente/Necesariamente menor o igual a)
MGT o F>>	NMGT o NF>>	Much Greater Than (Posiblemente/Necesariamente mucho mayor que)
MLT o F<<	NMLT o NF<<	Much Less Than (Posiblemente/Necesariamente mucho menor que)

En esta consulta, la etiqueta lingüística “alto” se distingue por estar precedida del signo \$. Un aspecto novedoso en la nueva versión de FSQL para la formulación de una consulta, consiste en que se puede especificar un umbral de cumplimiento (threshold, en inglés) para cada condición simple. Por ejemplo, en la consulta siguiente, se especifican dos umbrales:

```
SELECT * FROM Personas
WHERE cabello FEQ $Rubio THOLD 0.5
AND Edad FGT $Joven THOLD 0.8;
```

En FSQL para el cálculo el grado de cumplimiento de cada tupla a una condición global, se utiliza una función denominada CDEG(\*). La función CDEG se basa, por defecto, en los operadores típicos para la negación (1-x), conjunción (t-norma del mínimo) y disyunción (s-norma del máximo), pero pueden usarse otros operadores, si son definidos por el diseñador de la base de datos.

Adicionalmente, FSQL admite cuantificadores borrosos absolutos o relativos que se especifican como una restricción a una condición grupal en la sentencia SELECT del SQL tradicional. Por eso, ofrecen dos tipos de condiciones en la cláusula HAVING que sigue a una cláusula GROUP BY.

Otras de las propuesta más reconocidas en la tecnología de Bases de Datos para extender y flexibilizar el lenguaje SQL es el lenguaje SQLf propuesto por Bosc y Pivert. En este lenguaje no se incluyen operadores nuevos para extender el SQL, sino que el operador IS se sobrecarga (overriding, en inglés) para la evaluación semántica requerida (Bosc y Pivert, 1995). De acuerdo con la propuesta de estos autores una consulta se especifica mediante la sintaxis que se muestra enseguida.

```

SELECT {n|t|n,t} proyección
FROM relaciones
WHERE atributo comparador etiqueta

```

Donde  $n$  es el número máximo deseado de respuestas y  $t$  un umbral cualitativo. Una tupla estará en la respuesta si el grado de pertenencia de la clase derivada por la condición difusa, es mayor o igual al umbral  $t$ . Cuando no se especifica el umbral se considera un valor por defecto: 0.5.

En SQLf es posible hacer particiones difusas usando la cláusula GROUP BY acompañada con su operador de restricción HAVING, como lo admite FSQL, así como el uso de modificadores, cuantificadores y comparadores difusos.

Posteriormente, el lenguaje SQLf fue extendido para dar soporte a las características introducidas el segundo estándar SQL:1992 y posteriormente, para ajustarlo a las características definidas en el tercer estándar SQL:1999 que da el soporte al modelo objeto-relacional y de ahí su nombre SQLf3 (Gonçalves y Tineo, 2001). En estas extensiones la sintaxis del SQLf fue ligeramente modificada para formular una consulta buscando la ortogonalidad del modelo. La mejora que sufre consiste en especificar las condiciones para seleccionar las “mejores” tuplas después de las condiciones vagas, en la cláusula WHERE, en lugar de especificarlas en la cláusula SELECT:

```

SELECT proyección
FROM relaciones
WHERE atributo comparador etiqueta
WITH CALIBRATION {n|λ|n,λ}

```

Usando el lenguaje SQLf3, una consulta para mostrar las ciudades cuya temperatura se pueda considerar “alta” con un umbral de 0.5, se formularía así:

```

SELECT nombre FROM ciudades
WHERE temperatura IS alta WITH CALIBRATION 0.5

```

En las extensiones de SQLf, las definiciones de los conjuntos borrosos son proporcionadas por el usuario, por medio del sublenguaje DDL (Data Definition Language) y por esto, se evita la intervención del diseñador o administrador de la base de datos (Urrutia, Tineo y González, 2008).

En SQLf3 se pueden definir términos lingüísticos como: predicados, modificadores, conectivas, comparadores y cuantificadores, cada uno de ellos con su sintaxis particular. Para ello, se usa una declaración con la sintaxis:

```

CREATE FUZZY <definición del término>

```

Por ejemplo, en la sentencia siguiente, el predicado difuso “muchos” se define sobre números enteros en el dominio entre 0 y 10 y el grado de cumplimiento de este predicado queda determinado por la función trapezoidal con parámetros (7, 8, 10, 10) propuesta en (Zadeh, 1965):

```

CREATE FUZZY PREDICATE Muchos ON 0..10 AS (7, 8, 10, 10)

```

Hay dos aspectos destacables en SQLf y sus extensiones posteriores. El primero de ellos tiene que ver con la manera de incorporar los operadores o predicados difusos en la sintaxis, pues no se afecta la sencillez del lenguaje SQL, haciéndolo efectivamente más parecido al lenguaje natural que el FSQL. El segundo aspecto destacable es que no se altera el modelo de datos, sino el metamodelo de los datos (o los metadatos) para poder hacer el mapeo entre términos lingüísticos y valores numéricos.

Además de estas propuestas también existen otras como las descritas en (Kackrpyk y Zadrozny, 2001) y en (Ma y Wang, 2006) que mejoran la capacidad expresiva del lenguaje SQL, pero que también se caracterizan por la subjetividad para definir modelos de los conjuntos difusos que representan los términos vagos, bien sea porque éstos sean dados por el diseñador o administrador de la base de datos o por quien formula la consulta.

La debilidad de la subjetividad en la definición de la forma de los conjuntos borrosos que representan los términos lingüísticos, consiste en que un sistema de inferencia puede llegar a distintas respuestas ante una misma consulta, a pesar de tener la misma información disponible, generando posibles inconsistencias y por ende, restándole confiabilidad. La subjetividad en la asignación de los grados de pertenencia o en la elección de una forma de distribución de éstos, sólo sería admisible en los casos en que se tenga poca o ninguna información sobre un fenómeno dado como ocurre con las probabilidades subjetivas y éste no es el caso de un sistema de bases de datos.

Adicionalmente, al definir unos valores específicos y únicos para los grados de pertenencia, ya sea por el diseñador de una base de datos o por el propio usuario, no se está considerando que los términos vagos sean relativos o dependientes del contexto y que éste sólo se establece en el momento de la consulta. Para poder afirmar, por ejemplo, que un salario es “bajo” siempre se deben considerar los salarios del grupo de sujetos tomados como base de comparación y el momento en el tiempo en que se hace la consulta. Igualmente, una nota considerada “alta” en una asignatura no necesariamente debe considerarse así en otras asignaturas, incluso si son de una misma institución. Por esto, un sistema de inferencia que no tenga en cuenta el contexto y el momento histórico, no podrá capturar adecuadamente el significado de los términos vagos para ofrecerle al usuario respuestas confiables.

Una alternativa diferente a la asignación subjetiva de los grados de pertenencia de valores numéricos a conjuntos borrosos fue abordada en (Roux y Desachy, 1997) para el proceso de reconocimiento de patrones en imágenes satelitales. Ellos hicieron uso de la relación entre posibilidades y probabilidades establecidas en (Dubois y Prade, 2000), para la determinación de las posibilidades a partir de las probabilidades de ocurrencia de cada uno de los eventos, basándose en los datos disponibles, se usa la fórmula siguiente.

$$\Pi(E_i) = \sum_{j=1}^n \min(P(E_i), P(E_j)) \forall i \neq j$$

Donde  $\Pi$  representa la posibilidad y  $P$  la probabilidad de ocurrencia del conjunto etiquetado  $E_i$ , en un marco de cognición determinado, y  $n$  es el número de intervalos o categorías usadas en la construcción del histograma de frecuencias a partir de los datos disponibles. La fórmula propuesta significa que la posibilidad de un evento resulta de la sumatoria de la comparación de la probabilidad de ocurrencia del evento con la probabilidad de los demás eventos, escogiendo en cada caso, el valor mínimo.

Zapata (2002) también acoge la relación entre posibilidades y probabilidades establecida por Dubois y Prade para extender el lenguaje SQL con términos vagos. La propuesta de Zapata permite asociar atributos numéricos con términos vagos para admitir las reglas de tipo I y tipo II definidas en PRUF, ya antes descritas, permitiendo componer el criterio de búsqueda con las conectivas convencionales de la conjunción, la disyunción y la negación. Para la interpretación de los términos vagos sugiere un motor de inferencia que siga los pasos siguientes, en el momento de una consulta, usando los datos disponibles:

- a) La construcción de las distribuciones de frecuencia, usando un número fijo de diez intervalos.
- b) El cálculo de las posibilidades usando la fórmula de Dubois y Prade.
- c) Encontrar la distribución de posibilidades en forma acumulada para representar al conjunto de los valores mayores como “alto” o “grande”.
- d) Normalizar las posibilidades acumuladas para ninguna de ellas sea mayor a la unidad, dividiendo cada una por el total acumulado.
- e) Encontrar el valor en el *Universo de Discurso* cuya posibilidad sea mayor o igual al umbral mínimo especificado, por medio de la interpolación, para determinar cuáles tuplas satisfacen la etiqueta lingüística que representa la clase de los valores mayores. Este umbral parte en dos la distribución de posibilidades y, por defecto, su valor es 0.5.
- f) Hacer uso del concepto de concentración para la interpretación de términos con el calificativo “muy”. Para el cálculo de “muy alto”, por ejemplo, se usaría la fórmula  $CON(\mu_{alto}(x)) = \mu^2_{alto}(x)$ .
- g) Si se trata de la interpretación de términos como “bajo” o “poco”, se hace uso de la negación así:  $\mu_{bajo}(x) = \neg\mu_{alto}(x) = 1 - \mu_{alto}(x)$ .

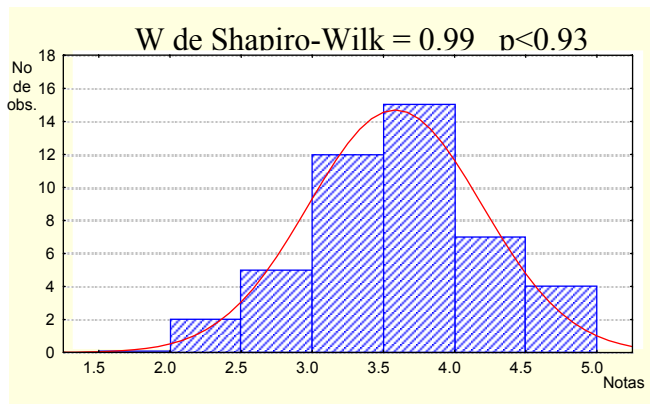
Para ilustrar el proceso de cálculo que debe llevar a cabo el motor de inferencia para hallar el significado de términos vagos, suponga que se pregunta por los estudiantes que obtuvieron una nota “alta” en un curso cualquiera. Para cumplir el primer paso del proceso, se subdivide el rango total de las notas definitivas de ese curso en intervalos de clase, pero para el ejemplo no se usarán los diez intervalos sugeridos, sino seis buscando simplicidad puesto que la cantidad de intervalos no es lo relevante del proceso.

**Tabla 7. Distribución relativa de las notas**

Intervalo de la calificación	Cantidad de elementos	Frecuencia relativa
$2.0 \leq x < 2.5$	2	4.4%
$2.5 \leq x < 3.0$	5	11.1%
$3.0 \leq x < 3.5$	12	26.6%
$3.5 \leq x < 4.0$	15	33.3%
$4.0 \leq x < 4.5$	7	15.6%
$4.5 \leq x < 5.0$	4	8.9%

A partir de 45 notas hipotéticas obtenidas por estudiantes en el curso X, se determina la cantidad de notas que caen dentro de cada intervalo definido y las correspondientes frecuencias relativas que se muestran en la Tabla 7. El paso siguiente de la técnica, consiste en encontrar las posibilidades acumuladas, lo cual hace necesario estimar las probabilidades de los eventos definidos por cada intervalo de clase y por lo tanto, se debe hacer un ajuste de los datos a una distribución teórica de probabilidad.

Realizando la prueba de bondad del ajuste a la distribución normal, basada en el estadístico W de Shapiro-Wilk, no se encontró alguna evidencia para rechazar normalidad. El histograma y el ajuste a la distribución normal, se muestran en la Figura 9, realizado con la ayuda de un paquete estadístico.



**Figura 9. Histograma de frecuencias para las notas definitivas.**

Por lo anterior, el modelo probabilista elegido para representar la frecuencia de ocurrencia de los eventos, es una distribución normal con media 3.6 y desviación estándar de 1.1. Los valores derivados de esta distribución para cada uno de intervalos, aparecen en la tabla siguiente.

**Tabla 8. Distribución de probabilidades de las notas**

Intervalo de clase	Cantidad de elementos	Frecuencia relativa	Probabilidad según la normal ( 3.6, 1.1)
$2.0 \leq x < 2.5$	2	4.4%	0.03
$2.5 \leq x < 3.0$	5	11.1%	0.13
$3.0 \leq x < 3.5$	12	26.6%	0.28
$3.5 \leq x < 4.0$	15	33.3%	0.31
$4.0 \leq x < 4.5$	7	15.6%	0.18
$4.5 \leq x < 5.0$	4	8.9%	0.06

Una vez obtenidas las probabilidades correspondientes a cada intervalo de clase se puede aplicar la fórmula para el cálculo de las posibilidades, según la propuesta de Dubois y Prade así:

$$\Pi(E_1) = 0.03 + 0.03 + 0.03 + 0.03 + 0.03 = 0.15$$

$$\Pi(E_2) = 0.03 + 0.13 + 0.13 + 0.13 + 0.06 = 0.48$$

$$\Pi(E_3) = 0.03 + 0.13 + 0.28 + 0.18 + 0.06 = 0.68$$

$$\Pi(E_4) = 0.03 + 0.13 + 0.28 + 0.18 + 0.06 = 0.68$$

$$\Pi(E_5) = 0.03 + 0.13 + 0.18 + 0.18 + 0.06 = 0.58$$

$$\Pi(E_6) = 0.03 + 0.06 + 0.06 + 0.06 + 0.06 = 0.27$$

Siguiendo con el proceso, es necesario encontrar las posibilidades acumuladas. Como son valores mayores que la unidad, se deben normalizar dividiendo cada valor por el valor total acumulado, que en este caso es 2.84 como se aprecia en la Tabla 9.

**Tabla 9. Posibilidades para las agrupaciones de las notas**

Intervalo de clase	Posibilidad	Posibilidades Acumuladas sin normalizar	Posibilidades Acumuladas normalizadas
$2.0 \leq x < 2.5$	0.15	0.15	0.05
$2.5 \leq x < 3.0$	0.48	0.63	0.22
$3.0 \leq x < 3.5$	0.68	1.31	0.46
$3.5 \leq x < 4.0$	0.68	1.99	0.70
$4.0 \leq x < 4.5$	0.58	2.57	0.90
$4.5 \leq x < 5.0$	0.27	2.84	1.00

Cuando ya se tiene la distribución acumulada de las posibilidades, se puede hallar el valor en el dominio que corresponde al umbral, por medio de interpolación. Por ejemplo, si se elige el valor de 0.8, la nota correspondiente es 4.2. Entonces, el sistema de inferencia asume que la clase de las notas “altas” está conformada por los estudiantes que hayan obtenido una nota superior o igual a 4.2 y las “bajas” son las que tienen un valor inferior a 4.2. De acuerdo con esto, el umbral define el punto de corte pero de una discriminación nítida, igual al utilizado en la Regresión Logística aunque en esta técnica estadística, el punto de corte siempre toma el valor 0.5. Por lo

tanto, este punto definido para la partición no corresponde al concepto de alfa corte de la Lógica Difusa.

La propuesta de Zapata presenta el problema de respuestas no discriminadas (Urrutia, Tineo y González, 2008) puesto que si el umbral especificado fuera cero, el soporte del conjunto difuso que representa a una etiqueta como “alto” o “grande” sería todo el *Universo de Discurso* y el soporte del conjunto difuso complemento sería el conjunto vacío. Es decir, no se haría ninguna discriminación.

Otra limitante de esta propuesta es la subdivisión del *Universo de Discurso* en dos categorías únicamente. Con una máquina de inferencia definida así se podría preguntar, por ejemplo, por los “jóvenes” o por los “viejos”, pero el sistema de inferencia no podría concluir quiénes se pueden considerar de una edad intermedia, que es lo mismo que pertenecer a la subclase borrosa de los “adultos”. El operador de la negación tampoco estaría bien definido para trabajar con términos intermedios. Debido a esto, dicha máquina no es adecuada para una variable lingüística con un número de términos o etiquetas lingüísticas diferente de dos.

Por otro lado, el proceso requerido en la discriminación de los objetos sería bastante complejo si se decide ser estrictos con la fórmula de derivación de las posibilidades a partir de las probabilidades, pues requiere el ajuste de las frecuencias empíricas a un modelo probabilista, que en cada caso puede ser diferente, además de una prueba de la bondad del ajuste.

Para evitar los inconvenientes mencionados, en este trabajo de investigación, se plantea una técnica alternativa donde la máquina de inferencia pueda resumir o aproximar, de manera dinámica y autónoma, la forma y localización de la distribución de los datos, sin necesidad utilizar la fórmula de Dubois y Prade, en el proceso de discriminación borrosa.

Adicionalmente, es una propuesta que no tiene la cobertura de las propuestas presentadas anteriormente, debido al reducido número de términos vagos que permite representar.

Resumiendo lo que se ha expuesto, se podría decir que a pesar de una buena cantidad de trabajo de investigación para la flexibilización de lenguajes para los sistemas de consulta con el manejo de la vaguedad, aún se encuentran deficiencias y limitaciones que se deben superar.

Finalmente, se puede afirmar que una completa flexibilización del lenguaje también debe incluir la dirección contraria. Esto es, que las respuestas a las consultas puedan ser vagas o dadas mediante valores de verdad. Esta posibilidad no es contemplada en los lenguajes de consulta a bases de datos relacionales o en sus extensiones; según lo estudiado.

## **5.2 Modelo Conceptual del Razonamiento Propuesto**

El proceso de interpretación de los términos vagos según el contexto lingüístico, no es un proceso trivial puesto que los modelos de los conjuntos difusos que representan cada uno de ellos no van a ser proporcionados por los humanos

expertos, sino que la máquina de inferencia deberá encontrarlos, por sí misma y en línea. Para ello, se basará en los hechos (los datos disponibles del contexto que delimita la consulta) y en las reglas semánticas que aquí se definan para cada uno de los patrones admisibles con los que pueden encajar las consultas, considerando las restricciones establecidas previamente.

La máquina de inferencia emulará a un experto calificado que puede razonar o descubrir nuevo conocimiento basándose en los datos contenidos en la base de datos. Por su autonomía para interpretar la vaguedad de una consulta, según el contexto, y por su capacidad de razonamiento podrá considerarse un *agente inteligente*, como se le denomina en la Ingeniería del Conocimiento (Studer, Benjamins y Fensel, 1998).

La actividad principal del agente inteligente, o del sistema de inferencia, será el proceso de concreción de la vaguedad de la consulta, que es un proceso de razonamiento analógico e inductivo. Pero deberá realizar también procesos de inferencia deductiva, para decidir cuáles elementos conformarán la relación derivada deseada por el usuario consultor. También deberá realizar este tipo de razonamiento para dar respuestas vagas, en el caso en que sea recomendable o suficiente, dar respuestas de este tipo.

El problema de encontrar el significado aproximado de los conceptos vagos dependientes del contexto puede ser visto como un problema de minería de datos, más concretamente como un problema de reconocimiento de los patrones que caracterizan a los elementos de una clase rotulada con alguna etiqueta lingüística, en un contexto dado. En esta investigación, se admite que dichos patrones se cumplen, al menos de manera parcial, por los miembros de una clase dada.

El modelo conceptual de razonamiento aproximado y adaptable a los distintos contextos, se presenta mediante dos vistas: el modelo estructural y el modelo funcional o de comportamiento. A continuación, se procede con la descripción del modelo estructural del dominio subyacente en el proceso de inferencia o razonamiento aproximado, el metamodelo, para luego presentar el modelo de comportamiento.

### **5.3 Modelo Estructural del Dominio del Razonamiento**

Un modelo estructural muestra las clases de objetos (los conceptos), concretos o abstractos, involucrados en el proceso de razonamiento, sus propiedades y relaciones. Primero se hará una especificación verbal del modelo conceptual y luego se presenta el modelo estructural mediante un Diagrama de Clases del lenguaje de modelado UML (Unified Modeling Language) propuesto en (Rumbaugh, Jacobson y Booch, 1998) y considerado como estándar por la OMG (Object Management Group).

#### **5.3.1 Supuestos del Modelo**

El primer supuesto del modelo es tener como soporte lógico al modelo objeto-relacional. Ello significa que todos los datos son estructurados y manipulados

haciendo uso del concepto de relación matemática. En una base de datos objeto-relacional pueden existir relaciones básicas y derivadas (denominadas “vistas” en la terminología de Bases de Datos), como también procedimientos almacenados y disparadores que permiten representar operaciones, reglas o restricciones de diferentes tipos, constituyéndose realmente en una base de conocimientos.

Los datos contenidos en la base de conocimientos se suponen ciertos. Es decir, se supone que ya han sido sometidos, en el momento de su registro en la base de datos, a reglas de validación para velar por su integridad.

El *Universo de Discurso*, o el dominio de una variable, unidimensional o multidimensional, se supone numérico y cerrado. Este último supuesto es conocido en la literatura como CWA (abreviatura de Closed World Assumption) que permite razonamientos revisables. Éstos tienen como característica principal el hecho que, habiendo obtenido una conclusión a partir de un conocimiento dado, esa conclusión puede ser refutada mediante la obtención de nuevo conocimiento. Para una explicación más detallada de esta característica, se sugiere consultar en (Dix, Furbach y Niemela, 1999).

Además de lo anterior, se supone conocido el conjunto de términos lingüísticos  $T(X)$  asociables a una variable cuantitativa  $X$  que deba ser transformada a valores cualitativos. En caso de no especificarse este conjunto de términos de manera explícita se considerará, por defecto, un marco de cognición con tres subconjuntos difusos. Éstos representarán a los conjuntos de los valores “bajos” o “pequeños”, los “medianos” y los “grandes” o “altos”, respectivamente, preservando el concepto de orden en un marco de cognición.

### 5.3.2 *Conceptos Básicos*

Los conceptos básicos involucrados en el modelo estructural del razonamiento que luego se presentarán de una manera esquemática en un Diagrama de Clases, del lenguaje de modelado UML, son los siguientes.

**Consulta.** En un sistema interactivo de consulta-respuesta, una *consulta* es una expresión o fórmula bien formada del lenguaje, empleada para solicitar información sobre los objetos contenidos en una base de datos. En una consulta se reconocen distintas componentes o atributos: la *proyección* que es una lista de los atributos básicos o derivados que se desean visualizar, la lista de relaciones básica o derivadas en la base de datos sobre las cuales se quiere realizar la proyección, que aquí se define como el *origen* de los datos y, de manera opcional, un agregado de condiciones vagas o concretas que deben cumplir los elementos o tuplas en la relación resultante. Así mismo, una consulta puede incluir una cláusula para realizar un agrupamiento de los datos o un ordenamiento de los datos, en la relación resultante.

**Contexto.** Es el entorno lingüístico del cual depende el sentido y el valor de las expresiones vagas. El contexto es el conjunto de objetos que sirven de base de comparación para realizar las inferencias. Las condiciones concretas especificadas en

los criterios de filtrado de una consulta, delimitan ese entorno lingüístico. De acuerdo con el alcance de este trabajo de investigación no se consideran factores extralingüísticos que puedan dar origen a otros contextos, como uno sociocultural o religioso, como se expresó en el primer capítulo.

**Variable Lingüística.** Una variable lingüística se distingue de otras variables porque sus valores pueden expresados mediante etiquetas lingüísticas. Es una variable cualitativa ordinal que puede ser transformada para que su dominio lo constituyan valores numéricos, a través de un proceso de concreción. El proceso inverso también es posible, una variable cuantitativa puede ser transformada en una variable lingüística.

**Etiqueta lingüística.** Una etiqueta un valor posible de una variable lingüística  $X$ , es un elemento del conjunto de términos  $T(X)$ . Un término o etiqueta lingüística se caracteriza por un valor sintáctico que es el símbolo o rótulo usado, para hacer referencia a él, y de un valor semántico o un significado (Delgado et al., 1997). Un término puede ser derivado de términos básicos que pueden ser modificados con la negación, con adverbios de cantidad o derivado de composiciones de éstos, mediante conectivas lógicas tales como la conjunción o la disyunción. En la presente Tesis Doctoral también se considera que una etiqueta lingüística puede derivarse de una combinación lineal de otros términos primarios, vagos o concretos.

**Reglas Semánticas o de Derivación.** La máquina de inferencia deberá basarse en un conjunto de modelos teóricos para la representación de los términos vagos incluidos en las condiciones de la consulta, de acuerdo con el patrón que se identifique en el análisis léxico. Una palabra o token incluido en un patrón puede ser un adjetivo calificativo, un adverbio de cantidad o puede ser un cuantificador vago relativo o absoluto. Otro patrón admisible es un compuesto o agregación de condiciones vagas.

**Modelos teóricos.** Estos modelos describen, de manera abstracta y genérica, los conjuntos difusos que representarán las etiquetas lingüísticas o los términos vagos usados en las consultas. Por lo tanto, el conjunto de modelos teóricos para la representación de los términos vagos, determina las reglas de discriminación borrosa de la base de conocimientos que usará el agente inteligente para hallar la semántica de las consultas vagas, de manera aproximada y adaptándose a los múltiples contextos. Para cada modelo de conjunto difuso que represente un término vago o un patrón, será necesario definir cuáles estadísticas o funciones de los datos deberán emplearse para la estimación de los parámetros.

### 5.3.3 *Restricciones del Modelo*

En el proceso de razonamiento, para la interpretación de una pregunta vaga según el contexto, la semántica de un término vago será representada por un modelo que se obtiene instanciando el modelo teórico y utilizando como estimadores de los parámetros algunas estadísticas o funciones de los datos disponibles del contexto que se delimita en la consulta.

Los gránulos de información o clases no disjuntas se representan generalmente usando la Teoría de los Conjuntos Difusos, pero también se podrían representar mediante la Teoría de los Conjuntos Rugosos que se basa en una lógica trivaluada (verdadero, falso e indiscernible). Puesto que esta última teoría admite que no se defina la función característica para ciertos valores en el *Universo de Discurso* (a los que se les asigna el valor de indeterminado o indiscernible), se incumpliría la propiedad de cobertura de un marco de cognición (Restricción Nro. 12, sección 4.5.5). Además, por definición, todo conjunto difuso es rugoso por lo que se podrían aprovechar las relaciones de equivalencia definidas en esa teoría, si fueran requeridas. Debido a esto, los modelos teóricos de los conjuntos que van a representar los términos vagos de las condiciones de las consultas provienen de una Teoría de Conjuntos Difusos.

De acuerdo con el alcance establecido para este trabajo de investigación, sólo se contempla el proceso de mapeo entre variables lingüísticas y cuantitativas o numéricas. Esta es una restricción impuesta al modelo de razonamiento, sobre los tipos de datos que admite un sistema gestor de bases de datos.

El concepto de partición difusa (Definición Nro.15, Capítulo 3) relaja el concepto de partición matemática pues admite algún grado de solapamiento entre los conjuntos resultantes en un marco de cognición, pero no admite que éstos sean vacíos. Por lo tanto, la técnica de discriminación difusa que se proponga debe garantizar que se cumpla esta restricción.

Otra restricción que se debe considerar sobre los términos vagos compuestos o derivados de una combinación de otros y que no se había especificado antes, es que no pueden derivarse de sí mismos. Esta restricción es obvia, pero que como va a ser controlada por una máquina debe ser especificada de manera explícita, en el sistema de inferencia.

Ya definidos los conceptos básicos y sus restricciones en la interpretación de las consultas con condiciones vagas, en la Figura 10 se presenta, de manera esquemática, el modelo estructural de razonamiento aproximado que debe complementar al metamodelo de la base de datos (metadatos) y que conforma la base de conocimientos del sistema de inferencia. Allí se muestra la clase "Atributo" de otro color para significar que es una clase ya existente en los metadatos de una base de datos, que sirve de vínculo con las nuevas clases propuestas. También se destacan las clases de objetos no persistentes, en el proceso de razonamiento aproximado, resaltadas con una caja en blanco en la parte superior. Entre ellas está el contexto lingüístico, que se deriva de las condiciones concretas especificadas en la consulta, y el significado del texto de la consulta que se infiere de este contexto y de las reglas semánticas definidas para el patrón con el que encaja.

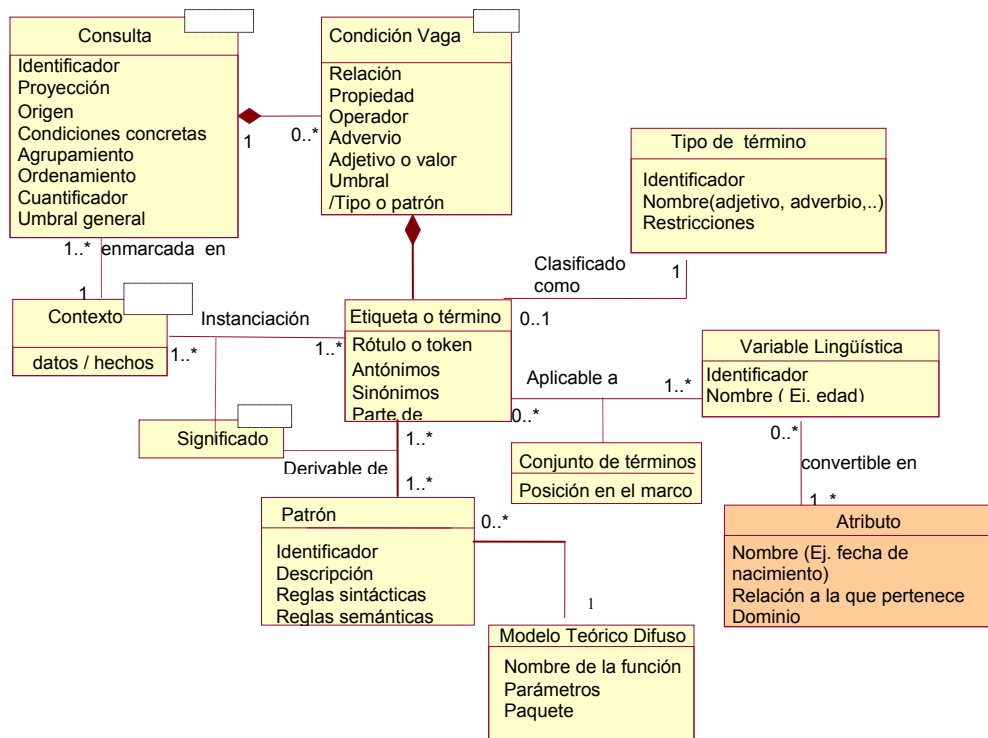


Figura 10. Modelo Estructural en la Interpretación de la Vaguedad

## 5.4 Modelo de Comportamiento del Sistema

Este modelo describe las tareas del sistema de inferencia difusa para estimar la semántica de los términos vagos especificados en las condiciones de las consultas.

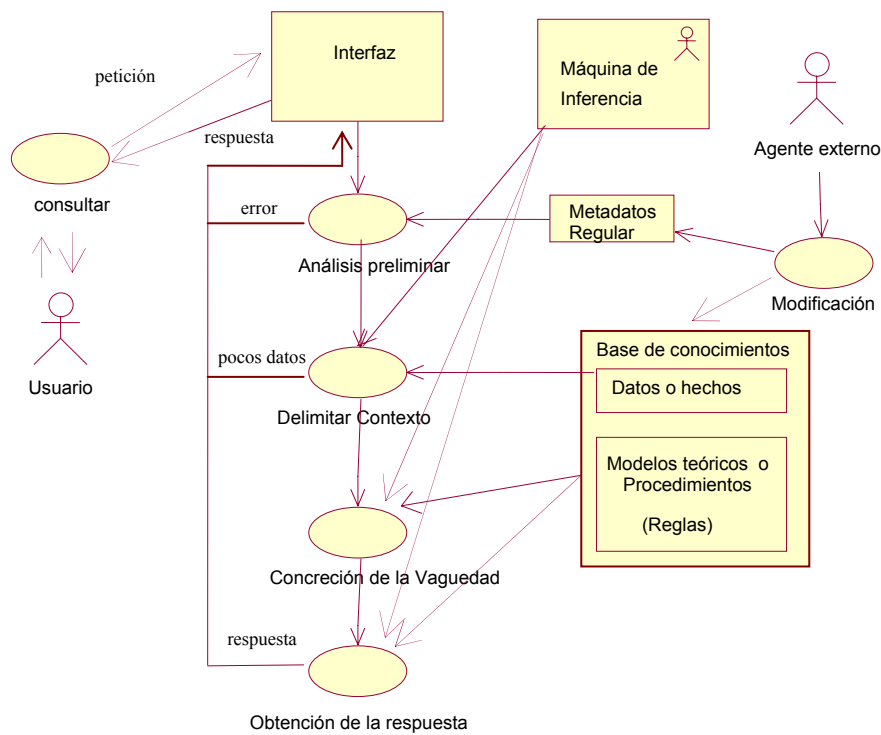
En la interpretación de una consulta con términos vagos, como en cualquier proceso de traducción, debe realizarse previamente un análisis léxico y sintáctico donde se validen los símbolos y la estructura gramatical de la sentencia emitida para poder proseguir con el proceso de concreción o cuantificación de la vaguedad.

Puesto que el significado de los términos vagos aquí tratados depende del contexto lingüístico, los modelos empleados para realizar las inferencias deben construirse a partir de los datos disponibles en la base de datos. Por lo tanto, antes de realizar el análisis semántico de la consulta, es necesario determinar el contexto lingüístico que enmarca la consulta.

En el análisis semántico de una consulta, también se debe considerar que el significado de cada término vago también depende de su tipo (si es un adverbio de cantidad, por ejemplo) o de la cantidad de términos admisibles para una variable lingüística, cuando se trate de la interpretación de un adjetivo calificativo. Por esta razón, en el modelo estructural del dominio del problema se definió una clase de objetos que representa los distintos modelos teóricos difusos que pueden ser instanciados en el proceso de concreción de la vaguedad, de acuerdo con el tipo de

término considerado. En esa clase se definen y mantienen las reglas del sistema de inferencia para el análisis semántico.

Realizando una descripción de alto nivel de abstracción, cuando el usuario formule una consulta vaga, se capturan y clasifican los términos vagos hallados en ella, mediante un análisis léxico. Estos valores son las variables de entrada que se transfieren a la máquina de inferencia para obtener la respuesta a la solicitud del usuario. El sistema de inferencia opera sobre el contexto específico, considerando los axiomas (los hechos o premisas) válidos en ese contexto, y usa las reglas semánticas que están definidas en el metamodelo de la base de datos para concretar la solicitud y entregar la información requerida por el usuario que consulta.



**Figura 11. Proceso General de Razonamiento en la Resolución de Preguntas Vagas**

En la Figura 11 se presenta el diagrama del proceso de razonamiento basado en conocimiento para el procesamiento de consultas vagas que dependen del contexto lingüístico. Una caja especial, con un icono de figura humana en su interior, está representando la máquina de inferencia como un agente inteligente: el experto en minar los datos y en obtener los modelos necesarios para estimar el grado de compatibilidad de los objetos del contexto, a las restricciones impuestas a la relación resultante.

Ahora se procede con la descripción más detallada de los pasos inmersos en el proceso de razonamiento aproximado y adaptable en el procesamiento de las consultas.

#### **5.4.1 Análisis Preliminar de la Consulta**

El análisis preliminar de la consulta es un proceso inicial para entender la solicitud del usuario y reconocer la existencia de vaguedad en la consulta. Este análisis consiste en las dos actividades previas de toda traducción: el análisis léxico y el sintáctico. Este análisis preliminar no es llevado a cabo por el sistema de inferencia, sino por la interfaz mediadora entre el usuario consultor de la base de datos y el agente inteligente (la máquina de inferencia). La interfaz, antes de solicitarle a este agente su ayuda para determinar la semántica de los términos vagos, debe realizar los pasos que se presentan a continuación.

##### **5.4.1.1 Análisis Léxico**

El objetivo de este análisis es comprobar que se han usado sólo *tokens* (símbolos o palabras del lenguaje) correctos. Más formalmente, un token se define como la unidad mínima de información con significado propio, dentro de un texto o secuencia ordenada de caracteres alfanuméricos (Harrison, 1987).

La variable de entrada al proceso de análisis léxico es el código fuente de la consulta, de donde se aíslan o separan los *tokens* que lo componen y se clasifican de acuerdo con su tipo. Los tipos de token pueden ser identificadores o nombres de tablas o atributos, funciones, adjetivos calificativos, adverbios de cantidad, operadores lógicos de comparación, calificadores o cuantificadores, entre otros.

La salida del proceso de análisis léxico, es un bloque de texto que queda clasificado y compuesto de cadenas de caracteres indivisibles conocidas como *lexemas*. Un ejemplo de un lexema es "Medellín", clasificado como un *token* de tipo literal y cuya definición es una cadena encerrada entre comillas dobles. Otro ejemplo es *cuadrado* que es un *token* de tipo función, que se define mediante una expresión regular de la forma *cuadrado(a) -> a\*a*. Cuando en el proceso de análisis, se encuentra algún carácter que no es un *token* válido, se produce un error léxico.

En el presente trabajo investigativo, el análisis léxico se centrará en reconocer el tipo de términos vagos inmersos en los criterios de la consulta, dado que de éste dependerá la técnica de discriminación borrosa que el motor de inferencia, el experto, usará para que el sistema de consulta-respuesta pueda responder a una consulta vaga.

Para evitar trabajos computacionales innecesarios, durante este análisis preliminar también se debe chequear si el usuario que consulta tiene privilegios (de lectura o ejecución) sobre los objetos (relaciones, atributos y funciones) de la base de datos, referidos en la consulta. Si no hay errores léxicos y el usuario está autorizado para acceder o ver las características deseadas de los objetos, se prosigue con el análisis sintáctico.

#### 5.4.1.2 Análisis Sintáctico

El análisis sintáctico tiene como objetivo comprobar que las construcciones o combinaciones de *tokens* usadas en el planteamiento de la consulta sean correctas. Para hacer esta comprobación, la interfaz que actúa como analizador, se debe apoyar en la gramática del lenguaje.

El lenguaje descrito por una gramática es el conjunto de todas las cadenas de caracteres que se pueden producir siguiendo unas reglas de producción o de reescritura. Una regla de producción informa que el símbolo al lado izquierdo debe ser reemplazado por alguna de las alternativas del lado derecho.

En las reglas de reescritura se distinguen dos tipos de elementos o símbolos: los símbolos terminales que corresponden a *tokens* identificados en el análisis léxico y los no terminales que corresponden a nombres de construcciones válidas. Las reglas de reescritura se pueden especificar usando diferentes notaciones. En este trabajo investigativo se empleará una de las más usadas: la notación BNF (Backus-Naur-Form) (Aho, Sethi y Ullman, 1990). Pero antes de presentar las reglas de reescritura de una consulta en SQL, de la cual se parte para las presentar las extensiones o modificaciones de este trabajo investigativo, es necesario especificar los formalismos o reglas generales que se acogen, para definir la gramática del lenguaje:

- El símbolo para especificar una regla de reescritura es '→'.
- Los símbolos no terminales aparecen en el lado izquierdo de una regla de reescritura.
- El símbolo no terminal se expande en el lado derecho de la regla de reescritura.
- La expansión de un símbolo puede estar conformada por símbolos terminales y no terminales.
- Un símbolo no terminal se declara con minúsculas.
- Un símbolo terminal (una palabra o token) se declara con mayúsculas. Si el símbolo terminal inicia con minúsculas entonces aparece entre comillas dobles o sencillas, pues en ese caso es un literal.
- '|' es el separador de selecciones entre dos o más símbolos. Cuando esta selección sea excluyente se encierra con '{' y '}'.
- ',' es el separador en una lista.
- '[' y ']' son usados para representar símbolos opcionales.
- Las comillas dobles se usan para representar literales o las etiquetas lingüísticas.
- Los paréntesis se usan para determinar el alcance y la precedencia de las operaciones o funciones.
- Otros identificadores que aparezcan en el código fuente de la consulta (como los nombres de tablas, atributos, funciones o de etiquetas lingüísticas) se resuelven mediante la consulta a los metadatos de la base

de datos. De modo parecido a la utilización de tablas de símbolos en otros lenguajes de programación (Aho, Sethi y Ullman, 1990).

- La repetición de un símbolo (una o más veces) se representa con puntos suspensivos.
- Una lista se encierra entre ' $<$ ' y ' $>$ ' y se separa con comas.
- Un comentario de línea es precedido por dos guiones.

Siguiendo estas convenciones para la especificación formal de las reglas de producción, el código fuente de una consulta en el lenguaje estándar SQL, sin ninguna modificación o extensión para admitir vaguedad, debe ajustarse a:

```
consulta → SELECT proyección
          FROM relaciones | (consulta)
          [WHERE condiciones]
          [GROUP BY atributo [HAVING condiciones]]
          [ORDER BY atributos [ASC | DESC]]
          [ { UNION [ ALL ] | INTERSECT | MINUS } (consulta) ]
proyección → * | relación.* | [DISTINCT] expr [AS nombre ] [, expr ...]
relaciones → relación [alias] [, relación...]
condiciones → condición_simple [conectiva condición_simple...]
condición_simple → {expr IS [NOT] NULL | expr comparador expr
                    | [expr] cuantificador (consulta)}
expr → {término | función(expr) | expr * expr | expr /expr
        | expr - expr | expr + expr}
término → identificador | literal | escalar
comparador → {> | < | >= | <= | <> | = | [NOT] LIKE}
conectiva → {AND | OR | +}
cuantificador → {ALL | EXISTS | SOME | ANY }
identificador → {relación.atributo | atributo | función(<expr> ) }
relación → {R1 | R2 ... }
atributo → {A1 | A2 ... }
función(expr) → F1 | F2 ...
literal → cadena de caracteres encerrada entre comillas
escalar → valor numérico simple
```

En las reglas de reescrituras recién presentadas, los nombres de las relaciones, atributos y funciones son variables que sólo se pueden resolver consultando los metadatos de la base de datos particular. Por eso, se representan como símbolos terminales R1 o A2 o FI.

Cuando se defina la nueva sintaxis del lenguaje de consulta, considerando los diferentes tipos de patrones de consultas vagas admisible en el nuevo lenguaje, las reglas de reescritura seguramente se modificarán, pero se evitará alterar la estructura gramatical del lenguaje SQL con el fin de preservar su simplicidad y su cercanía al lenguaje natural que le dio origen (en este caso, al inglés). También se velará porque las extensiones propuestas sean estrictamente aditivas, significando que la sintaxis previa se pueda considerar un caso particular de la nueva.

Con las reglas de reescritura que se definan, la interfaz mediadora entre el usuario consultor y el sistema gestor de bases de datos puede comprobar si el código fuente de una consulta es correcto. Esto es, que la cadena de texto que se recibe como entrada sea una fórmula bien formada. Para ello usará un árbol de expansión, donde se van colocando los símbolos terminales y no terminales que se detecten en el texto de la consulta. Si se puede construir un árbol, usando expansiones admisibles de las reglas, es que la entrada es sintácticamente correcta. Eso significaría que el texto de la consulta encaja con algún patrón o estructura gramatical de las definidas en el lenguaje interactivo que se propone.

Debe señalarse que el análisis preliminar de una consulta, se facilita mucho si al usuario se le ofrece una interfaz donde elija los valores apropiados para cada término de una consulta, y en las posiciones debidas, consultándolos previamente en los metadatos de la base de datos. Si el texto de una consulta no encaja con ninguno de los patrones, significa que la consulta no incluye términos vagos. Por lo tanto, lo que haría la interfaz en estos casos sería armar el texto completo de la consulta para remitírselo al sistema gestor de bases de datos para su ejecución y recibir la respuesta para mostrársela al usuario. En cambio, cuando el texto encaje con alguno de los patrones de consulta vaga admisibles, el sistema debe registrarla como pendiente de interpretación e invocar a la máquina de inferencia para que empiece su trabajo de delimitar el contexto lingüístico que enmarca la consulta.

De manera esquemática, en la Figura 12 de la página siguiente, se presenta el diagrama de actividades del análisis preliminar de la consulta, que debe llevar a cabo la interfaz, cuando el usuario formule una pregunta.

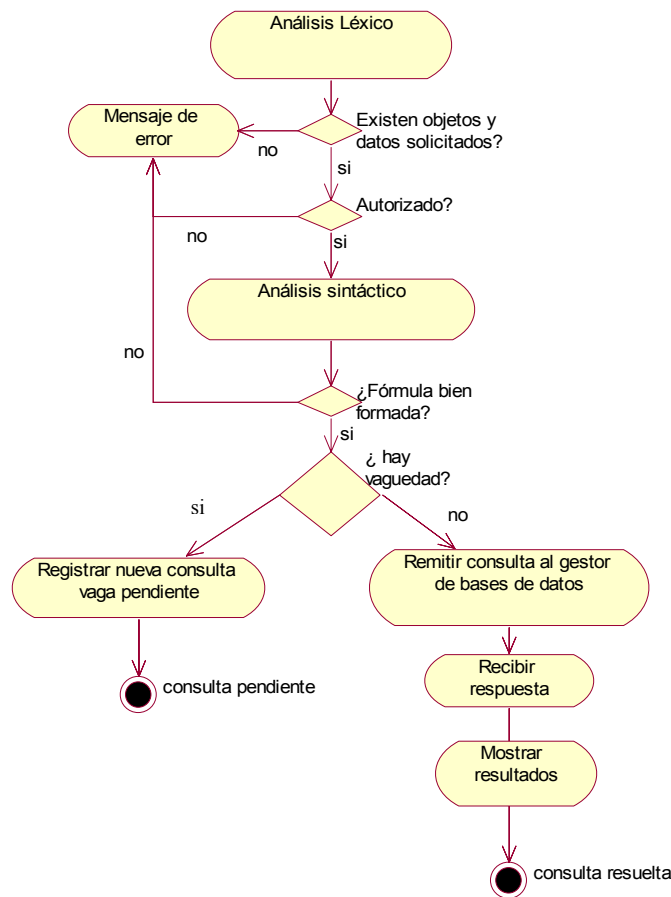


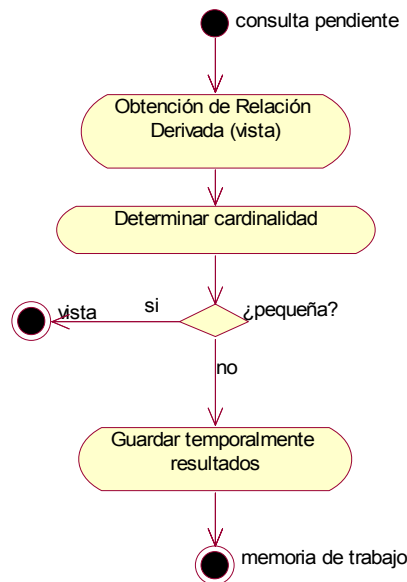
Figura 12. Diagrama de actividades en el análisis preliminar de la consulta

#### 5.4.2 Delimitación del Contexto Lingüístico de la Pregunta

Puesto que el significado de los términos vagos aquí tratados debe ajustarse al contexto lingüístico delimitado en la consulta, este paso es fundamental para que las respuestas de un sistema interactivo de consulta-respuesta tengan validez. Por lo tanto, es un paso previo obligatorio al análisis semántico de la vaguedad expresada en una consulta y consiste en determinar cuáles objetos conforman el dominio o *Universo de Discurso*, en un caso particular.

Como se observa en la Figura 8, el primer paso de la delimitación del contexto, consiste en la obtención de relación derivada o vista sobre la cual se considerarán los conceptos vagos. Dicha relación es una proyección de los atributos (simples o derivados) que se desean visualizar, especificados en la cláusula SELECT, cuyo origen es la relación o el producto cartesiano de las relaciones especificadas en la cláusula FROM, y restringida por las condiciones concretas (no vagas) especificadas en la cláusula WHERE. Condiciones que deben ser cumplidas estrictamente por

todas las tuplas de la relación resultante. Puesto que también se deben traer los valores cuantitativos de los atributos que deben ser transformados en etiquetas lingüísticas, la lista de la proyección debe estar complementada con los nombres de estos atributos.



**Figura 13. Diagrama de actividades de la delimitación del contexto**

Luego de generar la relación o vista de los objetos que hacen parte del contexto, es necesario considerar la cardinalidad de esta relación resultante. Si es pequeña (por ejemplo, 10 tuplas) que no hace necesario un resumen de los datos porque el usuario puede fácilmente realizar sus propias inferencias, el agente inteligente se la presenta como un listado o reporte tabular. En caso contrario se guardará, de manera temporal, convirtiéndola en una vista materializada. Esta información representa la “memoria de trabajo” del agente inteligente, esencial para interpretar los términos vagos, según el contexto.

### **5.4.3 Concreción de la Vaguedad**

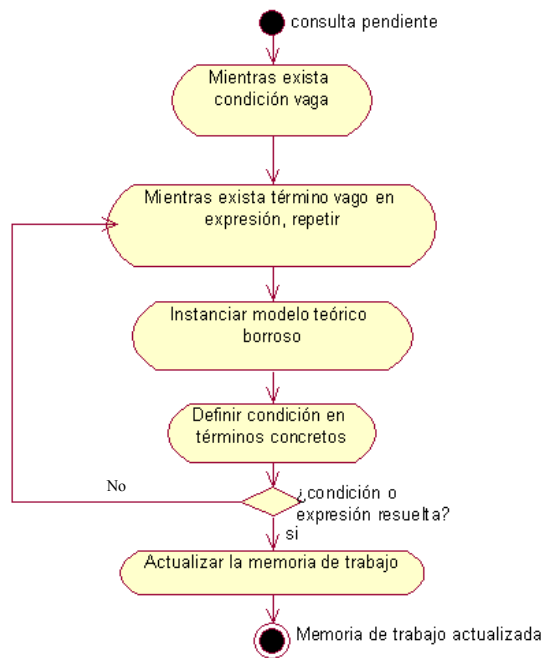
La concreción o cuantificación de la vaguedad es el corazón o proceso central en la interpretación de la vaguedad en una consulta pues consiste en el análisis semántico de cada término vago utilizado en la especificación de la misma. Este proceso también es conocido en la literatura como “Precision” y considerado un prerrequisito básico para la mecanización del lenguaje natural (Zadeh, 2006).

La concreción de la vaguedad es un proceso de inferencia no deductivo llevado a cabo por el agente inteligente (el sistema de inferencia) que consiste en ajustar dinámicamente los modelos de los conjuntos borrosos que representan las

condiciones vagas de las consultas, de acuerdo con el contexto establecido en el proceso anterior y del patrón sintáctico identificado. Si el patrón incluye una etiqueta lingüística, también dependerá del número de categorías consideradas en un marco de cognición particular. Por esto, el proceso de concreción o cuantificación de la vaguedad se representa como un conjunto de procesos de inferencia alternativos, en lugar de uno sólo.

Por la relevancia de este proceso, en el capítulo siguiente de esta Tesis Doctoral se analizarán y detallarán todas las técnicas que se consideren más apropiadas para hallar la semántica de cada tipo de término vago incluido en una consulta. Por el momento se muestra en la Figura 14 un diagrama de actividades, de alto nivel de abstracción, del proceso de concreción de la vaguedad.

Luego de estimar los parámetros particulares de los modelos difusos que representan los términos vagos de una condición especificada en la consulta, se guardan temporalmente en la base de datos. De este modo estarán disponibles para el siguiente paso que consiste en realizar un proceso de inferencia deductivo para dar la respuesta a las consultas.



**Figura 14. Diagrama de Actividades de la Concreción de la Vaguedad**

En el proceso de concreción de la vaguedad, puede ocurrir un proceso de actualización de la memoria de trabajo, para eliminar las filas que cumplen con absoluta certeza (grado de pertenencia = 0) una de las condiciones de filtrado y así lograr mayor eficiencia en el sistema.

#### 5.4.4 Obtención de la Respuesta

El sistema de razonamiento propuesto para dar sus respuestas, se basará en el grado de pertenencia o de encajamiento de una tupla a todas las condiciones especificadas en una consulta, tal como se hace en las extensiones del lenguaje SQL como SQLf y FSQL. Ese grado de encajamiento se deduce de los modelos difusos hallados para representar los términos vagos, según el contexto delimitado en la consulta. La relación resultante  $R^*$  contendrá aquellas tuplas que tengan un grado de encajamiento estrictamente superior al umbral o alfa corte especificado por el usuario final, que por defecto tomará el valor de cero. Así, solo se excluirían los objetos que no pertenezcan, en absoluto o con cierto grado definido mediante el umbral, a las clases difusas que resultan de restringir el Universo de Discurso (el contexto) según las condiciones de la consulta para lograr la flexibilidad deseada en un sistema interactivo de consulta-respuesta.

Con miras a que el sistema de consulta-respuesta sea aún más flexible por adaptabilidad a los distintos contextos que delimiten las consultas, se permitirá que la actividad principal del agente inteligente, o del sistema de inferencia, sea el proceso de concreción de la vaguedad de la consulta, que es un proceso de razonamiento analógico e inductivo. Pero deberá realizar también procesos de inferencia deductiva, para decidir cuáles elementos conformarán la relación derivada deseada por el usuario consultor. También deberá realizar este tipo de razonamiento para dar respuestas vagas, en el caso en que sea recomendable o suficiente, dar respuestas de este tipo.

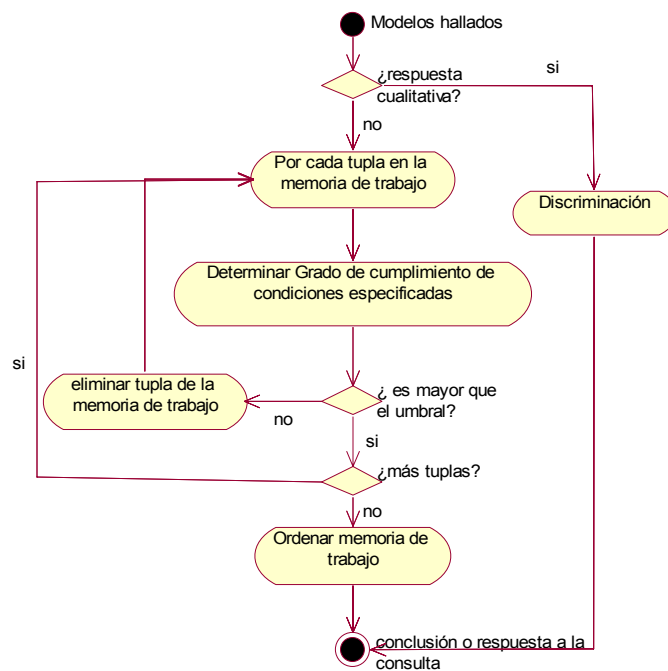
El usuario pueda limitar el número de tuplas en la tabla de resultados de una consulta, a una cantidad específica  $n$ . Esta cantidad y el umbral son conocidos como los calibradores cuantitativos y cualitativos de una consulta, respectivamente (Gonçalves y Tineo, 2007). Por lo tanto, dichos calibradores deben ser considerados para descartar algunas tuplas.

La forma deseable de presentar la relación resultante  $R^*$  es como una lista ordenada, en la que las tuplas con mayor grado de encajamiento a las condiciones de la consulta estén de primeras. Aunque se también se puede admitir que el usuario declare explícitamente otro orden diferente. Esto quiere decir que en una consulta prevalece la cláusula ORDER BY.

El valor o grado de cumplimiento de cada tupla con las condiciones o propiedades de la relación restringida  $R^*$ , también puede ser visible para el usuario final. Por lo tanto, la función puede ser especificada en la proyección de una consulta y para ello se la nombrará GC, como abreviatura de grado de cumplimiento o compatibilidad. La función GC tiene un argumento que puede ser el nombre de un atributo numérico de una relación, especificada en la cláusula FROM (que permite derivar el producto cartesiano de las relaciones involucradas), pero también el argumento puede ser un asterisco que se utiliza cuando se desee conocer el grado de cumplimiento de una tupla al agregado de todas las condiciones vagas establecidas en la consulta. Esta función se asemeja la propuesta por Galindo.

Cuando se trate de una consulta donde la respuesta esperada sea un valor de verdad cualitativo, se debe hacer la transformación o mapeo para pasar de un valor numérico, el grado verdad, a un valor lingüístico.

Por lo anterior, la conclusión es un proceso de inferencia deductivo que puede constar de dos pasos: el cálculo el grado de cumplimiento de las tuplas, del contexto, a todos los criterios establecidos en la consulta y la transformación o mapeo de valores cuantitativos a vagos, un proceso de difuminado, cuando la respuesta esperada sea un valor de verdad cualitativo. En la Figura 15 se muestra el proceso de concluir u obtener la respuesta a una pregunta vaga, por el sistema de inferencia.



**Figura 15. Diagrama de actividades de la obtención de la respuesta**

Cuando el sistema inteligente infiera la respuesta de una consulta, finaliza el proceso de interpretación, según el contexto lingüístico que la enmarca.

Ahora se procede a detallar el proceso de concreción de la vaguedad mediante el análisis sintáctico y semántico de los términos vagos que se van a representar o modelar.

# Capítulo 6

## 6 La Concreción de la Vaguedad

El proceso de la concreción o cuantificación de vaguedad de un término no sólo consiste en encontrar el intervalo de los valores numéricos que delimitan el dominio de una etiqueta o término vago según el contexto, sino también la función de pertenencia que permita evaluar el grado de cumplimiento de cada tupla, a las condiciones o criterios vagos de la consulta. El proceso, entonces, consiste en hallar los modelos de los conjuntos borrosos particulares que representan cada término vago incluido en una consulta. Este proceso de identificación o de ajuste se logra instanciando el modelo teórico de los conjuntos borrosos, definido para cada tipo de término vago admisible y calculando o estimando los valores de los parámetros de dichos modelos usando los datos del contexto.

A continuación se describen las técnicas que necesitará aplicar la máquina de inferencia, discriminadas por el tipo de preguntas vagas que puede plantear un usuario, según la clasificación realizada en el lenguaje teórico PRUF para la representación de variables lingüísticas. Para cada tipo de pregunta se analizarán las reglas sintácticas que se deben definir para completar la gramática, procurando no afectar la legibilidad y la simplicidad del lenguaje extendido, y las reglas semánticas necesarias para la interpretación de la vaguedad.

### 6.1 Preguntas de tipo I (con términos vagos simples)

Según el lenguaje teórico PRUF, las preguntas de tipo I incluyen las consultas con sólo un término primario o simple del conjunto de términos  $T(A)$ , dentro de las condiciones o restricciones impuestas a los objetos o tuplas de una relación. En este tipo de preguntas, el usuario usa una etiqueta lingüística  $E_j$  para calificar al atributo  $A$  de una relación de objetos  $R$ , en los criterios o en las condiciones de filtrado.

La forma genérica de la condición vaga simple es “ $A$  es  $E_j$ ”. Por ejemplo, en la consulta “¿Cuál es el nombre y la ubicación de las empresas con un “alto” número de empleados, en Colombia, y cuyas utilidades no hayan superado, el año pasado, los mil millones de pesos?”, la condición vaga está especificada sobre el atributo “número de empleados”, donde la etiqueta lingüística  $E_j$  es “alto” y el contexto está formado por las empresas colombianas.

Una sentencia o proposición de la forma “ $A$  es  $E_j$ ”, significa lo mismo que “es verdad que  $A$  es  $E_j$ ”. Por lo tanto, el valor de verdad de la proposición es realmente la diferencia entre una discriminación difusa y una convencional basada en una tabla de verdad con dos valores “cierto” o “falso”.

Como se había expresado en el Capítulo 3, una variable lingüística se caracteriza por la quintupla  $(X, T(X), U, G, S)$ , en la cual  $X$  es el nombre de la variable en cuestión,  $T(X)$  es el conjunto de términos o valores lingüísticos que toma la variable  $X$  del *Universo de Discurso*  $U$ ,  $G$  son las reglas sintácticas que generan nuevos términos de  $T(X)$  y  $S$  son las reglas semánticas que asocian cada valor o etiqueta lingüística  $E$  con su significado  $S(E)$ , por medio de un conjunto difuso en  $U$ , entre todos los posibles. Por eso, a continuación, se definirá la gramática  $G$  y la semántica  $S$ , en nuestro modelo de razonamiento aproximado para interpretar las condiciones vagas simples.

### 6.1.1 Reglas Sintácticas para Condiciones Vagas Simples

La forma sintáctica de las preguntas de tipo I en Álgebra Relacional, lenguaje teórico de bases de datos relacionales, se puede representar de la siguiente manera:

$$\Pi (\sigma_{R.A \text{ es } E_j} (R))$$

En esta expresión el símbolo  $\Pi$  representa el operador de la proyección que debe aplicarse para visualizar a  $P$ , que es una lista de atributos básicos o derivados, y  $\sigma$  es el operador de restricción para filtrar las tuplas de la relación  $R$  referida en la consulta, que de acuerdo con la propiedad  $A$ , puedan ser catalogadas con la etiqueta  $E_j$ , aunque sea de forma parcial. Ahora, se debe determinar la sintaxis que se considere más apropiada para representar las condiciones vagas simples en un lenguaje concreto de consulta a bases de datos, como es el lenguaje SQL.

De acuerdo con lo presentado en el capítulo anterior de este libro, sobre trabajos relacionados, las reglas sintácticas de FSQL para especificar una consulta le adiciona al lenguaje SQL un conjunto de palabras o tokens no habituales en el lenguaje natural como el símbolo  $\$$  que se le antepone a las etiquetas lingüísticas. Además, los operadores lógicos propuestos aunque permiten una alta flexibilidad para formular una consulta, como "FEQ" difícilmente son comprensibles para un usuario no conocedor de la terminología de la Lógica Difusa. Esto significa que el aumento de la riqueza expresiva del lenguaje se da a costa de su legibilidad y simplicidad.

En SQLf y sus extensiones se define una sintaxis más simple y comprensible, por ser más cercana al lenguaje natural gracias a la sobrecarga del operador IS. Esta es la razón para preferir esta sintaxis para representar las condiciones vagas simples en la extensión del lenguaje SQL propuesta. Sin embargo, en SQLf3 las etiquetas lingüísticas no son diferenciadas de los nombres de otros objetos en la base de datos. Esto significa que en el análisis léxico de la consulta, se tengan que examinar diferentes tablas de los metadatos y pudiendo originar problemas de ambigüedad con otros objetos de la base de datos. Por esto, se opta por especificar el rótulo entre comillas como es lo corriente en el lenguaje natural y en el lenguaje SQL para representar una constante de tipo varchar o literal. Esta alternativa no daría origen a una mala interpretación pues sólo puede calificar así a

un atributo de tipo cuantitativo cuando se incluya el operador IS en una condición de la consulta.

Adicionalmente, en SQLf se debe especificar al menos un calibrador de la consulta, pero aquí es opcional porque en la evaluación se incluye un alfa-corte estricto en lugar de uno débil, dado que el valor por defecto que se quiere considerar es el valor cero para mayor flexibilidad del sistema y ofrecer mayor comodidad al usuario final.

Por otro lado, en las propuestas estudiadas no se considera el token para definir el número de categorías admisibles en la caracterización de los objetos porque no fueron concebidos para discriminar en un número variable de categorías, lo cual constituye otro aporte para hacer los sistemas de consulta-respuesta más adaptables o flexibles.

De acuerdo con las razones expuestas, la sintaxis de una consulta en esta extensión del lenguaje SQL que se propone, es la siguiente:

```
SELECT proyección
FROM relaciones
WHERE expresión IS [m] E [j] [k]
      [WITH CALIBRATION n|threshold|n, threshold]
```

En dicho patrón, *proyección* es la lista de propiedades, básicas o derivadas, que el usuario quiere visualizar, *j* es la posición de una etiqueta en el marco de cognición, *k* es el número de etiquetas lingüísticas en el conjunto de términos definido para el atributo *A*, de alguna relación *R*, incluida en la cláusula FROM. Se supone que los valores *j* y *k* sólo se especificarían si se usa una etiqueta lingüística genérica como “alto(a)” que puede asociarse con múltiples variables o atributos y por eso se declaran como valores opcionales, encerrándolos con corchetes. El valor *n* es el calibrador cuantitativo, permite restringir a un número máximo “*n*” de las mejores respuestas a la consulta y el umbral o calibrador cualitativo (*threshold*), permite filtrar las tuplas cuyo grado de satisfacción a la consulta sea mayor a un nivel mínimo de tolerancia (el alfa-corte estricto). Este umbral es opcional para admitir que el usuario pueda especificar uno diferente de cero, que sería el valor por defecto.

Debe notarse, además, la ventaja de haber concebido una máquina de inferencia con capacidades de descubrir los modelos por su propia cuenta: una consulta también puede restringirse no sólo por el valor de un atributo, sino por medio de una función o expresión derivada. Por eso, en lugar de usar el término “atributo”, se empleó “expresión” en la sintaxis de la consulta de tipo I, con un término vago simple.

### 6.1.2 Reglas Semánticas para Condiciones Vagas Simples

En las propuestas para responder a consultas vagas a sistemas de bases de datos, generalmente no se delega al sistema de inferencia la responsabilidad de definir los modelos para catalogar a un objeto dentro de una categoría definida

vagamente (véase, por ejemplo, a Kackrpzyk y Zadrozny, 2001; Galindo, Urrutia y Piattini, 2005; Bosc, Kraft y Petra, 2005; Ma y Wang, 2006). Por esto, el diseñador o el mismo usuario de la base de datos debe definir, de antemano, los modelos de los conjuntos borrosos de cada variable lingüística que, generalmente, son definidos mediante la función trapezoidal.

### 6.1.3 Técnica de Discriminación para la Definición de Reglas Semánticas

Como el modelo de razonamiento aproximado de esta Tesis Doctoral propone discriminar, por defecto, en tres clases básicas, se iniciará el análisis de una partición difusa del *Universo de Discurso* considerando esta cantidad de etiquetas lingüísticas en el conjunto de términos  $T(X)$ . Aunque también debe ser admisible discriminar considerando otra cantidad diferente de clases o categorías, pero sujetos siempre a la restricción del número de elementos justificable sobre un marco de cognición (Restricción No. 9, sección 4.5.2) que limita a discriminar en un número de categorías menor que 7.

Se había expresado que la representación de la distribución de una variable en un contexto dado, por medio de la estadística de resumen de los cinco números, permite una partición natural del *Universo de Discurso*, en tres clases o categorías disjuntas y completamente exhaustivas. Dicha partición matemática sirve como punto de partida para la partición difusa requerida en la resolución del problema de la interpretación de los términos vagos, según el contexto. Así se asegura que se cumpla la propiedad de cobertura (Restricción No. 12, sección 4.5.5) y la preservación del orden en el marco de cognición (Restricción No. 8, sección 4.5.1).

La partición nítida de *Universo de Discurso* de una variable  $X$  origina un marco de cognición con una primera clase que contempla a los objetos con valores pequeños o menores, en la variable de interés, desde el valor mínimo hasta el primer cuartil  $Q_1$ , con el 25% de los datos. La clase intermedia, cubre el rango intercuartil (el 50% de los valores centrales) y la clase correspondiente a los valores mayores, va desde el tercer cuartil  $Q_3$  hasta el valor máximo, con el 25% de los datos restantes, con una forma parecida a presentada en la Figura 16.

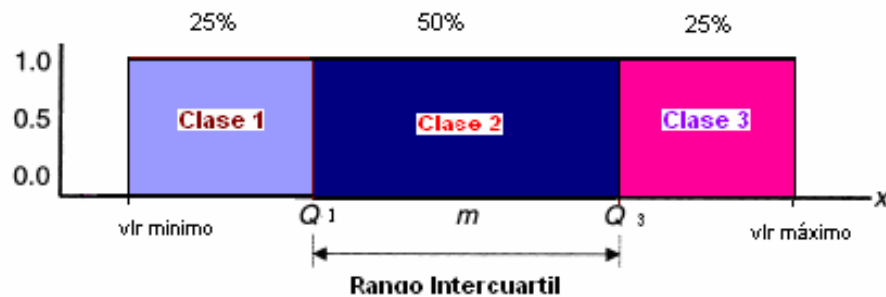


Figura 16. Partición nítida de la distribución de los datos.

Una vez definidos los límites de los conjuntos concretos o nítidos con la partición matemática realizada, se prosigue con un proceso de emborronamiento o difuminado para encontrar, ahora sí, los modelos de los conjuntos borrosos que corresponden a las tres posibles etiquetas lingüísticas definidas en el conjunto de términos  $T(X)$  de la variable de interés. Este proceso consiste en difuminar los bordes de separación entre las clases propuestas.

En el proceso de identificación de los modelos correspondientes a los tres conjuntos difusos que representan las etiquetas, se requiere primero determinar la forma de la función de pertenencia de cada uno de ellos. La forma inicial que se utilizará para representar los conjuntos difusos, por su simplicidad, será la función trapezoidal con parámetros  $(a, b, c, d)$ . Los límites inferior  $a$  y superior  $d$ , definen el soporte de la función de pertenencia (la base mayor del trapecio), y por los parámetros  $b$  y  $c$  determinan núcleo de la misma (la base menor del trapecio).

De acuerdo con las restricciones especificadas sobre los marcos de cognición en la discriminación borrosa que aquí se propone, se admitirá un área de solapamiento únicamente entre dos clases adyacentes (Restricción No. 11, sección 4.5.4) y se establece que los grados de pertenencia de cualquier objeto a los conjuntos borrosos adyacentes sean diferentes, exceptuando los puntos de cruce que, por definición, tienen un grado de pertenencia de 0.5 a dichos conjuntos. Esto también se requiere para que los gránulos sean distinguibles (Restricción No. 10, sección 4.5.3) o el grado de especificidad de la técnica de discriminación borrosa sea máximo, ya que de esta forma se estarían admitiendo sólo dos puntos de indecisión en la técnica de discriminación borrosa

Para determinar cuáles deberían ser los estimadores de los parámetros de las funciones trapezoidales que representan los conjuntos difusos, se propone utilizar nuevamente la estadística de resumen de los cinco números sobre la clase intermedia. Eso permite considerar como núcleo de la función de pertenencia, de esta clase intermedia, al 50% de sus datos centrales delimitados por sus dos cuartiles ( $Q_{N1}$  y  $Q_{N3}$ ), como se muestra en la Figura 17. Puesto que la clase intermedia, en la partición nítida, contiene el 50% de los datos originales, estos nuevos cuartiles cubren el 25% de los datos centrales de todo el *Universo de Discurso*.

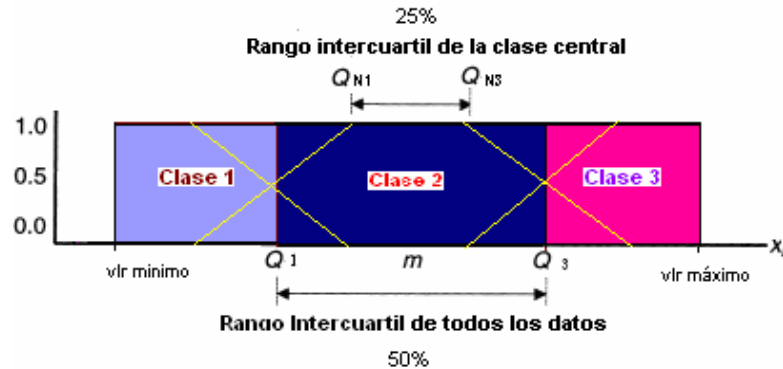


Figura 17. Explicación del proceso de difuminado

De acuerdo con lo anterior,  $Q_{N1}$  deja a su izquierda al 37.5% de los datos originales y  $Q_{N3}$  deja ese mismo porcentaje a su derecha. Luego, estos valores son los percentiles  $P_{37.5}$  y  $P_{62.5}$  de la distribución general de los datos, respectivamente. Ahora, el límite inferior del núcleo de la función de pertenencia de la clase intermedia  $P_{37.5}$  debe corresponder al límite superior de la clase difusa de los valores menores para cumplir con la restricción de complementariedad entre estas dos clases, en el área de solapamiento (Restricción No. 13, sección 4.5.6).

Además, puesto que entre  $Q_1$  y  $Q_{N1}$  (o entre  $Q_{N3}$  y  $Q_3$ ) se encuentran el 12.5% de los datos, se considera este mismo porcentaje de datos de las clases de los extremos, para definir el área total de solapamiento. Por lo tanto, el límite inferior del soporte de la clase intermedia corresponde al percentil  $P_{12.5}$  de los datos originales, que a su vez debe ser el valor máximo del núcleo de la función de pertenencia de la clase de los valores menores. Razonando de modo parecido, el límite inferior de la clase borrosa de los valores mayores es el percentil  $P_{62.5}$  porque debe coincidir con el valor máximo del núcleo de la clase intermedia y el límite superior de la clase borrosa intermedia corresponde el percentil  $P_{87.5}$  de los datos originales.

De acuerdo con la discriminación propuesta, en la Tabla 10, se presentan los modelos teóricos basados en la función trapezoidal para hallar las clases borrosas correspondientes a cada etiqueta lingüística, considerando tres clases en  $T(X)$ . En dicha tabla,  $P_q$  representa el percentil  $q$ -ésimo, con  $0 \leq q \leq 100$ , donde  $P_0$  es el valor mínimo y  $P_{100}$  es el máximo de todos los datos.

**Tabla 10. Modelos teóricos para etiquetas lingüísticas simples, considerando tres clases.**

Clase Borrosa	Dominio		Reglas semánticas para la definición de la función de pertenencia
	Límite Inferior	Límite Superior	
“Menor” “Bajo”	$P_0$	$P_{37.5}$	$\begin{cases} 1 & \text{si } P_0 \leq x \leq P_{12.5} \\ (P_{37.5} - x)/(P_{37.5} - P_{12.5}) & \text{si } P_{12.5} < x \leq P_{37.5} \\ 0 & \text{si } P_{37.5} \leq x \end{cases}$
“Mediano” “Medio”	$P_{12.5}$	$P_{87.5}$	$\begin{cases} 0 & \text{si } P_0 \leq x \leq P_{12.5} \\ (x - P_{12.5})/(P_{37.5} - P_{12.5}) & \text{si } P_{12.5} < x \leq P_{37.5} \\ 1 & \text{si } P_{37.5} < x \leq P_{62.5} \\ (P_{87.5} - x)/(P_{87.5} - P_{62.5}) & \text{si } P_{62.5} < x \leq P_{87.5} \\ 0 & \text{si } P_{87.5} < x \leq P_{100} \end{cases}$
“Mayor” “Alto”	$P_{62.5}$	$P_{100}$	$\begin{cases} 0 & \text{si } P_0 \leq x \leq P_{62.5} \\ (x - P_{62.5})/(P_{87.5} - P_{62.5}) & \text{si } P_{62.5} < x \leq P_{87.5} \\ 1 & \text{si } P_{87.5} < x \leq P_{100} \end{cases}$

Con propósitos ilustrativos, en la Figura 18, se muestra un caso donde los datos se distribuyen uniformemente sobre  $U$ . Por esto, la “perfección” y simetría de

los conjuntos difusos allí delineados. En este caso, los puntos de cruce corresponden a los cuartiles  $Q_1$  y  $Q_3$  de la distribución de los datos.

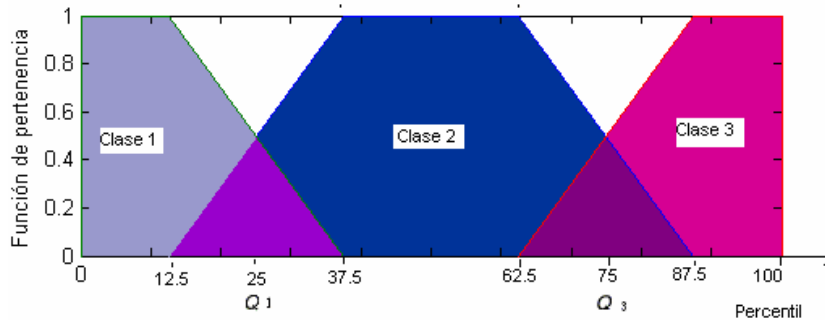
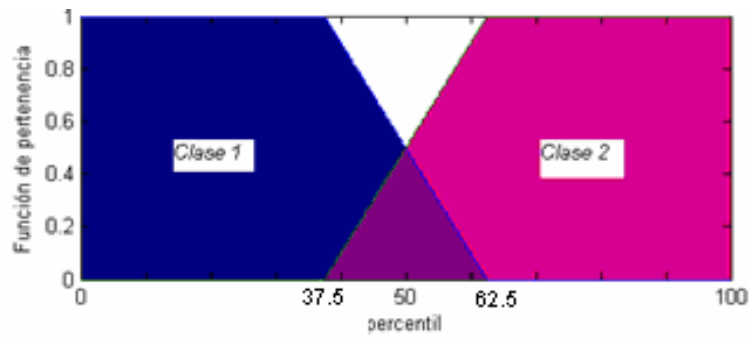


Figura 18. Representación gráfica de la partición difusa en tres clases

Como se había señalado antes, en un proceso de discriminación borrosa también sería plausible que se quieran considerar dos clases únicamente, por lo que también se debe proponer un modelo teórico para este caso. En la discriminación difusa en dos categorías, primero se haría una partición nítida haciendo uso de la mediana para que cada una de las clases contenga el 50% de los datos del *Universo de Discurso*. Luego, para cada clase generada, se calcula la estadística de resumen de los cinco números. El área de solapamiento está determinada por los valores máximos de la primera clase y los mínimos de la segunda. Esto significa que el 25% de los datos centrales constituyen esa área, dejando el 37.5% de los elementos restantes en cada una de las dos clases borrosas. La discriminación borrosa con dos conjuntos borrosos en el marco de cognición, se define formalmente en la Tabla 8 y se muestra gráficamente en la Figura 19.

Tabla 11. Modelos teóricos para etiquetas lingüísticas simples, considerando dos clases.

Clase Borrosa	Dominio		Reglas semánticas para la definición de la función de pertenencia
Etiquetas	Límite Inferior	Límite Superior	
“Menor” “Bajo”	$P_0$	$P_{62.5}$	$\begin{cases} 1 & \text{si } P_0 \leq x \leq P_{37.5} \\ (P_{62.5} - x)/(P_{62.5} - P_{37.5}) & \text{si } P_{37.5} < x \leq P_{62.5} \\ 0 & \text{si } P_{62.5} \leq x \end{cases}$
“Mayor” “Alto”	$P_{37.5}$	$P_{100}$	$\begin{cases} 0 & \text{si } P_0 \leq x \leq P_{37.5} \\ (x - P_{62.5})/(P_{87.5} - P_{62.5}) & \text{si } P_{37.5} < x \leq P_{62.5} \\ 1 & \text{si } P_{62.5} < x \leq P_{100} \end{cases}$



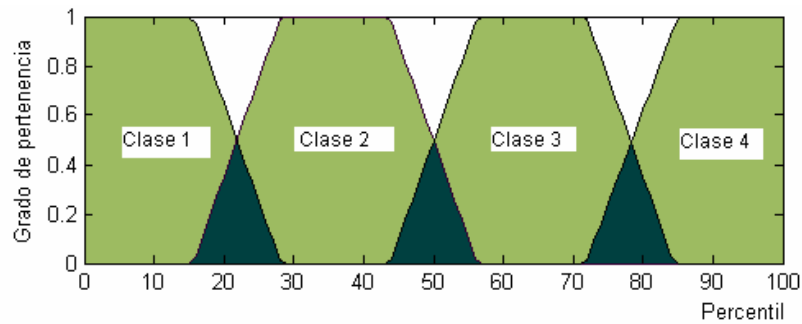
**Figura 19. Representación gráfica de la partición difusa en dos clases**

Ya descritas las técnicas para hallar los conjuntos difusos para representar etiquetas lingüísticas usadas como adjetivos calificativos considerando dos o tres clases, en un marco de cognición determinado, se puede emplear una técnica también basada en percentiles para una discriminación difusa de los objetos con una granularidad aún más fina. Una discriminación así se requiere en los casos donde el conjunto de términos primarios de una variable lingüística tenga un número superior a tres etiquetas, como en el caso de la estratificación socioeconómica de la ciudad de Medellín que permite clasificar a las viviendas en seis grupos o categorías distintas.

Cuando el conjunto de términos de una variable lingüística contenga un número de etiquetas igual a cuatro ( $K = 4$ ), se considerará un área de solapamiento entre las clases adyacentes igual a 12.5%, que es la mitad de esta área en la partición en dos clases o conjuntos difusos. Esto significa que como hay tres solapamientos que suman el 47.5% de la distribución total, el porcentaje restante se reparte entre las áreas que van a ser exclusivas de cada clase. Por lo tanto, cada una de esas áreas quedarán con el 15.6% de los datos. Considerando estos porcentajes y las restricciones impuestas a un marco de cognición, los modelos son:

- Clase 1 = hombro\_izquierdo( $P_0, P_{15.6}, P_{28.1}$ )
- Clase 2 = trapezoidal( $P_{15.6}, P_{28.1}, P_{43.8}, P_{56.3}$ )
- Clase 3 = trapezoidal( $P_{43.8}, P_{56.3}, P_{71.9}, P_{84.4}$ )
- Clase 4 = hombro\_derecho( $P_{71.9}, P_{84.4}, P_{100}$ )

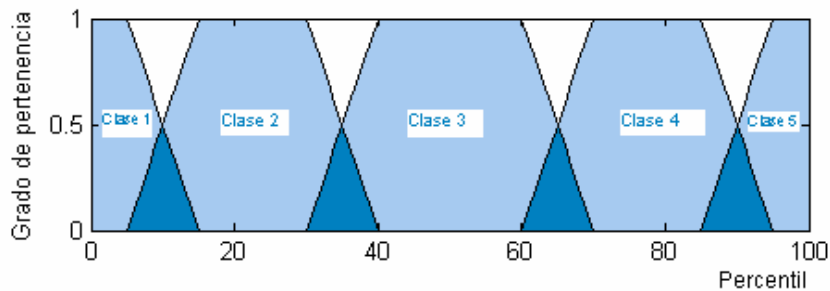
De acuerdo con estas reglas semánticas, la representación gráfica de la partición del *Universo de Discurso* en cuatro categorías es como se muestra en la Figura 20.



**Figura 20. Partición difusa en cuatro categorías**

Cuando se considera un número de clases  $K = 5$  en la partición difusa, el área de solapamiento debe reducirse con respecto a la partición anterior. Por eso, se fija un área de solapamiento igual al 10% del rango total de los datos. Como en este caso el número de clases es impar, se puede admitir que la clase de la mitad contenga un poco más de área, acorde con el efecto del difuminado esperado para los conceptos vagos.

- Clase 1 = hombro\_izquierdo( $P_0, P_5, P_{15}$ )
- Clase 2 = trapezoidal( $P_5, P_{15}, P_{30}, P_{40}$ )
- Clase 3 = trapezoidal( $P_{30}, P_{40}, P_{60}, P_{70}$ )
- Clase 4 = trapezoidal( $P_{60}, P_{70}, P_{85}, P_{95}$ )
- Clase 5 = hombro\_derecho( $P_{85}, P_{95}, P_{100}$ )



**Figura 21. Partición difusa en cinco categorías**

Para la partición difusa en seis clases se redujo a un 8% cada una de las cinco áreas de solapamiento, congruente con el aumento en la cantidad de clases en el marco de cognición. El 60% se reparte equitativamente entre las clases. Por tanto, los modelos de los conjuntos difusos en una partición en seis clases, son:

- Clase 1 = hombro\_izquierdo( $P_0, P_{10}, P_{18}$ )
- Clase 2 = trapezoidal( $P_{10}, P_{18}, P_{28}, P_{36}$ )
- Clase 3 = trapezoidal( $P_{28}, P_{36}, P_{46}, P_{54}$ )

Clase 4 = trapezoidal( $P_{46}, P_{54}, P_{64}, P_{72}$ )  
 Clase 5 = trapezoidal( $P_{64}, P_{72}, P_{82}, P_{90}$ )  
 Clase 6 = hombro\_derecho( $P_{82}, P_{90}, P_{100}$ )



**Figura 22. Partición difusa en seis categorías**

Como se ha podido observar, la elección de los límites de cada clase en las distintas las particiones difusas presentadas, ha sido algo arbitraria por lo que otra persona podrá definir diferentes modelos teóricos, según su criterio, pero sin olvidar las restricciones impuestas a los marcos de cognición para que cada partición tenga una buena estructura lógica. Estos cambios no implican mayor complejidad bajo el modelo de diseño concebido, pues las reglas semánticas para la discriminación difusa se guardan en los metadatos como valores de los atributos de la tabla denominada “patrones”. De manera pues, que sólo se tendrían que actualizar las tuplas correspondientes a un patrón determinado, con los nuevos percentiles que se usarán como parámetros o las nuevas formas de los modelos difusos que representan las etiquetas lingüísticas, permitiendo que los modelos puedan ser fácilmente modificados o complementados con otros.

Con esta posibilidad que se le otorga al sistema de inferencia para discriminar en un número variable de clases o conjuntos difusos, se aprecia la flexibilidad de la técnica propuesta no sólo por la posibilidad de adaptarse a diferentes de marcos de referencia o contextos, sino por admitir diferentes niveles de granularidad en la categorización difusa.

### ***6.1.3.1 Pruebas Experimentales con el Modelo Teórico Propuesto para la discriminación borrosa de los términos vagos simples***

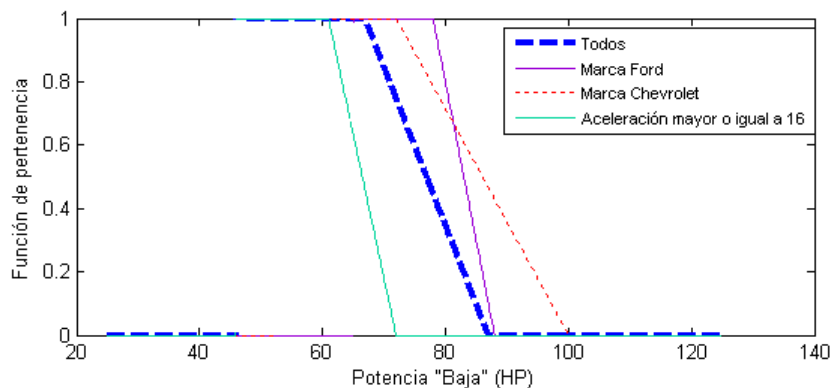
Una vez definido el modelo teórico para los conjuntos borrosos, que corresponde a cada etiqueta simple de una variable lingüística, y creados los algoritmos necesarios para el cálculo automático de los valores particulares de los mismos, se realizaron pruebas experimentales con datos reales de referencia. La base de datos tomada como problema de referencia (Benchmark) en esta Tesis Doctoral, contiene información real sobre algunas características de 398 autos.

**Tabla 12. Modelos para la potencia (HP) de los autos, considerando diferentes contextos**

Categoría	"Baja"	"Media"	"Alta"
Contexto	Hombro_izquierdo Parámetros	Trapezoidal Parámetros	Hombro_derecho Parámetros
Todos los autos (n=398)	( 46 67 87)	(67 87 105 150)	(105 150 230)
Marca Ford (n=48)	( 65 78 88)	(78 88 130 153)	(130 153 215)
Marca Chevrolet (n=46)	( 52 72 100)	(72 100 125 150)	(125 150 220)
Aceleración $\geq 16$ (n=166)	( 46 61 72)	(61 72 88 105)	(88 105 193)

Usando la base de datos de los autos, se hallaron los modelos de los conjuntos borrosos que representan la potencia de los autos, medida en caballos de fuerza (HP), considerando diferentes contextos y tres clases en el marco de condición. Los modelos hallados se presentan en la Tabla 12, donde las funciones de los extremos se denominan *hombro\_izquierdo* y *hombro\_derecho* (Cox, 1994), que son formas trapezoidales truncadas y por eso sólo requieren tres parámetros. En esta tabla, se puede observar cómo cambian los modelos de los autos considerando su potencia, según el contexto. Este comportamiento es el esperado para una variable lingüística, ratificándose además lo que se ha argumentado en la presente Tesis Doctoral: el significado de los términos vagos es relativo al contexto, exceptuando términos vagos subjetivos como la belleza o los que dependen de factores extralingüísticos como la bondad que depende de la creencia religiosa que se profese.

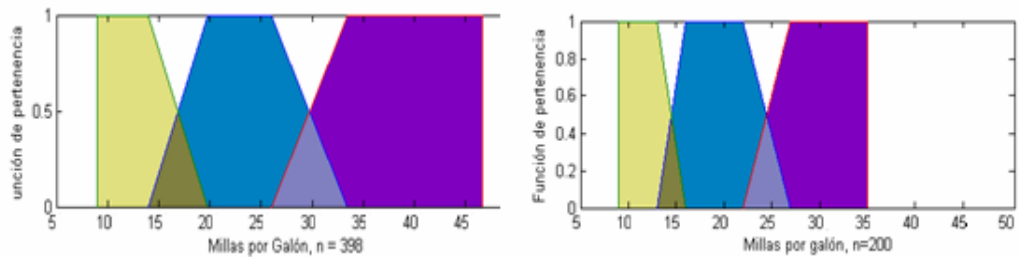
Para apreciar mejor las diferencias de los modelos ajustados a los distintos contextos, en la Figura 23 se muestran las clases difusas de los autos considerados con una potencia "baja".



**Figura 23. Modelos de potencia "baja", según diferentes contextos**

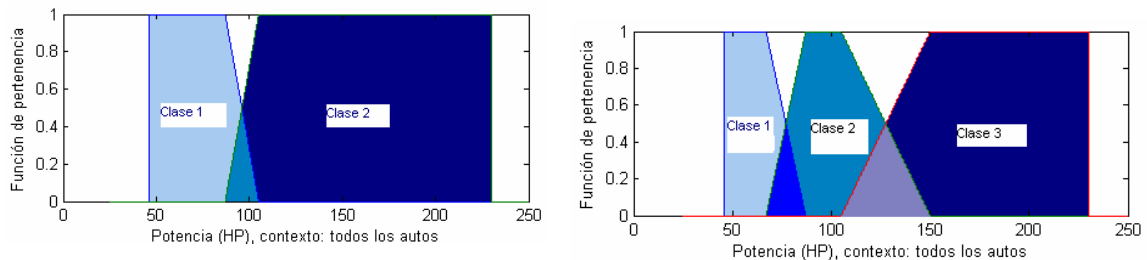
También se puede observar la dependencia de los términos vagos, del contexto que se puede delimitar en una consulta, en la Figura 24. A la izquierda, se graficaron

los conjuntos borrosos para el rendimiento de todos los autos y a la derecha, se muestra el comportamiento de la misma variable cuando los datos se restringen a los primeros 200 autos, considerando que éstos fueran los conocidos en un momento determinado de la evolución de la base de datos. Se observa que un auto con un rendimiento de 25 millas por galón se podría considerar de rendimiento “alto”, en el caso de restringir el contexto a los primeros 200, mientras no se puede considerar de esta manera en la población completa de autos.



**Figura 24. Conjuntos borrosos para el rendimiento de los autos**

Para mostrar como el sistema de inferencia, en la interpretación de los términos vagos simples de la condiciones de las consultas, puede ser adaptable no sólo a los cambios en el tiempo o en el espacio, en la Figura 25 se presentan dos marcos de cognición diferentes a los cuales se podría ajustar autónomamente dicho sistema. A la izquierda, se graficaron los conjuntos borrosos considerando dos clases, mientras que a la derecha se muestran los conjuntos borrosos considerando tres clases o etiquetas en el conjunto de términos  $T(Potencia)$  que corresponde al marco de cognición.



**Figura 25. Dos marcos de cognición para la potencia de los autos**

En la parte izquierda de la figura anterior se observa que un auto tiene potencia “baja” si el valor para este atributo se encuentra entre 50 y 100 caballos de fuerza aproximadamente, cuando se consideran dos clases en el marco de cognición. En cambio, cuando se consideran tres clases como se muestra en la parte derecha de la gráfica, este rango deja de cubrir los autos con una potencia entre 80 y 100 caballos de fuerza.

### 6.1.3.2 Funciones lineales versus no lineales

Las funciones de pertenencia propuestas aquí para la representación de conjuntos borrosos poseen la ventaja de su fácil definición y simplicidad en los cálculos. Usar una función más compleja no añadiría mayor precisión, pues se debe recordar que se están definiendo conceptos vagos. Sin embargo, parece existir una limitante en el uso de las funciones lineales, como la trapezoidal, para representar conjuntos borrosos de las etiquetas vagas, pues si se necesita aplicar una función de intensificación o dilatación, cuando un calificativo vago esté acompañado de un adverbio de cantidad o intensidad  $m$ , el núcleo de la función permanecerá igual, lo que implicaría casi ninguna diferencia entre los conjuntos borrosos de los “altos” y los “muy altos”, por ejemplo.

Para justificar lo que se acaba de expresar, sea el conjunto trapezoidal  $(a, b, c, d)$  la representación de una etiqueta lingüística  $E$ . Por lo tanto, el soporte está determinado por  $a$  y  $d$  y el núcleo de la función está delimitado por los parámetros  $b$  y  $c$ . Esto último significa que los valores del *Universo de Discurso* que se encuentren entre  $b$  y  $c$ , se les asigna el valor de 1 como grado de pertenencia al conjunto difuso. Por otro lado, las funciones de potencia usadas convencionalmente para intensificar (concentrar) o relajar el sentido de un término vago, como se había mencionado antes, son:

$$\text{“Muy } E\text{”} = \text{CON}(E) = E^2$$

$$\text{“Más o menos } E\text{”} = \text{DIL}(E) = E^{1/2}$$

$$\text{“Extremadamente } E\text{”} = \text{CON}(E) = E^3$$

De acuerdo con esas fórmulas y puesto que la raíz cuadrada de 1 es 1, y este valor elevado a cualquier potencia siempre da como resultado esta misma cantidad, los núcleos de las funciones derivadas de un conjunto difuso trapezoidal, por medio de la función de potencia, permanecen inalterables. Debido a esto, lo único que cambiaría en el caso de aplicar dichas funciones, sería la forma de las partes laterales del trapecio. Esto mismo sucede en las formas trapezoidales truncadas que sirven para representar las clases de los extremos. En la Figura 26, se puede apreciar el pobre efecto de la función de potencia, pues la forma punteada corresponde a la acentuación de las clases con el adverbio de cantidad “muy”.

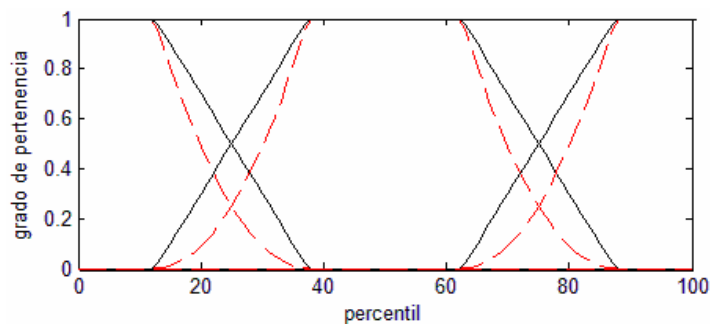
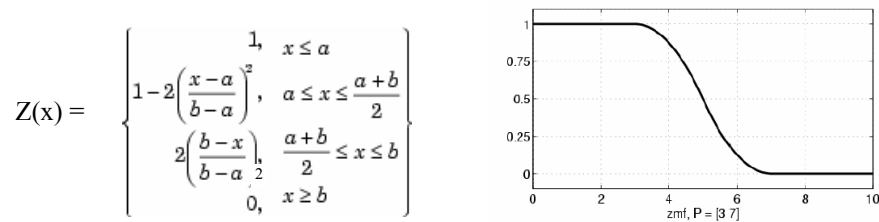


Figura 26. Formas trapezoidales y su acentuación con el adverbio “muy”

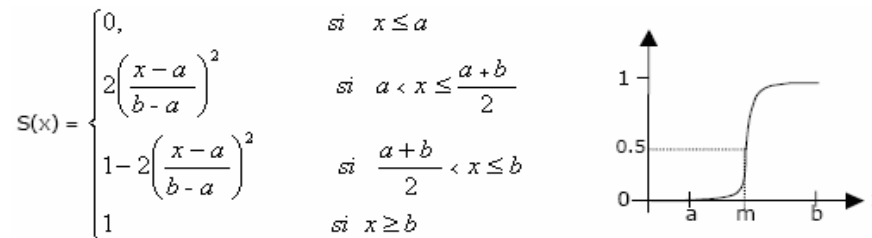
Por lo anterior, considerando un pobre efecto del operador de la acentuación definido como la función de potencia, se vio la necesidad de comparar las funciones lineales con funciones no lineales como la función Z, la función  $\Pi$  y la función S para la definición de los conjuntos borrosos, en el caso de considerar tres etiquetas o categorías diferentes. La función Z se usa para representar al conjunto de etiqueta “menor” o “pequeño”, la función  $\Pi$  sirve para representar los conjuntos intermedios y la S para representar el de los valores “altos” o “grandes”. La definición de la función Z y su forma, que depende de dos parámetros ( $a$  y  $b$ ), aparece en la Figura 27



**Figura 27. Función Z y su representación gráfica**

Los parámetros  $a$  y  $b$  de la función Z determinan los puntos de inflexión de la función Z. En el ejemplo presentado estos parámetros son  $a = 3$  y  $b = 7$ .

Para el modelo no lineal de la clase etiquetada con un término vago como “alto” o “grande”, se ha elegido la función S, que es el complemento de la función Z. La función genérica S se presenta en la Figura 28.



**Figura 28. Función S y su representación gráfica**

En el modelo no lineal para representar de manera suavizada la subclase borrosa de los “medianos” o “medios”, se considera apropiada la función  $\Pi$  (Matworks, 2006). Ésta se define como el producto de las dos funciones de pertenencia S y Z. Esto es,  $\Pi(a, b, c, d) = S(a, b) * Z(c, d)$ .

Como en el caso lineal cuando se consideran tres clases, se utilizará la estadística de resumen de los cinco números para definir los parámetros de las funciones de pertenencia no lineales, porque se pretende partir de la misma discriminación nítida inicial. De modo que, siendo congruentes con las definiciones dadas para las funciones de pertenencia lineales, los valores de  $a$  y  $b$  corresponderían, en la función Z, a los percentiles  $P_{12.5}$  y  $P_{37.5}$ , respectivamente. Sin

embargo, al proceder de este modo, el problema de la invariabilidad del núcleo de la función cuando se necesite enfatizar o relajar un término vago, persistiría. Esto mismo sucedería con los modelos de los conjuntos de los “medianos” y de los “altos”. Por esto, se requiere que las funciones no lineales utilizadas para representar los conjuntos borrosos sean unimodales para ver si es posible lograr un cambio significativo en los modelos originales cuando se necesite representar términos acentuados.

De acuerdo con lo anterior, en la discriminación borrosa usando formas no lineales de la función de pertenencia, en la clase de los “menores” o los “bajos” se fijará como aproximación de  $a$ , parámetro de la función  $Z$ , al valor mínimo de todos los elementos del *Universo de Discurso*. Es decir, el percentil  $P_0$ , el valor en el dominio, delimitado por el contexto, que no deja ningún porcentaje de valores a su izquierda. El valor de  $b$  de esta función debe ser la mediana  $P_{50}$ , con el fin cumplir la restricción de complementariedad (Restricción No. 13, sección 4.5.6) con la clase de los valores medios. Por lo tanto, el punto de inflexión de la función no lineal es  $(P_0 + P_{50})/2$ .

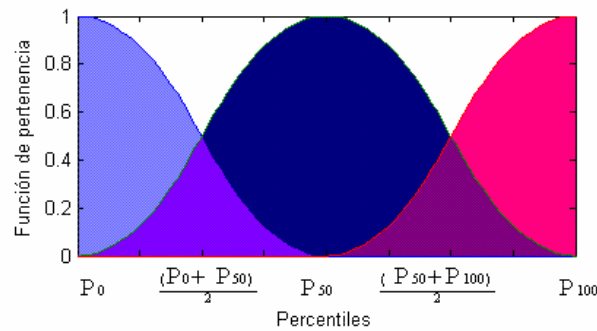
De forma análoga a la manera de proceder en el caso anterior, los parámetros de la función  $S$ , que determinan el soporte del conjunto borroso de los “altos” o “grandes”, se estimarán fijando a  $a = P_{50}$  y a  $b = P_{100}$ . Este último estimador es el valor máximo de todos los elementos del *Universo de Discurso* y el punto del inflexión se calculará, consecuentemente, con  $(P_{50} + P_{100})/2$ . Por otro lado, el conjunto borroso correspondiente a la etiqueta de los “medianos”, cuya forma es la función  $\Pi$ , debe tener como soporte a todo el *Universo de Discurso*, para lograr la complementariedad deseada con los otros dos conjuntos borrosos con los que se solapa. Por lo tanto, los puntos de inflexión deben ser  $(P_0 + P_{50})/2$  y  $(P_{50} + P_{100})/2$ . El parámetro adicional es el punto medio, que se estima con la mediana.

De acuerdo con lo anterior, en la Tabla 13, se presentan las reglas semánticas generales para la construcción de los modelos borrosos no lineales para la representación de los términos vagos simples, considerando tres categorías o clases en el marco de cognición.

**Tabla 13. Modelo teórico para las clases borrosas, basado en funciones no lineales**

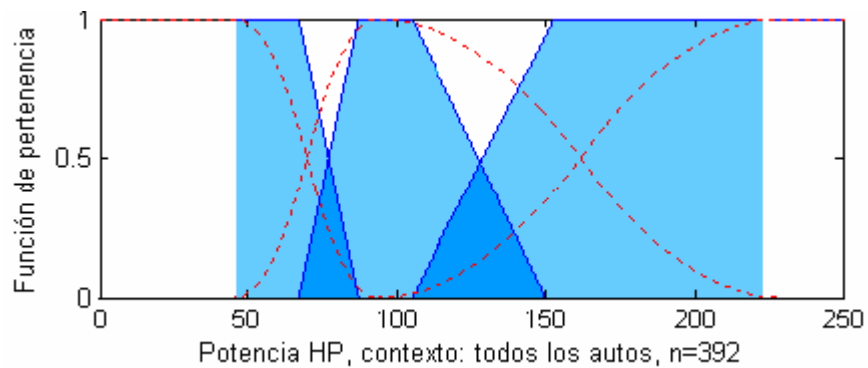
Clase Borrosa	Dominio		Reglas semánticas para la definición de la función de pertenencia
	Límite Inferior	Límite Superior	
“Menor” o “Bajo”	$P_0$	$P_{50}$	$Z(P_0, P_{50})$
“Mediano” o “Medio”	$P_0$	$P_{100}$	$\Pi(P_0, P_{50}, P_{100})$
“Mayor” o “Alto”	$P_{50}$	$P_{100}$	$S(P_{50}, P_{100})$

Ya definidos los estimadores de todos los parámetros de las funciones de pertenencia no lineales y unimodales de las tres clases borrosas, en la Figura 29 se presenta la partición borrosa propuesta. En ella, en el eje de las X se presentan los percentiles de la distribución de los datos. La simetría allí mostrada sólo se presentará en el caso de una distribución uniforme de los datos, en el contexto considerado.



**Figura 29. Modelos no lineales para la representación de conjuntos borrosos**

Para la comparación de los modelos lineales y no lineales, en la Figura 30 se muestran los modelos de los conjuntos borrosos ajustados para la potencia de los autos. Se aprecia cómo la forma sigue conservando una estructura parecida, pero también se ve cómo se amplían las zonas de solapamiento entre las clases definidas por los modelos no lineales. También se observa que la clase intermedia no es simétrica y que la clase con mayor variabilidad intragrupal es la clase de los autos con potencia "alta".



**Figura 30. Modelos lineales y no lineales, considerando tres clases**

Las diferencias más notables entre las distribuciones lineales y no lineales, se encuentran en el soporte y en el núcleo de las funciones de pertenencia de los conjuntos borrosos, como era de esperarse. Por ejemplo, el soporte de la clase intermedia contiene a los autos con una potencia entre 48 y 224 HP, usando el modelo no lineal, mientras que este rango se restringe a los autos con una potencia

entre 68 y 150, en el modelo lineal. Sin embargo, no se puede descartar la utilización de las formas no lineales unimodales para la representación de los conjuntos difusos, hasta no examinar la manera de acentuar o cambiar estos modelos en la interpretación de adverbios de cantidad. Y a pesar de las diferencias encontradas en las funciones de pertenencia, se puede verificar que tanto para los modelos lineales, como los no lineales se cumple que la suma de los grados de pertenencia de un elemento a todos los conjuntos difusos del marco de cognición, es uno (Restricción No. 13, sección 4.5.6). Esto es:

$$\mu^{\text{potencia\_baja}}(x) + \mu^{\text{potencia\_media}}(x) + \mu^{\text{potencia\_alta}}(x) = 1 \quad \forall x \in U$$

## 6.2 Preguntas de tipo II (con términos vagos simples acompañados de un modificador)

Un modificador lingüístico cambia los valores de verdad de una sentencia, dando origen a un nuevo elemento en el conjunto de términos  $T(x)$  de una variable lingüística, a partir de otro de los términos básicos o primarios considerados para esta variable. Para la variable edad, por ejemplo, su conjunto de términos  $T(X)$  está conformado por términos primarios como “joven”, “maduro” o “viejo” que pueden originar otro, por medio de la negación o con adverbios de cantidad como “muy joven” o “extremadamente viejo”. Todos estos modificadores se consideran operadores, pues cambian el significado de los operandos, que en este caso corresponden a los adjetivos calificativos.

### 6.2.1 Reglas Sintácticas para los Modificadores Lingüísticos

Para incorporar un modificador lingüístico, que puede ser denotado con la letra  $m$ , en el lenguaje de consulta, se debe ampliar la sintaxis de la consulta SQL para que permita anteponer el operador  $m$  al adjetivo o etiqueta  $E$ , en notación prefijo, así:

```
SELECT proyección
FROM relaciones
WHERE expresión IS [m] E [j] [k]
      [WITH CALIBRATION n|threshold|n, threshold]
```

Esta sintaxis propuesta, sigue las reglas sintácticas del lenguaje natural (en este caso, del inglés como en la expresión “where weight is very heavy”) al cual trata de aproximarse el SQL.

### 6.2.2 Reglas Semánticas para los Modificadores Lingüísticos

Regularmente, para hallar la semántica de los términos derivados en  $T(X)$ , primero se definen los conjuntos difusos de los términos básicos o primarios y a partir de éstos se calculan los conjuntos difusos de los términos compuestos. Por esto, la compatibilidad o concordancia de una tupla con una etiqueta lingüística modificada por un adverbio o por la negación, se infiere de manera deductiva de la función de pertenencia definida para la etiqueta lingüística primaria que le da origen. Convencionalmente, tal derivación se obtiene aplicando una función

matemática sobre el modelo de la etiqueta lingüística de interés (Cox, 1994). A continuación, se analizan los dos posibles modificadores: la negación y los adverbios de cantidad, de manera separada debido a sus diferencias semánticas, pero admitiendo que pueden coexistir en una sola sentencia. Esto significa que la sintaxis debe ser ampliada, así:

```
SELECT proyección
FROM relaciones
WHERE expresión IS [NOT][m] E[j][k]
      [WITH CALIBRATION n|threshold|n, threshold]
```

### 6.2.2.1 Reglas Semánticas para la Negación

La negación representa el complemento de una etiqueta lingüística. Por lo tanto, el grado de pertenencia de una tupla al conjunto borroso definido por esa etiqueta “negada” debe ser igual al grado de pertenencia al complemento de ese conjunto.

Existen varias propuestas para hallar el valor de pertenencia al complemento de un conjunto vago que cumplen con las restricciones de borde, monotonicidad e involución impuestas a los operadores de la negación. Tres de esas propuestas son el complemento de Sugeno o el de Yager, además de la definición clásica, donde  $NO(a) = (1 - a)$  (Yager, 1980). Sin embargo, los modelos teóricos para la construcción de los conjuntos borrosos en este trabajo de investigación cumplen, por definición, la restricción de complementariedad (Restricción No 13, sección 4.5.6) y cobertura (Restricción No. 12, sección 4.5.5). Por eso, también se cumple la ecuación:

$$\mu_E(t_i) + \mu_E^c(t_i) = \mu_E(t_i) + \mu_{NO(E)}(t_i) = 1, \forall E_j \in \text{Marco de cognición} \wedge \forall t_i \in R$$

$$\text{Luego, } \mu_E^c(t_i) = 1 - \mu_E(t_i) \forall t_i \in U$$

Por lo anterior, se ve conveniente emplear en esta Tesis Doctoral, la definición clásica de la negación para determinar el grado de pertenencia de una tupla  $t_i$ , de la relación  $R$ , al complemento de un conjunto difuso. Por ejemplo, si se necesita encontrar el grado de pertenencia de una persona al complemento del conjunto de los “jóvenes”, el agente inteligente lo puede obtener así:

$$\mu^{\text{“jóvenes”}^c}(t_i) = 1 - \mu^{\text{“jóvenes”}}(t_i) \forall t_i \in R$$

El operador clásico de la negación es estrictamente decreciente y es un tipo de negación fuerte pues cumple con la ley de involución definida para el álgebra de Bool (Fodor, 2004).

### 6.2.2.2 Reglas Semánticas para los Adverbios de Cantidad

Un adverbio de cantidad como “muy”, “poco” o “demasiado” se considera un operador unario que acentúa o relaja el significado de un operando, en este caso un adjetivo. Si la etiqueta  $E$  es un valor lingüístico caracterizado por un conjunto

borroso con función de pertenencia  $\mu_E(x)$ , entonces  $E^k$  ha sido interpretada como una versión modificada del valor lingüístico original y es definida por el grado de pertenencia  $\mu_E(x)^k$ .

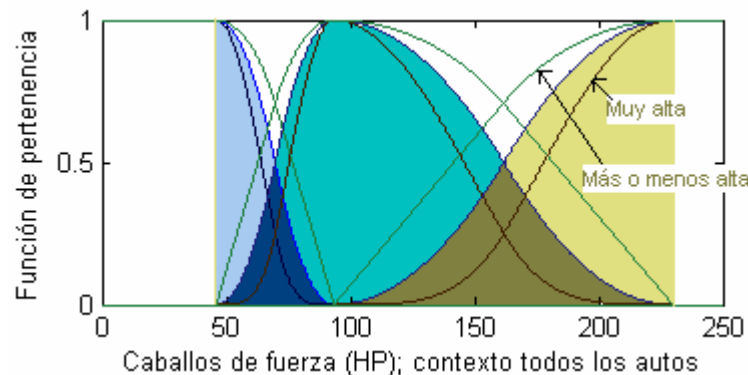
La función anterior, es conocida como la función de potencia de un conjunto difuso y dependiendo del valor del exponente  $k$  se denominan funciones de dilatación o de concentración. Convencionalmente, se utilizan las representaciones matemáticas para los adverbios de cantidad vagos (Jang, Sun y Mizutani, 1997; Cox, 1994):

$$\text{"Muy } E\text{"} = \text{CON}(E) = E^2$$

$$\text{"Más o menos } E\text{"} = \text{DIL}(E) = E^{1/2}$$

$$\text{"Extremadamente } E\text{"} = \text{CON}(E) = E^3$$

En la Figura 31 se muestra un ejemplo de aplicación de la función de concentración  $E^2$  y la función de dilatación  $E^{1/2}$ , para la potencia de todos los autos, medida en caballos de fuerza. A simple vista, se puede observar que las densidades (o las áreas bajo la curva) que representan a los subconjuntos enfatizados con el adverbio "muy" superan el 75% de la densidad del conjunto del cual fueron derivados, mostrando poco efecto del operador "muy" sobre las clases originales.



**Figura 31. La concentración y dilatación, basada en la función de potencia**

El efecto de aplicar el operador de dilatación  $E^{1/2}$ , para representar el conjunto con etiqueta "más o menos  $E$ " sería "inflar" un poco la densidad del conjunto borroso que representa una etiqueta lingüística, como se muestra con los datos de ejemplo. De forma similar al operador de concentración o acentuación, no genera cambios significativos en el soporte del conjunto, ni en el núcleo. Significando con ello que casi todos los elementos del conjunto que se quiere precisar, o acentuar, pertenecen también al subconjunto derivado. Los elementos con un alto grado de pertenencia a la clase original, siguen conservando un grado de pertenencia casi igual para la clase nueva.

Di Lascio, Gisolfi y Loia señalan que existe una discrepancia entre el uso intuitivo en el lenguaje natural de los adverbios de cantidad y los valores numéricos obtenidos con los operadores de dilatación y concentración definidos por Zadeh (Di Lascio, Gisolfi y Loia, 1996). Debido a esto, proponen tres parámetros,

independientemente de la función  $\mu(x)$ , para dar una dimensión cuantitativa a las características implícitas en un significado.

Para la posición espacial del prototipo en el conjunto derivado, en el *Universo de Discurso*, proponen como estimador de la tendencia central  $TC$  a:

$$TC = \left| \frac{b+a}{2} \right| - c, c < a \in U$$

Donde la constante  $c$  asociada con  $TC$  es el modificador de translación. De acuerdo con la fórmula, lo único que se exige a la constante de translación  $c$  es que sea menor que  $a$ . Es decir, no se la acota por la izquierda y por lo tanto, la translación podría originar conjuntos cuyo dominio esté por fuera del *Universo de Discurso*. Además, no se sugieren valores específicos para dicha constante.

En (Marín y Shen, 1999), se coincide en afirmar que las definiciones convencionales de los acentuadores o las aristas no originan cambios significativos. Para las funciones trapezoidales, con parámetros  $(a, b, c, d)$ , específicamente, proponen el siguiente cálculo para los parámetros del conjunto modificado:

$$m = (b + c) / 2$$

$$b' = m - ((m - b) * \beta)$$

$$a' = b' - ((b - a) * \beta)$$

$$c' = m + ((c - m) * \beta)$$

$$d' = c' + ((d - c) * \beta)$$

Donde  $\beta \in (0,1)$  es un parámetro que controla el grado de reducción debido a la aplicación del modificador de concentración. En particular, proponen utilizar para el modificador “más o menos  $E$ ” a  $\beta = 2/3$ , para “muy  $E$ ” a  $\beta = 1/2$  y para “extremadamente  $E$ ” a  $\beta = 1/8$ . Para ver el efecto de la aplicación de los distintos modificadores propuestos por Marín y Shen, se utilizó el conjunto de los autos con una potencia “alta” en la base de datos de referencia. El modelo del conjunto original de vehículos con potencia “alta” es una función hombro\_derecho(105 150 230) por lo cual para la obtener el modelo de los autos con potencia “muy alta” las fórmulas que se deben aplicar son:

$$m = (150 + 230) / 2 = 190$$

$$b' = 190 - (190 - 150) * 0.5 = 170$$

$$a' = 170 - ((150 - 105) * 0.5) = 147.5$$

$$c' = 190 + ((230 - 190) * 0.5) = 210$$

Por lo tanto, esta técnica dio origen al modelo hombro\_derecho(147.5 170 210) para los autos con potencia “muy alta”. Para la clase de autos con potencia “más o menos alta”, el modelo que se obtiene es hombro\_derecho(133 163 217) puesto que los valores para los tres coeficientes son los que aparecen a continuación.

$$b' = 190 - (190 - 150) * (2/3) = 163$$

$$a' = 163 - ((150 - 105) * (2/3)) = 133$$

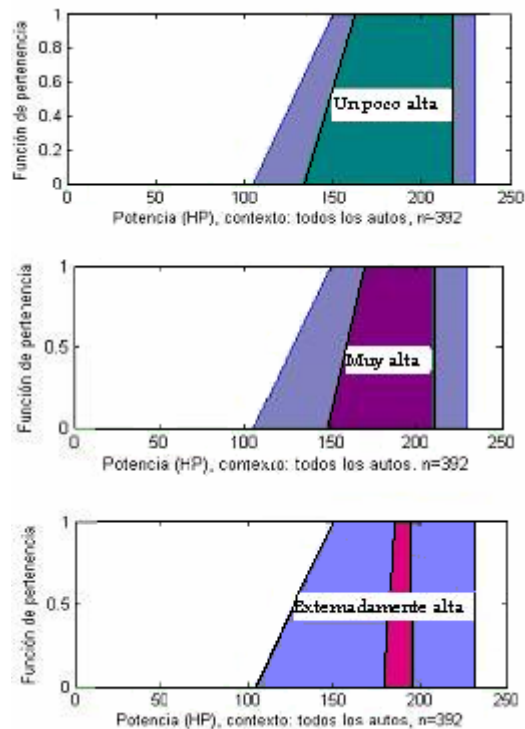
$$c' = 190 + ((230 - 190) * (2/3)) = 217$$

Para representar los autos con potencia "extremadamente alta" el modelo hombro derecho tiene los parámetros (179 185 195) puesto que:

$$b' = 190 - (190 - 150) * (1/8) = 185$$

$$a' = 185 - ((150 - 105) * (1/8)) = 179$$

$$c' = 190 + ((230 - 190) * (1/8)) = 195$$



**Figura 32. Efecto de los modificadores sobre la potencia "alta" de los autos**

Para apreciar mejor los efectos de los modificadores propuestos por Marín y Shen, se construyó la Figura 32. Observando los resultados en la figura, se evidencia que tienen mayor efecto que los producidos con el uso de la función de potencia, para encontrar los nuevos conjuntos borrosos de los adjetivos modificados con un adverbio. Sin embargo, se observa que los conjuntos derivados no llegan al límite superior del conjunto original y, paradójicamente, la clase más acentuada está más alejada del valor extremo del dominio de la variable de interés. No obstante, en esta propuesta, se cumplen las relaciones de subconjunto:

$$"Extremadamente E" \subseteq "muy E" \subseteq "E".$$

Esa interpretación de los conjuntos que representan las etiquetas derivadas, se adhiere a la interpretación inclusiva que se propone en (Zadeh, 1972) y que es acogida por otros autores. Pero como lo señalan De Cock y Kerre, y como se ha mostrado en este trabajo, las funciones de potencia o de translación que se han propuesto para la definición de los modificadores lingüísticos como “extremadamente” o “muy” son sólo herramientas técnicas que conservan la propiedad de inclusión entre los subconjuntos difusos obtenidos, con el conjunto original etiquetado  $E$ , pero que carecen de significado como propiedad inherente (De Cock y Kerre, 2004).

Por lo anterior, para representar los adverbios de cantidad como “Extremadamente” o “Muy”, en la presente Tesis Doctoral, se decide por una opción en la cual el razonamiento podría acercarse más a lo que haría un humano cualquiera para la interpretación de estos adverbios de cantidad. Esta estrategia consiste en volver a realizar el proceso de discriminación sobre el conjunto que representa la etiqueta de interés y sobre la cual se pretende aplicar un modificador de acentuación. A continuación, se describen las reglas semánticas para una interpretación concebida de esta manera.

### 6.2.2.3 Reglas Semánticas para el Adverbio “Muy”

El nuevo mecanismo de razonamiento aproximado que deberá realizar el agente inteligente, para encontrar la semántica del adverbio “muy” sigue el mismo proceso de discriminación de los términos vagos simples, pero no sobre todo el *Universo de Discurso*, sino sobre el dominio del conjunto que corresponde a la etiqueta lingüística  $E$ , que deba ser modificada con este adverbio.

Puesto que aquí se adoptaron las formas lineales para la representación de los términos vagos, por ser más sencillas en su cálculo que las no lineales, la partición del conjunto difuso que deba ser modificado con el adverbio “muy” también debe dar origen a formas trapezoidales o trapezoidales truncadas. La clase difusa que deba ser modificada por una etiqueta lingüística se subdividirá en tres subconjuntos borrosos utilizando como parámetros, estadísticas de posición. Se parte en tres para lograr buen pronunciamiento o acentuación de la clase.

Si se trata de la clase de los valores menores, y puesto que el límite inferior de ésta es  $P_0$  y el límite superior es  $P_{37.5}$ , entonces los límites del subconjunto borroso “muy bajo(a)” o “muy pequeño(a)” son el valor mínimo  $P_0$  y el percentil  $P_{37.5}$  de ese conjunto, que equivale al percentil  $P_{14}$  del conjunto de todos los datos del *Universo del Discurso*. El núcleo de la distribución de la función de pertenencia a este nuevo conjunto debe contener al 12.5% de los datos menores del 37.5% de la clase con etiqueta “baja” o “pequeña”, si existe congruencia con la primera discriminación realizada para hallar los modelos de los términos vagos simples. Por lo tanto el núcleo de la clase de los valores menores equivale al 4.7% de todos los datos y queda delimitado por  $P_0$  y  $P_{4.7}$ .

Un razonamiento similar es aplicable a clase de los “mayores” o “altos”. Por lo tanto, la cardinalidad de esta clase equivale al 14% de los elementos mayores en

todos los datos. Para la clase intermedia, el término “muy mediano”, que se puede considerar sinónimo de “muy típico” o “muy corriente”, debe filtrar también el mismo porcentaje de datos que en los dos casos anteriores pero incluyendo los elementos cercanos al punto medio de la distribución, bien sea que estén por encima o por debajo de este valor que será estimado con la mediana para tener un estimador más robusto del parámetro central (que no se deje influenciar por valores extremos o por sesgos en la distribución como la media aritmética). Por lo tanto, los valores del dominio del subconjunto acentuado estarán entre  $P_{43}$  y  $P_{57}$ . Para encontrar los valores que determinan el núcleo de la nueva función de pertenencia, se considerará que éste contenga el mismo porcentaje de datos que las clases de los extremos acentuadas. Es decir, el 4.7% de los valores centrales y por lo tanto, el núcleo queda delimitado por  $P_{48}$  y  $P_{52}$ .

De acuerdo con lo anterior, el modelo teórico para encontrar la semántica del adverbio “muy” sobre las distintas clases del marco de cognición puede derivarse de los datos del contexto, tal como se especifica en la Tabla 14.

**Tabla 14. Modelo teórico para el adverbio “muy”, basado en funciones lineales**

Clase Borrosa	Dominio		Reglas semánticas para la definición de la función de pertenencia
	Límite Inferior	Límite Superior	
“Muy bajo(a)” o “Muy pequeño(a)”	$P_0$	$P_{14}$	$\begin{cases} 1 & \text{si } P_0 \leq x \leq P_{4.7} \\ (P_{14} - x)/(P_{14} - P_{4.7}) & \text{si } P_{4.7} < x \leq P_{14} \\ 0 & \text{en otro caso} \end{cases}$
“Muy típico(a)” o “Muy corriente”	$P_{43}$	$P_{57}$	$\begin{cases} (x - P_{43})/(P_{48} - P_{43}) & \text{si } P_{43} < x \leq P_{48} \\ 1 & \text{si } P_{48} < x \leq P_{52} \\ (P_{57} - x)/(P_{57} - P_{52}) & \text{si } P_{52} < x \leq P_{57} \\ 0 & \text{en otro caso} \end{cases}$
“Muy Alto(a)” o “Muy grande”	$P_{86}$	$P_{100}$	$\begin{cases} 0 & \text{si } P_0 \leq x \leq P_{86} \\ (x - P_{86})/(P_{95} - P_{86}) & \text{si } P_{86} < x \leq P_{95} \\ 1 & \text{si } P_{95} \leq x \leq P_{100} \end{cases}$

En la Figura 33, se muestra con líneas punteadas la partición borrosa realizada sobre los conjuntos idealizados de las etiquetas vagas simples, para representar los subconjuntos acentuados con el adverbio “muy”, de acuerdo con los modelos teóricos recién definidos. Allí se aprecia claramente el efecto de la acentuación lograda con la técnica propuesta para este adverbio.

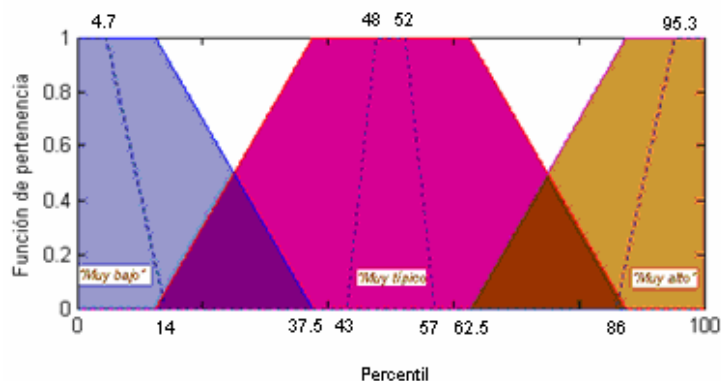


Figura 33. Partición borrosa para representar el adverbio “muy”

#### 6.2.2.4 Reglas Semánticas para el Adverbio “Extremadamente”

El adverbio “extremadamente” o “supremamente”, sinónimo de “altísimo”, “extremadamente común” o “bajísimo” según la clase considerada, se puede interpretar como la acentuación del término “muy E”. Por lo tanto, se particionaría el conjunto ya acentuado, en tres subconjuntos borrosos nuevamente.

Tabla 15. Modelos teóricos para la representación del adverbio “extremadamente”

Clase Borrosa	Dominio		Reglas semánticas para la definición de la función de pertenencia
	Límite Inferior	Límite Superior	
“Extremadamente bajo(a) o pequeño”	$P_0$	$P_2$	$\begin{cases} 1 & \text{si } P_0 \leq x \leq P_1 \\ (P_2 - x)/(P_2 - P_1) & \text{si } P_1 < x \leq P_2 \\ 0 & \text{si } P_2 \leq x \end{cases}$
“Extremadamente común o típico”	$P_{49}$	$P_{51}$	$\begin{cases} (x - P_{49})/(P_{50} - P_{49}) & \text{si } P_{49} < x \leq P_{50} \\ (P_{51} - x)/(P_{51} - P_{50}) & \text{si } P_{50} < x \leq P_{51} \\ 0 & \text{en otro caso} \end{cases}$
“Extremadamente Alto (Mayor)”	$P_{88}$	$P_{100}$	$\begin{cases} 0 & \text{si } x \leq P_{98} \\ (x - P_{98})/(P_{99} - P_{98}) & \text{si } P_{98} < x \leq P_{99} \\ 1 & \text{si } P_{99} \leq x \leq P_{100} \end{cases}$

Realizando un procedimiento similar al descrito para precisar el significado de una etiqueta lingüística modificada con el adverbio “muy”, se encuentran las reglas para determinar la semántica de un término vago simple acentuado “Extremadamente E”. Esta nueva opción constituye una alternativa a la función de potencia con un valor para el exponente de 3 que se ha propuesto convencionalmente en la literatura (Jang, Sun y Mizutani, 1997; Cox, 1994). En la

Tabla 15, se presentan los modelos teóricos propuestos para los conjuntos borrosos que representan a términos acentuados con el adverbio “extremadamente”.

### 6.2.2.5 Reglas Semánticas para el Adverbio “Más o menos”

El adverbio “más o menos”, considerado sinónimo de “un poco” o de “relativamente”, indica una interpretación un poco más relajada del concepto vago que modifica. Por lo tanto, para representar este adverbio vago se deben considerar los elementos que se encuentran más allá de la intersección del conjunto con las clases adyacentes.

Es por eso que para definir el soporte del nuevo conjunto modificado por el adverbio “más o menos” aquí se opta por incluir los elementos que superen a la mediana de la clase de la clase adyacente, cuando ésta sea menor o incluir los elementos que sean menores que la mediana de su categoría cuando la clase adyacentes sea mayor a la considerada. Y el núcleo de la función también se amplía hasta cubrir los elementos que pertenezcan a la intersección de la clase con las adyacentes.

Puesto que el valor de la mediana y los límites que fijan para la intersección de las clases, sólo se pueden determinar conociendo el número de categorías en la discriminación difusa, la fórmula para la interpretación del adverbio “más o menos” queda especificada como aparece a continuación.

$$\begin{aligned} \text{“Más o menos } E_i \text{”} = & \\ & \left\{ \begin{array}{l} \text{hombro\_izquierdo}(\text{mediana}(E_i), Li(E_{i+1}), \text{mediana}(E_{i+1}) \exists E_{i+1} \in \text{Marco} \wedge E_{i-1} \notin \text{Marco}) \\ \text{trapezoidal}(\text{mediana}(E_{i-1}), Ls(E_{i-1}), Li(E_{i+1}), \text{mediana}(E_{i+1}) \exists E_{i-1}, E_{i+1} \in \text{Marco}) \\ \text{hombro\_derecho}(\text{mediana}(E_{i-1}), Li(E_i), \text{mediana}(E_i) \exists E_{i-1} \in \text{Marco} \wedge E_{i+1} \notin \text{Marco}) \end{array} \right\} \end{aligned}$$

De acuerdo con la fórmula anterior, el conjunto que representa la relajación de un conjunto debido a la aplicación del modificador vago “más o menos” será hombro\_izquierdo, si se trata de la primera clase en el marco de cognición, una función trapezoidal si es intermedia y para el caso de la clase con los valores más altos, la clase modificada será un conjunto difuso hombro\_derecho.

### 6.2.3 Reglas de Interpretación de Composiciones

Dentro de las condiciones especificadas como criterios de filtrado en una consulta, también es posible que se planteen expresiones vagas más complejas que resultan del uso de las conectivas lógicas de la disyunción y de la conjunción. En el presente trabajo de investigación también se abordará el problema de representar y manejar términos vagos que pueden deducirse de una combinación lineal de condiciones vagas. Se tratarán los dos casos, separadamente.

#### 6.2.3.1 Reglas Sintácticas para composiciones con conectivas lógicas

Si una consulta contiene un criterio compuesto o agregado mediante conectivas lógicas, la respuesta dependerá de las funciones de pertenencia a cada

uno de los componentes del término agregado. El agente inteligente aplicará el modelo FITA que consiste en primero inferir y luego agregar, con el fin de ajustarnos a las reglas de interpretación de las preguntas tipo II, especificadas en el lenguaje teórico PRUF para sistemas de inferencia difusos.

Debe resaltarse que para incorporar las conectivas lógicas que permitan definir la expresión vaga que conforma una composición, no hay necesidad de extender la sintaxis del lenguaje original, pues en él ya existen las conectivas AND y OR. Por lo tanto, la sintaxis de la consulta SQL no tiene que modificarse o ampliarse. Luego, la sintaxis de una consulta que admita la composición de  $p$  criterios vagos o concretos es:

```
SELECT proyección
FROM relaciones
WHERE expr IS [NOT] [m] E [j] [k] {AND|OR}
      expr IS [NOT] [m] E [j] [k]...
[WITH CALIBRATION n|threshold|n, threshold]
```

Puesto que el grado de pertenencia de un objeto a un conjunto convencional también existe, aunque tome sólo dos valores (verdadero y falso), en una expresión vaga formada por la conjunción de etiquetas, también es posible incluir restricciones concretas sobre algunas de las variables que se quieran considerar en los criterios de filtrado. Para una condición concreta de una consulta, el grado pertenencia de una tupla cualquiera será 1 o 0, solamente. Por lo tanto, la sintaxis se puede generalizar así:

```
SELECT proyección
FROM relaciones
WHERE condición1 {AND|OR} condición2 [{AND|OR}...]
[WITH CALIBRATION n|threshold|n, threshold]
```

### 6.2.3.2 Reglas Semánticas para composiciones con conectivas lógicas

Sobre conjuntos difusos se admiten las operaciones tradicionales como la unión e intersección. Para este tipo de conjuntos estas operaciones se determinan, mediante las funciones de pertenencia, así:

$$\mu_{A \cup B}(x) = \oplus(\mu_A(x), \mu_B(x)) \quad \forall x \in U$$

$$\mu_{A \cap B}(x) = \otimes(\mu_A(x), \mu_B(x)) \quad \forall x \in U$$

Donde los símbolos  $\oplus$  y  $\otimes$  representan los operadores para las conectivas lógicas de la disyunción y la conjunción, respectivamente. El problema, entonces, para representar estas operaciones de conjunto, estriba en la selección apropiada de los operadores que las representan.

Como se señala en (Pradera et al., 2006), no sólo los grados de pertenencia a los conjuntos borrosos son dependientes del contexto, sino también las operaciones entre conjuntos. En algunos dominios será más importante el cumplimiento de unas propiedades matemáticas de los elementos que conforman una estructura algebraica, que otras.

En las teorías de conjuntos difusos, se han propuesto varias alternativas para representar la semántica de los operadores de la conjunción y la disyunción que cumplen con las restricciones impuestas como *s*-normas y *t*-conormas, pero que dejan de cumplir algunas propiedades deseables para las expresiones lógicas derivadas.

En la Tabla 16, se presentan algunos de los operadores sugeridos para las conectivas lógicas y las leyes o propiedades que incumplen del álgebra de Boole, según se presenta en (Dubois y Prade, 2000).

**Tabla 16. Algunos operadores propuestos para las conectivas lógicas**

Conjunción $\otimes (\mu_A(x), \mu_B(x))$	Disyunción $\oplus (\mu_A(x), \mu_B(x))$	Negación $\neg \mu_A(x)$	Leyes que se incumplen
$\min(\mu_A(x), \mu_B(x))$	$\max(\mu_A(x), \mu_B(x))$	$1 - \mu_A(x)$	<ul style="list-style-type: none"> <li>• Ley del medio excluido</li> <li>• Ley de no contradicción</li> </ul>
$\max(\mu_A(x) + \mu_B(x) - 1, 0)$	$\min(1, \mu_A(x) + \mu_B(x))$	$1 - \mu_A(x)$	<ul style="list-style-type: none"> <li>• Idempotencia</li> </ul>
$\mu_A(x) \times \mu_B(x)$	$\mu_A(x) + \mu_B(x) - (\mu_A(x) \times \mu_B(x))$	$1 - \mu_A(x)$	<ul style="list-style-type: none"> <li>• Ley del medio excluido</li> <li>• Ley de no contradicción</li> <li>• Idempotencia</li> </ul>

La ley de no contradicción, definida como  $\otimes (\mu_A(x), \neg \mu_A(x)) \equiv 0$  usando la notación de la Tabla 16, sólo debe ser una tautología en el álgebra booleana por la extensión que se pretende con las teorías de conjuntos difusos. Por eso, su incumplimiento se puede considerar menos grave que el incumplimiento otras leyes de la lógica de Boole.

El valor mínimo y el máximo como los operadores para representar la conjunción y la disyunción, respectivamente, son duales con respecto a la negación fuerte. Esto quiere decir que, en su álgebra son válidas las leyes de De Morgan y que un operador de estos puede derivarse o deducirse del otro, junto con el operador de la negación. Además, cumplen una propiedad que no lo hacen otros operadores, pues se verifica la relación siguiente (Hajek, 2005):

$$\max(\mu_A(x), \mu_B(x)) = \min(1 - (1 - \mu_A(x)), 1 - \mu_B(x)) = \min(\mu_A(x), \mu_{\neg B}(x)).$$

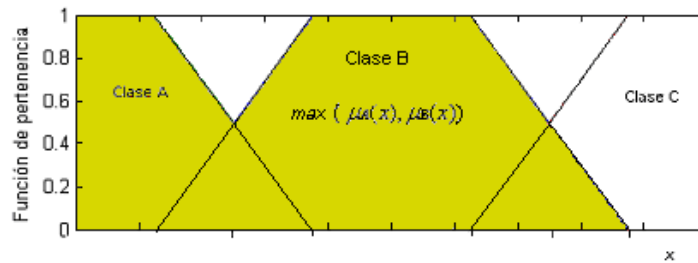
Por lo expuesto, es comprensible que la mayoría (sino todos) de los modelos difusos propuestos para flexibilizar el lenguaje de consulta a bases de datos, adopten la definición estándar de la conjunción y de la disyunción (el valor mínimo y el

máximo) para encontrar la semántica de expresiones vagas formadas con estas conectivas lógicas.

$$\mu_{A \cup B}(x) = \mu_A(x) \vee \mu_B(x) = \max(\mu_A(x), \mu_B(x)) \quad \forall x \in U$$

$$\mu_{A \cap B}(x) = \mu_A(x) \wedge \mu_B(x) = \min(\mu_A(x), \mu_B(x)) \quad \forall x \in U$$

No obstante, el uso de dichos operadores hace que se incumpla la ley del medio excluido, dado que  $\max(\mu_A(x), \neg\mu_A(x)) < 1$  cuando el objeto  $x$  se encuentre en la zona de solapamiento entre conjuntos. Esto quiere decir que un sistema de inferencia puede inferir que la posibilidad que tiene  $x$  de pertenecer al conjunto universal no sea 1, como se esperaría por la definición de este conjunto ( $\mu_U(x) \equiv 1 \quad \forall x \in U$ ). En el caso de un marco de cognición con tres conjuntos difusos, por ejemplo, y se preguntara por los que elementos que pertenecen al conjunto con etiqueta “baja” o al conjunto de los “medianos”, el sistema de inferencia le asignaría un grado de pertenencia menor que 1 a todos los elementos que estén en la intersección de dichos conjuntos, lo cual no tiene sentido lógico. La razón de esto, es que el uso del máximo como operador de la disyunción no da como resultado un conjunto convexo (Restricción Nro. 16, sección 4.5.9), como se muestra en la Figura 34. Por lo tanto, el nuevo conjunto carece de interpretabilidad.



**Figura 34. El valor máximo como operador de la disyunción**

Considerando otras de las restricciones impuestas al modelo de razonamiento aproximado en el procesamiento de consultas vagas, se incluyó la restricción de complementariedad en un marco de cognición  $\forall x \in U: \sum_{A \in \text{Marco}} \mu_A(x) = 1$  (Restricción

Nro. 13, sección 4.5.6) con el fin de preservar la cobertura y el concepto de conjunto universal. Consecuentemente, el operador de la disyunción que se observa más apropiado para aplicar sobre un marco de cognición específico, es la suma convencional. Este operador, bajo la restricción de complementariedad, es una s-conorma pues cumple con la ley conmutativa, asociativa, es una función monótona creciente y tiene como elemento neutro al cero y tiene como elemento absorbente al 1. Esto último significa que  $\mu_A(x) + 1 = 1$  es cierto porque el único caso en que se podría realizar una suma donde uno de los operandos tome el valor de 1 es cuando  $\mu_A(x) = 0$ . Por otro lado, la suma es una función continua ya que no produce grandes cambios en el conjunto difuso derivado de la disyunción de dos conjuntos, cuando los cambios son pequeños en alguno de los dos conjuntos que actúan como operandos, tal como lo exige la teoría difusa estándar. La suma, no cumple con la ley

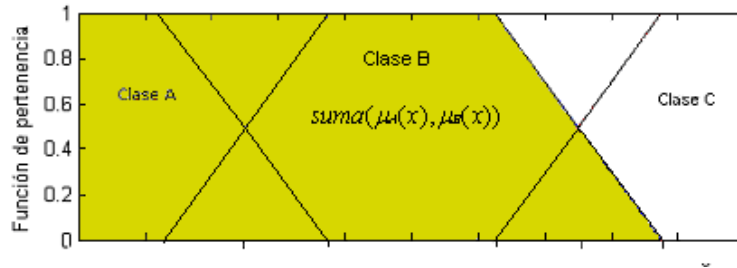
de idempotencia, pero sí con la superidempotencia, una característica que se había expresado anteriormente, puesto que:

$$\mu_A(x) + \mu_A(x) \geq \mu_A(x) \forall x \in U$$

Por lo anterior, el conjunto difuso derivado de la disyunción de conjuntos difusos, en un mismo marco de cognición, elegido en la presente investigación quedará determinado por:

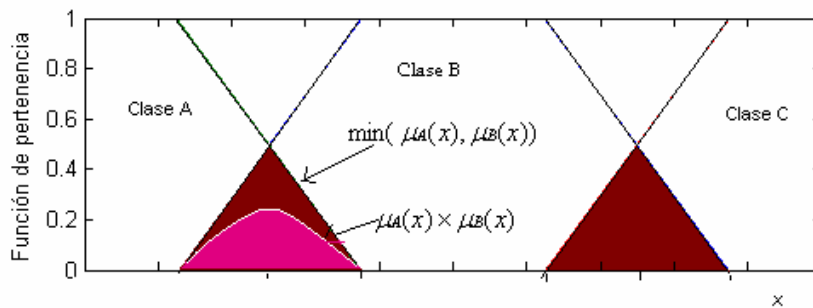
$$\mu_{A \cup B}(x) = \mu_A(x) \vee \mu_B(x) = \mu_A(x) + \mu_B(x) \forall x \in U \wedge A, B \in \text{Marco}$$

La representación gráfica del operador de la disyunción elegido se muestra en la Figura 35. Allí se puede observar que el conjunto derivado, al aplicarse esta operación es un conjunto convexo, condición fundamental para que sea un conjunto difuso interpretable.



**Figura 35. La suma como operador de la disyunción**

Cuando un usuario formule una consulta con una expresión de la forma “ $x$  es A o  $x$  es A” o que se llegara a ese tipo de expresión después de algunas operaciones algebraicas, que implicaría el cálculo  $\mu_A(x) + \mu_A(x)$ , se puede definir una regla general que fuerce al sistema de inferencia a dar como respuesta el valor esperado  $\mu_A(x)$ , antes de proceder con su evaluación. Con esto, no sólo se garantizaría que las respuestas del sistema de inferencia sean congruentes, sino que eviten cálculos innecesarios.



**Figura 36. Dos operadores para representar la conjunción**

Por otro lado, para la representación del operador de la conjunción, el mínimo de los operandos (los grados de pertenencia) se considera la norma triangular más apropiada, puesto que cubre toda la zona de solapamiento entre dos conjuntos. El

producto, por ejemplo, no la cubre en su totalidad. Esto se puede apreciar en la Figura 36.

El valor mínimo y la suma de los grados de pertenencia para representar las operaciones de la conjunción y de la disyunción, respectivamente, no cumplen la relación de dualidad con respecto al complemento. Pero este incumplimiento sólo implica que no se puede deducir uno de los operadores a partir del otro, por ser inválidas las leyes de De Morgan. Pero la inexistencia de una relación de dualidad tampoco se puede considerar una limitación tan grave como para abandonar la idea de usar el valor mínimo y la suma de los grados de pertenencia como las funciones más apropiadas para la interpretación de una expresión vaga formada con las conectivas lógicas de la conjunción o de la disyunción, en las condiciones de filtrado de una consulta.

Infortunadamente, en el caso donde las etiquetas lingüísticas correspondan a conjuntos borrosos sobre un espacio o *Universo de Discurso* multidimensional, la suma deja de ser un buen operador para definir la operación de disyunción porque no se cumpliría la ley de la clausura. Esto es, el sistema de inferencia podría concluir que el valor o grado pertenencia global, al conjunto difuso derivado por conjunción de varias características, sea mayor que 1. Este caso ocurre, por ejemplo, cuando se pregunte por los estudiantes que son “buenos” deportistas o de rendimiento académico “alto” y en la base de datos exista un estudiante que pertenece a esas categorías con un grado de 0.7 y 0.8, respectivamente. El sistema de inferencia sumaría estos grados de pertenencia y determinaría que el grado de pertenencia global de ese estudiante es de 1.5 a la clase derivada de la unión de estas dos clases.

Como una de las reglas fundamentales que debe cumplir cualquier operación aplicable a conjuntos, sean éstos concretos (los usuales) o difusos es la ley de la clausura, se tendrá que redefinir el operador de la conjunción para el caso multidimensional, donde el marco de cognición se defina en el producto cartesiano de diferentes dominios.

Para representar al operador de la disyunción, en el caso multidimensional, se podría reconsiderar al valor máximo del grado de pertenencia a uno de los conjuntos operandos, como la función con la cual reemplazar al operador “suma”. Sin embargo, el conjunto derivado puede no ser convexo y por lo tanto, no interpretable. Además, el valor máximo del grado de pertenencia a uno de los conjuntos difusos, es un operador que otorgaría un alto grado de compatibilidad o de encajamiento con la etiqueta del conjunto derivado de la expresión vaga, aunque el elemento sea poco compatible de manera marginal, con la mayoría de los términos vagos incluidos en una expresión (Pradera et al., 2006). Es decir, no existe un factor de *compensación* entre los valores altos y pequeños en los grados de pertenencia individuales, para determinar el grado de pertenencia global al agregado, tal como lo han señalado otros reconocidos autores (Herrera y Herrera-Viedma, 1997; Delgado et al., 1997; Yaguer, 1987).

Se había definido que la *agregación* (Definición 34, página 34) es una operación binaria aplicable a varios conjuntos difusos que siempre debe producir como

resultado, un nuevo conjunto de ese tipo. Por eso, se considera que la agregación es una clase de operación que generaliza a la unión o a la intersección de conjuntos. Y es así como otros autores hacen propuestas distintas a la estándar, en la Lógica Difusa, para encontrar los operadores que representan dichas operaciones. Por ejemplo, en (Herrera y Herrera-Viedma, 1997) se plantean dos operadores para la agregación: el operador LOWA (Linguistic Ordered Weighted Averaging) usado para combinar información lingüística y el operador LWA (Linguistic Weighted Averaging) utilizado para combinar información con distintos pesos, como una generalización del operador LOWA. El operador LOWA está basado en el operador OWA (Ordered Weighted Averaging) propuesto en (Yaguer, 1988) y una combinación de etiquetas genéricas, donde la forma de los conjuntos es predefinida buscando convexidad en los conjuntos derivados (Delgado, Verdegay y Vila, 1993).

Una operación OWA es una función  $+: [0, 1]^p \rightarrow [0, 1]$  de la forma  $\sum_{i=1}^p \beta_i \mu(x_i)$ ,

donde  $\beta_i \in [0, 1] \wedge \sum_{i=1}^p \beta_i = 1$ , que cumple las propiedades definidas para un operador de la agregación que se vuelven a enunciar.

Condiciones de Borde:

$$+(\beta_1 0, \beta_2 0, \dots, \beta_p 0) = \sum_{i=1}^p \beta_i \times 0 = 0$$

$$+(\beta_1 1, \beta_2 1, \dots, \beta_p 1) = \sum_{i=1}^p \beta_i \times 1 = \sum_{i=1}^p \beta_i = 1$$

La función que define el operador OWA es monótona creciente, dado que: si

$$\mu(x_i) \leq \mu(y_i) \forall i = 1, p \Rightarrow \sum_{i=1}^p \beta_i \mu(x_i) \leq \sum_{i=1}^p \beta_i \mu(y_i)$$

El orden de los operandos en la sumatoria es relevante, por motivos de eficiencia. Esto es, el factor de mayor peso va primero y los demás van en orden decreciente. Aunque esta reorganización que puede hacerse de manera transparente para el usuario final. La función OWA es continua, pues un ligero cambio en uno de los operandos produce un cambio aún menor el resultado de la agregación. Además,

el operador OWA cumple la idempotencia, Esto es,  $\sum_{i=1}^p \beta_i \mu(x) = \mu(x) \sum_{i=1}^p \beta_i = \mu(x)$

Compensación:

$$\min(\mu(x_1), \mu(x_2), \dots, \mu(x_p)) \leq \sum_{i=1}^p \beta_i \mu(x_i) \leq \text{media}(\mu(x_1), \mu(x_2), \dots, \mu(x_p))$$

La asociación:

$$+(\mu(x_1), +(\mu(x_2), \dots, \mu(x_p))) = +(+(\mu(x_1), \mu(x_2)), \dots, \mu(x_p))$$

Desagregación:

$$Si +(\mu(x_1), \mu(x_2), \dots, \mu(x_p)) = \mu'(x_s) \Rightarrow +(\mu'(x_s), \mu'(x_s), \dots, \mu'(x_s))$$

Si en el operador OWA, con la forma  $\sum_{i=1}^p \beta_i \mu(x_i)$ , se definen pesos iguales para las  $p$  etiquetas lingüísticas consideradas en la operación de la disyunción, esto es  $\beta_i = \frac{1}{p} \forall i = 1, p$ , entonces  $\sum_{i=1}^p \beta_i \mu(x_i) = \sum_{i=1}^p \frac{\mu(x_i)}{p}$ . Esto significa que el estimador OWA de la agregación, donde todas las restricciones sobre las variables lingüísticas tengan la misma importancia en el significado de la clase etiquetada resultante, es la media aritmética simple (no ponderada) de los grados de pertenencia individuales. La media aritmética, se caracteriza por su simplicidad en los cálculos y por ser altamente comprensible para cualquier usuario consultor de una base de datos. Estas cualidades permiten que con este operador se puedan derivar etiquetas lingüísticas, con significado, a partir de la unión de varias etiquetas simples. Consecuentemente, el grado de pertenencia global de una tupla  $t_i$  de la relación  $R$ , a la clase derivada de la unión de  $p$  etiquetas vagas simples, será:

$$\mu_{E1 \cup E2 \dots \cup Ep}(t_i) = \mu_{E1}(t_i) \vee \mu_{E2}(t_i) \dots \vee \mu_{Ep}(t_i) = \sum_{i=1}^p \frac{\mu_{Ei}(t_i)}{p} \forall t_i \in R$$

### 6.2.3.3 Reglas Sintácticas para una combinación lineal de restricciones simples

En las propuestas estudiadas para la flexibilización del lenguaje de interacción con bases de datos relacionales u objeto-relacionales, para admitir vaguedad en las consultas, no se incluye la posibilidad de representar restricciones derivadas de la combinación lineal de otras (véase, por ejemplo, a Kackrpzyk y Zadrozny, 2001; Galindo, Urrutia y Piattini, 2005; Bosc, Kraft y Petra, 2005; Ma y Wang, 2006; Goncalves y Tineo, 2007)). Aunque vale señalar que el lenguaje SQLf3 incorpora una estructura análoga para especificar atributos para tipos de datos definidos por el usuario, mediante el operador de la conjunción, pero no para restringir los objetos sobre los cuales se desean visualizar algunas de sus propiedades (Gonçalves y Tineo, 2001).

En un sistema interactivo de consulta-respuesta a bases de datos, es muy factible que se quiera saber cuál es el grado de compatibilidad que tienen los objetos de interés, con alguna etiqueta lingüística definida mediante una combinación lineal de restricciones, donde a cada una de ellas se le asigna un grado de relevancia o importancia  $\beta_i$ , en la interpretación del término vago derivado. A cada etiqueta lingüística involucrada se otorga un peso o ponderación distinta pero considerando que dichos pesos sumen la unidad (que equivale al cien por ciento). Esto es:

$$\sum_{i=1}^p \beta_i = 1, \text{ con } \beta_i \in [0,1] \forall i = 1, p$$

Bajo el supuesto del cumplimiento de esa restricción para los pesos asignables a cada condición simple de la combinación lineal, la forma genérica del modelo de un término vago agregado representado con la etiqueta  $E_j$  se puede especificar, en notación algebraica, así:

$$E_j = \beta_1 \text{Condición}_1 + \beta_2 \text{Condición}_2 + \dots + \beta_p \text{Condición}_p$$

Cada condición en la fórmula hace referencia a una propiedad básica o derivada de alguna de las relaciones especificadas en la cláusula FROM, y cada coeficiente  $\beta_i$  es la contribución o el peso que tiene esta condición en la explicación de la etiqueta  $E_j$ . Puesto que la combinación lineal puede dar origen a conjuntos nítidos o convencionales, entonces la fórmula anterior puede generalizarse así:

$$\text{Expresión} = \beta_1 \text{Condición}_1 + \beta_2 \text{Condición}_2 + \dots + \beta_p \text{Condición}_p$$

Esta generalización es útil para validar, por ejemplo, si el estado civil de una persona puede deducirse del grado de cumplimiento de ciertas características, definidas por las restricciones impuestas, a las cuales se les puede asignar un peso o ponderación diferente. Esto significa que la especificación de una expresión como una combinación lineal puede ser usada para validar modelos, si la propiedad especificada en la cláusula SELECT es un grado o valor de verdad, que puede ser calculado mediante una función predefinida en la base de datos, similar a la función CDEG del lenguaje FSQL (Galindo, 2008).

De acuerdo con lo anterior, la sintaxis de la consulta SQL extendida para que incluya términos agregados originados por una combinación lineal de condiciones simples, vagas o concretas, en los criterios de la consulta, se define como:

```
SELECT proyección
FROM relaciones
WHERE [expresión IS]  $\beta_1 * \text{condición}_1 + \beta_2 * \text{condición}_2 [+ \dots]$ 
      [WITH CALIBRATION n|threshold|n, threshold]
```

En esta extensión propuesta del lenguaje SQL, para admitir combinaciones lineales de restricciones vagas, se ha pensado que no es necesario especificar el nombre o etiqueta del conjunto derivado y por eso, el término “expresión IS” se ha fijado como opcional.

Para mostrar un ejemplo de una consulta que siga el patrón sintáctico descrito, suponga que se desea conocer las marcas de los autos “atractivos”, término derivado de la combinación lineal de un rendimiento “alto”, una potencia “alta” y un peso “bajo” considerando unos pesos de 0.4, 0.4 y 0.2, respectivamente. Acorde con la sintaxis, la consulta se expresaría de la siguiente manera, en el SQL extendido propuesto:

```
SELECT marca
FROM autos
WHERE "atractivo" IS 0.4*mpg IS "alto" + 0.4*hp IS "alto" + 0.2*peso IS "bajo"
```

El sistema de inferencia, para resolver una pregunta como la recién planteada, deberá saber cuántas categorías se definen en el marco de cognición y el orden de la etiqueta especificada de cada condición, para poder inferir lo que se considerará “alto” o “bajo” en ellas. Por esto, si no se especifican estas variables, como en el ejemplo presentado, y no se encuentra conjunto de términos asociado a cada variable lingüística involucrada, asumirá tres clases por defecto.

#### 6.2.3.4 Reglas Semánticas para una combinación lineal de restricciones simples

Para representar la importancia o relevancia de los predicados en otros sistemas difusos, distintos a los que tratan de flexibilizar un sistema de consultas a bases de datos, se han propuesto diversos operadores como el Max-min (Bellman y Zadeh, 1970; Yager, 1978), el Fuzzy AHP (Laarhoven y Pedrycz, 1983) y la media ponderada OWA (Baas y Kwakernaak, 1977; Dong et al., 1985; Dubois y Prade, 1980; Tseng y Klein, 1992). Dentro de ellos, el operador OWA se destaca por ser intuitivo y por haber sido ya elegido para representar a la disyunción, en el caso multidimensional, en la presente Tesis Doctoral.

Se había mencionado que el operador OWA y sus extensiones, son utilizados para combinar información con distinta relevancia o peso. De acuerdo con estudios previos, se consideran dos tipos de agregaciones distintas (Herrera y Herrera-Viedma, 1997).

- La agregación de los grados de importancia de la información, usada para reunir la opinión de los juicios expertos sobre los grados de pertenencia a los conjuntos difusos que representan etiqueta lingüísticas, con el uso del concepto del cuantificador *mayoría difusa*.
- La agregación de información ponderada que combina información y sus pesos. Este es el caso que se está aquí tratado, para las combinaciones lineales de restricciones simples.

Ya se había mostrado que un operador OWA cumple con las condiciones definidas para un operador de la agregación como las condiciones de borde, define una función monótona creciente, cumple con la compensación y las demás condiciones. En consecuencia, para determinar el grado de encajamiento global de una tupla  $t_i$  de una relación  $R$  con la expresión derivada de una combinación lineal de restricciones simples, impuestas sobre variables de cualquier tipo, se considerará como su operador a la media ponderada de los grados de pertenencia a las clases etiquetadas individuales. Esto es, el modelo que determina los grados de pertenencia a la combinación lineal es:

$$\mu_{expresión}(t_i) = \beta_1 \mu_{condición1}(t_i) + \beta_2 \mu_{condición2}(t_i) + \dots + \beta_p \mu_{condiciónp}(t_i)$$

Acorde con lo anterior, el grado de pertenencia global de una tupla  $t_i$  de la relación  $R$ , a la clase derivada de la unión de  $p$  condiciones simples, se calculará a partir de los grados de pertenencia marginales, mediante la fórmula siguiente.

$$\mu_{\text{expresión}}(t_i) = \sum_{i=1}^p \frac{\beta_i \mu_{\text{condición}_i}(t_i)}{\sum_{i=1}^p \beta_i} \quad \forall t_i \in R, \beta_i \in [0,1] \wedge \sum_{i=1}^p \beta_i = 1$$

En la especificación de una consulta con una restricción definida como una combinación lineal de otras, los valores de los pesos o las ponderaciones  $\beta_i$  son proporcionados por el usuario, dado que en muchos casos dependen de las reglas del dominio de la aplicación o de acuerdo con sus preferencias (su juicio personal). Los pesos que se le asignen a una entrevista, al examen de admisión y a la hoja de vida de un aspirante a un programa curricular determinado, para catalogarlo como “admisible”, dependerá la universidad en cuestión y por eso se tienen que especificar en la consulta. Si estas reglas del dominio para etiquetar una clase (concreta o difusa), no cambian con frecuencia, se puede pensar en su persistencia dentro de los metadatos de la base de datos. Por esto se requiere una sentencia DDL para la construcción o modificación de la expresión que se deriva de una combinación lineal, como:

```
CREATE [OR REPLACE] LABEL etiqueta AS
  LINEAR_COMBINACION(expresión, (beta_i, condición_i))
```

Conservar en los metadatos este tipo de etiquetas lingüísticas, haría muy cómoda la formulación de las preguntas, pues evitar que el usuario reescriba la combinación lineal, cada vez que la quiera incluir en una consulta. Por la persistencia de este tipo de objetos, sólo se tendría que especificar el nombre de la etiqueta derivada. Por esto, es necesario redefinir la sintaxis para que la especificación de combinación lineal pueda ser opcional en la cláusula WHERE de una consulta:

```
SELECT proyección
FROM relaciones
WHERE [expresión IS] { $\beta_1$ *condición1+ $\beta_2$ *condición2 [+...] |label}
  [{WITH CALIBRATION n|threshold|n, threshold}]
```

De acuerdo con esta nueva sintaxis, el usuario puede preguntar por las marcas de los autos “atractivos”, definidos previamente, así:

```
SELECT marca
FROM autos
WHERE "atractivo"
```

En otras situaciones el valor de una expresión definida como una combinación lineal, el juicio subjetivo es el que prima, como en la búsqueda de documentos. El usuario puede considerar sus propios pesos para cada una de las condiciones que conforman la combinación lineal, como la fecha de elaboración, el tamaño y el número de citas bibliográficas de los mismos y por eso pueden ser especificados, a su gusto.

### 6.3 Preguntas de Tipo III (con cuantificadores lingüísticos)

En un sistema interactivo de consulta-respuesta es posible que un usuario necesite usar cuantificadores lingüísticos en sus consultas, además de los términos

vagos ya considerados. Un cuantificador lingüístico es un término que representa una cantidad vaga como “la mayoría de...” o “unos pocos...”. Estos cuantificadores extienden los conceptos de cuantificador universal ( $\forall$ ) y existencial ( $\exists$ ) del cálculo de predicados de primer orden.

La inclusión de cuantificadores vagos en el lenguaje de consulta hace más poderoso un sistema de consulta-respuesta. Con ellos, los usuarios pueden formular preguntas como ¿Cuáles estudiantes contestaron correctamente “casi todas” las preguntas del cuestionario Y? o ¿Cuáles vendedores de la zona 1 tiene un total de ventas superior o igual al de la “mayoría” de los vendedores de la zona 2?.

Un cuantificador Q como “la mayoría” es relativo puesto que depende del total de elementos que conforman una relación básica o derivada e implican un mapeo o transformación del tipo  $Q_{relativo}: [0,1] \rightarrow [0,1]$ .

### 6.3.1 Reglas Sintácticas para los Cuantificadores Relativos

Para determinar cómo se tendría que extender el lenguaje SQL para la admisión de una pregunta con un cuantificador vago relativo, es necesario considerar el patrón de una consulta en SQL estándar que incluya un cuantificador existencial (EXISTS), el universal (ALL) y alguno (ANY o SOME) que hasta ahora son los cuantificadores admisibles (Blum, 2007). Según el estándar, se especifican en la cláusula WHERE, antecediendo a una subconsulta. Esto es:

```
SELECT proyección
FROM relaciones
WHERE {atributo|expresión} operador_lógico
      {EXISTS|SOME|ANY|ALL} (consulta)
```

Cuando en una consulta se especifica el cuantificador ANY o SOME, que son sinónimos, se obtienen las tuplas o registros que satisfacen la comparación de los valores de un atributo o expresión con alguno de los obtenidos en la subconsulta. Debe resaltarse que este cuantificador y el cuantificador ALL deben ir precedidos por un comparador lógico como “=” o “<”. En el ejemplo siguiente, se obtienen los nombres de los productos cuyo precio unitario es mayor que el de algún producto vendido con un descuento mayor del 25%:

```
SELECT nombre FROM Productos
WHERE Precio > ANY
      (SELECT Precio FROM Detalle_pedido
       WHERE Descuento > .25)
```

Por su lado, el cuantificador ALL se usa para obtener sólo los registros de la consulta principal que satisfagan la comparación con todos los valores obtenidos en la subconsulta. Si se cambia ANY por ALL en el ejemplo anterior, la consulta sólo devolverá los productos cuyo precio unitario sea mayor que el de todos los productos vendidos con un 25% o más de descuento. Por lo tanto, esta última condición es mucho más restrictiva.

En el lenguaje SQL extendido, considerando los cuantificadores vagos relativos como “la mayoría” o “casi todos” (most, en inglés) y “la minoría” o “unos pocos” (few, en inglés) es:

```
SELECT proyección
FROM relaciones
WHERE {atributo|expresión} operador_lógico
      {valor|expresión}| {EXISTS|SOME|ANY|ALL|MOST|FEW} (consulta)
      [{WITH CALIBRATION n|threshold|n, threshold}]
```

Como se observa en esta sentencia, la inclusión de los cuantificadores relativos en el lenguaje estándar SQL no altera la forma gramatical de las consultas, sino que se amplía el conjunto de palabras o “tokens”.

### 6.3.2 Reglas Semánticas para los Cuantificadores Relativos

Un cuantificador relativo permite especificar una restricción sobre la cantidad de elementos de una relación que satisfacen ciertas condiciones, bien sea plenamente o de manera parcial. Por lo tanto, una regla heurística que nos permita hallar la semántica de estos cuantificadores debe basarse, intuitivamente, en la proporción de casos en los que existe algún tipo de cumplimiento de las condiciones especificadas en la subconsulta, con respecto al total de elementos que conforman la relación considerada. Por lo tanto, el problema estriba en determinar entre cuáles valores de la proporción se puede considerar que los elementos cumplen con el cuantificador especificado en una consulta y en qué grado lo hacen.

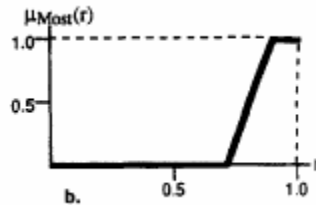
A diferencia de los términos vagos tratados hasta ahora para extender el lenguaje interactivo de consulta a sistemas de bases de datos, un cuantificador relativo como “unos pocos” o “casi todos” no depende del contexto, ni de las variables o atributos considerados, sino que depende de la percepción o acuerdo de lo que se pueda considerar “cercano” a la totalidad de los elementos que conforman un relación (al conjunto universal), en el caso de un cuantificador como “la mayoría” o “casi todos”, o al conjunto vacío para el cuantificador “casi ninguno..” o “la minoría”. Por esto, para la interpretación de las consultas con términos vagos lingüísticos de este estilo se buscará un modelo completamente instanciado de cada función de pertenencia. Esto es, los parámetros que determinan la forma y la posición de la función de pertenencia de un cuantificador relativo no serán variables, para determinar su valor en tiempo de ejecución a través de un proceso de minería de los datos, sino que tendrán valores particulares.

En (Rasmussen y Yager, 1999) se propone una extensión del lenguaje SQL para admitir vaguedad en las consultas, denominada SummarySQL. En esta propuesta, las cantidades relativas son caracterizadas indicando la proporción de los datos que satisfacen un resumen lingüístico. Allí se considera que un resumen lingüístico es una declaración de la forma “ $Q$  objetos en  $R$  son  $S$ ”, donde  $S$  se llama el descriptor (summarizer, en inglés),  $Q$  es la cantidad en acuerdo, y  $R$  es una relación o colección de datos. A cada resumen lingüístico se le asocia un valor de verdad en  $[0,1]$  llamado la medida de validez del resumen. La medida de validez denotada por  $\mu_Q(r)$ , o de

manera alternativa  $Q(r)$ , proporciona una indicación de cuán compatible es la proporción  $p$  de elementos en la relación  $R$  con la cantidad de referencia  $Q$ . De acuerdo con esto, proponen que la función de pertenencia para encontrar la semántica del cuantificador “la mayoría” sea definida mediante la fórmula:

$$\mu_{Most}(p) = \begin{cases} 0 & \text{si } p \leq 0.75 \\ (p - 0.75)/(0.9 - 0.75) & \text{si } 0.75 < p < 0.9 \\ 1 & \text{en otro caso} \end{cases}$$

Basados en esa definición, el conjunto difuso que representa al concepto de “mayoría” en SummarySQL tiene una representación gráfica como la que se ilustra en la Figura 37.



**Figura 37. El cuantificador “Most” en SummarySQL**

En el lenguaje FQUERY propuesto en (Kacprzyck y Zadrozny, 2001) se considera la misma forma para el conjunto difuso que representa el cuantificador “la mayoría”. Sin embargo, le permite al usuario definir los valores que determinan el soporte y el núcleo de la función de pertenencia, por medio de la interfaz gráfica. De forma parecida a las dos propuestas previas, la semántica del cuantificador “la mayoría” en (Tineo, 2000) la determina el usuario. Por otro lado, en el lenguaje SQLf, como puede verse en (Bosc y Liétard, 2004), se define con una función hombro\_derecho(0.5 0.75, 1).

La similitud de las propuestas recién presentadas, para la admisión del cuantificador difuso la “mayoría” o “casi todos” en las extensiones del lenguaje SQL, estriba en que la forma de la función de pertenencia es una función hombro\_derecho con parámetros  $(a, b, 1)$ . Una función monótona creciente y en la cual se cumple que el grado de verdad o medida de validez de la “mayoría” es 1, cuando una restricción es satisfecha por el 100% de los ejemplares de una relación dada. Esto quiere decir que, independiente de la forma de la distribución de los grados de pertenencia, se debe cumplir que  $\mu_{Most}(1) = 1$ . La diferencia entre las propuestas estudiadas está en los valores de los parámetros  $a$  y  $b$ . Esto porque no hay un criterio o soporte matemático para ayudar en la selección de ellos.

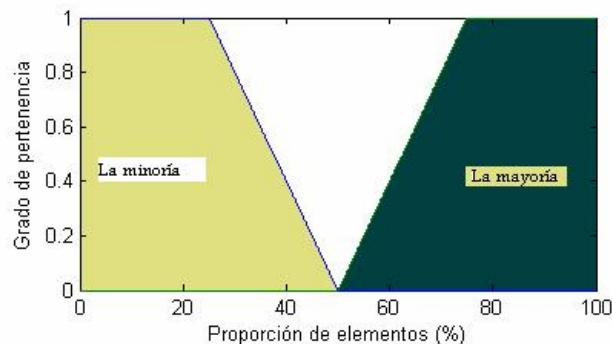
Por lo anterior, en esta propuesta, para que la máquina de inferencia pueda interpretar un cuantificador relativo vago, se determinará que cierta proporción  $p$  se aproxima al concepto de la “mayoría” mediante la función de pertenencia siguiente.

$$\mu_{\text{Most}}(p) = \begin{cases} 0 & \text{si } p \leq 0.5 \\ (p - 0.5)/(0.75 - 0.5) & \text{si } 0.5 < p < 0.75 \\ 1 & \text{si } 0.75 \leq p \leq 1 \end{cases}$$

Por otro lado, se considera que el cuantificador vago “la minoría” o “muy pocos” (*few*, en inglés) es el antónimo de la “mayoría” o “casi todos”. Por lo tanto, el término queda definido por la función hombro\_izquierdo  $(0, 0.25, 0.5)$ :

$$\mu_{\text{Few}}(p) = \begin{cases} 1 & \text{si } 0 \leq p \leq 0.25 \\ (0.5 - p)/(0.5 - 0.25) & \text{si } 0.25 < p < 0.5 \\ 0 & \text{en otro caso} \end{cases}$$

De acuerdo con lo anterior, el modelo de los conjuntos difusos que representan los cuantificadores vagos “la minoría” y “la mayoría” tienen una representación gráfica como la que se ilustra en la Figura 38.



**Figura 38. Cuantificadores relativos**

### 6.3.3 Reglas Sintácticas para Comparadores Vagos

La vaguedad en las condiciones o restricciones especificadas en una consulta también pueden referirse a cantidades concretas (valores escalares numéricos) como criterio de comparación, pero que son antecedidas por operadores lógicos de comparación vagos como “aproximadamente...” o “parecido a..”. Un ejemplo de este tipo de consulta es “¿cuáles marcas de autos tienen un rendimiento parecido al rendimiento promedio de todos los autos reportados?”. En esta pregunta se observa que la ubicación de un operador vago de comparación es la misma a la de un operador lógico concreto (exacto) de comparación como “igual que...” (=). De manera pues que la estructura gramatical de una consulta SQL no se verá afectada por la incorporación de este nuevo tipo de operadores vagos. Sólo será necesario

agregar una nueva palabra (token) al lenguaje por cada cuantificador vago que se quiera admitir y representar como “parecido a...” o “distinto a...”.

Puesto que ya existe en el lenguaje estándar SQL el operador LIKE, que ha sido exclusivo para comparar cadenas de caracteres con algún patrón definido por el usuario, se puede sobrecargar para que permita comparar valores numéricos escalares, en términos aproximados. Por lo tanto, la sintaxis de la sentencia SELECT, para especificar el operador de comparación es la que aparece, enseguida.

```
SELECT proyección
FROM relaciones
WHERE atributo|expresión [NOT] LIKE escalar
```

### 6.3.4 Reglas Semánticas para Comparadores Vagos

Las cantidades absolutas, precedidas de un comparador vago, han sido bien representadas por las distribuciones de números difusos, cuya forma es unimodal, ya que el prototipo es un valor único: la cantidad  $n$  que sirve de base de la comparación con los valores específicos que poseen los objetos de interés, en un atributo o propiedad. La forma usual elegida para la distribución de pertenencia de un número difuso es la distribución triangular  $(a, b, c)$ , donde  $a$  y  $c$  definen el soporte de la función de pertenencia y  $b$  es el prototipo que es un valor escalar. La forma de la distribución también pudo ser una distribución de tipo campana (bell), pero como antes se expuso, para los propósitos de este trabajo de investigación, los modelos lineales son suficientes para la aproximación al significado de la vaguedad.

En la

Figura 39 se puede apreciar su fórmula algebraica y la representación gráfica de un ejemplar de la distribución triangular.

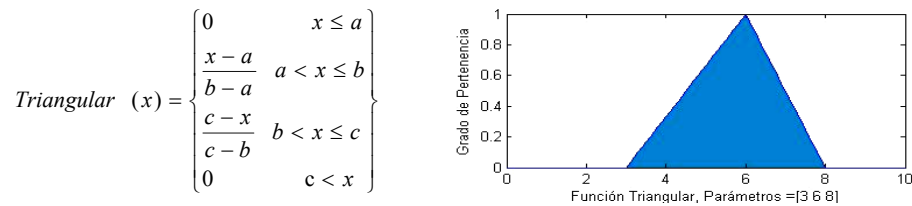


Figura 39. Distribución de un conjunto difuso triangular

En el caso de una función triangular, el problema consiste en determinar cuáles estadísticas de resumen se pueden emplear para el cálculo de los parámetros  $a$  y  $c$  que definen el soporte de la función de pertenencia de un conjunto difuso en particular. Cuando se trata del comparador “parecido a...” podemos imaginar un conjunto difuso unimodal con un soporte que cubra sólo una porción muy limitada del dominio de la variable o, de forma alternativa, que el conjunto tenga una cardinalidad muy pequeña, con los elementos o las tuplas más cercanas al prototipo del conjunto universal.

En este punto es importante recordar que cuando se consideró el caso de la interpretación de preguntas con modificadores lingüísticos, se observó que no sólo se podía acentuar una clase de los extremos, como cuando se califica algo como “muy pequeño” o “extremadamente alto”, sino también la clase del medio o de la mitad, si se consideran tres etiquetas lingüísticas en un marco de cognición. Con esta clase de la mitad, lo que causó algo de dificultad fue encontrar el rótulo o etiqueta para la clase acentuada generada, pero sabiendo que esa clase estaría conformada por los objetos que se aproximan al valor central (el valor medio o prototipo) de la distribución de los objetos, con respecto a una variable o propiedad y en un contexto dado, se determinó que “muy típico” o “extremadamente común” podrían ser los nombres apropiados. Como esa proximidad no tiene que ser únicamente con los valores escalares que representan al valor central o a los valores extremos, se puede generalizar para cualquier valor escalar posible dentro del dominio considerado.

De acuerdo con lo anterior, el concepto “parecido a  $n$ ” se asemeja al concepto de un modificador de acentuación. En consecuencia, el soporte que describe el número difuso  $n$  estará conformado por el mismo porcentaje de valores que representa un conjunto borroso con etiqueta  $E$ , acentuado con el adverbio “extremadamente”, dado que demanda mayor proximidad de los elementos con el valor escalar  $n$ . Y de acuerdo con la definición dada a este adverbio, que se basa en percentiles o estadísticas de posición, el conjunto difuso que representa la expresión “parecido al percentil  $n$ -ésimo” tiene una distribución triangular de la forma:

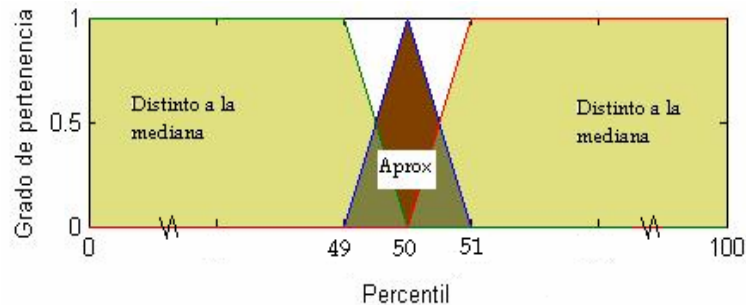
$$\text{“parecido al percentil } n\text{-ésimo”} = \begin{cases} (x - P_{n-1}) / (P_n - P_{n-1}) & \text{si } P_{n-1} < x \leq P_n \\ (P_{n+1} - x) / (P_{n+1} - P_n) & \text{si } P_n < x \leq P_{n+1} \\ 0 & \text{en otro caso} \end{cases}$$

Por lo tanto, para pasar a una expresión en términos absolutos como “parecido a  $n$ ”, siendo  $n$  un número escalar y no un porcentaje, primero se encontraría el percentil que le corresponde y luego se aplicaría la fórmula recién dada. Por ejemplo, para responder la pregunta: ¿Cuáles marcas de autos tienen un rendimiento aproximadamente de 23 millas por galón? primero se encontraría el percentil que le corresponde a esta cifra. Y según los autos registrados en la base de datos de referencia se obtiene que el valor de 23 millas que corresponde al percentil  $P_{50}$  (la mediana). De acuerdo con esto, se obtiene que  $P_{n-1} = P_{49} = 22.4$  y  $P_{n+1} = P_{51} = 24$  millas por galón, por lo que el conjunto difuso que representa “parecido a 23 millas por galón” está determinado por una función de pertenencia triangular con parámetros (22.4, 23, 24).

Por otro lado, el término “parecido a...” es el opuesto a “distinto a...” y por lo tanto, la función de pertenencia de este último término puede definirse como el complemento de la función del término “parecido a...” como se presenta a continuación.

$$\text{"distinto al percentil } n\text{-ésimo"} = \begin{cases} 1 & \text{si } P_0 < x \leq P_{n-1} \text{ ó si } P_{n+1} < x \leq P_{100} \\ (P_n - x)/(P_n - P_{n-1}) & \text{si } P_{n-1} < x \leq P_n \\ (x - P_n)/(P_{n+1} - P_n) & \text{si } P_n < x \leq P_{n+1} \\ 0 & \text{en otro caso} \end{cases}$$

Para mayor claridad, en la Figura 40 se muestra gráficamente como se representa el conjunto difuso "aproximadamente la mediana" o "más o menos el valor medio", en una distribución uniforme de los datos, junto con su complemento.



**Figura 40. Representación del término "parecido a..."**

Puesto que la elección de los parámetros para los conjuntos difusos que representan los cuantificadores relativos es en cierta forma arbitraria, como se ha visto en las propuestas presentadas, en un dominio de aplicación específico se podrían considerar otras definiciones de éstos u otros cuantificadores. Por esto, en nuestro modelo se debe considerar la posibilidad de cambiar la definición de este tipo de cuantificador, por medio de una interfaz gráfica o de una sentencia como la siguiente:

```
CREATE [OR REPLACE] QUANTIFIER nombre AS
conjunto_difuso(nombre_función, percentiles)
```

De forma parecida, se debería poder especificar un nuevo comparador lógico vago como "mayor que", o cambiar uno existente mediante una sentencia como:

```
CREATE [OR REPLACE] OPERATOR nombre AS
conjunto_difuso(nombre_función, percentiles)
```

No obstante, se piensa que dichas órdenes no deberían ser emitidas por los usuarios consultores de una base de datos. Pues ellos no tienen por qué conocer las nociones de una teoría difusa, por simples que éstas sean. Por lo tanto, se esperaría que la modificación del conjunto difuso que representa a un cuantificador sea efectuada por una persona calificada.

#### 6.4 Preguntas de Tipo IV (con calificadores lingüísticos)

Un calificador lingüístico es un adverbio que determina o trata de expresar el grado de cumplimiento de una cualidad o característica por un objeto o grupo de objetos. Este tipo de calificador trata de medir, en términos cualitativos, el grado de verdad, de probabilidad o de posibilidad de una proposición (Ribeiro y Moreira, 2003). Entre muchos ejemplos de preguntas con calificadores lingüísticos se tienen algunas como ¿Es cierto que los autos de marca X son “económicos”? y ¿Es “muy probable” que llueva mañana por la tarde?. De estos ejemplos se aprecia que la forma genérica de las preguntas con calificadores lingüísticos es “¿Es  $\tau$  que R son E?”, donde  $\tau$  un valor de verdad, una medida de probabilidad o posibilidad definida cualitativamente, R es una relación de la base de datos y E una etiqueta lingüística.

Para el lenguaje extendido de consulta a bases de datos que aquí se propone, se considerará únicamente el calificador lingüístico del valor de verdad de una proposición porque para representar las probabilidades se requiere un número suficiente de registros históricos de ocurrencia de los eventos considerados y un proceso de ajuste a modelos probabilistas que están por fuera del alcance de esta Tesis Doctoral.

El propósito principal de incorporar un calificador lingüístico es permitir que el sistema de consulta-respuesta pueda responder, vagamente, con un valor de verdad cualitativo como “muy cierto”, “poco cierto” o “absolutamente falso” cuando se le pregunte si ciertos objetos o grupos de objetos, en la base de datos, cumplen con los criterios especificados en la consulta.

Eso significa que las respuestas del sistema se parecerán más a las que contestaría un ser humano que a las que usualmente proporciona una máquina. Y eso precisamente es lo que se quiere, que el usuario piense que está interactuando con un verdadero experto, que no sólo le entiende la vaguedad expresada en las condiciones de las consultas, sino que también le puede responder en términos cualitativos o con etiquetas lingüísticas.

En los lenguajes teóricos, el Cálculo y el Álgebra Relacional que dan soporte al modelo relacional y al objeto-relacional, sólo se concibe que el resultado de una operación realizada sobre relaciones (básicas o derivadas) de una base de datos sea otra relación (Date, 2001). Esto es, se cumple la ley de la clausura y en consecuencia, en el lenguaje estándar SQL sucede lo mismo. Por eso, cuando no existen tuplas que cumplan las condiciones establecidas en una consulta, el resultado ofrecido por el sistema es una tabla o relación vacía, pero nunca se espera un resultado como “cierto”, “falso” o “nulo” que son los posibles valores de verdad considerados por los sistemas gestores de bases de datos actuales. Debido a esto, posiblemente, ninguna de las propuestas estudiadas propone cómo extender el SQL con el propósito que el sistema pueda responder con un valor o etiqueta lingüística asociada a un grado de verdad.

### 6.4.1 Reglas Sintácticas para Valores Lingüísticos de la Verdad

Como recién se expresó, en el lenguaje SQL no hay una sintaxis definida para una consulta que dé como resultado un valor de verdad, ni tampoco en las extensiones del lenguaje estudiadas. Por esto, se debe proponer una sintaxis particular del SQL para sentencias de la forma: "¿Es cierto que una relación  $R$  es  $E_j$ ?", donde  $E_j$  es una etiqueta lingüística para una variable o atributo (simple o compuesto).

Para el planteamiento de una pregunta cuya respuesta sea un valor lingüístico de la verdad se consideró que se puede definir una función similar a la función definida por Galindo y que aquí se contempla con el nombre de GC y que calcula el grado de pertenencia de un objeto a un conjunto difuso. Sin embargo, se observa que el uso de calificadores de la verdad, no tiene por qué restringirse a preguntas donde se quiera evaluar el grado de verdad del cumplimiento de una sola condición simple como *mpg IS 'alto'*, sino que puede ser admisible cualquier condición vaga o concreta, incluso un conjunto de condiciones agregadas por medio de las conectivas lógicas o mediante una combinación lineal. Por lo tanto, la función puede incluir como argumento cualquiera de las consultas consideradas, que se puede llamar  $VV(\text{consulta})$  que devuelva el valor lingüístico de verdad, a partir del texto de la consulta que se especifica como parámetro. Cuando se especifique esta función en la cláusula SELECT sobraría la cláusula FROM, pero en PostgreSQL esta última cláusula es opcional y en otras implementaciones del SQL como en Oracle, se podrá recurrir al artificio de una tabla ficticia como "Dual".

Cuando se quiera conocer el valor de verdad de una regla de asociación entre propiedades de los objetos, esta función tendría que ser redefinida como  $VV(\text{cuantificador}, \text{condiciones\_antecedente}, \text{condiciones\_consecuente}, \text{origen})$ . Es decir, una pregunta con esta forma o patrón equivale a la especificación de una proposición de la forma:  $\text{condiciones\_antecedente} \rightarrow \text{condiciones\_consecuente}$ . La respuesta esperada de esta función será una un registro con el valor de verdad cualitativo y los indicadores de la fortaleza de la asociación definidos en la literatura para las reglas de asociación como son la confiabilidad y el soporte que más adelante se detallan.

Para ilustrar el uso de la función  $VV$ , para saber el valor lingüístico de verdad de la pregunta si todos los estudiantes que perdieron el curso con código  $X$ , trabajan, la pregunta se formularía así:

```
SELECT VV(ALL, 'TRABAJA="SI"', 'estudiante.id=registro.idest
AND nota_def < 3
AND idcurso = "X"', 'registro, estudiante')
```

Como puede observarse, la admisión de cuantificadores en una pregunta y la admisión de respuestas con valores cualitativos del grado de la verdad, permite inferir nuevo conocimiento, a partir de los datos disponibles, de una manera sencilla.

Este conocimiento es útil para caracterizar las asociaciones existentes entre los objetos de la base de datos y es otro aporte de este trabajo de investigación.

#### 6.4.2 Reglas Semánticas para valores de verdad lingüísticos

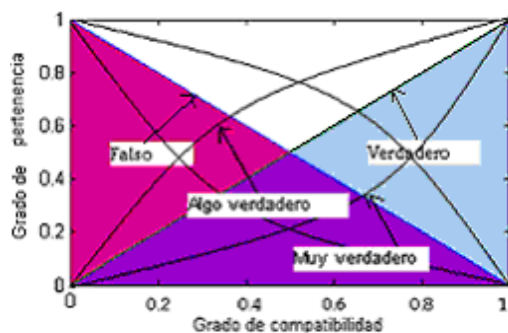
Para evaluar el grado de verdad de una sentencia de la forma “*R son E*” donde *R* es una relación y por tanto, un agregado de tuplas, es natural pensar en una función de agregación, un operador LOWA, que permita inferir ese grado a partir de los grados de pertenencia de cada tupla  $t_i$  de *R*, con  $i = 1$  hasta  $n$ , al conjunto difuso con etiqueta  $E_j$ . Puesto que el peso o contribución de cada tupla, con su grado de pertenencia individual, debe ser el mismo en el grado de compatibilidad del agregado, se usará el operador LOWA definido como la media aritmética simple. Esto significa que la regla de inferencia para determinar el grado de compatibilidad de la relación o tupla con una etiqueta lingüística *E*, considerando un alfa corte  $\alpha$  en  $[0, 1)$ , es:

$$“R \text{ son } E_j” \rightarrow \mu_{E_j}(R) = \sum_{i=1}^n \frac{\mu_{E_j}(t_i)}{n} > \alpha$$

La aplicación del operador LOWA así definido, daría como resultado, la *medida de validez* o de confianza de la sentencia “*R son E<sub>j</sub>*” puesto que este tipo de sentencias sería un caso particular de un *resumen lingüístico* de la forma “*QR son E*” cuando el cuantificador *Q* sea el cuantificador universal  $\forall$  (Rasmussen y Yager, 1999; Bosc y Liétard, 2004).

No obstante, dicha medida de validez es un decimal en  $[0, 1]$ . Por lo tanto, se necesita de una función de transformación para pasar de un número concreto a uno lingüístico y, así, el sistema pueda responder en los términos deseados. Esto es, el operador LOWA recién presentado define una función de mapeo  $+$ :  $[0,1] \rightarrow [0,1]$  por lo que habrá necesidad de otra transformación para pasar de grados de verdad en el intervalo  $[0,1]$  a valores como “muy cierto” o “falso”. Esta transformación  $\tau : [0,1] \rightarrow T$  es inversa a la cuantificación o la concreción de la vaguedad.

En (Baldwin, 1978) que se habla sobre sistemas de control difusos se propone una función de transformación  $\tau : [0,1] \rightarrow T$ , considerando un conjunto de términos  $T = \{“casi verdadero”, “verdadero”, “muy verdadero”, “absolutamente verdadero”, “absolutamente falso”, “casi falso”, “falso”, “muy falso”\}$ , que se basa en un conjunto difuso de forma triangular con parámetros  $(0, 1, 1)$  para representar la etiqueta lingüística “verdadero” y un conjunto triangular  $(0, 0, 1)$  para la etiqueta lingüística “falso”. Adicionalmente, para representar las etiquetas lingüísticas de los valores de verdad acentuados con el adverbio “muy”, propone aplicar la función de potencia, con un exponente de dos, sobre la distribución triangular correspondiente y para representar los valores de verdad relajados, como “casi cierto” o “casi falso”, aplicar la función de potencia con un exponente de  $1/2$ , como se aprecia en la Figura 41. Esta estrategia, por lo tanto, es la misma que se sugiere convencionalmente para la interpretación de etiquetas lingüísticas modificadas por adverbios de cantidad que no es guiada por el sentido común porque no es lo que haría el humano, de manera natural. Por eso, en esta propuesta se aborda otra estrategia.



**Figura 41. Funciones propuestas por Baldwin para calificadores de la verdad**

Con respecto a la forma triangular de los conjuntos borrosos sugerida para representar las etiquetas lingüísticas “verdadero” y “falso”, se considera que es apropiada pues lo que se está representando, realmente, son dos números difusos: el “absolutamente cierto” o “el absolutamente falso” y por lo tanto se espera que la forma de dichas funciones sea unimodal (Restricción No.17, sección 4.5.10). Pero como el grado de falsedad de un predicado es el complemento del grado de verdad, sólo sería necesario considerar un conjunto difuso triangular con parámetros (0, 1, 1) para representar la etiqueta lingüística “cierto” o “verdadero”. Del mismo modo, de este conjunto, también se derivan los términos lingüísticos acentuados para calificación de la verdad. Razón por la cual, lo único que se necesita definir es cómo realizar la discriminación para catalogar un real en  $[0,1]$  como “muy cierto”, por ejemplo.

Además, se debe tener presente que el proceso requerido para transformar los grados de verdad a términos lingüísticos, debe conducir a uno y sólo uno de ellos. Esto es imprescindible para que un sistema difuso pueda dar una respuesta utilizándolo. Por esto, la transformación  $\tau : [0,1] \rightarrow T$  sugiere una partición en clases mutuamente excluyentes del dominio de los reales, en el intervalo  $[0,1]$ . Y como esta partición nítida no depende del contexto, habrá que elegir los límites de las clases que representan cada calificador lingüístico de la verdad, con la ayuda del sentido común.

Considerando el conjunto de términos  $T = \{“absolutamente cierto”, “muy cierto”, “cierto”, “ni muy cierto, ni muy falso”, “poco cierto”, “falso”, “absolutamente falso”\}$ , se observa que tres de estos calificadores de la verdad se pueden considerar números difusos: el “absolutamente cierto”, cuando  $\mu_{E_j}(R) = 1$ , el “absolutamente falso” cuando  $\mu_{E_j}(R) = 0$  y “ni muy cierto, ni muy falso” cuando  $\mu_{E_j}(R)$  sea igual a 0.5. La razón de ello es que si  $\mu_E(R) = 1$  indica un encajamiento total de todas las tuplas de  $R$ , a la clase etiquetada. Un valor  $\mu_{E_j}(R) = 0$  indicaría que ninguna de ellas encaja con la etiqueta  $E_j$ , mientras que  $\mu_{E_j}(R) = 0.5$ , representa un punto de indecisión. Para la representación de los cuatro calificadores restantes se pueden determinar intervalos de clase del mismo tamaño porque los valores en el intervalo  $[0,1]$  se distribuyen uniformemente.

Por lo anterior, para determinar el calificador de la verdad de una sentencia de la forma “ $R$  son  $E_j$ ”, identificando la etiqueta lingüística apropiada para el calificador de verdad en el conjunto de términos  $T$ , se propone una transformación o mapeo  $\tau : [0,1] \rightarrow T$  basada en las siguientes reglas de inferencia.

Si  $\mu_{E_j}(R) = 1 \rightarrow$  "absolutamente cierto"

Si  $0.75 \leq \mu_{E_j}(R) < 1 \rightarrow$  "muy cierto"

Si  $0.5 < \mu_{E_j}(R) < 0.75 \rightarrow$  "cierto"

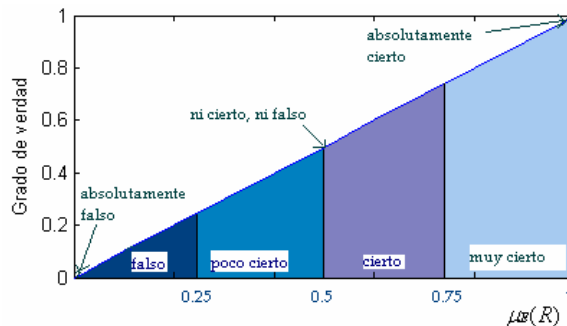
Si  $\mu_{E_j}(R) = 0.5 \rightarrow$  "ni muy cierto, ni muy falso"

Si  $0.25 \leq \mu_{E_j}(R) < 0.5 \rightarrow$  "poco cierto"

Si  $0 < \mu_{E_j}(R) < 0.25 \rightarrow$  "falso"

Si  $\mu_{E_j}(R) = 0 \rightarrow$  "absolutamente falso"

La representación gráfica de la partición matemática propuesta en esta Tesis Doctoral para determinar el calificador de la verdad de una sentencia de la forma “ $R$  son  $E$ ”, se muestra en la Figura 42.



**Figura 42. Partición para los calificadores lingüísticos de la verdad**

Ya definido el modelo de inferencia para los calificadores de verdad, éste se puede utilizar también para evaluar expresiones que incluyan cuantificadores, dado que la expresión “ $R$  son  $E$ ” implícitamente lleva consigo el cuantificador universal. Para ello, entonces, sólo es necesario considerar en la evaluación, el grado de pertenencia del cuantificador elegido:

$$“QR \text{ son } E_j” \rightarrow \mu_Q\left(\frac{1}{n} \sum_{i=1}^n \mu_{E_j}(t_i)\right) \geq \alpha$$

Por otro lado, el tipo de técnicas que permiten el hallazgo de asociaciones entre variables es conocido, en la Minería de Datos, desde hace buen tiempo bajo la denominación de *reglas de asociación* (Agrawal, Imielinski, Swami, 1993) cuya aplicación se conoció como “Análisis de las Canastas del Mercado” (Market-Basquets Analysis). Se dice que una *regla de asociación* está conformada por dos conjuntos, los que cumplen la premisa y los que cumplen la conclusión. Por esto, una regla permite inferir una concordancia entre la premisa y la conclusión que se suele medir por medio de estadísticas basadas en las frecuencias relativas de los eventos.

Las reglas de asociación de la forma  $X \rightarrow Y$ , también puede establecerse para los atributos o variables lingüísticas. Como ejemplo,  $X$  puede ser la relación de los aspirantes con un puntaje “alto” o “medio” en el examen de admisión de un programa académico y la relación  $Y$  puede estar formada por los aspirantes que están trabajando. La regla  $X \rightarrow Y$  significa que si encontramos todas las características definidas por  $X$  tenemos buena posibilidad de encontrar a  $Y$ . Por lo tanto, lo usual es considerar que la implicación  $X \rightarrow Y$  no se basa únicamente en una lógica bivaluada (cuando  $X$  es cierta, entonces  $Y$  también es cierta), sino que denota una implicación o asociación difusa. Debido a esto, la fuerza de la regla de asociación se cuantifica mediante las estadísticas (Kotsiantis, Kanellopoulos, 2006):

- **Soporte:** una regla de asociación  $X \rightarrow Y$  tiene un soporte  $S$  si un  $s\%$  de los objetos, en el dominio  $R$ , pertenecen al conjunto  $X \cap Y$ . Cuando el soporte en un caso particular es bajo, está indicando que la ocurrencia simultánea de las características definidas por  $Y$  o  $X$  es poco probable, según los datos existentes en  $R$ .

$$\text{soporte}(\text{regla}) = \frac{\text{soporte}(X)}{\text{cardinalidad}(R)} \approx P(Y \cap X)$$

- **Confianza:** una regla  $X \rightarrow Y$  es válida, con una confianza  $c$ , si el  $c\%$  de los objetos en la relación caracterizada por  $X$ , también cumplen la condición  $Y$ . Esto es, la confianza de la asociación es una estimación de la probabilidad condicional de  $Y$  dado que ha ocurrido  $X$ :

$$\text{conf}(\text{regla}) = \frac{\text{soporte}(\text{regla})}{\text{soporte}(X)} = \frac{\text{soporte}(X \cap Y)}{\text{soporte}(X)} \approx P(Y / X)$$

Puesto que las dos estadísticas presentadas proporcionan información valiosa sobre la fortaleza de una asociación, el sistema de inferencia al dar su respuesta a este tipo de preguntas, en términos cualitativos, también la acompañará de estos indicadores. Sin embargo, como el soporte, así definido, no informa sobre los valores particulares existentes en el denominador y en el numerador de estos indicadores, se considera conveniente adicionarlos a la respuesta.

De acuerdo con lo anterior, la respuesta esperada para el tipo de preguntas sobre asociaciones difusas es un registro que contiene el calificador de la verdad, la cantidad de elementos que cumplen el primer grupo de condiciones (el soporte de  $X$ ), la extensión de la relación considerada (la cardinalidad de  $R$ ), el soporte de la regla y la confianza de la misma.

Con la explicación del proceso de inferencia o de razonamiento aproximado para que un sistema interactivo de consulta-respuesta pueda responder con calificadores de la verdad, no sólo se cubren todos los tipos de preguntas que establece el lenguaje PRUF para el manejo de variables lingüísticas, sino que se cubrieron otros. Además de cubrir el que se acaba de presentar, también se cubre el caso de una condición formada por combinaciones lineales de condiciones simples, vagas o concretas.

A manera de síntesis, a continuación, se presentará la descripción formal del proceso de la concreción de los términos vagos incluidos en las condiciones de la consulta, de acuerdo el contexto lingüístico que la enmarca.

### **6.5 Especificación Formal de la Concreción de la Vaguedad**

Aquí se han propuesto unas modificaciones y extensiones al lenguaje de consulta estándar SQL a bases de datos objeto-relacionales, para admitir vaguedad en ellas. De acuerdo con el modelo propuesto, el usuario ahora puede usar etiquetas lingüísticas para restringir los valores que toman las variables o atributos cuantitativos de cualquier relación, en la base de datos, incluidos los que tienen un dominio de tipo fecha. Los nombres o rótulos para las etiquetas lingüísticas pueden ser genéricas como “alto(a)” o estar predefinidas en los metadatos de la base de datos, como parte del conjunto de términos de algunas variables, a manera de una tabla de símbolos. Para poder usar adjetivos calificativos en las consultas, se ha sobrecargado el operador existente IS, que es usado convencionalmente para saber si el valor específico de un atributo o expresión es nulo o no.

En el caso de especificar un adverbio de cantidad, se ha definido que éste se debe anteponer a la etiqueta que actúa operando, siguiendo las reglas gramaticales definidas para los operadores unarios en SQL y acordes con el lenguaje natural al cual se trata de aproximar (el inglés). Para especificar agregaciones con las conectivas lógicas de la disyunción y de la conjunción (los operadores AND y OR) la sintaxis no cambia, sólo se adiciona un nuevo operador con el símbolo “+” para especificar combinaciones lineales de condiciones vagas o concretas. También, se sobrecargó al operador lógico de comparación “LIKE” para no sólo aplicarlo a cadenas de caracteres, sino para especificar el comparador vago con un número cuantitativo.

El nuevo lenguaje propuesto permite el uso de cuantificadores relativos. Por esto, fue necesario ampliar el léxico del lenguaje SQL con palabras o tokens como “MOST” y “FEW”. De esta manera se permite que el usuario pueda conocer el valor de verdad del cumplimiento de ciertas regularidades entre los objetos de la base de datos y el sistema le responda en términos cualitativos. En el caso de un cuantificador relativo, la consulta empieza con la especificación de ese cuantificador y con las condiciones que se quieran evaluar, en reemplazo de la cláusula SELECT que en este caso se hace innecesaria.

De acuerdo con los tipos de términos vagos considerados en el modelo propuesto de razonamiento, para su representación y manejo, se han propuesto las reglas sintácticas que definen la nueva gramática  $G$  del lenguaje SQL extendido. De igual modo, se establecieron las reglas para estimar la semántica  $S$  que asocian cada valor o etiqueta lingüística  $E$  con su significado  $S(E)$ , por medio de un conjunto difuso en el marco de cognición, entre todos los posibles. A continuación se presenta la especificación formal de las reglas sintácticas y luego se prosigue con las semánticas.

### 6.5.1 Reglas Sintácticas de la Orden de Consulta

En el análisis preliminar de una consulta, se debe validar que la fórmula declarada por el usuario, se ajuste a uno de los patrones posibles derivables de la fórmula general de una orden de consulta.

Para la especificación formal de las reglas de reescritura, que determinan la fórmula general de la orden de una consulta, se empleará la notación BNF (Backus-Naur-Form) que es la metasintaxis comúnmente usada para expresar gramáticas libres de contexto (Aho, Sethi y Ullman, 1990). Es decir, una manera formal de describir lenguajes formales como el SQL que es el lenguaje que se propone extender.

En las nuevas reglas de producción que conforman el metalenguaje para la admisión de los nuevos términos vagos que se proponen en la presente Tesis Doctoral, se muestran las extensiones en negrilla. También se muestran así, los operadores que han sido sobrecargados para la flexibilización del lenguaje de consulta SQL:

```
consulta → SELECT proyección
          FROM relaciones | (consulta)
          [WHERE condiciones]
          [GROUP BY atributos [HAVING condiciones]]
          [ORDER BY atributos [ASC | DESC]]
          [ { UNION [ ALL ] | INTERSECT | MINUS } (consulta) ]
proyección → * | relación.* | [DISTINCT] expr [AS nombre ] [, expr ...]
atributos → atributo [, atributo...]
relaciones → relación [alias] [, relación...] | (consulta)
condiciones → condición_simple [conectiva condición_simple...]
             [WITH CALIBRATION { n | umbral | n, umbral } ]
condición_simple → {expr IS [NOT] NULL | expr operador expr
                    | condición_vaga | [expr] cuantificador (consulta)}
condición_vaga → expr IS [NOT] [modificador] etiqueta [j] [s]
expr → {término | función(expr) | expr * expr | expr /expr
        | expr - expr | expr + expr}
término → identificador | literal | etiqueta | escalar
relación → {R1 | R2 ... }
atributo → {A1 | A2 ... }
```

**etiqueta** → {E | E1 | literal ...}  
 función(<expr>) → F1 | F2 ...  
**modificador** → {VERY | EXTREMELY}  
**operador** → {> | < | >= | <= | <> | = | [NOT] LIKE}  
**conectiva** → {AND | OR | +}  
**cuantificador** → {ALL | EXISTS | SOME | ANY | FEW | MOST}  
 identificador → cadena de caracteres válida para un nombre  
 literal → cadena de caracteres encerrada entre comillas  
**escalar** → valor numérico simple  
**umbral** → real en (0, 1] - límite inferior para el grado de pertenencia  
**j** → entero\_positivo -- posición de una etiqueta o categoría  
**s** → entero\_positivo -- número de categorías consideradas  
**n** → entero\_positivo -- número máximo de filas por retornar

Como se puede observar en el lenguaje extendido no se usan palabras técnicas propias de la lógica difusa o símbolos extraños que afecten la legibilidad del mismo. Se ha logrado que el sistema de inferencia sea transparente para el usuario consultor de una base de datos.

Además debe notarse que si se eliminan los tokens o palabras nuevas, en las reglas de producción del metalenguaje, se vuelve a las reglas originales del lenguaje estándar SQL. Esto significa que la extensión propuesta es estrictamente aditiva.

Las reglas de producción generan diversos árboles de expansión o patrones válidos. Por esto, en el análisis preliminar de una consulta se debe identificar el patrón al cual se ajusta y si no se encuentra ninguno, entonces es porque ha ocurrido un error sintáctico. Se debe resaltar que cuando se ofrece una interfaz gráfica de usuario para formar el texto de las consultas, los errores sintácticos se minimizan.

Si no ha ocurrido un error sintáctico en la formulación de una consulta, y como las reglas de producción cobijan las solicitudes usuales, se debe chequear si el patrón de la consulta incluye términos vagos para proseguir con su análisis semántico.

### 6.5.2 Reglas Semánticas de la Orden de Consulta

En la interpretación de una consulta vaga, el sistema de inferencia (el agente inteligente) debe determinar las reglas semánticas aplicables a cada caso particular. En el modelo de razonamiento propuesto, se valdrá de los modelos teóricos propuestos para cada patrón y de las estadísticas de resumen definidas como parámetros de los modelos particulares.

**Tabla 17. Modelos propuestos para la representación de etiquetas lingüísticas**

Etiqueta	Patrón Sintáctico en SQL extendido	Modelo Heurístico Propuesto	Reglas semánticas
Etiqueta Simple (considerando 2 clases en el marco)	expr IS E j	simple_fdp (expr , E, j, 2)	Si $j = 1 \rightarrow$ hombro_izquierdo( $P_0, P_{37.5}, P_{62.5}$ ) Si $j = 2 \rightarrow$ hombro_derecho( $P_{37.5}, P_{62.5}, P_{100}$ )
Etiqueta Simple (considerando 3 clases en el marco)	expr IS E j	simple_fdp (expr, E, j, 3)	Si $j = 1 \rightarrow$ hombro_izquierdo ( $P_0, P_{12.5}, P_{37.5}$ ) Si $j = 2 \rightarrow$ trapezoidal ( $P_{12.5}, P_{37.5}, P_{62.5}, P_{87.5}$ ) Si $j = 3 \rightarrow$ hombro_derecho ( $P_{62.5}, P_{87.5}, P_{100}$ )
Etiqueta acentuada con el Adverbio "Muy"	expr IS VERY E J	muy_fdp(expr, E, j) = simple_fdp( simple_fdp(expr,E, j, 3), expr, E, j, 3)	Si $j = 1 \rightarrow$ hombro_izquierdo ( $P_0, P_{4.7}, P_{14}$ ) Si $j = 2 \rightarrow$ trapezoidal ( $P_{43}, P_{48}, P_{52}, P_{57}$ ) Si $j = 3 \rightarrow$ hombro_derecho ( $P_{86}, P_{95}, P_{100}$ )
Etiqueta acentuada con el Adverbio con el "Extremadamente"	expr IS EXTREMELY E J	extrem_fdp (expr, E, j) = simple_fdp( simple_fdp( simple_fdp( expr, E, j, 3) ,expr, E, j, 3) ,expr,,E,j, 3)	Si $j = 1 \rightarrow$ hombro_izquierdo( $P_0, P_1, P_2$ ) Si $j = 2 \rightarrow$ triangular ( $P_{49}, P_{50}, P_{51}$ ) Si $j = 3 \rightarrow$ hombro_derecho ( $P_{98}, P_{99}, P_{100}$ )
El complemento de una etiqueta	expr IS NOT [m] E j k	1- f(expr, m, E, j, k)	Si $m$ es nulo $\rightarrow$ 1- simple_fdp(expr, E, j, k) Si $m = VERY \rightarrow$ 1- muy_fdp(expr,, E, j, k) Si $m = EXTREMELY \rightarrow$ 1- extrem_fdp (expr, E, j, k)

En la Tabla 17, se presentan las reglas definidas para hallar los modelos de las etiquetas lingüísticas que pueden ser usadas, de acuerdo con una variable o una expresión deducible a partir de una variable lingüística. La expresión puede ser considerada otra variable de este estilo, aplicando el principio de extensión de Zadeh que nos dice que si la entrada de una función es borrosa, las salidas de la función serán también borrosas. Por esto, el patrón sintáctico de una condición vaga de una consulta se ha generalizado para admitir cualquier expresión, en lugar de restringirla a una variable o atributo de una relación. En dichos patrones sintácticos se omitieron los posibles calibradores que se pueden especificar en una consulta,

puesto que éstos sólo son considerados por el sistema de inferencia en el proceso deductivo de filtrar las tuplas de la memoria de trabajo y dar su respuesta.

En la tabla anterior, se presentan las reglas semánticas para las etiquetas lingüísticas considerando dos y tres clases o categorías difusas en un marco de cognición, aunque el proceso se generalizó para considerar más clases, haciendo uso de otros percentiles en la distribución de los datos, tal como aparece en el Capítulo 6. También se puede apreciar que los modelos heurísticos propuestos para las etiquetas modificadas por un adverbio de cantidad, son versiones diferentes de los modelos computacionales. Estos últimos modelos son más convenientes por su rapidez en el cálculo. Por lo tanto, las versiones computacionales son las recomendables para implementar en un sistema gestor de bases de datos objeto-relacional.

**Tabla 18. Modelos para agregaciones, cuantificadores y calificadores de verdad**

Nombre de Patrón	Patrón Sintáctico en SQL extendido	Reglas semánticas
Agregación de etiquetas con la conjunción	condición_simple1 AND condición_simple2	Si condición_simple1 $\neq$ condición_simple2 $\rightarrow$ $\min(\text{gc}(\text{TUPLA}, \text{condición\_simple1}), \text{gc}(\text{TUPLA}, \text{condición\_simple2}))$
Agregación de etiquetas con la disyunción	condición_simple1 OR condición_simple2	Si la propiedad es la misma en ambas condiciones $\rightarrow$ suma (gc(TUPLA, condición_simple1), gc (TUPLA, condición_simple2)) En caso contrario $\rightarrow$ media (gc(TUPLA, condición_simple1), gc (TUPLA, condición_simple2))
Combinación lineal de etiquetas	[ETIQUETA IS] B1*condición_simple1 + B2* condición_simple2 +...	Si B1+B2+...+BP =1 $\rightarrow$ $(B1*\text{gc}(\text{TUPLA}, \text{condición\_simple1}) +$ $B2*\text{gc}(\text{TUPLA}, \text{condición\_simple2}) , \dots)$
Cuantificador relativo "La mayoría"	condición_simple MOST (consulta)	hombro_derecho ( $P_{50}, P_{75}, P_{100}$ )
Cuantificador relativo "La minoría"	condición_simple FEW (consulta)	hombro_izquierdo( $P_0, P_{25}, P_{50}$ )
Cuantificador absoluto "similar a"	expr LIKE escalar	triangular( $P_{n-1}, P_n, P_{n+1}$ )
Cuantificador absoluto "distinto a"	expr NOT LIKE escalar	1- triangular( $P_{n-1}, P_n, P_{n+1}$ )

Nombre de Patrón	Patrón Sintáctico en SQL extendido	Reglas semánticas
Calificador de verdad	SELECT VV(consulta) FROM...  SELECT VV(cuantificador, condiciones_antecedente, condiciones_consecuente, origen) FROM...	<i>Si <math>\mu_E(R) = 1 \rightarrow</math> "absolutamente cierto"</i> <i>Si <math>0.75 \leq \mu_E(R) &lt; 1 \rightarrow</math> "muy cierto"</i> <i>Si <math>0.5 &lt; \mu_E(R) &lt; 0.75 \rightarrow</math> "cierto"</i> <i>Si <math>\mu_E(R) = 0.5 \rightarrow</math> "ni muy cierto, ni muy falso"</i> <i>Si <math>0.25 \leq \mu_E(R) &lt; 0.5 \rightarrow</math> "poco cierto"</i> <i>Si <math>0 &lt; \mu_E(R) &lt; 0.25 \rightarrow</math> "falso"</i> <i>Si <math>\mu_E(R) = 0 \rightarrow</math> "absolutamente falso"</i>

Además de las reglas para admitir condiciones vagas con etiquetas lingüísticas, usadas como adjetivos calificativos, se han definido otras reglas para representar las etiquetas lingüísticas que se derivan de la agregación o combinación de otras. Esta propuesta incluye también las reglas sintácticas y semánticas para representar cuantificadores vagos relativos y las reglas para responder a las consultas con términos lingüísticos de la verdad. Estas reglas se presentan en la Tabla 18, de forma resumida. En esta misma tabla, la función denominada *gc* calcula el grado de pertenencia o cumplimiento, por parte de una tupla, de la condición que se recibe como parámetro.

## 6.6 Extensiones adicionales del lenguaje SQL, considerando el estándar SQL:1999

El lenguaje SQL de los sistemas gestores de bases de datos objeto-relacionales, está basado en el tercer estándar definido por la norma 9075 de la ISO/IEC de 1999. Allí aparecen los conceptos de la programación orientada por objetos y las componentes dinámicas que deben caracterizar al lenguaje de consulta SQL (ISO/IEC International Standard, 1999).

Puesto que el aporte esperado de este trabajo de investigación está orientado a los sistemas de consulta a bases de datos cuyo soporte lógico es el modelo objeto-relacional, aún falta describir cómo pueden ser usadas las extensiones propuestas al lenguaje SQL para admitir vaguedad, al definir los objetos o procedimientos del paradigma objeto-relacional bajo el estándar SQL:1999.

A continuación se describen los nuevos componentes u objetos de una base de datos que define el estándar que pueden ser modificados para incorporar términos vagos en ellos, de forma análoga a la extensión denominada SQLf3 propuesta a partir del lenguaje SQLf para dar soporte a las características que define el tercer estándar SQL:1999 (Goncalves y Tineo, 2001).

### 6.6.1 Funciones y Procedimientos Almacenados

Un procedimiento almacenado o una función es un programa que es compilado y almacenado físicamente en una base de datos. La ventaja de un procedimiento almacenado es que al ser invocado, es ejecutado directamente en el

motor de bases de datos, accediendo directamente a los datos que necesita manipular y sólo necesita enviar los resultados, evitando la sobrecarga resultante de comunicar grandes cantidades de datos salientes o entrantes, entre la base de datos y las aplicaciones. También facilitan el mantenimiento puesto que evitan la incorporación de la misma lógica en distintas aplicaciones.

Con la extensión imperativa del lenguaje SQL, admitida con el tercer estándar SQL:1999, se pueden construir procedimientos que pueden ser ejecutados intencionalmente o automáticamente cuando ocurre un evento. También, es posible construir funciones propias e incluso, en algunos casos, bibliotecas de funciones y procedimientos comúnmente llamados paquetes.

En un procedimiento o función construido con lenguaje imperativo se puede especificar una consulta SQL. Por lo tanto, esa consulta puede ajustarse a los patrones vagos admisibles para la representación de vaguedad. Como ejemplo, el procedimiento siguiente califica un vendedor como “excelente” cuando cumpla los criterios vagos enunciados.

```
CREATE PROCEDURE calificar_vendedor(id numeric)
AS
BEGIN
UPDATE vendedor
SET calificación = "Excelente"
WHERE identificación= id AND
acumulado_ventas IS "alto(a)" AND
(hrs_trabajadas/hrs_programadas) IS "alto(a)";
END;
```

Cabe señalar que como la cláusula WHERE puede estar incluida en cualquier sentencia de actualización o de borrado, las condiciones vagas también pueden especificarse en estas otras sentencias DML (Data Manipulation Language) del SQL.

### 6.6.2 *Cursores*

Los cursores son usados en un procedimiento o función para procesar más allá de la primera fila retornada por una consulta. Por esto, cuando se declara un cursor también se pueden incluir términos vagos, como en el ejemplo:

```
DECLARE
CURSOR aumento_por_empleado IS
SELECT nombre, sueldo*0.1 AS aumento
FROM empleados WHERE sueldo IS "bajo"
OR sueldo IS "medio";
```

### 6.6.3 *Disparadores*

Los disparadores son procedimientos que se ejecutan de manera automática o implícita, ante una operación de mantenimiento de datos (borrado, inserción o modificación) sobre cierta tabla. Y como cualquier otro procedimiento, puede incluir

una consulta a la base de datos. Por lo tanto, también se pueden incluir condiciones vagas en un disparador, como en el ejemplo:

```
CREATE OR REPLACE TRIGGER verif_sueldo
BEFORE INSERT OR UPDATE OF sueldo, cargo ON empleados
FOR EACH ROW
DECLARE
    v_min empleados.sueldo%TYPE;
BEGIN
    SELECT min(sueldo) INTO v_min
        FROM empleados
        WHERE fecha_ing IS "reciente";
    IF :new.sueldo < v_min THEN
        RAISE_APPLICATION_ERROR (-20501,
            'Se detectó un error en el sueldo');
    END IF;
END;
```

#### **6.6.4 Tipos de Datos definidos por el Usuario**

Un tipo de dato definido por el usuario es un tipo de dato abstracto (ADT, por sus siglas en inglés), representado mediante una clase. Por ello, se describe mediante sus propiedades: sus atributos y métodos. Estos últimos permiten especificar operaciones aplicables al ADT. También permiten definir atributos derivados. Por lo tanto, dentro de la especificación de un tipo de datos se pueden definir métodos que incluyan condiciones vagas o que generen el grado de compatibilidad de los objetos definidos por el tipo, con alguna etiqueta vaga. Por ejemplo, el método siguiente encuentra el grado de compatibilidad *cg* de una persona a un conjunto rotulado con una etiqueta lingüística *E*, con la posición *J* en el marco de cognición y considerando el número de categorías especificado por *Nro\_cat*, como un atributo derivado:

```
CREATE or REPLACE TYPE BODY tipo_persona AS
MEMBER FUNCTION gc_grupo_edad (Fecha_Nac DATE, E VARCHAR, J
numeric, Nro_cat NUMERIC) RETURN NUMERIC
IS
BEGIN
RETURN gc((SysDate - Fecha_Nac)/365, E, J, Nro_cat);
END;
END;
```

#### **6.6.5 Vistas**

Una vista no es más que una consulta que se ejecuta cada vez que se hace referencia a ella en otra consulta. Por lo tanto, una vista puede definirse por medio de una consulta vaga que siga alguno de los patrones admisibles en la extensión del lenguaje propuesto, como en el ejemplo siguiente:

```
CREATE VIEW autos_atractivos AS
```

```
SELECT * FROM autos
WHERE 0.4*(mpg IS "alto")+ 0.6*(precio IS "bajo")
```

Esto mismas especificaciones son válidas para una vista materializada, antes llamada instantánea (snapshot, en inglés).

### 6.6.6 La Cláusula FROM de una Consulta

Bajo el estándar SQL:1999 se otorga mayor flexibilidad al lenguaje permitiendo que en la cláusula FROM de una consulta, se especifique una subconsulta para indicar la procedencia de los datos. Por lo tanto, una consulta vaga también se podría especificar en ese lugar, con la extensión del lenguaje propuesto. Por ejemplo, la consulta siguiente sería válida:

```
SELECT marca, precio
FROM (SELECT * FROM autos WHERE mpg IS "alto")
WHERE precio IS "bajo"
```

### 6.6.7 La Declaración CASE en una Consulta

Otra de las novedades del estándar SQL:1999 es la declaración CASE que se admite en la cláusula SELECT. Se usa para crear variables o atributos a partir de otros y su sintaxis es:

```
SELECT CASE WHEN condiciones THEN resultado1
           [WHEN ...]
           [ELSE resultadoj]
END
```

Puesto que en una declaración CASE, el símbolo no terminal "condiciones" puede incluir predicados vagos, con la extensión del lenguaje propuesto, la sentencia siguiente sería válida:

```
SELECT marca,
       CASE WHEN (mpg IS 'alto' AND precio IS 'bajo')
            THEN 'atractivo'
            WHEN (mpg IS 'medio' and precio IS 'bajo')
            THEN 'algo atractivo'
            ELSE 'nada atractivo'
       END
       AS concepto
FROM autos
```

## 6.7 Nuevas Sentencias de Definición de Datos

Además de las modificaciones en la orden de consulta SQL para admitir vaguedad, también se ha considerado que en cualquier momento se pueden agregar o cambiar los operadores o cuantificadores definidos. Por lo tanto, se debe enriquecer el lenguaje con nuevas sentencias de definición de datos. Siguiendo la forma gramatical para la creación o modificación de objetos de la base de datos, la sintaxis propuesta es:

```
CREATE [OR REPLACE] QUANTIFIER identificador AS  
conjunto_difuso(nombre_función, estimadores_de_parámetros)
```

```
CREATE [OR REPLACE] OPERATOR identificador AS  
conjunto_difuso(nombre_función, estimadores_de_parámetros)
```

En dichas órdenes, la cantidad de estimadores de los parámetros deben corresponder al nombre de la función. Para eliminar alguno de los objetos creados, se puede hacer por medio de la orden usada convencionalmente:

```
DROP OPERATOR identificador o  
DROP QUANTIFIER identificador
```

De forma parecida, se deben poder agregar o eliminar etiquetas lingüísticas para complementar el conjunto de términos de una variable. Cuando éstas no sean relativas al contexto lingüístico que delimite una consulta, también se deberá poder especificar el modelo difuso correspondiente, junto con el valor de los argumentos:

```
CREATE [OR REPLACE] LABEL identificador AS  
conjunto_difuso(variable, nombre_función, [parámetros],  
posición_marco)
```

# Capítulo 7

## 7 Evaluación del Modelo Propuesto

Este Capítulo se dedica a la evaluación del modelo de razonamiento aproximado propuesto, presentado en el capítulo precedente. Esta evaluación tratará dos aspectos. El primero de ellos, es la evaluación de su factibilidad técnica mediante la continuación del ciclo de desarrollo a partir del modelo conceptual propuesto. Concretamente, se describe la arquitectura de un sistema de consulta-respuesta basado en el modelo y se describe el modelo relacional correspondiente a la base de conocimientos requerida en la interpretación de la vaguedad y se acompaña de las interfaces Web que muestran como se pueden materializar las nuevas construcciones del modelo para la interpretación de los términos vagos admisibles, y el segundo aspecto de evaluación tiene que ver con la confirmación teórica del cumplimiento de las propiedades y restricciones a las cuales debe ajustarse el modelo de razonamiento propuesto, definidas previamente en el Capítulo 4.

### 7.1 *Factibilidad técnica de sistemas basados en el modelo propuesto*

Para determinar la factibilidad técnica de sistemas basados en el modelo propuesto se diseñó un sistema WEB interactivo de consulta-respuesta, usando el lenguaje PHP y el JavaScript para la creación de las interfaces humano-máquina y como sistema gestor de bases de datos se utilizó a PostgreSQL, cuyo modelo lógico está basado en el modelo objeto-relacional. El software utilizado tiene la ventaja de ser de distribución libre que, junto a la base de datos de referencia, brinda la posibilidad de comprobar más fácilmente los resultados para los ejemplos presentados.

Inicialmente la idea era construir un sistema de consulta-respuesta sobre componentes o material educativo en una plataforma de educación virtual concebida para una red académica de la cual forma parte la Universidad Nacional de Colombia, pero posteriormente se pensó que el prototipo podía ser general o independiente del dominio de aplicación para que se apreciara mejor su utilidad y generalidad.

#### 7.1.1 *Arquitectura General del Sistema Flexible de Consulta-Respuesta*

La arquitectura que se consideró más conveniente para un sistema flexible de consulta-respuesta, basado en el modelo propuesto, está constituida por una red de computadores en la que se distribuyen las tres capas lógicas o módulos de

programación que conforman el sistema. Dicha arquitectura se muestra gráficamente en la

Figura 43.

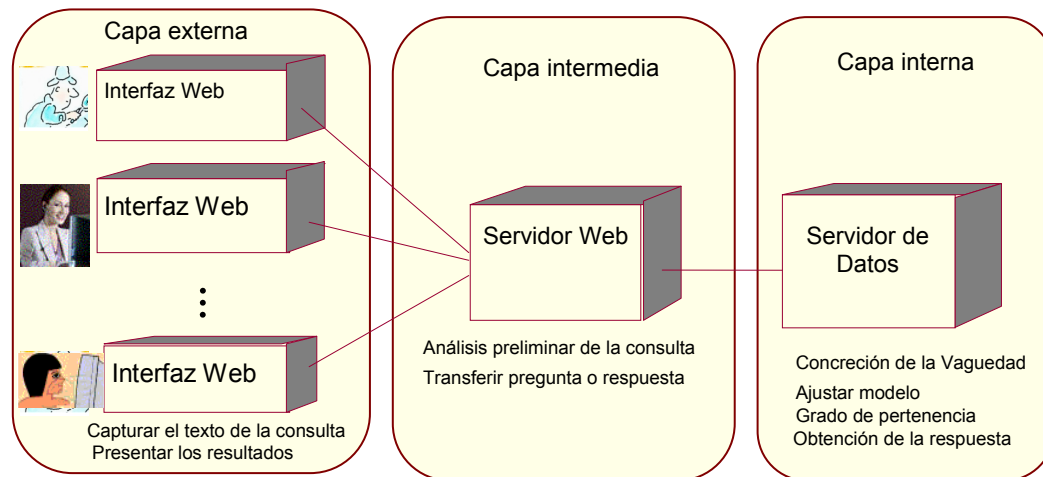


Figura 43. Arquitectura para el sistema de consulta-respuesta

El nodo principal de la red, la capa interna, lo constituye el servidor de datos. Allí residen no sólo los datos, sino las reglas sintácticas y semánticas para representar y operar con términos vagos. Esta decisión de diseño se tomó pensando en la reutilización de las reglas de inferencia difusa, las funciones de los conjuntos difusos o las funciones estadísticas por otros aplicativos o en otros análisis más profundos de los datos y considerando otro tipo de usuario más calificado.

Otra razón para guardar de manera centralizada las reglas y procedimientos del sistema de inferencia en la base de datos, es facilitar la escalabilidad y el mantenimiento del sistema, puesto que las modificaciones o extensiones sólo afectarían la capa interna, logrando independencia entre los datos y los programas que los usan. Por otro lado, si se hubiera optado por definir dichas reglas en la capa intermedia, como se ha hecho tradicionalmente en las propuestas de flexibilización del lenguaje para admitir imprecisión, se tendría la desventaja del tráfico de datos entre el servidor Web y el servidor de datos, afectando el rendimiento o la eficiencia del sistema. Además, para la actualización o complemento de las reglas, se tendría que hacer uso de un lenguaje de más bajo nivel de abstracción, no declarativo como el SQL, lo que haría más difícil las mejoras o extensiones futuras.

La capa lógica externa del sistema, que reside en el computador del usuario, es una capa liviana constituida por un conjunto de páginas HTML que permiten formular las preguntas deseadas y presentar los resultados formateados, a través de la Web. Con este diseño, el sistema de consulta-respuesta se hace independiente del

sitio de acceso, de la máquina y del sistema operativo que se tenga instalado, lo que implica su disponibilidad universal y su facilidad de acceso, sin necesidad de instalación de ninguna componente cliente. El usuario sólo requerirá un programa navegador de Internet ofreciendo la mayor portabilidad posible al sistema flexible de consulta-respuesta.

La capa intermedia (middleware, en inglés) la constituye el servidor Web que sirve de mediador entre los nodos cliente y el servidor de datos, además de encargarse del análisis léxico y sintáctico de la fórmula de la consulta. Esta componente maneja los protocolos de comunicación entre el sistema gestor de bases de datos y las interfaces Web, que se necesitan para abrir o cerrar la conexión con la base de datos, para enviar las solicitudes o recibir las respuestas desde el servidor de datos. Este servidor Web posee el intérprete de los lenguajes de programación que están embebidos en las páginas Web que constituyen las interfaces de usuario. Aquí se optó por un servidor Apache de Apache Software Foundation.

### **7.1.2 *Diseño Detallado del Sistema de Consulta-Respuesta***

En esta sección se detallará la funcionalidad del prototipo construido de un sistema flexible de consulta-respuesta basado en el modelo propuesto de razonamiento, adaptable a los distintos contextos lingüísticos.

En las interfaces gráficas de usuario concebidas para el planteamiento de los distintos tipos de consultas admisibles, usando el lenguaje extendido en esta Tesis Doctoral, se podrá apreciar cómo los conceptos o símbolos de la Lógica Difusa son transparentes para el usuario final, que en nuestro caso, es un usuario cualquiera.

Después de la descripción de la capa externa del sistema, se presentará el modelo lógico de los datos y el modelo funcional requerido en la interpretación de la vaguedad de las consultas, bajo el esquema objeto-relacional.

#### **7.1.2.1 *Formulación de la Consulta***

El primer paso que se lleva a cabo en un sistema interactivo de consulta-respuesta es la formulación de una consulta por parte del usuario final. Puesto que los datos que deben ser proporcionados son diferentes dependiendo del tipo de pregunta que se formule, el prototipo desarrollado consta de dos interfaces diferentes. En la primera de ellas, se pueden formular preguntas acerca de las características de los objetos, imponiéndoles ciertas restricciones vagas o concretas, que a su vez pueden ser simples o compuestas, dependiendo del uso de los operadores de agregación. La segunda interfaz se diseñó con el objetivo de encontrar o verificar la existencia de relaciones o asociaciones entre variables o atributos, gracias a la inclusión de los cuantificadores lingüísticos en el lenguaje de consulta.

#### **7.1.2.2 *Interfaz para Consultas con Condiciones Simples o Compuestas***

La interfaz diseñada para este tipo de preguntas se muestra en la Figura 44. En ella se aprecian cuatro secciones principales que tienen como propósito ofrecer la

flexibilidad requerida para permitir una gran variedad de preguntas que pueden incluir condiciones vagas, pero que no se restringen sólo a ellas.

Consulta de características de objetos en la base de datos

**Mostrar:**

Origen/tabla	Propiedad	Nombre temporal	Ordenar
Autos	modelo		asc
	mpg	millas por galón	desc
	marca		

**Condiciones que deben cumplir los elementos buscados:**

Origen	Propiedad	No	Condición	Adverbio	Adjetivo o Vlr	Ponderación	Conectiva
Autos	mpg	<input checked="" type="checkbox"/>	es	muy	alto(a)		
		<input checked="" type="checkbox"/>					
		<input checked="" type="checkbox"/>					
		<input checked="" type="checkbox"/>					

**Grado mínimo de cumplimiento de todas las condiciones**  (valor entre cero y uno)

**Agrupar por:**

Origen/tabla	Propiedad

**Condiciones para los grupos**

Origen	Propiedad	Condición	Valor	Conectiva

**Máxima cantidad de registros que se desean ver**

Figura 44. Interfaz de consultas con condiciones vagas simples o compuestas

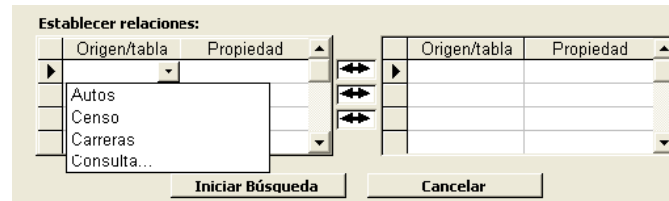
El actor o usuario final de esta interfaz es un usuario cualquiera, al que no se le pide conocer el lenguaje SQL, un lenguaje de programación que a pesar de ser declarativo demanda un alto dominio de las herramientas informáticas para plantear las solicitudes requeridas. Para los usuarios expertos en SQL, se puede crear una interfaz más sencilla donde escriban y editen en un sólo cuadro de texto, la fórmula bien formada de su consulta.

En la interfaz diseñada como ejemplo, los únicos datos obligatorios corresponden a la sección “Mostrar”, una tabla con cuatro columnas y un número variable de filas que permiten especificar las características (atributos básicos o derivados) de los objetos que se desean visualizar y el origen o procedencia de los mismos. La columna “Origen” puede ser nula exceptuando ese valor para la primera fila. Si en una fila existe un valor nulo para esa columna, significa que el atributo procede del origen especificado, en la última de las filas anteriores con algún valor para este atributo.

Con ese diseño, las consultas pueden involucrar varias relaciones o vistas de las existentes en la base de datos, aunque las seleccionadas deben tener alguna relación

directa o indirecta. Por esto, cuando se especifiquen orígenes diferentes, aparecerá la ventana emergente que se muestra en la

Figura 45, concebida para establecer las relaciones entre las tablas o vistas especificadas.



**Figura 45. Ventana emergente “Establecer relaciones”**

Para facilitar la entrada de datos y evitar posibles errores léxicos, el sistema ofrece listas con los valores posibles para los nombres de los objetos. Estas listas se usan para especificar las relaciones (básicas o vistas) existentes en la base de datos, los nombres de los atributos o propiedades de las tablas seleccionadas y los valores posibles para definir las condiciones de la consulta (operadores vagos de comparación, adverbios, calificativos, entre otros). Por lo tanto, para garantizar que el sistema interactivo se mantenga actualizado sobre los cambios en el lenguaje del dominio de aplicación, se pensó guardar en los metadatos de la base de datos toda esta información. Esto hace que después de abrir la conexión con la base de datos, se deban ejecutar algunas consultas que formen las listas de valores requeridas. Entre ellas, la correspondiente a la columna con la etiqueta “Origen/tabla” de la sección “Mostrar”. En PostgreSQL, la orden requerida para consultar las tablas y vistas existentes en la base de datos es:

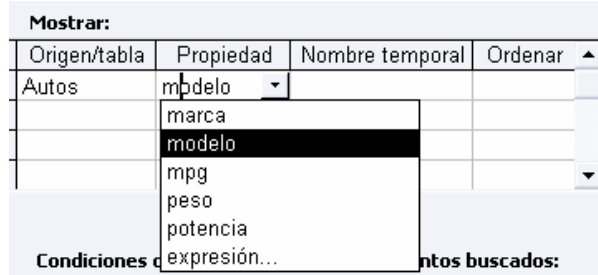
```
SELECT tablename AS origen
FROM pg_tables
WHERE schemaname = 'public'
ORDER BY tablename;
```

La lista desplegable con las relaciones o vistas disponibles, permite seleccionar una de ellas, por cada atributo o propiedad que se desee visualizar. Cuando el usuario ya haya hecho una selección, se debe formar dinámicamente otra lista con los nombres de los atributos de la relación seleccionada. Puesto que más adelante, cuando se esté detectando la presencia de vaguedad en las condiciones de la búsqueda, se necesita conocer el tipo de los atributos especificados, se debe aprovechar la consulta al catálogo del sistema para traer no sólo el nombre de todos los atributos de la relación elegida, sino su tipo. Supongamos, por ejemplo, que se ha seleccionado la relación “Autos”. Por lo tanto, se debe formar dinámicamente la siguiente orden.

```
SELECT column_name AS propiedad, data_type AS tipo
FROM columns
WHERE table_name = 'autos'
```

`ORDER BY column_name;`

Como el usuario final no tiene por qué conocer los tipos de los atributos, sólo se muestran los nombres en la lista de valores correspondiente a la columna "Propiedad" de la sección "Mostrar", como se aprecia en la Figura 46.



**Figura 46. Sección "Mostrar" de Interfaz de Consulta**

En la lista desplegable ofrecida para entrar los datos a la columna "propiedad" también aparece, al final, el ítem "expresión...". Esta opción tiene como propósito ofrecerle al usuario una manera amigable de formar expresiones. Si se selecciona esta opción, aparece la ventana emergente que se muestra en la Figura 47. En dicha interfaz, se pueden seleccionar y usar las funciones definidas en la base de datos, para ser aplicables a los atributos de la tabla elegida.



**Figura 47. Ventana emergente para formar expresiones**

La columna "Nombre temporal" que aparece en la sección "Mostrar" de la interfaz, es el alias que se le quiere dar a una de propiedades en la tabla resultante y la columna "Ordenar", cuyos dos valores posibles son "ASC" y "DESC", indican un orden lexicográfico ascendente y descendente, respectivamente. Si no se especifica un ordenamiento determinado, los resultados se presentan de manera descendente, empezando con las tuplas que cumplan las condiciones de la búsqueda con mayor grado.

En la sección de la interfaz rotulada con "Condiciones que deben cumplir los elementos buscados", que se muestra en la Figura 48, cada fila representa una condición simple de la cláusula WHERE. Estas condiciones simples se unen por medio del operador de la conjunción, de la disyunción o de la combinación lineal.

Puesto que la cantidad de condiciones no se puede determinar de antemano, esta sección también debe contener un número variable de filas o permitir adicionar nuevas filas cuando se agoten las que aparecen en pantalla.

En la sección diseñada para especificar las condiciones de la búsqueda, también se ofrecen listas desplegables de valores para todas las columnas excepto para la ponderación o el peso de una condición, en una combinación lineal. La columna "Origen" puede ser nula, pero desde la segunda fila, como en el caso de la sección "Mostrar". Esto significará que el atributo procede del origen especificado previamente en la última de las filas anteriores. El valor específico para la columna "Adjetivo o Vlr" es entrado por el usuario o seleccionado de una lista que se forma dinámicamente con las etiquetas que aparezcan en el conjunto de términos para el atributo considerado, si se trata de uno cuantitativo. Si el atributo es de este tipo, pero no tiene un conjunto de términos asociado, se ofrece una lista de etiquetas con valores genéricos como "alto(a)", "pequeño(a)", entre otros. Cuando se escoge una etiqueta genérica, se le pregunta al usuario por el número de categorías que quiere considerar en la clasificación de los datos y se guarda ese valor en otra columna no visible. Si no se especifica ninguno, se suponen tres categorías para la discriminación o clasificación difusa de los objetos.

Condiciones que deben cumplir los elementos buscados:							
Origen	Propiedad	No	Condición	Adverbio	Adjetivo o Vlr	Ponderación	Conectiva
Autos	mpg	<input type="checkbox"/>	es		alto(a)		
	peso	<input type="checkbox"/>	mayor que		2000		
		<input type="checkbox"/>					
		<input type="checkbox"/>					

Grado mínimo de cumplimiento de todas las condiciones  (valor entre cero y uno)

Figura 48. Sección "Condiciones" de la Interfaz de Consulta de Características

En la sección donde se especifican las condiciones o restricciones que deben cumplir las tuplas, el usuario puede optar por especificar una expresión tanto para la columna "Propiedad", como para la columna "Adjetivo o vlr", del mismo modo que en la sección "Mostrar" de la interfaz. Para evitar errores sintácticos también se realizan algunos chequeos o acciones. Por ejemplo, si el atributo o propiedad especificado no es numérico (int4, int8, int2, float4, float8 o numeric) o de fecha (date), se deshabilitan las columnas "Adverbio" y "Ponderación". Si se especifica un adverbio, sólo es posible emplear una etiqueta para la columna "Adjetivo o Vlr". También se chequea que cada condición quede especificada completamente, de acuerdo con alguno de los patrones aceptables en el lenguaje. Si alguna condición no encaja con alguno de los patrones, se le informa al usuario del error y se le permite reintentarlo. Por otro lado, se verifica que el umbral o el grado de cumplimiento mínimo que se imponga sobre la condición general, sea un valor en el intervalo [0,1]. En caso de no especificar ningún umbral se supone un alfa corte estricto de cero.

Después del análisis léxico, considerando el patrón formado por los valores entrados por el usuario, si al menos una condición de la consulta es vaga se debe guardar temporalmente en la base de datos, para que el sistema de inferencia halle la semántica de los términos u operadores incluidos. En este caso, también se incluye la propiedad o atributo implicado, como parte de la proyección de la orden de consulta, si antes no se había especificado en la sección “Mostrar”. Cuando una condición especificada sea concreta y no forma parte de una combinación lineal de condiciones, que se identifica por incluir la conectiva “+”, se agrega a la cadena de texto correspondiente a las condiciones concretas que luego sirven para armar el texto completo de la consulta que debe ejecutarse para delimitar el contexto y formar la memoria de trabajo (los hechos) del sistema de inferencia, antes de iniciar el proceso de análisis semántico.

La sección de la interfaz de consulta rotulada con la etiqueta “Agrupar por”, corresponde a la especificación de grupos de los cuales se quieren consultar algunas características. Por lo tanto, esta sección sólo podrá tener valores cuando se especifiquen funciones de grupo en la sección “Mostrar”, como la función que calcula el promedio de un conjunto de valores (la función AVG) o el valor máximo (la función MAX). Si la sección “Agrupar por” está vacía, no se puede especificar ninguna restricción en la sección “Condiciones para los grupos” de la interfaz.

Consulta de características de objetos en la base de datos

**Mostrar:**

Origen/tabla	Propiedad	Nombre temporal	Ordenar
Autos	modelo		▲
	hp	caballos de fuerza	
			▼

**Condiciones que deben cumplir los elementos buscados:**

Origen	Propiedad	No	Condición	Adverbio	Adjetivo o Vlr	Ponderación	Conectiva
Autos	mpg	<input checked="" type="checkbox"/>	es	muy	alto(a)	0,5	+
	marca	<input checked="" type="checkbox"/>	=		Ford	0,3	+
	peso	<input checked="" type="checkbox"/>	es		bajo	0,2	
		<input checked="" type="checkbox"/>					

**Grado mínimo de cumplimiento de todas las condiciones**  (valor entre cero y uno)

**Agrupar por:**

Origen/tabla	Propiedad
	▲
	▼

**Condiciones para los grupos**

	Origen	Propiedad	Condición	Valor	Conectiva
▶					▲
					▼

**Máxima cantidad de registros que se desean ver**

Figura 49. Especificación de una Combinación Lineal de Condiciones Simples

En la interfaz de consulta diseñada para visualizar algunas propiedades de los objetos, también es posible plantear preguntas con una condición formada por una combinación lineal de restricciones simples vagas o concretas. En este caso, se usa la conectiva '+' y el valor del peso o ponderación de la condición sobre las demás, se define en la columna "Ponderación". En este caso, se debe verificar que la suma de esta columna, para las condiciones simples que forman parte de la combinación, dé como resultado el valor de 1 que representa al 100% de los pesos. En el ejemplo que se ilustra en la Figura 49, se quiere visualizar el modelo y los caballos de fuerza de los autos cuyo rendimiento sea "muy alto", la marca sea igual a "Ford" y su peso sea "bajo" y a cada uno de estos factores se le están asignando unas ponderaciones de 0.5, 0.3 y 0.2, respectivamente.

Dentro de las etiquetas lingüísticas almacenadas en los metadatos de la base de datos, es posible que haya algunas definidas como la combinación lineal de ciertas condiciones vagas o concretas. Por lo tanto, si el usuario especifica la etiqueta deseada, en la interfaz, el sistema puede traer y mostrar las condiciones predefinidas para catalogarlo así. Por ejemplo, si el usuario selecciona el término "atractivo" para calificar a los autos y éste está guardado como una etiqueta del sistema, con una consulta simple y en línea al catálogo de la base de datos, se pueden traer y mostrar en la interfaz, los calificativos y los pesos que se le hayan otorgado a cada uno de ellos.

Con respecto a los cuantificadores relativos admisibles en las condiciones de las consultas, cuando el usuario final incluya un término como "la mayoría", el valor que debe entrarse en la columna "Adjetivo o Vlr" debe ser una subconsulta para tener una fórmula bien formada, acorde con la sintaxis del lenguaje estándar. Por lo tanto, se puede usar otra instancia de la misma interfaz de consulta para especificar la subconsulta deseada. Y Apenas el usuario termine de formular la subconsulta, se cierra su interfaz y regresa a la interfaz de la consulta original.

El último caso a considerar sobre los tipos de consulta que se pueden formular a través de la interfaz diseñada, son los cuantificadores absolutos representados por los operadores vagos "LIKE" y "NOT LIKE". Como este operador ya existía para las cadenas de caracteres, sólo se necesita detectar si una condición de la búsqueda está haciendo referencia a un cuantificador absoluto. Esto se detecta, si el operando es un valor escalar numérico o una subconsulta.

Cuando el usuario haya terminado de plantear su consulta, deberá presionar el botón "Iniciar búsqueda". En ese momento se forman las cadenas de texto que corresponden a las componentes principales de la consulta. De la sección "Mostrar", se forma una cadena con la proyección o lista de las propiedades de los objetos que se desean visualizar (la cláusula SELECT de la orden de consulta en SQL), otra cadena con la lista de las tablas involucradas (la cláusula FROM) y otra cadena de texto que describe el orden para la visualización de los resultados (la cláusula ORDER BY). En el ejemplo presentado en la Figura 40, como no hay condiciones concretas se arman las cadenas de texto que deben ser usadas o guardadas

temporalmente, como parte de la consulta. Las cadenas de texto formadas, son las siguientes:

```

proyección_original ← 'SELECT modelo,
                        mpg AS "millas por galón", marca '
proyección ← 'SELECT modelo, mpg, marca '
origen ← 'FROM autos '
join ← ''
condiciones ← ''
condiciones ← concatenación(join, condiciones)
columnas ← ''
ordenamiento ← 'ORDER BY modelo ASC, mpg DESC'
group_by ← ''
having ← ''
agrupamiento ← concatenación( group_by, having)
proyección ← concatenación (proyección, columnas)
memoria ← concatenación (proyección, origen, condiciones,
                        agrupamiento)

```

De acuerdo con lo anterior, la memoria de trabajo se forma a partir de la ejecución de la cadena de texto denominada "memoria". Debe notarse que en esta cadena no se incluyó el nombre de la columna "mpg", incluida en la condición vaga especificada en el ejemplo tratado, porque ya estaba incluida en la cadena de la proyección.

Las condiciones vagas simples de la consulta o concretas pero que formen parte de una combinación lineal, deben quedar como pendientes de interpretación en otra tabla temporal. Cada una de ellas, se va desambiguando, desde el extremo derecho hasta el izquierdo, colocando el resultado de una evaluación intermedia, al lado derecho de la fórmula. Cuando en la fórmula de una condición, todos los términos vagos hayan sido concretados, mediante el modelo semántico ajustado a su patrón, se guarda el texto traducido como otra componente de la condición de una consulta.

**Tabla 19. Resultados de la ejecución de la consulta vaga de ejemplo**

modelo	millas por galón	marca
71	35	datsum 1200
75	33	honda civic cvcc
76	33	honda civic
77	36	renault 5 gtl
77	33.5	dodge colt m/m
77	33.5	datsum f-10 hatchback
78	43.1	volkswagen rabbit custom diesel
78	39.4	datsum b210 gx
78	36.1	honda civic cvcc
78	36.1	ford fiesta
79	37.3	fiat strada custom
79	35.7	dodge colt hatchback custom
79	34.5	plymouth horizon tc3
79	34.2	plymouth horizon
79	34.1	maxda glc deluxe

Después de la concreción de todas las condiciones vagas, el sistema de inferencia debe considerar los calibradores especificados en la consulta (el alfa-corte y el número máximo de filas que se desean visualizar para filtrar las tuplas). De esta manera, el sistema debe responder a la pregunta planteada en la interfaz con la presentación de los resultados, en forma tabular. Para el ejemplo que se está tratando, la Tabla 15 muestra los resultados que generaría el sistema, considerando un alfa corte estricto de cero y estableciendo que sólo se quieren visualizar los primeros 15 autos, cuyo rendimiento se pueda considerar “muy alto” y ordenados, según se especificó (por modelo ascendentemente y por millas por galón, de manera descendente).

### 7.1.2.3 Interfaz de Consultas para Validar Reglas de Asociación

El modo de plantear una consulta cuando se quiera conocer el valor lingüístico de verdad de una regla de asociación, es diferente al caso anterior. Este tipo de preguntas permite saber si los objetos que cumplen ciertas condiciones, también cumplen otras y por eso, es necesario especificar tanto las condiciones que conforman el antecedente, como la consecuencia o conclusión de la regla de asociación. Además, la respuesta a este tipo de consultas no contiene una proyección o vista de atributos y por eso se puede aprovechar la misma interfaz para mostrar los resultados. Por lo tanto, para este tipo de preguntas, se ha concebido la interfaz que aparece en la Figura 50.

Validación de Reglas de Asociación

Dentro de la colección de

Los elementos cumplen con las condiciones:

Propiedad	No	Condición	Adverbio	Adjetivo o Vlr	Ponderación	Conectiva
mpg	<input checked="" type="checkbox"/>	es		alto(a)		
	<input type="checkbox"/>					
	<input type="checkbox"/>					
	<input type="checkbox"/>					

¿Es cierto que  también cumplen las condiciones siguientes?:

Propiedad	No	Condición	Adverbio	Adjetivo o Vlr	Ponderación	Conectiva
	<input type="checkbox"/>					
	<input type="checkbox"/>					
	<input type="checkbox"/>					
	<input type="checkbox"/>					

Respuesta

Elementos que cumplen el primer grupo de condiciones:  de   %

Elementos que cumplen ambos grupos de condiciones:  de   %

Figura 50. Interfaz para validación de reglas de asociación

La interacción del usuario con esta interfaz es parecida al caso anterior. De una lista de las posibles tablas o vistas en la base de datos, se elige aquella que contenga las características de los elementos que sirven de base para la comparación. También

se otorga la posibilidad de usar los diferentes cuantificadores lingüísticos definidos en el lenguaje extendido, mediante otra lista de valores.

Para ilustrar el uso de la interfaz para la validación de reglas de asociación, suponga que se quiere saber si los autos con valores “altos” para el atributo mpg (las millas por galón) también son “livianos”. El planteamiento de pregunta y la respuesta del sistema se muestran en la Figura 51.

Validación de Reglas de Asociación

Dentro de la colección de Autos

**Los elementos cumplen con las condiciones:**

Propiedad	No	Condición	Adverbio	Adjetivo o Vlr	Ponderación	Conectiva	
mpg	<input checked="" type="checkbox"/>	es		alto(a)			▲
	<input type="checkbox"/>						
	<input type="checkbox"/>						
	<input type="checkbox"/>						▼

¿Es cierto que todos también cumplen las condiciones siguientes?:

Propiedad	No	Condición	Adverbio	Adjetivo o Vlr	Ponderación	Conectiva	
peso	<input checked="" type="checkbox"/>	es		bajo			▲
	<input type="checkbox"/>						
	<input type="checkbox"/>						
	<input type="checkbox"/>						▼

Comprobar
Limpiar
Cerrar

**Respuesta** Muy cierto

**Elementos que cumplen el primer grupo de condiciones:**

136 de 398

37 %

**Elementos que cumplen ambos grupos de condiciones:**

108

78 %

Figura 51. Ejemplo de uso de la interfaz para validación de reglas

#### 7.1.2.4 Modelo Lógico de Datos

El modelo lógico de datos describe los datos necesarios en los metadatos para la interpretación de las preguntas vagas. Resulta de la transformación del metamodelo estructural concebido en la fase de análisis de este proyecto de investigación a un modelo lógico. Este paso, significa bajar de nivel de abstracción en el proceso ingenieril, pasar de un modelo conceptual a un modelo de diseño. Para la presentación del modelo de diseño se ha optado por el modelo relacional, significando que se especificará mediante relaciones o tablas. En ellas, se describirán sus propiedades y restricciones.

En las tablas o relaciones requeridas para la interpretación y operación con términos vagos por el sistema de consulta, el término *CP* significa una clave primaria, que declara una restricción o regla de integridad de entidad, y *CF* es una

clave foránea que determina una regla de integridad referencial. El orden de presentación de las relaciones está restringido por este último tipo de reglas, que debe tenerse en cuenta para la creación de las relaciones en una base de datos.

**Tabla 20. Relación "Tipo\_termino".**

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
Id	Identificador del término	CP	U	No	entero	Puede ser un consecutivo	8
Nombre			U	No	texto		Adverbio
Restricciones				Si	texto	Pueden chequearse mediante programas o disparadores	Ubicado después del operador IS pero antes de un adjetivo calificativo

**Tabla 21. Relación "Termino".**

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
Id	Identificador del término	CP	U	No	entero	Puede ser un consecutivo	2
Nombre	Etiqueta o lexema de un término vago		U	No	texto		Costoso
Tipo	Tipo de token o de término	CF		No		Referencias en "Tipo_terminos"	Etiqueta
Antónimos				Si	texto		Barato
Sinónimos				Si	texto		Caro

**Tabla 22. Relación "Modelo".**

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
Id	Identificador del modelo	CP	U	No	entero	Puede ser un consecutivo	2
Nombre	Nombre del modelo implementado		U	No	texto		Triangular
Descripción	Características descriptivas del modelo			Si	texto		Unimodal
Nro_Param	Número de parámetros			No	entero		3
paquete	Nombre del paquete que contiene la función			Si	texto		

**Tabla 23. Relación "Patron".**

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
Id	Identificador del patrón sintáctico	CP	U	No	numérico		6
Patron_sintáctico	Expresión regular en SQL extendido			No	texto		MUY E
Antecedente	Condiciones o premisas que cumple el patrón	CP		No	texto		Operador = IS, Negación es nulo, Adj_vlr= Etiqueta, Adverbio = Muy J= 2 y k = 3
Id_modelo	Identificador de un modelo difuso que le corresponde			No	texto	Referencias en la tabla "Modelos"	2
Param1	Parámetro 1 del modelo			No	numérico	Valor en [0, 100]	49
Param2	Parámetro 2 del modelo			No	numérico	Valor en [0, 100]	50
Param3	Parámetro 3 del modelo			No	numérico	Valor en [0, 100]	51
Param4	Parámetro 4 del modelo			Si	numérico	Valor en [0, 100]	NULL

**Tabla 24. Relación "Consulta"**

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
Id	Identificador de la consulta	CP	Si	No	numérico		121255
Proyeccion	Lista de expresiones para crear la memoria de trabajo			Si	texto		SELECT MPG, MARCA, MODELO
Proy_original	Lista de expresiones o propiedades especificadas			Si	texto		SELECT MPG as "millas por galón", modelo
Cond_concretas	Condiciones concretas y enlaces de múltiples tablas			Si	texto		WHERE modelo = 82
Agrupamiento	Texto de una cláusula GROUP BY			Si	texto		GROUP BY modelo

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
Cond_grupo	Es el texto de una cláusula HAVING			Si	texto	Es nulo si 'agrupamiento' es nulo	
Ordenamiento	Cadena con lista de atributos y su tipo de orden			Si	texto		ORDER BY mpg ASC
Cuantificador	Nombre del cuantificador usado			Si	texto	Valor en ('MOST', 'SOME', 'FEW', 'ANY', 'EXISTS', 'ALL', )	
Umbral	Umbral			Si	decimal	$0 \leq \text{umbral} \leq 1$	
Max_filas	Número máximo de filas que se quieren ver			Si	entero		

**Tabla 25. Relación "Cond\_vaga".**

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
Id_consulta	Identificador de la consulta	CP, CF	U <sub>1</sub>	No		Referencias en relación 'Consultas'	121255
Cond_nro	Número de la condición (fila)	CP	U <sub>1</sub>	No	entero	Cond_nro >=1	3
Lugar	Lugar de la condición vaga, en la orden de consulta	CP	U <sub>1</sub>	No	texto	Valor en (INICIO, WHERE)	WHERE
Conectiva	Conectiva lógica con la anterior condición simple		No	Si	texto	Valor en ('AND', 'OR', '+')	AND
Origen	Tabla o vista de donde procede la propiedad		No	No	texto		autos
Propiedad	Expresión o nombre del atributo		No	Si	texto		mpg
Negación			No	Si	texto	Valor en ('NOT', NULL)	NOT
Operador	Operador lógico de comparación		No	No	texto	Valor en ('IS', 'IS NOT', '=', '>', '>=', '<', '<=', '!=', '<>', 'LIKE', 'NOT LIKE')	IS
Adverbio	Adverbio de cantidad para acentuar un adjetivo		No	Si	texto	Valor en ('MOST', 'SOME', 'FEW', 'ANY', 'EXISTS', 'ALL', )	VERY
Adj_vlr	Etiqueta, valor numérico o literal		No	No		Referencias en tabla "Terminos"	Alto(a)

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
peso	Peso o ponderación si se trata de una combinación		No	Si	real	Valor en (0,1)	0.3
k	Número de categorías que se considerarán en la clasificación		No	Si	entero	$2 \leq \text{nro\_cat} \leq 7$	3
j	Posición de una etiqueta en el marco		No	Si	entero	$1 \leq j \leq \text{nro\_cat}$	1
Id_patron	Código del tipo o patrón de la condición.		No	No		Referencias en tabla "Patron"	4
modelo	Modelo ajustado para la condición		No	Si			Hombro_de recho(25, 29, 34)
concreta	Restricción especificada en términos concretos			Si	texto		Between 25 AND 34

**Tabla 26. Relación "Var\_ling".**

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
Id	Identificador	CP	U	No	numérico		1
Nombre	Nombre		U	No	texto		Edad
Atributo	Nombre tabla y del atributo numérico relacionado			No	texto	Referencia en la tabla "columns" de los metadatos	Emp.fecha_nac

**Tabla 27. Relación "Conj\_Terminos".**

Atributo	Descripción	Clave	¿Valor Único?	¿Valor nulo?	Tipo	Otras restricciones de columna	Ejemplos
Idtermino	Identificador del término o etiqueta	CP, CF <sub>1</sub>	U <sub>1</sub>	No		Referencias en "Terminos"	2
Idvariable	Identificador de la variable lingüística	CP, CF <sub>2</sub>	U <sub>1</sub>	No		Referencias en "Var_ling"	1
Posicion	Posición de la etiqueta en un marco de cognición				Entero pequeño		3
Idagregado	Identificador del otro término vago al cual pertenece como componente	CF		Si		No puede ser igual a Idtermino	1
Peso	Ponderación o peso, si se hace parte de una combinación lineal			Si	Valor en (0, 1)		0.8

### 7.1.2.5 Modelo Lógico Funcional

Este modelo describe las tareas o funciones involucradas en la interpretación de la vaguedad de las consultas, según el contexto, sus responsables y los objetos requeridos.

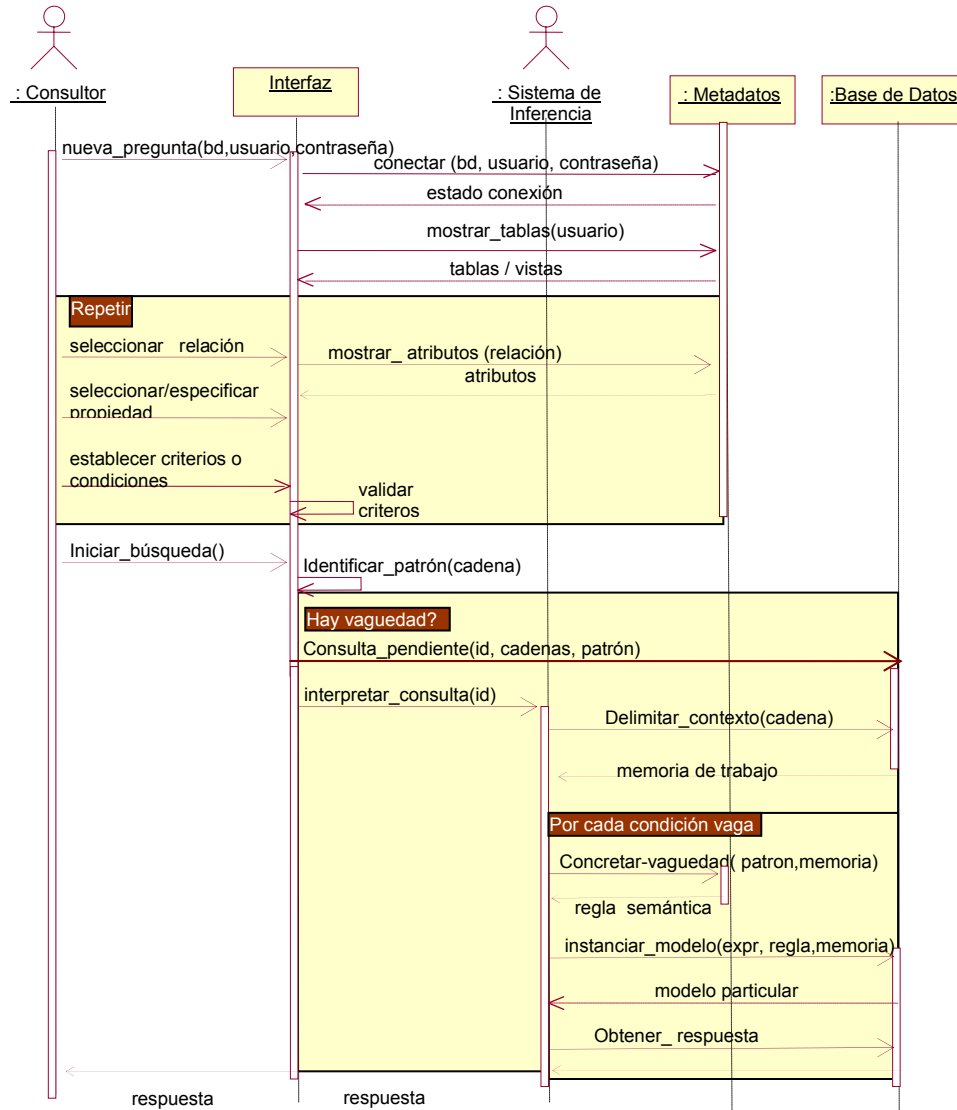


Figura 52. Diagrama de secuencias para la resolución de preguntas vagas

La representación de los aspectos funcionales del sistema se presenta de una manera esquemática en la Figura 52. Este diagrama de secuencias, bajo la notación del lenguaje UML (Unified Modeling Language), describe la colaboración entre los objetos de un sistema flexible de consulta-respuesta, basado en el modelo de razonamiento propuesto.

La funcionalidad descrita en el diagrama es general, pues cobija todos los tipos de pregunta que el usuario puede formular en el sistema de consulta-respuesta usando el modelo de razonamiento propuesto.

En el diagrama recién presentado se observa que antes de empezar a interactuar con el sistema, se requiere abrir una conexión con la base de datos. Para ello, se necesitaría una interfaz previa donde se le pregunte al interesado a cuál base de datos se quiere referir en sus consultas, cuál es su nombre de usuario y su contraseña. Si la base de datos es pública este paso puede ser transparente para el usuario.

Para la definición de las funciones requeridas en la interpretación de la vaguedad de las consultas, se hizo uso de las características dinámicas del lenguaje no procedimental inmerso en los sistemas gestores de bases de datos objeto-relacionales actuales y en los lenguajes de programación Web. Dichas características permitieron que no se tuvieran que conocer los nombres de las tablas o de los atributos contenidos en una base de datos y referidos en una consulta. También posibilitan, en entre otras cosas, manipular los datos guardados de manera temporal en las memorias de trabajo, cuya estructura es completamente desconocida en el momento de la programación.

#### a) Módulos principales

Nombre	Responsable	Funcionalidad
Nueva_preg_tipo1	Interfaz	Capturar los datos de la consulta, proporcionar listas de los objetos y sus propiedades para armar el texto de la consulta e identificar el patrón de la misma. Si se trata de una consulta vaga, asentar los datos en la base de datos e invocar al agente inteligente para la interpretación de la consulta. En caso contrario, envía la cadena SQL para su ejecución por el servidor de datos. Por último presenta los resultados en forma tabular.
Nueva_preg_tipo2	Interfaz	Capturar los datos de la consulta cuyo resultado esperado es un valor de verdad y proporcionar listas de valores para los objetos referidos en la interfaz. Guardar los datos temporalmente, Presentar los resultados.
Interpretar_consulta	Agente	Proceso general de la interpretación de una consulta, de acuerdo con el patrón identificado y considerando el contexto lingüístico que enmarca la consulta.
Delimitar_contexto	Agente	Con la cadena SQL que recibe como argumento, la ejecuta para crear la memoria de trabajo de la consulta vaga. Si existe una igual, para la resolución de otra consulta, sólo la identifica para reutilizarla.
Concretar_vaguedad	Agente	De acuerdo con los patrones sintácticos identificados en la consulta, elegir los modelos teóricos difusos para ajustarlos al contexto.
Obtener_respuesta	Agente	Calcular el grado de cumplimiento de las tuplas a las condiciones especificadas y eliminar las tuplas, según los calibradores especificados, cuando se trate de preguntas de tipo 1. Luego de esto, enviar los resultados al servidor Web para su presentación al usuario.

## b) Módulos auxiliares

Nombre	Invocado por	Funcionalidad
Identificar_patrón	Nueva_preg_tipo1	Identificación del patrón de cada pregunta o condición vaga, entre los posibles.
Consulta_pendiente	Nueva_preg_tipo1 Nueva_preg_tipo2	Asiento de una consulta vaga pendiente. Incluye la inserción de las componentes de la consulta y sus condiciones, en las tablas temporales correspondientes.
simple_fdp	fdp	Proporciona el grado de cumplimiento de una tupla a restricción establecida mediante el uso de una etiqueta lingüística, impuesta sobre un atributo o una expresión derivable de éste.
muy_fdp	fdp	Proporciona el grado de cumplimiento de una tupla a restricción establecida mediante el uso de una etiqueta lingüística, acentuada con el adverbio “muy”, impuesta sobre una propiedad.
extrem_fdp	fdp	Proporciona el grado de cumplimiento de una tupla a una clase con una etiqueta lingüística, acentuada con “extremadamente”.
cuant_like	fdp	Proporciona el grado de cumplimiento de una tupla a restricción establecida mediante el uso del cuantificador absoluto LIKE.
cuant_minoria	fdp	Proporciona el grado de cumplimiento de una tupla a restricción establecida mediante el uso del cuantificador relativo FEW.
cuant_mayoria	fdp	Proporciona el grado de cumplimiento de una tupla a restricción establecida mediante el uso del cuantificador relativo MOST.
hombro_izquierdo	simple_fdp, muy_fdp, extrem_fdp, cuant_minoria	Función de pertenencia del conjunto difuso con parámetros a, b, c.
hombro_derecho	simple_fdp, muy_fdp, extrem_fdp, cuant_mayoria	Función de conjunto difuso con parámetros b, c, d.
trapezoidal	simple_fdp, muy_fdp,	Función de pertenencia del conjunto difuso con parámetros a, b, c, d.
triangular	extrem_fdp, cuant_like	Función de conjunto difuso con parámetros a, b, c, d.
gc	obtener_respuesta	Determina el grado de cumplimiento de una tupla, a una condición particular o a un conjunto de condiciones.
percentil	Concretar_vaguedad	Retorna el valor particular de una variable o atributo que deja a su izquierda, un porcentaje $q$ de los datos ordenados.

## 7.2 Confirmación teórica del modelo propuesto

Con el objeto de validar el cumplimiento de las propiedades o restricciones impuestas al modelo de razonamiento aproximado para la admisión de términos vagos en las consultas, se detallará cómo se cumplió cada una de ellas, agrupándolas de la manera cómo se especificaron las restricciones.

### 7.2.1 *Confiabilidad*

La vaguedad aquí tratada es dependiente del contexto lingüístico que enmarca una consulta y por eso, puede ser variable en el tiempo o en el espacio. Puesto que el modelo de razonamiento incluye un proceso de minería de los datos, que se ejecuta automáticamente y en línea, para hallar los modelos de los conjuntos difusos que representan cada término vago incluido en una consulta, el modelo propuesto se adapta dinámicamente al contexto lingüístico delimitado en la misma, preservando la validez de las inferencias. Esta validez no se podría garantizar si los modelos fueran especificados por los humanos. Por lo tanto, el modelo de razonamiento aquí propuesto supera las propuestas previas que proponen flexibilizar el lenguaje de consulta, basándose en la Lógica Difusa (véase, por ejemplo, a Kackrpyk y Zadrozny, 2001; Galindo, Urrutia y Piattini, 2005; Bosc, Kraft y Petry, 2005; Ma y Wang, 2006; Gonçalves y Tineo, 2007).

Puesto que los modelos de los conjuntos difusos definidos en la presente Tesis Doctoral para representar los términos vagos de la consulta, no dependen de los juicios de expertos, el motor de inferencia, siempre producirá la misma respuesta cuando tenga la misma información de un contexto determinado. Por estas razones, la validez de las inferencias y la consistencia en sus respuestas, el modelo de razonamiento propuesto en esta Tesis Doctoral se considera confiable.

### 7.2.2 *Extensibilidad*

El modelo propuesto extiende modelos previos, pues incorpora un nuevo tipo de agregación: la combinación lineal de condiciones vagas o concretas, donde a cada una de ellas se le puede dar un peso o ponderación diferente en la explicación del término vago derivado. Este caso no es contemplado en el lenguaje teórico PRUF de Zadeh, ni en las propuestas estudiadas para la construcción de sistemas flexibles de consulta-respuesta. Tampoco en ellas se contempla la posibilidad de responder con valores lingüísticos de la verdad como “muy cierto” o “absolutamente cierto”. Por tanto, la presente propuesta extiende todas las propuestas estudiadas para la extensión del lenguaje estándar de consulta a bases de datos relacionales y objeto-relacionales, el SQL y al lenguaje PRUF.

Además, se logró que las extensiones propuestas fueran estrictamente aditivas puesto que el modelo que se toma como el original, el lenguaje de consulta SQL a bases de datos, es un caso restringido o particular del lenguaje propuesto. Esto se aprecia claramente en las reglas de reescritura o de producción presentadas en el capítulo anterior. Allí se resaltaron las reglas o palabras nuevas del lenguaje. Si éstas fueran eliminadas, las reglas de reescritura corresponderían a las presentadas en el Capítulo 6, definidas para el SQL estándar.

Pensando en la extensibilidad futura del modelo de razonamiento se ha optado por definir los términos vagos, con sus reglas semánticas y sintácticas, como nuevas relaciones del catálogo o metadatos de la base de datos, obteniendo independencia lógica entre los datos y los programas. Si en cualquier momento de deseara agregar un nuevo término o cambiar el modelo difuso que representa a un patrón sintáctico

determinado, sólo habría que agregar una nueva tupla o hacer una modificación en alguna existente, con órdenes sencillas del SQL o a través de una interfaz concebida para estos propósitos. Esto simplifica sustancialmente los cambios o las adiciones futuras que se quieran realizar al modelo de razonamiento propuesto, por no tener que alterar los programas y tener que recompilarlos.

### **7.2.3 Generalidad**

El modelo conceptual aquí concebido es independiente del dominio de aplicación pues no se circunscribe a uno sólo, en particular. Como ejemplo se usó una base de datos sobre autos y otra del Censo de los Estados Unidos de 1994. Ello muestra que el sistema es factible para cualquier dominio con variables o atributos de tipo cuantitativo que puedan ser mapeadas a términos lingüísticos. También es independiente del sistema gestor de bases de datos porque como modelo conceptual no está atado a un sistema gestor de bases de datos determinado. Un sistema interactivo de consulta-respuesta basado en el modelo propuesto puede ser construido usando a Oracle, SQL Server o PostgreSQL, entre aquellos que tienen como soporte lógico el modelo objeto-relacional. Este tipo de sistema gestor de bases de datos permite definir y guardar en la bases de datos las funciones o procedimientos requeridos por el sistema de inferencia, mediante un lenguaje procedimental como el PL/SQL de Oracle Corporation o el PgPLSQL de PostgreSQL. Significa que la base de datos, en dichos gestores, es realmente una base de conocimientos, pues no sólo se pueden almacenar los datos o los hechos, sino las reglas del sistema de inferencia.

No obstante lo anterior, si un gestor de bases de datos no permite el almacenamiento de reglas necesarias para la interpretación de la vaguedad, el modelo de razonamiento propuesto también podría implementarse definiendo la lógica del sistema de inferencia en el lenguaje de programación de las interfaces para la interacción humano-máquina. Aunque esto traería una pérdida de eficiencia del sistema debida el tráfico de, posiblemente, grandes cantidades de datos desde la base de datos hasta el servidor Web..

Por otro lado, un sistema interactivo y flexible de consulta basado en el modelo propuesto puede ser usado por cualquier tipo de usuario, sin requerir que sea experto en lógica difusa o en el dominio de aplicación.

De todo lo expuesto, se puede afirmar que el modelo propuesto es aún más general de lo esperado.

### **7.2.4 Corrección**

La corrección del modelo de razonamiento se deduce de su corrección interna y externa. La corrección interna se satisface dependiendo de cómo haya sido concebido el sistema de inferencia difusa requerido en la interpretación de vaguedad. Por eso, este tipo de corrección depende de la lógica que le da soporte al proceso de razonamiento aproximado y de las técnicas para la representación

automática de los términos y operadores vagos que admite el lenguaje SQL extendido.

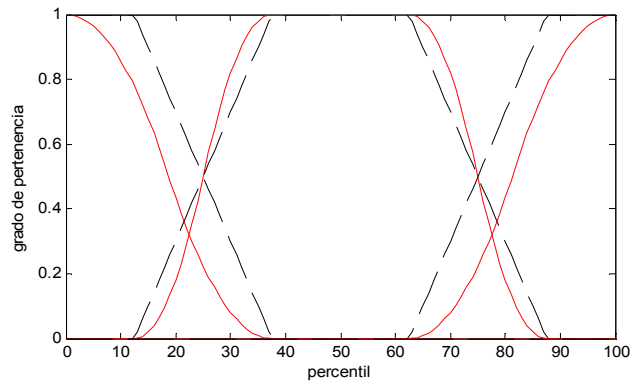
Para velar por la corrección interna deseada en el nuevo modelo de razonamiento, se definieron una serie de restricciones, en el Capítulo 4, que el sistema de inferencia debía cumplir para que emulara el modo de proceder de un experto humano, usando lógica de sentido común en el proceso de concreción de la vaguedad y que fuera coherente con alguna teoría que no se basara en la lógica bivaluada tradicional para permitir grados de pertenencia diferentes al cero o al uno, a los conjuntos rotulados con alguna etiqueta lingüística y también permitir algún solapamiento entre las clases.

En este proyecto se descartó el uso de la Teoría de los Conjuntos Rugosos (Pawlack, 1982) para la representación de los términos vagos porque la forma de estos conjuntos no es continua y se incumple la restricción de cobertura (Restricción No.12, sección 4.5.5). Dicho incumplimiento genera el grave inconveniente no poder clasificar a los elementos u objetos que se encuentren en las zonas de indecisión. Por esto, se eligieron funciones de pertenencia continuas basadas en una Teoría de Conjuntos Difusos, para los conjuntos que representarán las etiquetas lingüísticas y sus agregaciones y así, garantizar el cumplimiento de las restricciones de cobertura y convexidad. Además, de un conjunto difuso que represente una etiqueta lingüística se puede derivar el conjunto rugoso correspondiente, asignando el valor de indiscernible al grado de pertenencia de los elementos que no forman parte del núcleo, pero sí del soporte de la función. Esta característica permite realizar otras operaciones algebraicas sobre conjuntos basados en una teoría difusa, que se requieran en el futuro con otros propósitos, usando las relaciones de equivalencia definidas para los conjuntos rugosos.

Cuando se presentaron las formas lineales elegidas para la representación de las etiquetas lingüísticas correspondientes a adjetivos calificativos, se realizó una comparación de éstas con formas de conjuntos difusos no lineales como la función  $S$ , la función  $Z$  y la función  $\Pi$ . En la comparación se consideró el caso extremo donde las distribuciones no lineales fueran unimodales para ver si el efecto de la función de potencia para acentuar o relajar una categoría difusa con el modificador “muy” o “extremadamente”, era mayor que el producido en una función trapezoidal o trapezoidal truncada. Sin embargo, la aplicación de la función de potencia sobre los modelos no lineales se observó que tampoco produce el efecto esperado para acentuar o relajar el significado de una etiqueta lingüística porque el tamaño los soportes de los nuevos conjuntos difusos son idénticos a los originales y los elementos que pertenecían a la clase básica conservaban casi el mismo grado de pertenencia a la clase derivada. Por esto, se optó por otra manera de representar los modificadores “muy” y “extremadamente”, a partir de particiones sucesivas del conjunto que debe ser modificado, más parecido a lo que haría un humano ante un caso de estos.

De acuerdo con el nuevo procedimiento planteado para representar los modificadores, las formas no lineales vuelven a ser otra alternativa para la representación de las clases y por eso se hace necesaria su evaluación.

En primer lugar, si se eligieran formas no lineales unimodales, entonces sólo un valor en el dominio sería considerado como el prototipo de la clase (aquel cuyo grado de pertenencia es 1) y esto sólo se ve apropiado para la representación de un número difuso. Por esto es más recomendable formas no lineales donde el núcleo admita más de un valor, más parecidas a las trapezoidales. En la Figura 53 se muestran las dos formas descritas, donde se puede observar que los conjuntos no lineales y los lineales, representados con líneas punteadas, no difieren significativamente si se consideran los mismos percentiles usados como parámetros en las formas lineales.



**Figura 53. Formas lineales versus no lineales**

En la figura anterior se observa que los modelos no lineales para los conjuntos difusos no permiten asegurar que el punto de indecisión sea 0.5 puesto que las funciones  $Z$  y  $S$  sólo reciben como argumentos dos parámetros, que son el valor mínimo y máximo del conjunto difuso, con los cuales se calcula su promedio para definir el tercer parámetro que es el punto de inflexión. Además, la partición difusa con los modelos no lineales considerados tampoco aseguraría que la suma de los grados de pertenencia de cualquier elemento, a todos los conjuntos difusos en el marco de cognición, sea 1 y esta restricción se fijó para garantizar que los elementos pertenezcan al conjunto universal con ese grado, que es lo natural. Como prueba de ese incumplimiento, se calcularon los grados de pertenencia o compatibilidad del percentil 20% a los conjuntos difusos que conforman un marco de cognición:

$$\mu^{\text{"valores menores"}}(P_{20}) = \text{gc}(20, Z(0, 37.5)) = 0.4356$$

$$\mu^{\text{"valores medianos"}}(P_{20}) = \text{gc}(20, \Pi(12.5, 37.5, 62.5, 87.5)) = 0.18$$

$$\mu^{\text{"valores altos"}}(P_{20}) = \text{gc}(20, S(62.5, 100)) = 0$$

$$\mu^{\text{"valores menores"}}(P_{20}) + \mu^{\text{"valores medianos"}}(P_{20}) + \mu^{\text{"valores altos"}}(P_{20}) = 0.6156$$

Por lo anterior, se haría necesario definir funciones no lineales propias para poder cumplir con algunas de las restricciones impuestas al marco de cognición en la discriminación difusa, como se hizo en este trabajo cuando se hizo la comparación. En ese caso, las formas serían aún más parecidas que las presentadas en la Figura 53 y por esta similitud no se justificaría emplear modelos con curvas suavizadas como representación de la forma de un conjunto difuso, puesto que en el momento de la

evaluación de las tuplas para determinar si pertenecen a los resultados, se deben hacer más preguntas para saber cuál de los cálculos o fórmulas se debe aplicar para determinar el grado de pertenencia y cada fórmula es más compleja que la correspondiente a los modelos no lineales. Esta complejidad podría degradar el desempeño del sistema cuando el número de elementos considerados sea grande, sabiendo que la ganancia es poca o ninguna.

De lo expuesto, se puede concluir que las funciones de pertenencia con formas lineales para la representación de conjuntos borrosos son las apropiadas porque se ajustan a las restricciones impuestas al modelo de razonamiento aproximado, que además poseen la ventaja de su fácil interpretación y simplicidad en los cálculos. Posiblemente por estas características, también han sido las elegidas para representar los términos vagos en todas las propuestas estudiadas de extensión del lenguaje SQL para admitir vaguedad en las consultas.

Una característica relevante sobre la función o modelo de los conjuntos difusos que representan términos lingüísticos, es su convexidad (Restricción Nro. 16, sección 4.5.9). El modelo de un conjunto difuso que no posea esta característica no tiene sentido lógico y carece de interpretación. Por lo tanto, para asegurar la convexidad de los conjuntos difusos, se optó por una técnica de discriminación “arriba-abajo” (top-down) bajo el supuesto de formas trapezoidales (o trapezoidales truncadas) para los conjuntos difusos que representan etiquetas lingüísticas y por formas triangulares para representar escalares o cantidades concretas, acogiendo la misma estrategia empleada en la Estadística de trabajar con supuestos sobre formas de las reglas discriminantes y siguiendo la estrategia sugerida en (Herrera y Herrera-Viedma, 1997) para encontrar los operadores LOWA que representan las agregaciones. Las funciones de pertenencia elegidas en el modelo de razonamiento propuesto son funciones que permiten cumplir la restricción de normalidad de los conjuntos difusos (Restricción No. 15, sección 4.5.8), donde se esperan que los prototipos de las clases etiquetadas sean varios y no uno solamente, en los casos diferentes a la representación de números difusos.

Para lograr que los modelos de los términos vagos además de interpretables fueran coherentes, el marco de cognición debía estar bien estructurado, con un orden apropiado entre las clases, entre otras características. Se puede observar en las reglas semánticas definidas para determinar los conjuntos difusos en un marco de cognición y en los ejemplos presentados que se preservó el concepto de orden (Restricción Nro. 8, sección 4.5.1). En nuestro modelo, un conjunto borroso con etiqueta “baja” siempre estará situado, en el marco de cognición, primero que el que representa la clase de los “medianos” y éste antes que el conjunto con la etiqueta “alta”, cuando se consideran tres conjuntos en el marco. Además, se cumplió la restricción de considerar un número reducido de conjuntos difusos en un marco de cognición (Restricción No.9, sección 4.5.2), aunque esta característica está determinada por el conjunto de términos de una variable lingüística.

Con miras a lograr una buena especificidad de la técnica de discriminación difusa, se impuso la restricción de considerar que los conjuntos difusos, aunque solapados, fueran distinguibles (Restricción No.10, sección 4.5.3). Por esto, los

modelos de los conjuntos borrosos que representan las etiquetas lingüísticas se diseñaron de tal forma que un elemento cualquiera del dominio pertenece, como máximo, a dos conjuntos adyacentes en el marco (Restricción No.11, sección 4.5.4). El diseño de los conjuntos difusos en un marco de cognición también se ajustó a la restricción de complementariedad (Restricción No.13, sección 4.5.6) y a la restricción de aceptación completa (Restricción No.14, sección 4.5.7). Esto significa que la suma de los grados de pertenencia de un elemento a esos dos conjuntos siempre será uno y que a medida que éste deja de ser compatible con una clase etiquetada, aumenta su compatibilidad con la clase adyacente, en el marco de cognición. Por estas dos razones, en nuestra propuesta, sólo habrá un valor en el dominio que tendrá como grado de pertenencia de 0.5 a dos conjuntos adyacentes, convirtiéndose en el único punto de indecisión de la técnica de discriminación.

Como fue necesario implementar un conjuntos funciones que no vienen predefinidas en un sistema gestor de bases de datos, se tuvo la precaución de comparar los resultados obtenidos, con los arrojados por herramientas ampliamente reconocidas en el ámbito mundial. Por ejemplo, las funciones de pertenencia difusa, se contrastaron con las que produce el paquete Matlab versión R2006a de la compañía MathWorks y la función para el cálculo de los percentiles se comparó con la definida en Excel de Microsoft Corporation.

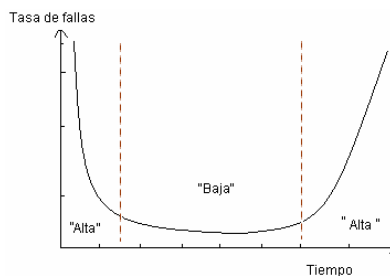
Por otro lado, para velar por la corrección externa que es determinada por la estructura gramatical de nuevas expresiones vagas admisibles en el lenguaje SQL extendido, se fijó que la posición de los términos vagos estuviera acorde con la estructura gramatical definida para el inglés, idioma al que trata de aproximarse el lenguaje SQL. También se puso especial cuidado para que los nuevos símbolos o palabras del lenguaje extendido no forzaran al usuario consultor a emplear los tecnicismos de la Lógica Difusa considerando que el tipo de usuario que va a realizar las consultas a la base de datos, es un usuario cualquiera. Es decir, con las extensiones propuestas el modelo de razonamiento aproximado será “caja-negra” para el usuario final, contribuyendo así con la corrección externa del modelo de razonamiento.

#### ***7.2.4.1 Comportamiento del método de razonamiento con distribuciones atípicas***

El método de razonamiento propuesto para hallar la semántica de una clase rotulada con alguna etiqueta lingüística, se basa en la Lógica Difusa que supone que las categorías se traslapan pues no hay bordes claramente definidos para discriminar a los objetos; donde además se considera que el paso de una categoría a otra es gradual y no abrupto. Sin embargo, pueden existir situaciones donde estos supuestos no sean válidos y el sistema de inferencia debe ser capaz de detectarlas para lograr mayor corrección del método de razonamiento.

En los ejemplos presentados de la representación de una colección de datos por medio de modelos no paramétricos, se hizo evidente que la forma de las distribuciones puede ser muy variada: unas distribuciones presentan formas

simétricas y otras no, en unas de ellas los datos están muy concentrados y en otras hay gran dispersión. Como los términos vagos aquí tratados son dependientes del contexto lingüístico que enmarca una consulta, tampoco sería raro que los datos estuvieran a un extremo del dominio de la variable considerada y en otros casos, estuvieran en el opuesto. Pero como la técnica empleada para hallar los parámetros de los conjuntos difusos que representan adjetivos calificativos vagos, está basada en percentiles que son medidas relativas, el sistema haría la partición difusa, en los  $k$  conjuntos difusos especificados, sin importar si los datos se encuentran a un extremo o al otro. Sin embargo, si en la distribución de los datos se identifican dos grupos de datos ubicados a los extremos, cuando se consideren tres clases o categorías en el marco de cognición, la técnica debe variar su comportamiento pues está ante un caso de no vaguedad.

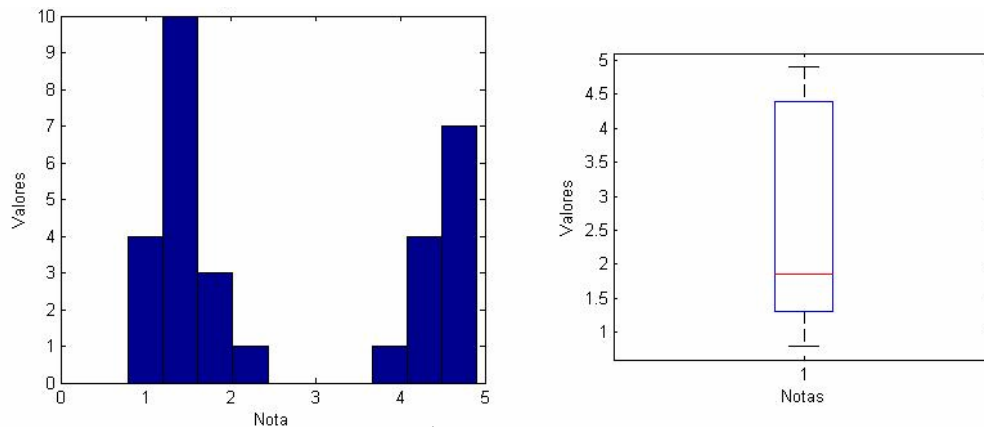


**Figura 54. Distribución de la tasa de fallas de un aparato electrónico**

Un tipo de distribución donde la mayoría de datos están a los extremos tiene una forma conocida como la curva de la bañera, en Control Estadístico de la Calidad (Oakland, 2000). En este ámbito, esta forma de distribución representa la tasa de fallas de los aparatos electrónicos durante su vida útil, así como representa la tasa de mortalidad de los humanos. Como se puede apreciar en la Figura 54, la curva de la bañera por ser cóncava, no tiene interpretación en un marco de cognición puesto que en éste existe el concepto de orden. Los aparatos con riesgo “alto” de fallas se encuentran en los extremos, mientras que los aparatos de “bajo” riesgo se encuentran en la fase intermedia de su vida útil.

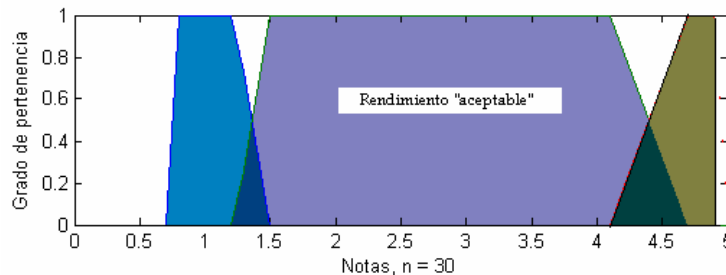
Debe observarse que en la curva de la bañera presentada, la tasa de fallas no llega al valor de cero en ningún punto de la distribución. Por eso, si se quisiera agrupar en tres clases etiquetadas como “equipos con fallas tempranas”, “con fallas intermedias” y “con fallas al final de la vida útil”, el sistema de inferencia se comportaría apropiadamente porque ninguna de ellas es vacía. Pero si en la distribución todos los valores están en los extremos, lo más conveniente fuera que el sistema de inferencia determinara que la clase central no existe, cuando se le solicite representarla.

Para ilustrar el caso de una distribución con los valores de un atributo sólo en los extremos, se quiso emplear datos reales pero en los examinados esta situación atípica no se encontró. Debido a esto, se usó un conjunto de notas hipotéticas (1.2, 1.6, 1.3, 1.5, 1.3, 1.8, 2.1, 1.9, 4.3, 4.2, 1.5, 4.6, 4.7, 4.8, 4.9, 1.3, 4.8, 1.6, 2, 0.8, 4.4, 4.6, 1.4, 1.3, 1, 1, 4.5, 1.5, 4 y 4.2), obtenidas por un grupo de estudiantes en un curso X.



**Figura 55. Distribución hipotética de las notas de un curso**

En el histograma de la Figura 55, claramente se aprecia el distanciamiento entre los dos subconjuntos que se forman en la distribución, aunque en el Diagrama de Cajas y Bigotes esta separación no es evidente. Sin embargo, la caja del DCB cubre buena parte del rango total de los datos, pues debe cobijar a la mitad de las notas con los valores centrales. Se observa que la longitud de la caja cubre más del 50% que es lo esperado para una distribución uniforme, como la correspondiente a la distribución de los autos, según su modelo o año de fabricación (véase la Figura 8). Consecuentemente, el sistema de inferencia propuesto para la discriminación difusa, considerando tres clases en el marco de cognición, generará una clase para los valores medianos con una alta variabilidad.



**Figura 56. Discriminación difusa de las notas hipotéticas**

En la Figura 56 se muestran los modelos ajustados a las notas hipotéticas cuando se discrimina en tres clases. Considerando un alfa corte estricto igual a cero, se deduce que los estudiantes con notas "aceptables" son aquellos que obtuvieron una nota definitiva superior a 1.2 pero inferior a 4.7.

Puesto que las notas tienen un referente absoluto, que puede fluctuar entre 0 y 5, es fácil detectar la alta variabilidad en la clase intermedia; pero en la generalidad de los casos no se puede juzgar cuán variables son los datos dentro de una clase únicamente con la amplitud del rango que la determina, sino que es necesario considerar el rango total de la variable en cuestión.

Por lo anterior, con fines comparativos se construyó la Tabla 28. Allí se puede apreciar que la proporción de la longitud del soporte para la clase de las notas

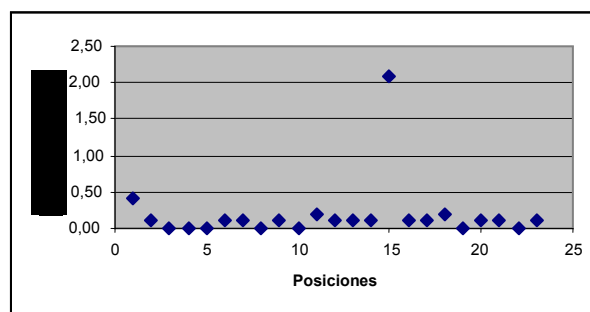
“aceptables” sobre el rango total es del 84%, mientras que para una distribución uniforme se esperaría que esa proporción fuera cercana al 75% (la diferencia entre 87.5 y 12.5), como ocurre en la distribución de los autos según su año de fabricación. También se puede observar que esa proporción es cada vez menor cuando la distribución de los datos tiende a la distribución normal como ocurre con los autos considerando su desplazamiento.

**Tabla 28. Estadísticas de resumen, considerando diferentes variables**

Estadísticas	Rendimiento (mpg)	Potencia (HP)	Modelo (año)	Desplazamiento (metros/segundo)	Notas
Tamaño de la muestra	398	394	398	398	24
Valor Mínimo	9	46	70	8	0.8
Valor Máximo	46.6	230	82	28.4	4.9
Rango total	37.6	184	12	16.8	4.1
Límite inferior del soporte de la clase intermedia ( $P_{12.5}$ )	14	67	72	12.5	1.23
Límite superior del soporte de la clase Intermedia ( $P_{87.5}$ )	33.5	150	81	18.95	4.68
Longitud soporte de la clase Intermedia	19.5	83	9	6.45	3.45
Longitud del soporte/Rango total	52%	45%	75%	38%	84%

De acuerdo con lo anterior, se puede establecer que un valor mayor del 80% para la proporción entre la longitud del soporte de la clase intermedia y el rango total de todos los valores, es un indicativo de una alta variabilidad los valores dentro de esta clase. Sin embargo, esta alta variabilidad intragrupal no implica necesariamente la existencia de únicamente dos clases polarizadas cuando se realiza una discriminación en tres clases, y por eso se debe proseguir buscando la existencia de un hueco o cambio abrupto en la distribución de los datos.

Para la detección de un hueco en la distribución se pueden calcular las diferencias entre las parejas de valores consecutivos de los datos ordenados y determinar si la máxima diferencia se puede considerar extrema o atípica.



**Figura 57. Diferencias entre dos notas consecutivas**

En una gráfica como la que aparece en la Figura 57, la detección de una diferencia extrema o atípica se hace fácilmente, pero si se quiere que el sistema de inferencia detecte las diferencias extremas, por medio de un procedimiento o una función se necesita alguna técnica matemática para hallarlos. Comúnmente se usa la fórmula de Tukey que toma como referencia el rango intercuartil y considera un valor como “extremo” aquel que se encuentre por fuera de 1,5 veces esa distancia desde la mediana, bien sea que se encuentre por encima o por debajo de este valor (Trosset, 2001).

En el ejemplo de las notas, la máxima diferencia entre dos consecutivas es 2.1, el rango intercuartil es 0.1 y la mediana es 0.1. Por lo tanto, la distancia límite que se debe considerar es  $0.1 * 1.5 = 0.15$  y un valor por fuera de  $0.1 \pm 0.15$  se catalogaría como extremo, como realmente ocurre en el ejemplo. Por consiguiente, se puede inferir que hay un cambio abrupto o un gran distanciamiento entre dos subgrupos de valores del conjunto total de datos.

De lo expuesto, cuando una condición de una consulta se refiera a la clase intermedia, el sistema de inferencia primero determinará si existe una alta variabilidad intragrupal por medio de la proporción entre el soporte de la clase y el rango total. Si esto ocurre, se prosigue con el chequeo de la máxima diferencia entre dos valores consecutivos para determinar si se puede considerar atípica. Si es así, el sistema de inferencia deberá asignar el valor de cero al grado de pertenencia a todos los elementos que formen parte de la memoria de trabajo, a dicha clase.

Como se describió, el chequeo de la existencia de una alta variabilidad de la clase intermedia es muy sencillo pues sólo requiere del valor mínimo y el máximo como datos adicionales para calcular la proporción de la longitud del soporte de la clase con respecto al rango total. Además, estos dos datos son fácilmente asequibles cuando los datos están ordenados. Por lo tanto, la estrategia para minimizar el tiempo de procesamiento, en el caso de representación de una etiqueta lingüística para la clase intermedia, consiste en calcularlos dentro procedimiento llamado “concretar\_vaguedad” que devuelve todos los parámetros del conjunto difuso de acuerdo con el patrón identificado y basándose en los datos disponibles. Como el modelo de la clase central o intermedia es un conjunto trapezoidal que tiene 4 parámetros, el procedimiento ahora deberá devolver seis.

Aunque se espera que en pocos casos sea necesario chequear si existe una clara separación entre dos grupos de datos, por medio de la máxima diferencia hallada entre dos valores consecutivos, el procedimiento descrito para detectar un distanciamiento significativo en los datos, debe crearse como una función almacenada dentro de la base de datos para ser invocada en los casos requeridos. De esta manera, se logra que el comportamiento del sistema sea el apropiado, aún en los casos atípicos donde los datos se encuentren a los dos extremos de la distribución.

### **7.2.5 Riqueza Semántica o Potencia Expresiva**

Tomando como referencia al lenguaje teórico PRUF, que ha servido de base teórica para las propuestas que permiten vaguedad en las consultas a bases de datos,

el modelo propuesto en la presente Tesis Doctoral, se puede considerar rico semánticamente puesto que no sólo cubre las reglas de interpretación de las cuatro categorías enunciadas en el lenguaje PRUF, sino que permite especificar términos vagos expresados como una combinación lineal de condiciones simples, vagas o concretas.

La presente propuesta, con respecto a otras extensiones del lenguaje de consulta a bases de datos como SQLf o FSQl, también tiene mayor capacidad expresiva al permitir que el sistema de inferencia incluya en sus respuestas calificadores lingüísticos de la verdad, correspondientes a las preguntas de tipo IV en el lenguaje PRUF descrito en la sección 3.1.4 de este libro .

Los calificadores lingüísticos junto con la representación de cuantificadores existenciales relativos vagos, ofrecen la posibilidad de realizar tareas de Minería de Datos a los usuarios no expertos, al permitirles conocer el valor de verdad de posibles relaciones entre variables o propiedades de los objetos de una manera sencilla, junto con los indicadores de la fortaleza de la asociación.

### **7.2.6 Formalidad o Rigor**

Se puede afirmar que el modelo del lenguaje extendido propuesto es formal porque fue posible especificar las reglas sintácticas y semánticas para determinar el significado de cada término vago simple o agregado contemplado en el presente trabajo de investigación, de una manera precisa y con una interpretación única. Esta especificación formal, se realizó en el Capítulo 7, por medio de reglas de producción que deben ser consideradas durante el análisis sintáctico del texto de una consulta y por medio de fórmulas algebraicas para la especificación de las reglas semánticas que deben aplicarse según el patrón identificado.

La especificación formal de los términos vagos se validó cuando se efectuaron las consultas de ejemplo, en el análisis de los tipos de preguntas contemplados en el Capítulo 5, y en la transformación del modelo conceptual a un modelo de diseño, cuando se realizó el estudio de factibilidad técnica de sistemas basados en el modelo de razonamiento propuesto que se describe un poco más adelante.

### **7.2.7 Robustez**

Se puede afirmar que el modelo de razonamiento propuesto es robusto porque las respuestas del sistema de inferencia no se van a ver afectadas por ligeras variaciones en los datos. La técnica de discriminación borrosa empleada para hallar los modelos de las etiquetas o términos lingüísticos, se basa en medidas de posición (los percentiles) para determinar los valores de los parámetros de las funciones de pertenencia. Estas medidas no sólo son robustas a los valores extremos, sino que permiten representar cualquier tipo de distribución de los datos. Además, si se realiza un ligero cambio en un valor en el dominio que corresponde a uno de los percentiles que deben calcularse, la forma del conjunto difuso que representa un término vago puede variar ligeramente o incluso no cambiar nada, si el cambio en el valor no afecta su posición, dentro del conjunto de datos del contexto.

### 7.2.8 Consideraciones sobre la eficiencia de un sistema basado en el modelo propuesto

Una característica fundamental de cualquier sistema informático es conocida en la Ingeniería de Software, como su *usabilidad*. Esta propiedad es definida como: “la bondad de utilizar un producto por los usuarios especificados para alcanzar unos objetivos determinados con efectividad, satisfacción y eficiencia, en un dominio específico” (Snyder, 2003). Por lo tanto, no hay duda que además de velar por el cumplimiento de características esenciales que se especificaron como restricciones para los sistemas flexibles de consulta que se basen en el modelo de razonamiento propuesto, es necesario velar por la oportunidad de las respuestas.

Por lo anterior, se hizo necesario identificar los distintos mecanismos que pueden implementarse para ahorrarle esfuerzos a la máquina de inferencia, procurando en todo caso, no afectar la validez de las respuestas y por ende, la confiabilidad del sistema.

Un sistema interactivo y flexible de consulta a bases de datos, al cual se le proporcionen los modelos o conjuntos difusos que representan una etiqueta vaga, evitaría buena parte del trabajo requerido para determinar cuáles objetos o tuplas cumplen con las restricciones impuestas en una consulta vaga, pero se sacrificaría la confiabilidad del sistema porque dichos modelos pueden no ser válidos en el contexto o entorno que delimite la consulta.

Como la característica distintiva de la presente propuesta es su adaptabilidad tanto para considerar los múltiples contextos lingüísticos, como de niveles de granularidad y de ahí el título del trabajo de investigación, el centro de atención en esta investigación para velar por un buen desempeño del sistema, es el trabajo requerido en la obtención de los modelos de cada patrón de consulta vaga admisible, usando los datos disponibles.

De acuerdo con lo anterior, se prestó especial cuidado en la selección de los procedimientos o técnicas requeridas en el proceso de ajuste de los datos a los modelos teóricos para que fueran lo más sencillas posibles pero sin desaprovechar la información contenida en los datos de manera individual. Es por ello que para hallar el valor de los parámetros de los conjuntos difusos que representan los términos vagos, se propuso una técnica basada en percentiles pues otras técnicas alternativas como la estimación del núcleo de las densidades de probabilidad no sólo son más complejas, lo que retardaría innecesariamente el tiempo de respuesta, sino que pueden generar modelos no convexos que demandan otros pasos en el proceso, como también ocurriría con un histograma o polígono de frecuencias relativas. Sin embargo, los percentiles, como son medidas de posición, son funciones que requieren ordenar el conjunto de los datos delimitado por la propia consulta.

En la arquitectura concebida como parte del modelo de diseño, el sistema de inferencia (las reglas de inferencia y la base de conocimiento) está inmerso en el metamodelo de la base de datos y por eso, no sólo se evita el tráfico de los datos entre el servidor de datos y el servidor Web o el sitio donde resida la capa lógica del sistema flexible de consulta, sino que el ordenamiento de los datos se le puede delegar al sistema gestor de bases de datos que está altamente equipado para esta

tarea<sup>4</sup>. Consecuentemente, se evita que se deba programar, en un lenguaje no declarativo, uno o varios métodos de ordenamiento que por lo menos igualen a los ya implementados en los sistemas gestores de bases de datos y de esta forma, no degradar la eficiencia del sistema flexible en la resolución de consultas formuladas vagamente.

Adicionalmente, para buscar mayor rapidez en los cálculos, se diseñó una función especial llamada “concretar\_vaguedad” que devuelve las estimaciones de todos los parámetros de un modelo difuso de una sola vez, aprovechando las nuevas características del lenguaje procedimental que hoy ofrecen los sistemas gestores de bases de datos y que permiten devolver como respuesta un dato agregado o varios valores independientes, no únicamente un valor escalar o atómico. De esta manera sólo se necesita realizar un ordenamiento y no varios, dependiendo del número de parámetros del modelo del conjunto difuso que representa una etiqueta lingüística.

Otra ventaja de mantener y aplicar las reglas de inferencia en el servidor de datos, como aquí se propone, es el aprovechamiento de un instrumento esencial de los sistemas gestores de bases de datos que busca aumentar la velocidad de respuesta: los archivos de índices que se crean de manera automática o intencionada. Con los archivos de índices, el optimizador de consultas tiene una alternativa para mantener los datos ordenados, evitando realizar esta operación. Por lo tanto, se recomendaría indexar las relaciones de acuerdo con los atributos cuantitativos, convertibles en variables lingüísticas, más referenciados en las condiciones vagas de las consultas.

Por otro lado, pensando en un mecanismo propio de optimización en la resolución de las consultas vagas, se puede definir un método parecido al basado en costos que suele usar un sistema gestor de bases de datos para decidir por el plan de ejecución de una consulta que considere más conveniente, puesto que generalmente hay varias alternativas. Oracle, por ejemplo, mantiene en los metadatos una serie de estadísticas como la cardinalidad de un objeto de datos, la cantidad de valores distintos, el valor máximo y el valor mínimo de cada atributo de una relación, que utiliza en el proceso de optimización. Sin embargo, si estas estadísticas no están actualizadas el optimizador puede fallar y elegir una opción inapropiada. Dichos metadatos se actualizan explícitamente mediante la sentencia ANALYZE, cuya sintaxis es:

```
ANALYZE [TABLE, INDEX] nombre [COMPUTE, ESTIMATE] STATISTICS;
```

En la orden anterior se aprecia que se tienen dos opciones para calcular las estadísticas requeridas en el proceso de optimización. Especificando la palabra “compute”, el cálculo de las estadísticas se basaría en todos los elementos que forman parte de la tabla o el índice referido en la sentencia. Cuando la tabla o índice sea demasiado grande se puede especificar la opción “estimate”, para usar una muestra con las primeras filas del objeto de datos mencionado y que el sistema se encarga de elegir, de manera transparente, para el usuario (Colgan, 2007).

---

<sup>4</sup> Usualmente, un sistema gestor de bases de datos tiene predefinido un algoritmo de ordenamiento con un orden de complejidad  $O(n \log n)$

Siguiendo una estrategia parecida para evitar repetir el trabajo de minería de los datos cuando se realizan las mismas preguntas vagas frecuentemente, sería que se especificara explícitamente el texto de una consulta vaga y ésta se almacenara junto con la especificación de la consulta en términos concretos, después de realizar el proceso de concreción. Estos metadatos se complementarían con la cantidad de datos en las tablas referidas en la consulta, el número de ellos considerados en el proceso de concreción y la fecha del ajuste, mediante la sentencia siguiente, en la cual se utiliza el término que Zadeh (2006) define para la concreción de la vaguedad:

```
PRESITATION consulta [COMPUTE, ESTIMATE];
```

En esta sentencia, el término “consulta” hace referencia al texto de una consulta bien formada en el lenguaje SQL extendido. Puede observarse que también se pueden ofrecer dos opciones para hallar los modelos concretos: un censo o un muestreo aleatorio<sup>5</sup>. Esta última opción podría ser elegida si el volumen de datos restringido por la consulta es demasiado grande y se define como aleatorio para preservar la confiabilidad de los modelos.

Debido a la robustez de los estimadores de los parámetros, por ser medidas de posición que pueden incluso mantenerse invariables con ligeros cambios en los datos, se piensa que los modelos de los conjuntos difusos que representan las etiquetas lingüísticas que puedan guardarse como otros metadatos, por medio de la sentencia anterior, pueden ser confiables si no ha transcurrido mucho el tiempo entre una consulta y otra.

Independientemente a que se use el método de optimización de las consultas vagas recién planteado, siempre se revisará si la consulta que entra como pendiente se realiza sobre los mismos datos y se imponen los mismos criterios de filtrado de una consulta que se está concretando en esos momentos. Esto se hace en el momento de delimitar el contexto lingüístico, antes de crear otra memoria de trabajo temporal, innecesariamente.

También se debe considerar que el tamaño de la memoria de trabajo del sistema de inferencia sea, en algunos casos, muy grande debido al volumen de datos que hoy pueden guardarse en las bases de datos y no se tenga el modelo concreto de la pregunta vaga, guardado en los metadatos. Cuando esto ocurra, al sistema de inferencia se le puede agregar una funcionalidad para que seleccione, de manera aleatoria, un subconjunto de los datos para estimar los modelos de los conjuntos difusos. Este procedimiento también emula lo que haría el ser humano cuando les es imposible, no tiene el tiempo o el dinero para realizar un censo o estudio completo de una población de objetos.

Otra alternativa para evitar el ordenamiento de los datos, cuando el volumen de los datos sea muy grande, consiste en estimar los percentiles por medio de la función de distribución empírica acumulada. Este proceso es de menor complejidad algorítmica porque para calcular las frecuencias acumuladas solo se requiere contar los elementos que se encuentran en cada intervalo de clase, sin ordenarlos, y como

---

<sup>5</sup> En realidad, es pseudoaleatorio pues los números aleatorios generados por una máquina es una serie que, a largo plazo, se repite dependiendo del algoritmo utilizado y de la semilla inicial.

este procedimiento sólo depende del volumen de datos su orden de complejidad es  $O(n)$ , aunque esta ventaja en eficiencia sea a costa un poco de la entropía de la técnica.

Por último, otro aspecto que se puede considerar para aumentar la eficiencia del sistema repartir el trabajo en múltiples agentes en lugar de uno sólo. Cada uno de ellos puede estar especializado en un patrón sintáctico de las consultas vagas admisibles en el lenguaje extendido para realizar el razonamiento aproximado. Así, podrían estar colaborando simultáneamente en la interpretación de una consulta vaga.

### 7.3 Conclusiones

Con este trabajo de investigación se ha verificado la hipótesis de que era posible mejorar los sistemas de consulta-respuesta a bases de datos. Se llega a esta conclusión por las razones que se enuncian, a continuación:

- Se logró la confiabilidad requerida en un sistema de consulta a bases de datos, con su adaptabilidad a los distintos contextos en la interpretación de los adjetivos calificativos que restringen a los objetos de la base de datos y a los diferentes niveles de granulación admisibles en una partición difusa. Se cumplió este objetivo principal, con la definición de las reglas de inferencia requeridas para representar cada patrón de consulta vaga admisible para que, por su propia cuenta y dinámicamente, pueda hallar los modelos particulares a partir de la información disponible en la base de datos.
- Se ha mostrado que el sistema de inferencia difuso inmerso del sistema gestor de una base de datos puede por sí mismo, hallar los modelos requeridos para la discriminación y clasificación de los objetos, evitando la necesidad de expertos. Por lo tanto, se ha concebido un agente inteligente encargado de tareas de Minería de Datos que usualmente sólo realiza personal técnico, altamente calificado. Se muestra, entonces, que un sistema informático puede hacer labores complejas y ofrecer respuestas más confiables al eliminar los juicios subjetivos en la elaboración de los modelos.
- Se ha mejorado el lenguaje de los sistemas de consulta a bases de datos, con la redefinición de la representación de los adverbios como “muy” o “extremadamente”, realizando otro proceso de discriminación sobre el conjunto difuso que representa el adjetivo vago de interés. Así, se propone un método de razonamiento aproximado, que emula mejor el modo de proceder de los humanos.
- El lenguaje de consulta a bases de datos se ha enriquecido semánticamente con la representación de combinaciones lineales de condiciones simples, vaga o concretas.
- Se ha logrado que las extensiones y modificaciones en el lenguaje SQL no afectaran su legibilidad y su proximidad con el lenguaje natural que le dio origen: el inglés. Ni en estas interfaces, ni en el lenguaje SQL extendido se

incorporan tecnicismos de la Lógica Difusa, lo cual significa que el modelo de razonamiento propuesto se puede considerar “Caja Negra” bajo la óptica de los usuarios finales, pero “Caja Blanca” para los técnicos que deseen hacer mejoras o complementos futuros en el sistema de inferencia. También, se logró pasar de un lenguaje simbólico SQL a un lenguaje orientado al usuario, mediante las interfaces gráficas diseñadas.

- La representación y manejo de cuantificadores relativos en el lenguaje SQL extendido permite identificar relaciones o reglas de asociación entre variables. Esto amplía significativamente las capacidades del lenguaje y le permite, a los consultores no expertos, realizar tareas de Minería de Datos, de una manera sencilla.
- Por otro lado, se ha ampliado las posibilidades de plantear consultas a sistemas de bases de datos, donde los resultados esperados incluyan los valores de verdad rotulados con alguna etiqueta lingüística. Caso no contemplado en los lenguajes de bases de datos como el SQL.
- Se ha mostrado que los sistemas de inferencia difusos pueden hacer uso de la Estadística para el ajuste de los datos a los modelos teóricos difusos que representan la semántica los términos vagos o imprecisos. Por esto, se concluye que la Lógica Difusa y la Estadística nos son disciplinas o técnicas alternativas, sino que se pueden complementar para dar solución a problemas de incertidumbre que se vuelven cada vez más cotidianos: el Descubrimiento de Conocimiento o la Búsqueda Inteligente en Bases de Datos.
- El modelo conceptual de razonamiento para la interpretación de la vaguedad en las consultas es independiente del dominio de aplicación y de ahí, su utilidad. En el modelo de diseño se aprecia claramente, la generalidad de modelo propuesto.
- Mediante el diseño de un aplicativo Web al cual se le han incorporado todas las características definidas en el modelo conceptual, se ha mostrado que la construcción de sistemas flexibles de consulta-respuesta es factible, sin tener que depender de las adiciones o cambios que deban ser realizados por las casas productoras de los sistemas gestores de bases de datos. Esto se logró gracias a las características dinámicas del lenguaje procedimental que hoy ofrecen los sistemas gestores de bases de datos objeto-relacionales.
- Debido a la flexibilidad lograda en el lenguaje, los sistemas de consulta-respuesta basados en la presente propuesta no sólo son factibles, sino que inciden positivamente en su “usabilidad”. En la Ingeniería del Software esta característica fundamental de calidad, se define como un agregado de amigabilidad y del provecho que podamos obtener de los servicios que nos ofrece un sistema informático.
- La calidad de un modelo conceptual no se circunscribe a un solo factor como su extensibilidad, sino que depende de otros factores como la confiabilidad y la generalidad, entre otros. Debido a esto, determinar la calidad de un modelo

no es asunto sencillo, pero haber definido previamente una serie de restricciones que debía cumplir el modelo conceptual de razonamiento aproximado en la interpretación de la vaguedad, facilitó su concepción y evaluación.

- Se ha definido una arquitectura para el sistema de inferencia, donde las reglas no son definidas en el servidor de aplicaciones Web o en los programas de navegación de los usuarios, separadas de los hechos o la base de datos. Esta arquitectura no sólo facilita el mantenimiento o mejora del sistema de inferencia, en la interpretación de las consultas vagas, sino la reutilización de las reglas o funciones en otras aplicaciones orientadas a otros tipos de usuarios más calificados, como los analistas o mineros de los datos.

## 7.4 Trabajo Futuro

El uso creciente de los computadores ha dado origen a grandes volúmenes de datos que sabiéndolos aprovechar se pueden convertir en conocimiento invaluable. Y de este trabajo de investigación se desprenden otros trabajos que conducen a facilitar las tareas de Descubrimiento de Conocimiento en Bases de Datos, realizadas por los analistas o los mineros de datos, un tipo de usuario más calificado.

Entre los trabajos futuros que se vislumbran, tanto para fortalecer los sistemas flexibles de consulta-respuesta, así como para que las organizaciones puedan aprovechar mejor los datos almacenados en la toma de decisiones, se mencionan los siguientes:

- Aumentar la potencia expresiva del lenguaje con la inclusión de otros comparadores vagos como “mucho mayor que  $x$ ” o “mucho menor que  $x$ ”, siendo  $x$  un valor cuantitativo cualquiera. La representación de estos términos vagos requiere de pruebas estadísticas que permitan encontrar diferencias significativas entre valores observados en los datos.
- Aumentar el alcance de esta propuesta con su aplicación a los sistemas semi-estructurados o aplicativos Web que no tengan acceso a bases de datos relacionales u objeto-relacionales, sino que los datos estén guardados en otros formatos. Esto es posible porque la máquina de inferencia se puede implementar empleando el modelo conceptual propuesto y usando un lenguaje de consulta a datos semi-estructurados, como el XQuery y lenguajes de programación Web como el PHP o el Javascript.
- La representación de cambios o tendencias en los conceptos vagos, debidos al tiempo o al contexto histórico o al contexto físico considerado. Con ello, se podría preguntar si el rendimiento académico de los estudiantes, en una asignatura dada, ha “mejorado” con respecto a los estudiantes del período académico anterior, por ejemplo.
- La ampliación de las capacidades para que un sistema interactivo permita realizar tareas de predicción en la Minería de Datos que incluyan algún tipo de vaguedad derivable de múltiples variables o atributos, basadas en muestras de aprendizaje.

- Determinar la manera de poder inferir cuándo un término vago deja de ser dependiente del contexto lingüístico y se vuelve universal o absoluto. Esto con el fin de ahorrar esfuerzos al sistema de inferencia porque el modelo de un concepto cuya semántica no depende del contexto lingüístico, se puede agregar a la base de conocimientos como un hecho o axioma mediante una orden de definición de datos.
- Concebir sistemas que permitan operar con términos lingüísticos vagos que representen probabilidades. Esto permitiría la resolución de preguntas de investigación o de descubrimiento de nuevo conocimiento donde exista incertidumbre debida a fenómenos aleatorios, de una manera más sencilla y cercana a un usuario no experto en Estadística.

En el futuro inmediato, esta propuesta se aplicará en el trabajo de investigación llamado “Descubrimiento de Conocimiento sobre la Innovación en Colombia a partir de las Encuestas de Innovación y Desarrollo Tecnológico, La Encuesta Anual Manufacturera y la base de datos Scienti” proyecto financiado por la Universidad Nacional de Colombia y Colciencias que agrupa investigadores de varias universidades e instituciones gubernamentales del país y con vigencia 2007-2009. Esta aplicación real del modelo propuesto permitirá no sólo responder preguntas de investigación que ya fueron planteadas de manera vaga, sino refinar el modelo de razonamiento propuesto y efectuar las extensiones que se pueden realizar a corto plazo.

## 8 Referencias Bibliográficas

- Aho, A; Sethi, R. y Ullman, J. (1990). *Compiladores. Principios, Técnicas y Herramientas*. Addison-Wesley Iberoamericana S.A.
- Agrawal, R; Imielinski, T. y Swami, A. (1993) Mining association rules between sets of items in large databases. *Proc, ACM SIGMOD, Conf. on Management of Data*, págs 207-216. Washington, D.C.
- Álvarez, H. y Peña, M. (2004). Modelamiento con Sistemas de Inferencia Borrosa Tipo Takagi-Sugeno. *Revista Dyna*.
- Amo, J. et al.. (2004). Fuzzy classification systems. *Computing, Artificial Intelligence and Information Technology. European Journal of Operational Research* 156. 495-507.
- Apt, K. (2003). *Principles of Constraint Programming*. Ed. Cambridge University Press.
- Bach, K. (2005). *Speech Acts and Pragmatics*. Devitt, Michael and Richard Hanley (eds.) The Blackwell Guide to the Philosophy of Language. Oxford.
- Baeza, R. y Ribeiro, B. (1999). *Modern Information Retrieval*. Ed ACM Press. Addison Wesley. Inglaterra. Capítulo 1.
- Baldwin, J. (1978). Fuzzy Logic and Fuzzy Reasoning, Intl. *Journal of Man-Machine Studies*.
- Bally, Ch. (1977). *El lenguaje y la vida humana*. Editorial Losada Argentina.
- Bargiela, A. y Pedrycz, W. (2003). *Granular Computing: An Introduction*. Kluwer Academic Publisher. Londres.
- Barker, R (1991). *Case Method: Entity-Relationship Modelling*. Ed Oracle Press.
- Betranpetit, J. y Junyent, C. (2000). *Viaje a los orígenes. Una historia biológica de la especie humana*. Ed. Península. Barcelona
- Berkran, R. y Trubatch, S. (1997). *Fuzzy Systems Design Principles. Building Fuzzy IF-THEN Rules Bases*. IEEE Press.
- Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithm*. Plenum Press, Nueva York.
- Bloom, B. (1990). *Taxonomía de los Objetivos de la Educación*. Décima edición. Editorial El Ateneo. Buenos Aires.

- Bloom, R. (2007). *PostgreSQL 8 for Windows (Database Professional's Library)*. McGraw-Hill Osborne Media.
- Bosc, P. (1979). *Vagueness, Ambiguity, and all the Rest. An explication and an intuitive test*. W. van de Velde, W. Vandeweghe (eds.): Sprachstruktur, Individuum und Gesellschaft. Niemeyer. Tübingen. págs. 9-19.
- Bosc, P. y Pivert, O. (1995) *SQLF: A Relational Database Language for Fuzzy Querying*. IEEE Transactions on Fuzzy System. Vol 3 Nro.1 , págs 1-17
- Bosc, P., Motro A. y Pasi G. (2001) *Report on The fourth International Conference on Flexible Query Answering systems (FQAS)*.
- Bosc, P., Kraft, D. y Petry, F. (2005). Fuzzy sets in database and information systems: Status and opportunities. *Fuzzy Sets and Systems* Vol 156, págs 418-426
- Cámara de La Fuente, L. (2004). *La representación lingüística del conocimiento y su relevancia en la ingeniería lingüística* [disponible en línea]: <<http://www.hipertext.net>> [Consulta: 23/06/2007]. ISSN 1695-5498
- Casillas J., Cordon, O., Herrera F. y Magdalena L. (2003). Interpretability improvements to find the balance interpretability accuracy in fuzzy modeling: an overview. *Interpretability Issues in Fuzzy Modeling*. Springer-Verlag, Alemania. págs 3-22.
- Cerezo, M. (1994). *Texto, contexto y situación*. Ed. Octaedro. Barcelona.
- Codina, C. et al. (2001). "Sobre los elementos a considerar en la representación del conocimiento de cara a la recuperación de información (el punto de vista cognitivo)". Editorial Cabré, T. Barcelona.
- Colgan, M (2007). SQL plan management in Oracle Database 11G. Oracle Corporation World Headquarters. EEUU. Disponible en [http://www.oracle.com/technology/products/manageability/database/pdf/w07/spm\\_white\\_paper\\_ow07.pdf](http://www.oracle.com/technology/products/manageability/database/pdf/w07/spm_white_paper_ow07.pdf). Consulta en 9/7/2008.
- Comisión Europea (2002). DG XIII-E4, EUFO 0-176, [disponible en línea] [http://www.hltcentral.org/usr\\_docs/project-source/en/index.html](http://www.hltcentral.org/usr_docs/project-source/en/index.html) [Consulta: 23/06/2007]
- Cox, E. (1994). *The fuzzy systems handbook: a practitioner's guide to building, using and maintaining fuzzy systems*. Academic Press. Estados Unidos.
- Czogala, E. y Leski, J (2000). *Fuzzy and Neuro-Fuzzy Intelligent Systems*. Ed Springer-Verlag Nueva York; 1ra ed.
- Chen, X., Ender, P., Mitchell, M. y Wells, C. (2003). *Regression with Stata*, from <http://www.ats.ucla.edu/stat/stata/webbooks/reg/default.htm> .

- Cherkassky, V y Mulier, F (1998). *Learning from Data: Concepts, Theory, and Methods*. Ed Wiley, Nueva York.
- Date, C. (2001). *Introducción a los Sistemas de Bases de Datos*. Ed Addison Wesley. 7ª Edición.
- De Cock, M. y Kerre E. (2004). *Fuzzy modifiers based on fuzzy relations*. Information Sciences 160.
- Delgado, M. et al. (1997). *Aggregation of Linguistic Information Based on a Symbolic Approach*. ETS de Ingeniería Informática. Universidad de Granada.
- Delgado, M., Verdegay, J. y Vila, M. (1993). On Aggregation Operations of Linguistic Labels. *International Journal of Intelligent Systems*. Vol 8 págs 351-370.
- Devillez, A. 2004. Four fuzzy supervised classification methods for discriminating classes of non-convex shape. *Fuzzy Sets and Systems*, Vol. 141, No. 2, pp.219-240, 2004
- Diez, F. (2006). *Introducción al Razonamiento Aproximado*. Dpto. Inteligencia Artificial UNED. Disponible en:  
<http://www.ia.uned.es/~fjdiez/libros/razaprox.html>.
- Dix, J., U. Furbach y Niemela, I. (1999). *Nonmonotonic Reasoning: towards efficient calculi and Implementations*. Elsevier Science Publishers.
- D'Urso, P y Giordani, P (2006). A weighted fuzzy c-means clustering model for fuzzy data. En *Computational Statistics & Data Analysis* Volume 50, Issue 6, , Págs 1496-1523.
- Dubois, D. y Prade, H. (2000). *Fundamentals of Fuzzy Sets*. The Handbooks of Fuzzy Sets Series, Kluwer Academic Publishers. Holanda
- Dubois, D., H. Prade y Testemale, C. (1988). Weighted Fuzzy Pattern Matching, *Fuzzy Sets and Systems*, 28, pags 313-331.
- Duda, R, Hart, P y Stork, R (2001). *Pattern Classification*. Segunda edición, Wiley Interscience. ISBN 0471056693.
- Dyer, C. (2003). *Machine Learning*. Lecture Notes. Universidad de Wisconsin. Capítulos 18.1 - 18.3. Disponible en <http://www.cs.wisc.edu/~dyer/cs540/notes/learning.html>. Revisión: marzo 2008.
- Evsukoff, A., A.Branco y Ebecken, N. (2006) "Generating Weighted Fuzzy Queries from Fuzzy Classifier Rules". The 11th IPMU International Conference, , Paris.
- Espinosa, J. y Vandewalle, J. (2000). Constructing fuzzy models from numerical data- AFRELI algorithm. *IEEE Transactions on Fuzzy Systems*, 8(5) págs. 591-600.

- Fisher, R. (1936). *The use of multiple measurements in taxonomic problems*. *Eugen*. 7 págs 179-188.
- Fodor, J. (1991). Strict preference relations based on weak t-norms. *Fuzzy Sets and Systems*, Vol. 43. Págs 327-336.
- Fodor, J. (2004), "Left-continuous t-norms in fuzzy logic: An overview". *Acta Polytechnica Hungarica* 1(2), ISSN 1785-8860. Disponible en: [http://www.bmf.hu/journal/Fodor\\_2.pdf](http://www.bmf.hu/journal/Fodor_2.pdf). Última consulta en línea septiembre, 2007
- Frigg, R. y Hartmann S. (2006) *Models in Science*. Disponible en <http://plato.stanford.edu/entries/models-science/>. Última consulta en línea: abril de 2007.
- Galindo, J. et al. (1998). *A Server for Fuzzy SQL Queries, in Flexible Query Answering Systems*, eds. T. Andreasen, H. Christiansen and H.L. Larsen, Lecture Notes in Artificial Intelligence (LNAI) 1495, pp. 164-174. Ed. Springer.
- Galindo, J. (2001). *Conjuntos y Sistemas Difusos (Lógica Difusa y Aplicaciones)*. Universidad de Málaga (España). Disponible en: <http://www.lcc.uma.es/~ppgg/FSS>. Última consulta en línea: abril de 2007.
- Galindo J., A. Urrutia y Piattini M. (2005), "Fuzzy Databases: Modeling, Design and Implementation". Idea Group Publishing Hershey. EE UU.
- Galindo, J. (2008). Introduction and Trends in Fuzzy Logic and Fuzzy Databases. En *Handbook of Research on Fuzzy Information Processing in Databases*. José Galindo. Idea Group Inc (IGI).
- Grabish M., Orlovski y Yager (1998). Fuzzy aggregation of numerical preferences. *The Handbook of Fuzzy Sets Series, Vol. 4: Fuzzy Sets in Decision Analysis, Operations Research and Statistics*, R. Slowinski (ed), Kluwer Academic, págs 31-68.
- Gonçalves, M. y Tineo, L. (2001). *SQLf Flexible querying language extension by means of the norm SQL2*. FUZZY – IEEE'2001 Conference. Australia.
- Gonçalves, M. y Tineo, L. (2001). SQLf3: An extensión of SQLf with SQL3 features. *IEEE Internacional Fuzzy Systems Conference*. Págs 477-479.
- Gonçalves, M. y Tineo, L. (2007). "A New Step towards Flexible XQuery". *Avances en Sistemas e Informática*. Vol 4. No. 3. págs 27-34.
- Guillaume, S. (2001). Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE Transactions on Fuzzy Systems*. 9(3):426-443.
- Han, J y Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publisher. Academic Press.

- Hajek, P. (2005) "Arithmetical complexity of fuzzy logic – a survey-". *Soft Computing* 9: 935-941.
- Harrison, M. (1987). *Introduction to Formal Language Theory*. Addison-Wesley, Reading.
- Hahsler, M. (2006). "A model-based frequency constraint for mining associations from transaction data". *Data Mining and Knowledge Discovery*, 13(2) págs137-166.
- Hawthorne, J. (2005). "Inductive Logic", The Stanford Encyclopedia of Philosophy (<http://plato.stanford.edu/entries/logic-inductive/>).
- Hinrichsen, D. y Pritchard, A. (2005). *Mathematical Systems Theory I - Modelling, State Space Analysis, Stability and Robustness*. Springer Verlag. ISBN 0-978-3-540-44125-0
- Hobbs, J. (1985). "Granularity". En *Proceeding of the 9<sup>th</sup> International Joint Conference on Artificial Intelligence*. Págs 432-435.
- Huang, G., P. Saratchandran y Sundararajan, N (2004). *An efficient sequential learning algorithm for growing and pruning RBF (GAP-RBF) networks*. Systems, Man and Cybernetics, Part B, IEEE Transactions on Volume 34, Issue 6, Págs: 2284 - 2292.
- Huber, P. (2004). *Robust Statistics*. Wiley Series in Probability and Statistics.
- Information Technology Laboratory Itl (2006) . *Statistical Reference Data Sets Archives*. Disponible en: <http://www.itl.nist.gov/div898/strd/general/dataarchive.html>
- Insfrán, P. y Wieringa, C (2002). *Requirements Engineering*. Capítulo Requirements Engineering-Based Conceptual Modelling. Ed Springer Londres. Pág 61 - 72.
- ISO/IEC International Standard (1999). *Information Systems -Database Language SQL*. Disponible en <http://www.ncb.ernet.in/education/modules/dbms/sql99index.html>. Última consulta en línea: octubre de 2008.
- Hair, J. et al. (1999). *Análisis Multivariante*. Madrid: Prentice Hall, 5 ed.
- Hastie, T., R. Tibshirani, y Friedman, J. (2001). *The Elements of Statistical Learning*. Data Mining, Inference and Prediction. Springer series in statistics. Págs 468-480.
- Harris, R. (2001). *A Primer of Multivariate Statistics*. Lawrence Erlbaum Associates. 3ra ed.
- Herrera, F. y Herrera-Viedma, E. (1997). Aggregation Operators for Linguistic Weighted Information. *IEEE Transactions on Systems, Man and Cybernetics*. 27. Págs 268-290.

- Jamei et al. (2001). "Elicitation and fine tuning of mandani-type fuzzy rules using symbiotic evolution". En *Proceedings of European Symposium on Intelligent Technologies. Hybrid Systems and their Implementations on Smart Adaptive Systems* (EUNITE 2001). España.
- Jain, A., M. Murty y Flynn, P. (1999). *Data Clustering: A Review*. ACM Computing Surveys, Vol 31, No. 3 pp. 264-323.
- Jang J., C. Sun y Mizutani, E. *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Ed Prentice Hall. Págs. 56-60.
- Jiménez, H. et al. (2001). "A multi-objective evolutionary algorithm for fuzzy modelling". En *Proceedings of NAFIPS*, Págs. 1222-1228. Nueva York.
- Johnson, R. y Wichern, D. (2001). *Applied Multivariate Statistical Analysis*. Prentice Hall. 5ta ed. EEUU.
- Joslyn, C. (1994). *A Possibilistic approach to qualitative Model-Based Diagnosis*. disponible en <http://calculus.math.utep.edu/~andrzej/papers/cliff94possibilistic.pdf>.
- Jovell, A. (1995). *Análisis de regresión logística*. Madrid: CIS.
- Jurafsky, D. (2005). *Pragmatics and Computacional Linguistics*. Handbook of pragmatics. Oxford: Blackwell.
- Kackprzyk, J. y Zadrozny, S (2001). Computing with words in intelligent database querying: standalone and Internet-based applications. *Information Sciences* 134 págs 71-109.
- Kallenberg, O. (2002). *Foundations of Modern Probability*, Segunda Ed . Springer Series in Statistics. 650 pp. ISBN 0-387-95313-2.
- Kandel, A. And Langhols, G. (1994). *Fuzzy Control Systems*. CRC Press.
- Klement, E., Mesiar, R. y Pap, E. (2000), *Triangular Norms*. Dordrecht: Kluwer. ISBN 0-7923-6416-3.
- Klir, G. y Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic. Theory and Applications*. Prentice Hall, Nueva York.
- Large, A. (1999). *Information seeking in the online age: principles and practice*. Londres. Bowker-Saur,.
- Lofti, A., Andersen, H y Tosi, A (1996). Interpretation preservation of adaptive fuzzy inference systems. *International Journal of Approximated Reasoning*. 15:379-394.
- Llisterri, J. (2003). Lingüística y Tecnologías del Lenguaje, Linx. *Panorámica de Estudios Lingüísticos*. 2:9-71. Disponible en: [http://liceu.uab.es/~joaquim/publicacions/Fonetica\\_TecnolHabra.pdf](http://liceu.uab.es/~joaquim/publicacions/Fonetica_TecnolHabra.pdf)
- Llopis, Ferrández y Viñedo (2001). *Question Answering Tracks*. Universidad de Alicante.

- Lomas, C. et al. (1997). *Ciencias del lenguaje, competencia comunicativa y enseñanza de la lengua*. Ed Paidós. Barcelona.
- Lohninger, T (1999). *Teach/Me Data Analysis*, Springer-Verlag, Berlin-New York-Tokyo.
- Loh W. (2004) *CRUISE Classification Tree*. Disponible en: <http://www.stat.wisc.edu/~loh/cruise.html>.
- Ma, Z. y Wang, H. (2006). STEP implementation of imperfect EXPRESS model in fuzzy object-oriented databases. *Fuzzy Sets and Systems*. Vol 157 págs. 1597-1621.
- Mac Queen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Págs. 281-297.
- Martín Del Brio, B. y Sanz Molina, A. (2002). *Redes Neuronales y Sistemas Difusos*. 2ª edición ampliada y revisada. Ed. RA-MA.
- MathWorks Inc. (2006). *Statistics Toolbox User's Guide*. Disponible en <http://www.mathworks.com/access/helpdesk/help/toolbox/stats/>. Consultado en: 20/09/2008.
- Melton, J. (2003). *Advanced SQL: 2003. Understanding object-relational and other features*. Morgan Kaufmann Publishers. Elsevier Science. USA..
- Medina, J., D. Pons y Vila, M. (1994). GEFRED: A generalized model of Fuzzy Relational Data Bases. Ver 1.1 . *Information Sciences*.
- Mencar, C. (2004). *Theory of Fuzzy Information Granulation: Contributions to Interpretability Issues*. Tesis doctoral. Universidad de Bari. Italia.
- Miller, G. (1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information". *The Psychological Review* 63:81-97. Disponible en <http://www.well.com/user/smalin/miller.html>
- Mineau y Godin (1995). Automatic Structuring of Knowledge Bases by Conceptual Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 7(5).
- Mitra, S. y Acharya, T. (2003). *Data Mining. Multimedia, Soft Computing and Bioinformatics*. John Wiley & Sons.págs 5-7.
- Montgomery, D. y G. Runger (2003). *Applied Statistics and Probability for Engineers*. John Wiley & Sons, Inc. 3ra ed. Págs. 204-209
- Morales, G. (2000). *Elementos de lógica difusa*. Centro de Investigación y Estudios Avanzados del IPN. Disponible en <http://delta.cs.cinvestav.mx/~gmorales/ldifl/ldifl.html>. Fecha de último acceso en línea: Junio de 2007.

- Moreno et al. (1999). *Introducción al Procesamiento del Lenguaje Natural*. Editorial Universidad de Alicante. España.
- Negnevitsky, M. (2004). *Artificial Intelligence: A Guide to Intelligent Systems*. Addison Wesley. ISBN 0321204662.
- Neter, J. y Wasserman W. (2001). *Applied Linear Regression Analysis*. New York: Wiley.
- Novak, J. (1990). *Concept maps and veer diagrams: Two metacognitive tools for science and mathematics education*. *Instructional Science*, 19, 29-52.
- Nozaki, K. H. Ishibuchi y Tanaka, H. (1996) Adaptive fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems*, Vol. 4, No. 3, págs 238 – 250.
- Parsons, S. (1996). Current approaches to handling imperfect information in data and knowledge bases. *IEEE Transactions on Knowledge and Data Engineering*. 8 (3) 353-372.
- Oakland, J.S. (2000). *Total Quality Management*. The route to improving performance. 3ra Ed. Butterworth Heinemann. Oxford.
- O'Hagan, A. (1994) *Bayesian Inference*. Ed Edward Arnold, London.
- Oracle Corporation. (2006). *Oracle® Database SQL Reference 10g. Release 1 (10.1)* Parte No. B10759-01. Disponible en [http://download.oracle.com/docs/cd/B14117\\_01/server.101/b10759.pdf](http://download.oracle.com/docs/cd/B14117_01/server.101/b10759.pdf)
- Pawlak, Z. (1994). *Vagueness and uncertainty: a rough set perspective*, Technical Report ICS Research Report 1994, Institute of Computer Science, Warsaw University of Technology, Warsaw, Polonia.
- Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*, Kluwer Academic.
- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Ed. Cambridge University Press, ISBN 0-521-77362-8. Inglaterra.
- Peirce, C. (1992). Reasoning and the Logic of Things. *The Cambridge Conferences Lectures de 1898*. K. L. Ketner (ed). Cambridge, MA: Harvard University.
- Peña-Reyes, C. y Sipper, M. (2003). "FUZZY CoCo: Balancing accuracy and interpretability of fuzzy models by jeans of coevolution". En *Accuracy Improvements in Liguistic Fuzzy Modeling*. Págs 119-146. Springer-Verlag.
- Peña, M. (2001). *Fuzzy model based control*. Ph.D. thesis. INAUT, Fac. de Ingeniería, U.N.S.J. ISBN 950-605-278-6. EFU. Argentina.
- Pérez, C. (2004). *Técnicas de Análisis Multivariante de Datos*. Ed Pearson Educación, Madrid.

- Perlich, C. y Provost, F. (2006). *Distribution-based aggregation for relational learning with identifier attributes*. Ed Springer Science y Business Media, Inc. Disponible en: [http://pages.stern.nyu.edu/~efprovost/Papers/perlich\\_provost\\_mlj.pdf](http://pages.stern.nyu.edu/~efprovost/Papers/perlich_provost_mlj.pdf)
- Pradera A, Trillas E., Guadarrama y Renedo E. (2006). *On Construction Imprecise Fuzzy Set Theories*. European Centre for Soft Computing.
- Quinlan, J. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publisher, San Mateo. CA.
- Radhakrishna R y Toutenburg, H. (1999) *Linear Models: Least squares and alternatives*. Ed Springer series in Statistics. 2da ed.
- Rasmussen, D. y Yager, R. (1999) Finding fuzzy and gradual functional dependencies with Summary SQL. *Fuzzy Sets and Systems*, 106, pág. 131-142.
- Rumbaugh J., I. Jacobson y Booch G. (1998). *The Unified Modeling Language Reference Manual*, Addison-Wesley, ISBN 020130998X.
- Ruspini, E (1969). "A new approach to clustering". *Information and Control*, 15:22-32.
- Schuermann, J (1996). *Pattern Classification: A Unified View of Statistical and Neural Approaches* Wiley&Sons, ISBN 0471135348.
- Schreiber et al. (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press.
- Snyder, C. (2003). Paper Prototyping. The Fast and Easy Way to Design and Refine.
- Soong, T. (2004). *Fundamentals of Probability and Statistics for Engineers*. John Wiley and Sons, Inc. Págs 249-252
- Studer R., R. Benjamins y Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering*. 25 págs 161-197.
- Sicilia, M. (2002). *Maze: Un Modelo de Hipermedia Adaptativa con Soporte de Información imperfecta*. Tesis Doctoral. Universidad Carlos III. Madrid.
- Scott, J. y Freese, J. (2001). *Regression Models for Categorical Dependent Variables Using Stata*. Texas: Stata Press.
- Trillas, E. Alsina, C. y Pradera, A. (2007) *On a Class of Fuzzy Set Theories*. IEEE en el 2007, Fuzzy Systems Conference,. IEEE International Disponible en <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04295351> Última consulta en línea: octubre de 2008.
- Trosset, M., (2001). *An Introduction to Statistical Inference and Data Analysis*. John Wiley and Sons, Inc. pp. 135-137.

- Tsekouras, G. (2005). On the use of the weighted fuzzy c-means in fuzzy modeling. *Advances in Engineering Software* 36 págs 287–300.
- Urrutia, A. M. Varas y Galindo, J. (2003). *Diseño de una Base de Datos Difusa Modelada con UML. IDEAS. Disponible en:* <http://www.inf.udec.cl/~mvaras/papers/2003/fuzzy-IDEAS-2003.pdf> .
- Urrutia A., J. Galindo y Piattini, M. (2003). *Propuesta de un Modelo Conceptual Difuso*. IX Jornadas Iberoamericanas de Informática. Ritos2. Colombia.
- Urrutia, A., Tineo, L. y González, C. (2008). FSQL and SQLf: Towards a Standard in Fuzzy Databases. En *Handbook of Research on Fuzzy Information Processing in Databases*. José Galindo. Idea Group Inc (IGI).
- US Census Bureau (2007). The census bureau database 1994. Disponible en <http://www.census.gov/ftp/pub/DES/www/welcome.html>
- Valverde, LI (1996). *L.A. Zadeh: Del Control Analítico al Control Borroso*. Entrevista. Disponible en <http://dmi.uib.es/people/valverde/laz.html>.
- Walpole, R., R. Myers y Myers, S. (1999). *Probabilidad y Estadística para Ingenieros*. 6ª ed. Prentice-Hall. México.
- Wasserman, L. (2005). *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics.
- Webb, A. (2002) *Statistical Pattern Recognition*. Ed. John Wiley and Sons. 2da edición. Inglaterra
- Wechsler, D. (1944). *The Measurement of Adult Intelligence*. Baltimore: The Williams & Wilkins Company.
- Weigand, H. (1997). "Multilingual Ontology-Based Lexicon for News Filtering –The TREVI Project", en K. Mahesh. Págs 138-159.
- Yager, R. (1980), "On a general class of fuzzy connectives". *Fuzzy Sets and Systems* vol 4 págs 235-242.
- Yager, R. (1988), "On Ordered Weighted Averaging Aggregation Operator in Multicriteria Decision Making". *IEEE Transactions on Systems, Man and Cybernetics*, págs:183-190.
- Yager, R. y Kacprzyk, J. (1999). "Linguistic data summaries: A perspective". *Proc. of 8th Int. Fuzzy Systems Association World Congress (IFSA '99)*, págs 44–48.
- Yao, Y. (2000). Granular Computing: Basic Issues and Possible Solutions. *Proceedings of the 5th Joint Conference on Information Sciences*. págs 186-189.
- Yee, T. y Hastie, T. (2003). *Reduced-rank vector generalized linear models*. *Statistical Modelling* 3, 15–41.
- Zadeh, L. (1965). Fuzzy Sets. *Information and Control*. Vol 8. Págs. 338-353.

- Zadeh, L. (1975). "The Concept of Linguistic Variable and its Application to Approximate Reasoning". *Information Sciences*, 8, Págs. 199-249, Págs 301-357, (parte II), 9, Págs. 43-80, (parte III).
- Zadeh, L. (1978). "PRUF - A Meaning Representation Language for Natural Languages". *Man-Machine Studies* 10. Págs. 395-460.
- Zadeh, L. (1979). "A theory of approximate reasoning". *Machine Intelligence*, vol. 9, Hayes, D. Michie, and L.I. Mikulich, Eds. New York: Wiley, pp. 149-194.
- Zadeh, L. (2006). "Chapter 9. From Search Engines to Question Answering Systems- The Problems of World Knowledge, Relevance, Deduction and Precisation". *Capturing Intelligence*, Volumen 1, Págs 163-210.
- Zaïane, O. (1999). "Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering". Disponible en <http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html>
- Zapata, C. (2002). *Incorporación de operadores borrosos en Manejadores de Bases de Datos Relacionales*. Tesis de Grado de Maestría en Ingeniería de Sistemas. Universidad Nacional de Colombia.
- Zimmermann, H. (1993). *Fuzzy set theory and its applications*, Dordrecht: Kluwer, segunda edición.
- Zloof, M. (1981). "QBE/OBE: A Language for Office and Business Automation". *IEEE Computer* 14(5): 13-22.

