



UGR

Universidad  
de Granada



TESIS DOCTORAL

Data Warehousing con procesamiento de datos textuales

Elizabet Tejeda Ávila

*Granada, 2010*

Editor: Editorial de la Universidad de Granada  
Autor: Elizabet Tejeda Ávila  
D.L.: GR 3468-2010  
ISBN: 978-84-693-5222-9





*ugr*

Universidad  
de Granada



Data Warehousing con procesamiento de datos textuales

memoria que presenta

Elizabet Tejeda Ávila

para optar al grado de

Doctora en Informática

*2010*

DIRECTORAS

Dra. María Amparo Vila Miranda

Dra. María José Martín Bautista

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN  
E INTELIGENCIA ARTIFICIAL

La memoria titulada " Data Warehousing con procesamiento de datos textuales ", que presenta Dña. Elizabet Tejeda Ávila para optar al grado de Doctora, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección de las Doctoras Dra. María Amparo Vila Miranda y Dra. María José Martín Bautista.

Granada, 2010.

La Doctoranda

Las Directoras

Dña. Elizabet Tejeda Ávila

Dra. María Amparo Vila Miranda

Dra. María José Martín Bautista

## Dedicatoria

Poder al fin, dedicar esta tesis, es un sueño hecho realidad. Lo hago con inmensa alegría y consciente de que ha sido posible gracias al cariño y apoyo incondicional de las personas a las que va dedicada.

*A mi amado esposo Sandro que es el Sol de mis días. Porque su cariño y apoyo incondicional todo este tiempo me ha ayudado a ser mejor persona en todos los sentidos. Puedo decir sin duda alguna que él ha sido la persona que en mayor medida ha hecho posible este logro.*

*A los hijos que esperamos vengan pronto y coronen todos los éxitos que hemos alcanzado.*

*A mi querida madre, que es mi mayor tesoro. Por tantos años de amor infinito que me ha llevado hasta aquí.*

*A mi hermano Rafael, que desde que nació me llenó de alegría y lo llevo presente siempre en mi corazón.*

*A Fernan por su cariño y ayuda durante todos estos años a quien quiero como a un padre.*

*A mis queridos abuelos María, Eligio, Idelino y Nérida, por ser raíces de amor para mis dos familias.*

*Al resto de la familia, en especial a mis queridos cuñados Rodney y Guille.  
También a mis tías Mima, Carmita y Nena.*

*A mis suegros queridos Mary y papá, por ser las personas más nobles que he conocido y por todo el cariño que les profeso.*

*A mis familiares españoles Juanfra, Manuela, José Luis, Elvira y familias. Porque son personas de gran corazón que he tenido la suerte de conocer.*

*A mis amigos: Yumilka, Yaimara, Erich, Yuni, Sutil y Omarita por sus amistades sinceras.*

*A mis profesoras Flora y Rosa porque con cariño infinito me enseñaron el buen camino a seguir.*

## Agradecimientos

Desde niña he vivido con una máxima: “A las estrellas no se sube por caminos llanos”, frase de nuestro apóstol José Martí. Con mucho esfuerzo he logrado todo en mi vida, pero también con mucha ayuda, y con el cariño de muchos familiares y amigos. Por ello, este logro además de ser el fruto de muchos años de estudio, lo es del apoyo de muchas personas, a todas ellas lamentablemente por falta de espacio no podré mencionar.

Quiero agradecer de forma muy especial a dos mujeres extraordinarias que admiro sinceramente, mis tutoras. Hoy he llegado hasta aquí gracias a ellas. Este trabajo se ha nutrido del talento admirable de ambas.

Gracias a Amparo Vila por brindarme su confianza, sus maravillosas ideas y valiosa guía durante todo este tiempo. Ha sido un inmenso orgullo tenerla como tutora. Le agradezco por su invaluable apoyo en las oportunidades que me ha dado de venir a Granada.

Mi agradecimiento sincero también, a mi otra tutora María José Martín, por su amistad, por su constante y decisivo apoyo en todo. Ella ha sido un ejemplo a seguir como profesional y persona.

Agradecer el cariño y ayuda de Miguel Delgado. Contar solo con su presencia me ha dado ánimos en los momentos más difíciles.

Gracias a la generosa colaboración del profesor Verdegay he podido llegar hasta aquí. La terminación de este doctorado también es fruto de su trabajo y buena voluntad. Mis agradecimientos a él, a nuestra querida Rosa, los profesores que nos impartieron docencia, así como a las autoridades de la Universidad de Granada, Universidad de Holguín y Universidad de Camagüey por darme esta oportunidad.

Mi agradecimiento infinito y especial a mi amigo Karel, porque su ayuda ha sido decisiva en la terminación exitosa de ésta tesis.

En ese sentido también ha sido fundamental la ayuda de la doctora Manuela, gracias e ella por el interés y la dedicación prestados en la realización de parte importante de la tesis. De igual forma gracias a Juanfra por dedicarme su tiempo y elevar la calidad expresiva y estética de esta memoria.



A mis compañeros del Departamento de Computación de la Universidad de Camagüey, gracias, porque ellos han tenido que trabajar aún más para que yo pudiera estar acá y lograr este sueño. De forma especial a Karel, Lizardo, Jorge, Julito y Yailé.

Gracias también a todas las personas del Departamento de Ciencias de la Computación e Inteligencia Artificial, particularmente a Miguel Prados, Requena, Carlos Molina, Jesús, Nacho y Raúl y al resto de los profesores.

Al profesor Víctor Herrero de la facultad de Comunicación y Documentación de la Universidad de Granada, por su gentil ayuda para la evaluación del modelo.

Me gustaría agradecer el apoyo de mis compañeros de doctorado y amigos cubanos: Yoel y Enrique. De igual forma agradecer a mis compañeros becarios: María, Clara, Aida, Miguel y Sergio, en especial a los queridos amigos Javi y Mariola.

Agradecer sinceramente a los amigos José Luis Villena y Elvira, Rocío y Dani, a Naima y Arthur por todo el cariño y ayuda prestados.

Por último y de manera especial gracias a Juanfra, Manuela y familias. Agradezco la acogida que me han dado. Entre sus familiares en especial a María, María Rosa e Higinio. Son personas de gran bondad y altruismo.

A todos lleguen mis más sinceros agradecimientos.

*Granada, Julio de 2010.*

# Índice general

<b>1. Introducción</b>	<b>21</b>
1.1. Planteamiento del problema . . . . .	23
1.2. Marco de trabajo . . . . .	25
1.3. Objetivos . . . . .	27
1.4. Aportaciones . . . . .	28
1.5. Contenidos de la memoria . . . . .	29
<b>2. Data warehousing con fuentes de información heterogéneas</b>	<b>31</b>
2.1. Data warehousing y OLAP . . . . .	32
2.1.1. El modelo de datos multidimensional . . . . .	34
2.2. Estudios previos sobre data warehousing y procesamiento textual . . . . .	39
2.2.1. Estudios sobre data warehousing que utilizan XML para el procesamiento textual . . . . .	40
2.2.2. Estudios sobre data warehousing que utilizan técnicas de descubrimiento de conocimiento para la exploración de textos . . . . .	46
2.2.2.1. Propuestas de DW de documentos . . . . .	47
2.2.2.2. Resumen de otras propuestas que utilizan técnicas de descubrimiento de conocimiento para la exploración de textos . . . . .	53

2.2.2.3. Estudios relacionados con fuentes de datos médicos . . . . .	56
2.3. Discusión . . . . .	59
2.4. Conclusiones . . . . .	61
<b>3. Propuesta de modelo multidimensional con manejo semánti- co de datos textuales</b>	<b>65</b>
3.1. Antecedentes . . . . .	66
3.1.1. Concepto de estructura-AP y operaciones asociadas (Martínez, 2008) . . . . .	66
3.1.2. Definición y propiedades de los conjuntos-AP . . . . .	67
3.1.3. Definición y propiedades de una estructura-AP . . . . .	69
3.1.3.1. Acoplamiento de conjuntos de términos con las estructuras-AP . . . . .	72
3.1.4. Proceso de transformación de un atributo textual a un atributo-AP . . . . .	74
3.2. Definición de una dimensión asociada a un atributo-AP . . . . .	75
3.3. Operaciones sobre una estructura-AP asociada a consulta . . . . .	79
3.3.1. Definición de jerarquía de consultas . . . . .	80
3.3.2. Relaciones entre la forma de acoplamiento y las opera- ciones de Drill-Down y Roll-Up definidas previamente . . . . .	86
3.4. Operación Dice para una dimensión-AP . . . . .	88
3.5. Ejemplos que ilustran las definiciones anteriores . . . . .	88
3.6. Conclusiones . . . . .	91
<b>4. Sistema data warehousing con el uso de dimensiones-AP</b>	<b>93</b>
4.1. Descripción del sistema . . . . .	94
4.1.1. Arquitectura general del sistema . . . . .	95

---

4.1.2.	Preprocesamiento de la BD Inicial con atributos textuales	97
4.1.3.	Arquitectura de Wonder v3.0 . . . . .	98
4.2.	Definición de un cubo de datos en Wonder con el uso de dimensiones-AP . . . . .	101
4.2.1.	Formación del cubo de datos . . . . .	101
4.2.2.	Obtención de una dimensión-AP a partir de un atributo-AP . . . . .	102
4.3.	Data warehousing de publicaciones científicas . . . . .	103
4.3.1.	Origen y descripción del modelo de datos a utilizar . . . . .	106
4.3.2.	Transformación del modelo relacional al modelo dimensional . . . . .	107
4.3.3.	Ejemplos implementados . . . . .	109
4.4.	Data warehousing con datos médicos . . . . .	122
4.4.1.	Origen y descripción del modelo de datos a utilizar . . . . .	123
4.4.2.	Transformación del modelo relacional al modelo dimensional . . . . .	125
4.4.3.	Ejemplos . . . . .	126
4.5.	Conclusiones . . . . .	131
<b>5.</b>	<b>Evaluación del modelo</b>	<b>133</b>
5.1.	Utilidad en bases de datos médicas . . . . .	134
5.1.1.	Criterios técnicos de especialistas . . . . .	134
5.1.2.	Gestores de datos médicos . . . . .	138
5.2.	Utilidad en bases de datos de publicaciones científicas . . . . .	142
5.2.1.	Bibliometría . . . . .	143
5.3.	Conclusiones . . . . .	147

<b>6. Conclusiones y trabajos futuros</b>	<b>151</b>
6.1. Conclusiones . . . . .	151
6.2. Trabajos futuros . . . . .	155
<b>A. Manual de usuario de Wonder 3.0</b>	<b>157</b>
A.1. Características de Wonder 3.0 . . . . .	157
A.1.1. Requerimientos para su ejecución . . . . .	158
A.2. Guía de explotación . . . . .	158
A.2.1. Conexión al servidor de datos . . . . .	158
A.2.2. Vistas de la aplicación . . . . .	160
A.2.3. Crear un cubo . . . . .	163
A.2.4. Mostrar un gráfico . . . . .	167
A.2.5. Mostrar un informe . . . . .	168
A.2.6. Consultar un cubo . . . . .	170
A.2.7. Operaciones Roll-up y Drill-Down sobre una dimensión	171
<b>Bibliografía</b>	<b>179</b>

# Índice de figuras

2.1. Arquitectura del proceso data warehousing . . . . .	33
2.2. Sistema general de la arquitectura . . . . .	42
2.3. Visión general del prototipo del sistema . . . . .	44
2.4. Arquitectura de un WoW . . . . .	45
2.5. Arquitectura de un warehousing de documentos . . . . .	48
2.6. El número de veces que un autor es citado en una particular conferencia. . . . .	50
2.7. El mismo análisis de la figura 2.6 pero con las mejores palabras claves de las publicaciones donde ha sido citado cada autor. . . . .	51
2.8. Ventana OLAP que muestra un R-Cubo . . . . .	52
3.1. Retículo de un Conjunto-AP . . . . .	68
3.2. Estructura-AP global . . . . .	70
3.3. Ejemplo básico del uso de las operaciones Roll-Up y Drill-Down	85
3.4. Ejemplo del uso de las operaciones Roll-Up y Drill-Down en una jerarquía de consultas . . . . .	85
4.1. Arquitectura general del sistema . . . . .	95
4.2. Proceso de transformación del atributo textual a la dimensión- AP . . . . .	97

4.3. Arquitectura en capas de Wonder . . . . .	99
4.4. Proceso de obtención de cubos de datos con dimensiones-AP .	101
4.5. Relación de un atributo-AP con una dimensión-AP asociada .	103
4.6. Modelo relacional de tablas de la base de datos de publicaciones	106
4.7. Modelo dimensional de los datos de publicaciones . . . . .	110
4.8. Interfaz principal de Wonder . . . . .	111
4.9. Resultado de un Dice por revista al cubo del Ejemplo 7 . . . .	113
4.10. Cantidad de artículos que contienen en sus títulos, acoplamiento fuerte con determinadas frases de consultas . . . . .	114
4.11. Cantidad de artículos que contienen en sus títulos, acoplamiento débil con determinadas frases de consultas . . . . .	115
4.13. Resultados obtenidos con el uso de la Jerarquía de consultas con frases menos finas y acoplamiento fuerte . . . . .	115
4.12. Jerarquía correspondiente a la consulta (COMPUTATIONAL, MODEL), (DATABASES, DEDUCTIVE, FUNCTIONAL, LOGIC) . . . . .	116
4.14. Resultados obtenidos con el uso de la Jerarquía de consultas con frases menos finas y acoplamiento débil . . . . .	117
4.15. Resultados del Ejemplo 9 . . . . .	117
4.16. Resultados de aplicar acoplamiento débil en el Ejemplo 10 . .	119
4.17. Resultados de aplicar acoplamiento fuerte en el Ejemplo 10 . .	120
4.18. Resultados del uso de la cantidad promedio de páginas como medida, en el Ejemplo 10 . . . . .	120
4.19. Jerarquía de consulta asociada a las frases del Ejemplo 10 . .	121
4.20. Resultados de realizar una consulta más detallada con el uso de la jerarquía de consulta asociada . . . . .	122
4.21. Resultados Ejemplo 11 . . . . .	122
4.22. Modelo relacional de los datos médicos . . . . .	124

---

4.23. Modelo dimensional de los datos . . . . .	126
4.24. Resultados del Ejemplo 12 . . . . .	127
4.25. Resultados del Ejemplo 12 aplicando acoplamiento débil . . . . .	128
4.26. Jerarquía asociada a la consulta del Ejemplo 12 . . . . .	129
4.27. Resultados del uso de la jerarquía de consulta del Ejemplo 12 . . . . .	129
4.28. Resultados del uso de la jerarquía de consulta del Ejemplo 12 con frases menos detalladas . . . . .	129
4.29. Resultados del Ejemplo 13 donde se muestra el acoplamiento con los tipos de anestias . . . . .	131
4.30. Resultados del Ejemplo 13 después de usar la jerarquía de consulta para obtener la frase final CLAVOS ENDER OS-TEOSÍNTESIS. . . . .	131
5.1. Tratamientos quirúrgicos relacionados con el cáncer de mama . . . . .	135
5.2. Tipos de amputaciones relacionados con el diagnóstico pie diabético . . . . .	135
5.3. Cantidades de intervenciones relacionadas con los diagnósticos melanoma y espinocelular . . . . .	136
5.4. Cantidad de implantes en las intervenciones de catarata . . . . .	137
5.5. Ventana de Wonder donde se realiza la operación Dice . . . . .	138
5.6. Ventana de Wonder de la opción Consulting Cube donde se puede corroborar el Dice . . . . .	139
5.7. Número de mujeres y hombres con el diagnóstico fractura . . . . .	140
5.8. Comportamiento de las urgencias en los días de la semana de Enero-Marzo . . . . .	140
5.9. Promedio de espera, de atención en urgencias asociado a las causas: Agresión, Accdte Tráfico y Accdte de Trabajo . . . . .	141
5.10. Promedio de espera, de atención en urgencias en los días de la semana y el primer trimestre del año . . . . .	142



5.11. Cantidad de urgencias en una semana relacionadas con las urgencias que tienen las frases: DOLOR TORÁCICO, VÓMITOS y FRACTURA . . . . .	143
5.12. Jerarquía de consulta asociada a la consulta del Ejemplo 21 .	144
5.13. Resultados de usar la jerarquía del Ejemplo 21 con frases más detalladas . . . . .	145
5.14. Cantidad de artículo publicados en los últimos 5 años del Ejemplo 22 . . . . .	145
5.15. Cantidad de publicaciones del grupo Soft Computing y Sistemas de Información Inteligentes . . . . .	146
5.16. Cantidad de artículos de los últimos 5 años que tienen acoplamiento débil con las frases FUZZY DATABASES y PATTERN RECOGNITION . . . . .	147
5.17. Jerarquía asociada al Ejemplo 23 . . . . .	148
5.18. Resultado del uso de la jerarquía de consulta del Ejemplo 23 .	149
5.19. Autores que han publicado exactamente sobre el tema (DATABASES, FSQL, FUZZY) . . . . .	149
A.1. Icono para realizar la conexión con el servidor de datos . . . .	159
A.2. Ventana de conexión con el servidor de datos . . . . .	160
A.3. Interfaz principal de Wonder . . . . .	161
A.4. Vista OLAP de la aplicación . . . . .	162
A.5. Vista OLAP con su menú contextual . . . . .	162
A.6. Vista de propiedades . . . . .	163
A.7. Primera ventana para crear un cubo: Establecer conexión con la fuente de datos . . . . .	164
A.9. Definición y mapeo de las dimensiones . . . . .	164
A.8. Definición del nombre del cubo y mapeo con la fuente de datos	165
A.10. Definición de una dimensión-AP . . . . .	166

---

A.11. Definición de las medidas del cubo . . . . .	167
A.12. Definición del título y tipo de un gráfico . . . . .	168
A.13. Configuración de los parámetros de un gráfico . . . . .	168
A.14. Definición inicial de un informe . . . . .	169
A.15. Configuración de las dimensiones por las que estará com- puesto el informe . . . . .	170
A.16. Selección de la medida que se mostrará en el informe . . . . .	171
A.17. Ejemplo de informe . . . . .	172
A.18. Definición inicial de la opción Consulting Cube . . . . .	173
A.19. Configuración de los parámetros necesarios para consultar un cubo . . . . .	173
A.20. Consultando las dimensiones del cubo . . . . .	174
A.21. Resultados que se obtienen de consultar un cubo . . . . .	174
A.22. Uso de una jerarquía clásica . . . . .	175
A.23. Uso de una jerarquía de consulta . . . . .	175
A.24. Obtención de los niveles de la jerarquía de consulta . . . . .	176
A.25. Elección del tipo de acoplamiento a aplicar . . . . .	177



# Índice de tablas

2.1. Medida de conteo con diferentes jerarquías . . . . .	37
2.2. Medida de conteo con diferentes jerarquías . . . . .	37
2.3. Medida de agregación: media aritmética . . . . .	38
2.4. Resumen de estudios sobre data warehousing y OLAP con procesamiento textual . . . . .	63
3.1. Tabla que muestra una porción de una base da datos corre- spondiente a un servicio de urgencias médicas . . . . .	67
3.2. Tabla con atributo-AP correspondiente al Ejemplo 4 . . . . .	77
3.3. Ejemplo de acoplamiento débil . . . . .	78
3.4. Ejemplo de acoplamiento fuerte . . . . .	78
3.5. Ejemplo de cubo bidimensional con acoplamiento débil . . . . .	78
3.6. Ejemplo de cubo bidimensional con acoplamiento fuerte . . . . .	79
3.7. Ejemplo de cubo bidimensional con acoplamiento fuerte y agre- gación tiempo medio . . . . .	79
3.8. Ejemplo de acoplamiento fuerte con $C^1$ . . . . .	89
3.9. Ejemplo de jerarquía-AP . . . . .	89
3.10. Ejemplo de consulta más detallada que $C^1$ con acoplamiento fuerte . . . . .	89
3.11. Ejemplo de consulta menos fina que $C^1$ con acoplamiento fuerte	90

3.12. Ejemplo de consulta menos fina que $C^1$ con acoplamiento fuerte	90
3.13. Ejemplo de acoplamiento débil con $C^1$ . . . . .	90
4.1. Dimensiones y sus jerarquías . . . . .	108

# Capítulo 1

## Introducción

La gestión de la información y el conocimiento es actualmente una actividad estratégica para el éxito de las empresas. Optimizar el manejo de la información se ha convertido en un elemento competitivo de gran poder. Si además, se convierte dicha información en conocimiento, sería la clave para la toma de decisiones.

La gran acumulación de datos en sistemas transaccionales, o simplemente en archivos de oficina constituye un problema, debido a su creciente volumen y diversidad. La dificultad en su aprovechamiento aumenta cuando no se dispone de las herramientas necesarias que posibiliten su consulta y, sobre todo, es difícil decidir cuáles son realmente útiles y reúnen los requisitos de calidad necesarios.

Para poder afrontar estos inconvenientes, surgen los Sistemas de Gestión de Información, que pueden ser muy efectivos cuando son el resultado de contar por un lado con una infraestructura de información bien diseñada, y por otro, con la tecnología adecuada. O sea, no basta con tener toda la información localizada y bien distribuida, si no se tienen las herramientas adecuadas para explotar al máximo el conocimiento que ella encierra.

Los Sistemas de Gestión de Información se caracterizan por su diversidad. En ellos se aprecia la gestión de diferentes tipos de datos, incluidos los textuales, que provienen de fuentes de información heterogéneas y, por lo tanto, varían

también las tecnologías que deben emplearse en cada caso. Esta idea se expresa claramente en la clasificación propuesta en (Raghavan y García-Molina, 2001), donde se tiene en cuenta la gestión con tipos de datos estructurados, semi-estructurados y no estructurados:

1. Sistemas de Recuperación de Textos sobre SGBD Relacionales u Orientadas a Objetos: dentro de éstos se enmarcan los relacionados con los temas de recuperación de información (Hedlund et al., 2001). La información contenida en documentos presenta una gran heterogeneidad y se necesita una formalización de sus estructuras (Llavori et al., 1999). Estos sistemas agrupan la gran variedad de propuestas de motores de recuperación de información (Martínez-Méndez y Rodríguez, 2003), donde se proponen métodos de representación de estos documentos, para luego poder gestionar la información obtenida desde bases de datos relacionales convencionales (Cosijn et al., 2004).
2. Sistemas de Recuperación de Textos en modelos semi-estructurados: de forma similar a la anterior clasificación, estos sistemas se encargan de la gestión de datos que se logran estructurar con un esquema bien definido, como es el formato del Lenguaje Extensible de Marcas (XML) (Berstel y Boasson, 2000). XML brinda la posibilidad de estructurar datos y documentos (Fuhr, 2003), ventaja que se combina con las facilidades de los SGBD que utilizan.
3. Sistemas de Gestión de Base de Datos Semi-estructurados (XML): se refieren a los SGBD Relacionales u Orientados a Objetos que se caracterizan por un manejo eficiente de datos poco estructurados, explotando así grandes posibilidades en la gestión de documentos en formato XML (Shanmugasundaram et al., 1999).

El reto de poder procesar disímiles tipos y fuentes de datos se ve reflejado claramente en la clasificación mostrada. Fuentes de datos como Internet, los crecientes informes de las empresas, los correos electrónicos y ficheros textuales o de otros tipos, reclaman nuevos métodos que garanticen su mejor aprovechamiento, teniendo en cuenta que hay conocimiento implícito en dichas fuentes que no puede extraerse de forma trivial.

Para completar la clasificación anterior, el uso de nuevas tecnologías como Data Warehousing (DW)<sup>1</sup> (Galhardas et al., 2001) , OLAP (Procesamiento Analítico en Línea)(E.F. Codd y Salley, 1993) como paradigma en el campo de la Inteligencia de Negocio, y por otro lado, la Minería de datos (Delgado et al., 2002) o Minería de textos (Justicia, 2004), etc, se requieren dentro de los Sistemas de Gestión de Información, para que éstos puedan adaptarse a diferentes situaciones y necesidades. Todas estas tecnologías están encaminadas a desarrollar los ámbitos de la Gestión del Conocimiento e Inteligencia del Negocio (Nemati et al., 2002). Ambos conceptos han aflorado en las empresas como una ayuda para la eficiencia.

Los datos de tipo texto forman parte de la mayoría de la información que se gestiona en las organizaciones. Como ejemplo de este tipo de datos, tenemos los atributos que contienen texto en las bases de datos convencionales. Por este motivo, el desarrollo de herramientas que combinen tecnologías avanzadas como las que anteriormente relacionamos, se hace cada vez más necesario. Se han alcanzado avances significativos en este campo, como por ejemplo sistemas de Minería de texto (Feldman y Sanger, 2007) y/o Minería de datos (Han y Kamber, 2000) que brindan valiosas alternativas de búsqueda de información en este tipo de datos sin estructura definida. Si estas soluciones se insertan dentro de sistemas de ayuda a la toma de decisiones, como OLAP, las prestaciones de estos últimos aumentan considerablemente. Todavía queda por descubrir un amplio número de soluciones, donde se combinen las tecnologías necesarias para extraer de la información textual todo el conocimiento posible.

## 1.1. Planteamiento del problema

La proliferación de los sistemas DW y OLAP nos abre las puertas al mundo del análisis multidimensional de los datos, más rico en información útil para la toma de decisiones. Este análisis aporta grandes ventajas cuando se trata

---

<sup>1</sup>Este término tiene su equivalente en español como "Almacenamiento de datos", pero hemos optado por dejarlo en inglés debido a que Data Warehousing es un término más acuñado en el campo de la Informática, con un significado más específico que "Almacenamiento de datos".



de los atributos valor contenidos en las bases de datos relacionales, pero no está previsto de igual manera para el caso de los atributos textuales que no cuentan con valores discretos, sino que se caracterizan por la falta de estructura y homogeneidad.

Por ejemplo, supóngase que se tiene un sistema para procesar encuestas dedicadas a recoger la opinión de los estudiantes universitarios, en cuanto a un tema determinado. En este caso, son manejados con facilidad todos los campos tradicionales que, por lo general, son de selección; pero no sucede así con los que sean textuales. Los campos tales, como la opinión del estudiante o las observaciones sobre determinada elección son cualitativamente valiosos y, al estar formados por textos se hace difícil su procesamiento automático. No obstante, se pueden encontrar trabajos como el de Sandro Martínez (Martínez, 2008) donde se logra procesar automáticamente el conocimiento asociado a textos libres, pero no valoran la utilización de dicho conocimiento dentro de un sistema data warehousing. El presente trabajo está dedicado a estudiar esta dificultad, asociada a los sistemas data warehousing y OLAP, con el fin de proponer una solución.

Se ha alcanzado un gran número de soluciones donde se procesan documentos textuales que provienen de fuentes externas (Jensen et al., 2001). En varios de estos casos, las fuentes de datos suelen estar en formato XML (Rusty, 2001) o las estructuras extraídas de los documentos se modelan en XML, para facilitar su manejo (Niemi et al., 2003). Además, se involucran indistintamente técnicas de análisis de datos o procesamiento sintáctico de textos, sobre las estructuras que se extraen de los documentos (Keith et al., 2006).

En algunos trabajos se ha logrado procesar textos cortos como correos electrónicos o publicaciones, definiendo sus partes en modelos multidimensionales para realizar consultas sobre ellas (Tseng y Chou, 2006). Pero, en ningún caso se ha analizado el resultado de dicho procesamiento textual, dentro de sistemas data warehousing, de igual forma que los datos que manejan valores discretos. El análisis del conocimiento implícito en este tipo de atributos, en sistemas data warehousing con el uso de procesamiento OLAP es la propuesta del presente trabajo.

## Problema a resolver

*El tratamiento de atributos textuales en el entorno data warehousing y OLAP, para realizar operaciones conjuntamente como con cualquier otro tipo de dato. Estas operaciones pueden incluir desde consultas sobre el conocimiento obtenido de los textos hasta resúmenes.*

Para darle solución a este problema de investigación, partimos de la formulación de la siguiente hipótesis.

## Hipótesis

*Se puede obtener un nuevo modelo multidimensional que sea capaz de modelar el conocimiento que contiene los atributos textuales en cubos de datos, a través del uso de técnicas de Minería de Textos que conducen a una nueva Forma Intermedia de Representación: la estructura-AP (Martínez, 2008). Dicho modelo se puede implementar en un sistema data warehousing, donde las operaciones OLAP sobre dimensiones textuales (dimensiones-AP) se realicen conjuntamente y de la misma manera que las dimensiones no textuales.*

Una vez planteada nuestra hipótesis, en la próxima sección se detalla el marco de trabajo en el que se desarrolla la solución al problema planteado.

## 1.2. Marco de trabajo

Este trabajo se enmarca dentro de los temas de data warehousing, OLAP y procesamiento textual. De estos tres campos, el data warehousing y el OLAP mantienen una estrecha relación. Ambas tecnologías surgieron con un interés común: lograr un mejor aprovechamiento de grandes y diversas acumulaciones de información. Existen productos DW que proveen servicios OLAP y utilizando herramientas OLAP, los usuarios pueden acceder al DW mejorando así la comprensión del negocio para la toma de decisiones. En

nuestro caso nos centramos en mayor medida en la parte de las herramientas OLAP y en el modelo multidimensional que implementan.

Por otro lado el procesamiento textual es un campo más específico, referido sólo a textos. El mismo se identifica con variadas técnicas, entre las que se encuentran la Minería de datos y Minería de textos (Weiss et al., 2004). Estas técnicas, en la mayoría de los casos, no están ligadas a procesos data warehousing, a pesar de que comparten un objetivo general común: la extracción de conocimiento.

En esta tesis se han logrado combinar las utilidades de procesos de minería con data warehousing, de la siguiente manera:

Primero se realiza el preprocesamiento necesario, para obtener parte el conocimiento implícito contenido en atributos textuales de bases de datos con el uso del algoritmo Apriori. Este preprocesamiento se realiza de forma automática, con la herramienta Text Mining Tool (Martínez, 2008) y obtiene como resultado nuevas estructuras de conocimiento correspondientes a los atributos textuales de la base de datos, como son la estructura-AP global y las subestructuras-AP inducidas correspondientes a cada tupla del atributo textual original que se procesó; estas subestructuras no son más que los posibles valores del atributo-AP. En este atributo-AP ya se tiene la parte de dicha estructura de conocimiento correspondiente a cada tupla del atributo textual original.

Contando con dicha información, se pasa a definir un nuevo modelo multidimensional que brinde soporte a textos con el uso del atributo-AP obtenido. Se transforma en una dimensión-AP tras haber comprobado que un atributo-AP cumple con las condiciones de una dimensión: tener definidos un dominio y una partición. El nuevo modelo, además de definir un nuevo tipo de dimensión textual como dimensión-AP, implementa las operaciones OLAP clásicas para este tipo de dimensiones, de forma tal que se puedan relacionar con las del tipo clásico sin ninguna dificultad.

Por último se implementa una herramienta OLAP, llamada Wonder OLAP Server v3.0, en adelante Wonder v3.0, que lleva a la práctica el nuevo modelo. Este sistema se utiliza en el desarrollo de sistemas data warehousing en dos entornos diferentes, uno el de publicaciones científicas y otro el entorno médico. Los data warehousing implementados muestran la utilidad del modelo y

el buen funcionamiento de Wonder.

### 1.3. Objetivos

Como se comentó anteriormente, el objetivo fundamental de este trabajo es la definición e implementación de un nuevo modelo multidimensional, con soporte a atributos textuales en sus dimensiones. Para ello, se utilizarán algunas de las definiciones relacionadas con una estructura extraída de los textos, llamada estructura-AP.

Todo lo anteriormente expuesto se concreta en una serie de objetivos específicos o tareas a realizar que son las siguientes:

- *Realizar un estudio bibliográfico para conocer el estado del arte sobre data warehousing y OLAP, relacionados con el procesamiento de datos textuales.* Con ello pretendemos situarnos en el estado actual del problema a resolver y estudiar las técnicas necesarias dentro de la solución planteada.
- *Formalizar matemáticamente el nuevo modelo multidimensional, que facilite el procesamiento del conocimiento implícito de atributos textuales en bases de datos en entorno OLAP.* De este modo se pueden establecer las estructuras y operaciones necesarias para realizar el procesamiento de los cubos que contienen dicha información.
- *Realizar la implementación de un sistema data warehousing que sea capaz de modelar y consultar dichos cubos multidimensionales, con soporte a atributos textuales.* Con ello se demostraría que el modelo propuesto es viable en su implementación.
- *Prueba experimental del funcionamiento de sistemas data warehousing en los entornos bibliográficos y médicos, con datos reales.* Ésto permitiría llevar a la práctica los resultados teóricos previamente obtenidos.
- *Evaluación y refinamiento del sistema y el modelo obtenido con datos reales, utilizando el criterio de expertos.* Por último, para corroborar

la calidad del modelo propuesto y las ventajas que introduce para la realización de consultas sobre atributos textuales en bases de datos en entorno OLAP, se tendrán en cuenta consultas propuestas por expertos, que demuestran la utilidad práctica del modelo y que corroboran la veracidad de la información obtenida.

Un resumen del cumplimiento de estos objetivos específicos se puede ver en el capítulo 6 de esta memoria.

## 1.4. Aportaciones

*Desde el punto de vista teórico* este trabajo realiza las siguientes aportaciones:

1. Se ha dado las definiciones formales de dominio y partición de un atributo-AP, para transformar el mismo en una dimensión-AP. Este dominio se obtiene a partir de la estructura-AP que recoge un conocimiento asociado al atributo textual correspondiente y agrupa todos los posibles valores que puede tener la dimensión-AP.
2. Se ha obtenido un nuevo modelo multidimensional que brinda soporte en sus dimensiones a atributos textuales. Dicho modelo implementa dimensiones-AP referentes a dimensiones textuales, que facilitan la extracción de conocimiento útil para la toma de decisiones.
3. Se redefinieron las operaciones clásicas OLAP para usarlas sobre la dimensión-AP. Entre ellas el Dice y una nueva jerarquía, llamada jerarquía de consulta.

*Desde el punto de vista tecnológico* se han desarrollado las siguientes aportaciones:

1. Se ha desarrollado una herramienta OLAP utilizando software libre, Wonder OLAP Server 3.0 que implementa el nuevo modelo multidimensional propuesto.

2. Se han implementado sistemas data warehousing experimentales para evaluar la utilidad del modelo en diferentes entornos de datos. Dichos entornos son el relacionado con las publicaciones científicas y el hospitalario.

## 1.5. Contenidos de la memoria

La memoria está compuesta por 6 capítulos que cumplen los objetivos antes expuestos. En primer lugar, tras esta introducción, se valoran estudios previos relacionados con el tema que nos ocupa. En el **Capítulo 2 Data warehousing con fuentes de información heterogéneas** se discutirá sobre el entorno que caracteriza el problema abordado y los tipos de soluciones que aparecen en la literatura.

En el **Tercer Capítulo: Propuesta de modelo multidimensional con manejo semántico de datos textuales**, nos centramos en la definición de los nuevos dominios de datos para el caso de dimensiones textuales, con el uso de las estructuras ya obtenidas en (Martínez, 2008). También se introduce un conjunto de operaciones que permitirá realizar consultas sobre dichas dimensiones textuales. A lo largo del capítulo se ejemplifica cada definición, propiedad u operación que se proponen.

El **Cuarto Capítulo: Sistemas Data warehousing con el uso de dimensiones-AP** comienza con la presentación de la arquitectura del sistema y luego se desarrolla la implementación del modelo propuesto a través de dos sistemas data warehousing, relacionados con dos entornos de datos diferentes, el entorno de las publicaciones científicas y el hospitalario.

En el **Quinto Capítulo: Evaluación del modelo**, se efectúa la validación del modelo teniendo como premisa su utilidad práctica en sistemas data warehousing. Para ello se realizan consultas en los data warehousing obtenidos, propuestas por expertos.

Por último, en el **Sexto Capítulo: Conclusiones y trabajos futuros**, se encuentran las conclusiones obtenidas con la realización de este trabajo y las líneas de investigación a seguir en el futuro.

Para facilitar la mejor comprensión de este trabajo se ha incluido un apéndice donde se recogen detalles de las funcionalidades del sistema. En el **Apéndice A: Manual de usuario de Wonder**, se detallan todas las prestaciones del sistema OLAP implementado Wonder V3.0, de forma gráfica y escrita.

## Capítulo 2

# Data warehousing con fuentes de información heterogéneas

En el presente capítulo se realiza un estudio de las principales tendencias de data warehousing y OLAP, relacionados con el procesamiento de datos textuales. Con ello pretendemos situarnos en el estado actual del problema que se pretende resolver y estudiar las técnicas y soluciones afines propuestas en los estudios más recientes.

Se comienza exponiendo las principales características de los procesos de data warehousing y OLAP, haciendo énfasis en la estrecha relación entre ambos. En la siguiente sección se describe brevemente el modelo multidimensional clásico que implementan ambas tecnologías y más adelante se comienza con la exposición de los estudios previos encontrados.

Los trabajos analizados se han agrupado en dependencia de las técnicas que usan para procesar los textos, un grupo se conforma por los estudios sobre data warehousing que utilizan XML para el procesamiento textual. El otro grupo está compuesto por los trabajos sobre data warehousing que utilizan técnicas de descubrimiento de conocimiento para la exploración de textos.



## 2.1. Data warehousing y OLAP

En la actualidad existen nuevas premisas en la forma de utilizar los sistemas computacionales. Ya no son suficientes los sistemas transaccionales que procesan y gestionan de grandes cantidades de datos; se han hecho necesarios sistemas informáticos que incluyan Gestión del Conocimiento e Inteligencia del Negocio que ayuden a la toma informada de decisiones (Shim et al., 2002). Por este motivo las empresas han concentrado grandes recursos en un nuevo concepto tecnológico: Data warehousing.

Data warehousing (Trujillo y Song, 2008) se define como el proceso por el cual las empresas extraen sentido y significado de sus datos, a través del uso de un repositorio de datos o data warehouse (Wolff, 2002). OLAP es una tecnología que ayuda a los trabajadores del conocimiento a hacer con rapidez sus procesos empresariales y la toma de decisiones; permite a analistas y ejecutivos analizar los datos rápidamente, de forma interactiva y teniendo en cuenta varias entidades del negocio (Vassiliadis y Sellis, 1999), (Thomsen, 2002).

Los datos existentes en un sistema DW por lo general se manejan por medio de uno o varios servidores OLAP, como se puede apreciar en la figura 2.1; esta es una idea que ilustra claramente la fuerte relación entre ambos conceptos. Estos servidores presentan vistas multidimensionales de los datos a una gran variedad de interfaces: de consulta directa, herramientas para generar informes, herramientas de análisis exploratorio (gráficos, estadística descriptiva, etc.) y herramientas de minería de datos propiamente dichas. Todo esto hace que con los servidores OLAP y sus robustas máquinas de cálculo, los datos históricos almacenados en el DW sean mucho mejor utilizados.

La capacidad que OLAP tiene de poder integrar metadatos con una base de datos relacional, la hace candidata a ser utilizada en sistemas que necesitan extraer los datos de fuentes externas, modelarlos con el uso de metadatos y combinarlos con otros tipos de datos 2.1, para al final obtener un modelo de datos multidimensional. En general, todos los sistemas OLAP cumplen

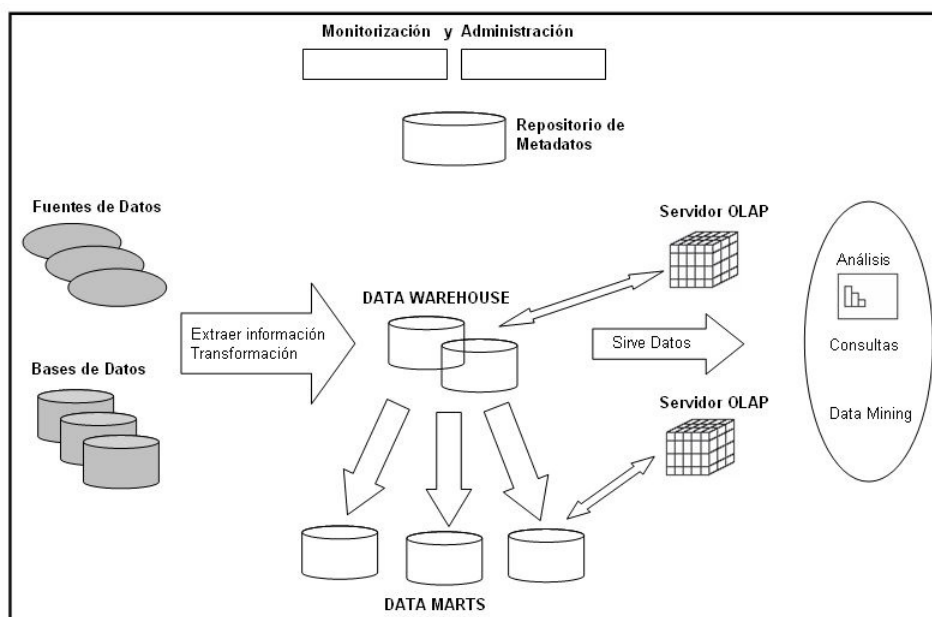


Figura 2.1: Arquitectura del proceso data warehousing

con la capacidad de tratar gran cantidad de datos modelados con un número ilimitado de dimensiones, y sus tiempos de respuesta a consultas muy rápidos los hace una buena opción tecnológica.

Si al proceso de extraer conocimiento de los datos, se le relaciona con técnicas de Minería de Datos y/o Minería de Textos (Delgado et al., 2002), y además se le incluyen procesamientos OLAP, podría lograrse un sistema data warehousing de importantes prestaciones; esta idea forma parte de la motivación del presente trabajo.

A continuación repasamos las características del modelo de datos multidimensional y seguidamente revisamos varios trabajos sobre data warehousing y OLAP relacionados con procesamiento textual. Al final del capítulo se incluye una discusión sobre las propuestas de la literatura y las conclusiones.

### 2.1.1. El modelo de datos multidimensional

Recordemos el modelo de datos multidimensional "clásico" (Pedersen y Jensen, 2001). Éste es utilizado tanto en los sistemas data warehousing como en la tecnología OLAP. En dicho modelo se contemplan los siguientes elementos.

- Un conjunto  $d_1, \dots, d_n$  de "dimensiones" definidas en una base de datos. Es decir, atributos con dominio discreto pertenecientes al esquema de dicha base de datos, a partir de los cuales se van a agrupar los datos. Cada dimensión  $d_i$  tiene asociados:
  - Un dominio básico  $D_i = \{x_1 \dots x_{m_i}\}$ , de valores discretos tal que, cada tupla  $t$  de la base de datos toma en el atributo  $d_i$  un valor único y bien determinado  $x_i$  notaremos  $d_i[t] = x_i$
  - Una jerarquía de agrupamiento que permite considerar valores de análisis distintos. Dicha jerarquía  $\mathcal{H}_i = \{\mathcal{C}_{i1} \dots \mathcal{C}_{il}\}$  está formada por particiones de  $D_i$  de forma que:

$$\forall k \in \{1, 2 \dots l\} \mathcal{C}_{ik} \subseteq \mathcal{P}(D_i) \quad \mathcal{C}_{ik} = \{X_{ik}^1, \dots, X_{ik}^h\}$$

siendo

$$\forall j, r \quad X_{ik}^j \cap X_{ik}^r = \emptyset \quad \text{y} \quad \bigcup_{j=1}^h X_{ik}^j = D_i$$

La jerarquía  $\mathcal{H}_i$  es un retículo de inclusión cuyo elemento minimal es  $D_i$ , considerado elemento a elemento y el maximal es  $D_i$  considerado como una partición de un solo elemento.

- $V$  una medida numérica asociada a estas dimensiones, de tal manera que siempre se pueda obtener  $V = f(Y_1, Y_2, \dots, Y_n)$  donde  $Y_1, \dots, Y_n$  son "valores" de las dimensiones antes consideradas. Hay que tener en cuenta que dichos valores pueden no ser exactamente los del dominio sino los de alguna partición de la jerarquía. Es decir, si en la dimensión  $d_i$  consideramos el nivel  $\mathcal{C}_{ik}$ , entonces  $Y_i \in \mathcal{C}_{ik}$ . Esta medida  $V$  puede ser:

- Una medida de "conteo" que nos da el número de tuplas de la base de datos que verifican:  $\forall i \in \{1, \dots, n\} \quad d_i[t] \in Y_i$
- Cualquier otro atributo numérico que semánticamente se asocie a las dimensiones consideradas.

La estructura  $(V, Y_1, Y_2, \dots, Y_n)$  se denomina cubo de datos y en el caso de que  $n = 2$  se representa como una tabla, si  $n = 3$ , se representa a veces como un cubo tridimensional, de ahí el nombre.

- Existe también un criterio de agregación de  $V$ ,  $AGG$ , que se aplica cuando se consideran valores "de conjunto" en alguna de las dimensiones. Es decir:

$$V = f(x_1, \dots, Y_k, \dots, x_n) = AGG_{x_k \in Y_k} f(x_1, \dots, x_k, \dots, x_n)$$

$AGG$ , puede ser sumatorio o alguna función de tipo estadístico como, la media  $AVG$ , desviación típica  $STD$ , etc; obviamente si la medida es de conteo, la función de agregación es  $SUM$ .

El siguiente ejemplo muestra como se construye un modelo multidimensional a partir de una determinada base de datos.

**Ejemplo 1** Consideremos la siguiente tabla de una base de datos médica:

n-enfermo	edad	fecha-ingreso	procedencia
1	25	12/02/2009 V	Granada
2	20	15/03/2009 J	Córdoba
3	10	10/06/2009 D	Albacete
4	50	08/02/2009 L	Álava
5	3	10/10/2009 X	Madrid
6	12	05/03/2009 L	Badajoz
7	65	07/12/2009 V	Madrid
8	45	06/11/2009 M	Murcia
9	70	05/01/2009 V	Jaén
10	18	05/05/2009 S	Sevilla
11	8	01/06/2009 V	León
12	70	25/10/2009 J	Salamanca
13	30	30/03/2009 V	Toledo
14	20	31/12/2009 L	Madrid
15	26	04/11/2009 D	Valencia
16	44	11/07/2009 X	Castellón
17	35	13/08/2009 L	Sevilla
18	75	14/06/2008 X	Málaga
19	8	08/08/2008 M	Córdoba
20	12	12/12/2008 M	Alicante
21	15	12/11/2008 D	Barcelona
22	10	10/03/2008 V	Granada
23	17	15/04/2008 S	Jaén
24	20	15/02/2008 D	Cádiz
25	30	12/07/2008 X	Coruña

Consideremos también que las dimensiones son fecha-ingreso y procedencia, siendo las jerarquías para éstas las siguientes:

- Para fecha-ingreso:
  - Día-semana={L,M,X,J,V,S,D} que a su vez se agrupa según Situación={laborable={L,M,X,J,V}, festivo{S,D}}
  - Mes={0,1,2,...,12} que se agrupa a su vez en Estación={Primavera{03,04,05}, Verano{06,07,08}, Otoño{09,10,11}, Invierno{12,01,02}}
  - Año={2008, 2009}
  
- Para procedencia:
  - Comunidad autónoma
  - Situación={Sur-este, Sur-oeste, Centro, Levante, Nor-este, Nor-oeste}

	Sur-este	Sur-oeste	Centro	Levante	Nor-este	Nor-oeste	Total
01	1						1
02	1	1			1		3
03	1	1	2				4
04	1						1
05		1					1
06	1		1			1	3
07				1		1	2
08		2					2
09							
10			2				2
11				2	1		3
12			2	1			3
Total	5	5	7	4	2	2	25

Tabla 2.1: Medida de conteo con diferentes jerarquías

	Sur-este	Sur-oeste	Centro	Levante	Nor-este	Nor-oeste	Total
Primavera	2	2	2				6
Verano	1	2	1	1		2	7
Otoño			2	2	1		5
Invierno	2	1	2	1	1		7
Total	5	5	7	4	2	2	25

Tabla 2.2: Medida de conteo con diferentes jerarquías

Las medidas a considerar van a ser edad y conteo de tuplas. En estas condiciones, si suponemos que la fecha se agrupa en meses y la procedencia en situación, el cubo de conteo quedaría como la tabla 2.1, si agrupamos ahora los meses por estaciones, el cubo de conteo quedaría como en la tabla 2.2.

Si consideramos ahora como variable dependiente la edad y como medida de agregación la media aritmética, agrupando como en el caso anterior, obtendríamos el siguiente cubo de datos que se refleja en la tabla 2.3.

A partir del concepto de cubo de datos se definen una serie de operaciones que corresponden a las distintas posibilidades de análisis sobre las dimensiones. Las operaciones más comunes son:

- Roll-up** resume los datos del cubo "ascendiendo" en las jerarquías de una determinada dimensión de una partición más específica a otra más genérica. Un ejemplo de roll-up, aparece en las tablas 2.1 y 2.2 del ejemplo anterior, donde se pasa de meses a estaciones en la dimensión

	Sur-este	Sur-oeste	Centro	Levante	Nor-este	Nor-oeste	Total
01	70						70
02	25	20			1		42.5
03	10	20	21				18
04	17						17
05		18					18
06	75		10			8	31
07				46		30	38
08		21.5					21.5
09							
10			36.5				36.5
11				35.5	15		28.6
12			42.5	12			32,3
Total	19.5	20.2	30	31.7	32.5	19	

Tabla 2.3: Medida de agregación: media aritmética

tiempo.

- **Drill-down** es la operación inversa, pasando de una partición más genérica a otra más específica. Un ejemplo de drill-down sería pasar de situación a comunidad autónoma en la dimensión lugar en la tabla 2.3 del ejemplo anterior.
- Existen otras operaciones como **Slice** y **Dice** que lo que hacen es restringir la información a una parte de la base de datos. Por ejemplo, estudiar la información referida a un año concreto o a Andalucía.

En la definición de cubo y sus operaciones se considera que las dimensiones están definidas en dominios de valores discretos bien establecidos sobre los que se construyen las jerarquías. Si se consideran otros tipos de dominios es necesario extender este concepto.

Teniendo en cuenta un dominio textual, dada la falta de estructura de los textos, es preciso primeramente obtener una estructura y sobre esta dimensión textual definir además del dominio, la partición y crear nuevas formas de obtener la jerarquía asociada.

A continuación exponemos algunos trabajos sobre data warehousing que están relacionados con el procesamiento textual. En ellos se puede apreciar la importancia del uso del modelo de datos multidimensional para extraer información de datos textuales.

## 2.2. Estudios previos sobre data warehousing y procesamiento textual

Un alto porcentaje de la información que se gestiona en las organizaciones es de tipo textual. Sin embargo este tipo de información es difícil de procesar debido a la falta de homogeneidad y de estructura; muchas veces es texto libre que, por ejemplo, puede formar parte de una base de datos como un atributo textual. Debido a los inconvenientes antes mencionados estos atributos textuales cuentan con formas de consulta limitadas.

Como ya comentamos en la introducción, una de las debilidades de los sistemas data warehousing y puede decirse, de todos los sistemas de gestión de información, es la falta de funcionalidades para el procesamiento de los datos de tipo textual, especialmente desde el punto de vista semántico. Es importante resaltar que la mayoría de los trabajos plantean el procesamiento textual, no en atributos, sino en documentos textuales como fuente de datos externa. Algunos trabajos combinan técnicas de Minería de Datos o Minería de Textos con los sistemas OLAP (Cody et al., 2002), de manera que los datos son preprocesados antes de ser modelados en los cubos.

Realizar una clasificación para agrupar las propuestas de sistemas data warehousing no es una tarea sencilla, debido a la gran diversidad de soluciones que se pueden encontrar. Se aprecia la falta de estándares de manejo de datos y de herramientas generales que sean capaces de enriquecer los procesos clásicos de data warehousing en lugar de modificarlos para obtener resultados específicos.

Los estudios revisados siguen dos estrategias generales para procesar información textual. Una estrategia consiste en usar fundamentalmente el lenguaje XML para estructurar los documentos textuales y facilitar su manejo; la otra consiste en el uso de técnicas de extracción de conocimiento, como Minería de datos o IR para extraer de los datos textuales algo más que su información explícita. Nuestra propuesta sigue la segunda estrategia.

Aún así hemos agrupado los trabajos analizados en los grupos siguientes.

- *Estudios sobre data warehousing que utilizan XML para el procesamien-*



*to textual.*

Esta clase agrupa aquellos trabajos sobre data warehousing que no extraen de los textos más información que la que explícitamente pueden ofrecer. Éstos emplean recursos como los lenguajes XML y UML para estructurar y modelar las fuentes de datos textuales externas o internas que son manipuladas, y así poder procesarlas.

- *Estudios sobre data warehousing que utilizan técnicas de descubrimiento de conocimiento para la exploración de textos.*

A esta clase pertenecen los trabajos donde se vinculan los temas de data warehousing con técnicas de descubrimiento de conocimiento como Minería de Datos, Minería de Textos (Ahonen et al., 1998) o Recuperación de Información (Salton y McGill, 1983). Toman como fuentes de datos externas los documentos textuales o parten de Bases de Datos textuales multidimensionales. Estas fuentes son procesadas con técnicas de minería para analizar de ellas el conocimiento que encierran.

### **2.2.1. Estudios sobre data warehousing que utilizan XML para el procesamiento textual**

A continuación presentamos trabajos sobre data warehousing que combinan el uso de lenguajes como XML y UML para analizar documentos textuales. El formato XML puede ser usado como formato de los datos de entrada y para dar estructura a varios tipos de documentos textuales, facilitando así su procesamiento. Es importante señalar que en este grupo los documentos textuales tienen que tener alguna estructura, como XML, por lo tanto se puede decir que el procesamiento textual que realizan se basa en la extracción de las partes del documento que forman parte de la estructura predefinida.

En (Jensen et al., 2001) se presenta un acercamiento a la especificación de las bases de datos OLAP basadas en datos Web, usando el Lenguaje de Modelado Unificado (UML) para la captura multidimensional de los datos, y una arquitectura que permite la integración de XML con las fuentes de datos que usa la herramienta OLAP.

Los autores de (Jensen et al., 2001) presentan una arquitectura para integrar los datos de XML a nivel conceptual que además soporta las fuentes de los datos relacionales, ayudando así la construcción de almacenes de datos (Data Warehouse). Estos almacenes están basados en parte, por datos relacionales y en parte por datos XML disponibles en la red. Actualmente, está desarrollándose un prototipo de sistema que soporte esta arquitectura, como parte de un proyecto que investiga las tecnologías necesarias para construir DW de datos web. Luego, propone una especificación precisa de una Base de datos OLAP (BD OLAP) basada en múltiples XML y/o fuentes de datos relacionales. Esta especificación se modela en un diagrama tipo copo de nieve usando el UML para luego obtener el modelo multidimensional de la base de datos. Usar diagramas de UML para describir y visualizar la estructura lógica de documentos XML, es un método que se usa para facilitar el diseño de las BD OLAP. Y por último, describen las consideraciones especiales que necesitan ser tomadas en cuenta al diseñar una BD OLAP sobre los datos de XML, tales como las dimensiones con las jerarquías, asegurando así la correcta agregación de datos.

Lo anteriormente explicado se describe en un caso de estudio que trata sobre vendedores y suministradores de componentes electrónicos.

En la figura 2.2 se describe el sistema general de la arquitectura, donde se integran los XML y las fuentes de datos relacionales a un nivel conceptual.

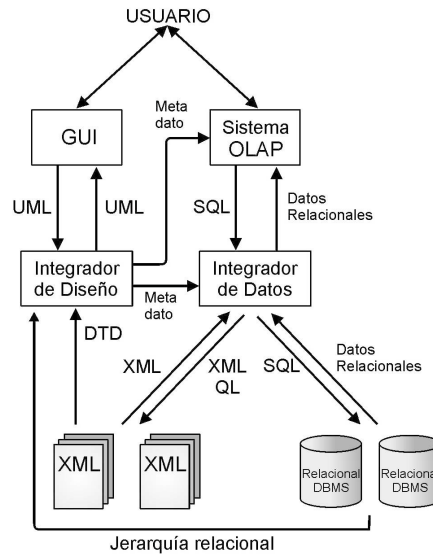


Figura 2.2: Sistema general de la arquitectura

Este trabajo es una muestra del creciente uso de las herramientas OLAP y su interacción con el formato XML para lograr dar una estructura a los documentos textos que se quieren analizar. Además tiene como novedad la combinación de XML con UML en la búsqueda del modelo multidimensional. Pero se limita a proponer un modelo, ya que plantea que la implementación de esta propuesta forma parte de los trabajos futuros. El uso que hacen de XML brinda la posibilidad de procesar sintácticamente diversos tipos de documentos y/o datos textuales, no de aprovechar el conocimiento implícito que hay en ellos. De forma similar ocurre en (Niemi et al., 2003) y los trabajos que siguen.

El trabajo de (Niemi et al., 2003) también trata sobre la relación de XML con cubos OLAP. En el mismo argumentan sobre una aplicación de la tecnología distribuida para la construcción de cubos basados en XML. Tienen en cuenta que, por lo general, los datos de una base de datos OLAP son recolectados desde varios repositorios de datos, como por ejemplo desde bases de datos relacionales que pueden ser muy extensas, aunque en ocasiones estas fuentes de datos son necesarias solo en determinados momentos. Sin embargo, se realiza el almacenamiento de todos esos datos en la base de datos

OLAP y también todos esos datos son actualizados constantemente, lo que supone una tarea de alta demanda que en la práctica afecta en gran medida el funcionamiento del sistema OLAP. A esto se suma el hecho de que la construcción de cubos OLAP puede ser un proceso lento y costoso.

Para resolver estos inconvenientes este trabajo presenta un prototipo de sistema, que aplica la tecnología de grid (malla) o distribuida, para compartir el procesamiento necesario en el proceso de construcción de los cubos OLAP. Como los datos suelen tener diversas fuentes proponen usar XML como formato intermedio de la colección de datos. La definición de nuevos cubos OLAP que en ocasiones necesita de selecciones y agregaciones de datos está distribuída en computadoras que almacenan los datos originales.

Este trabajo es una alternativa de sistemas DW donde se utiliza la tecnología distribuída que ofrece el lenguaje Java, en combinación con XML, para lograr mayor eficiencia en los procesos involucrados en el procesamiento analítico.

En (Turkka et al., 2008) presentan una herramienta para la construcción de cubos de datos desde documentos XML estructuralmente heterogéneos. Este artículo introduce una nueva forma de extracción de datos primarios de alto nivel, que utiliza la estrategia de evaluación de consulta SPC (Smallest Possible Context) o Contexto más pequeño posible. Se usa un prototipo de sistema OLAP de construcción de cubos de datos y se muestra a través de un ejemplo de infométrica. Para la construcción del cubo de datos desde un documento XML, el usuario especifica la operación CREATE CUBE. La figura 2.3 muestra a continuación una visión general del prototipo del sistema.

Según la figura 2.3 primero el usuario especifica su consulta, la misma es compilada por el sistema y si es correcta se analizan los componentes de la expresión. En el punto 3 se genera el conjunto de la expresión y cada componente de la expresión es ejecutada en la colección de documentos XML. Luego en el punto 4 la información extraída se recoge en la memoria en tiempo de ejecución, como un resultado intermedio. En el punto 5 se obtienen los valores de las medidas aplicando las funciones de agregación y por último en el 6 se muestra el cubo de datos resultante.

Este trabajo (Turkka et al., 2008) se destaca por su originalidad al permitir la integración de datos de documentos XML heterogéneos. Además brinda

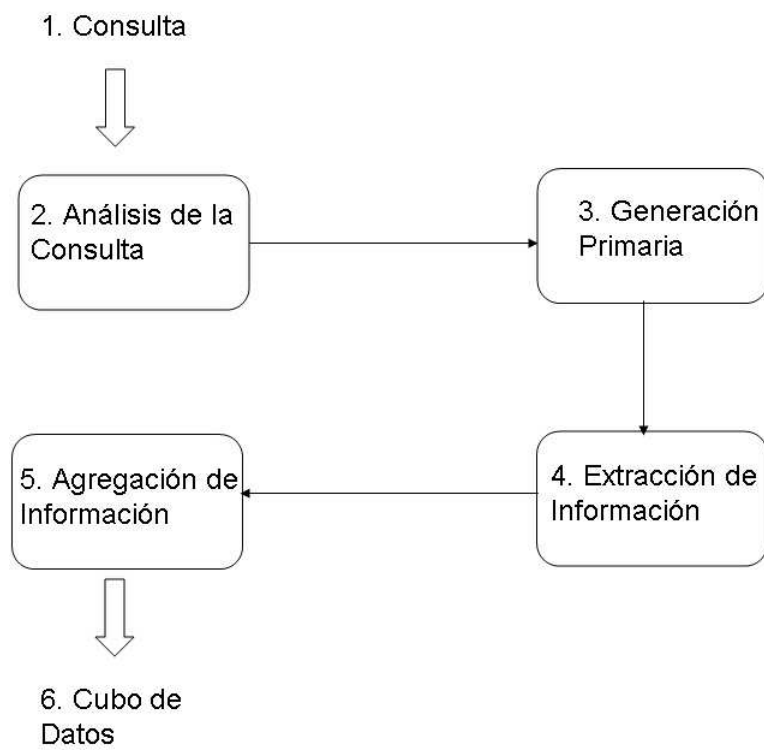


Figura 2.3: Visión general del prototipo del sistema

una valiosa contribución, al evitar que el usuario tenga que conocer de qué forma se representa la información en el XML.

En (Keith et al., 2006) se propone la creación de un Data Warehouse of Words (WoW) o almacén de palabras, para que sea utilizado por una herramienta de usuario OLAP y así poder analizar con rapidez y mayor detalle los grandes archivos de texto electrónicos. El WoW se construye con procesos ETL<sup>1</sup> (Extracción, Transformación y Carga), donde la extracción involucra texto plano y documentos XML que luego son almacenados en cubos de datos, como se muestra en la arquitectura de un WoW, en la figura 2.4.

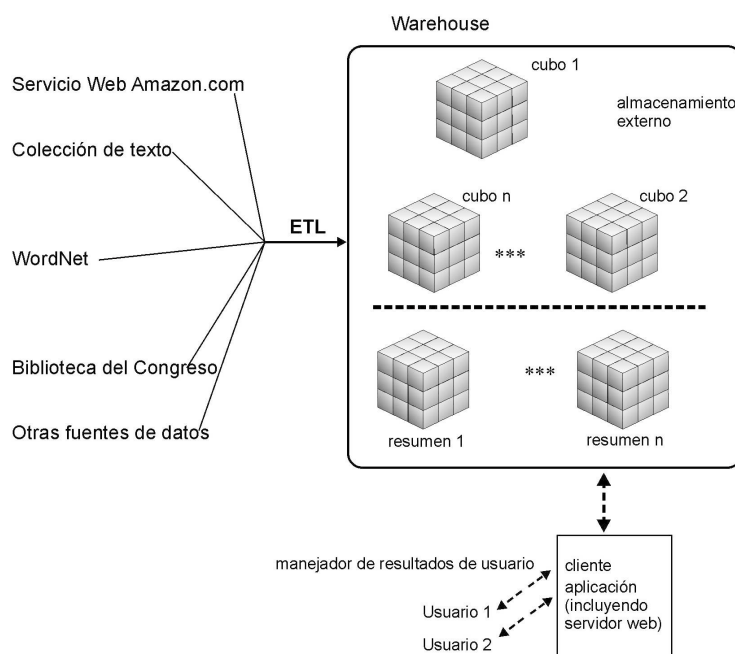


Figura 2.4: Arquitectura de un WoW

Algunas de las aplicaciones que WoW puede soportar son, por ejemplo, herramientas analíticas orientadas al usuario en el campo de las humanidades para la referencia del autor, de análisis léxico, y de análisis estilo-métrico.

<sup>1</sup>ETL son las siglas en inglés de Extraer, Transformar y Cargar (Extract, Transform and Load). Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

Para dar soporte al estudio estilo-métrico inicial, la analogía, y uso de frases de consulta, el WoW contiene varios cubos. Aquí se mencionan dos.

1. El estilo oración (Libro  $\times$  Palabra  $\times$  CantidadPalabras  $\times$  CantidadComas  $\times$  CantidadPuntoycoma  $\times$  CantidadPuntoFinal  $\rightarrow$  CantidadOcurrencias). Cada count es un entero, y la dimensión Word representa la primera palabra en una frase. Muchas preguntas prácticas involucran rollups de este cubo. Por ejemplo, la longitud promedio de las frases por cada siglo, o el uso de comas por autores que escriben las frases largas.

2. La frase corta (Libro  $\times$  Palabra  $\times$  Palabra  $\times$  Palabra  $\times$  Palabra  $\rightarrow$  CantidadOcurrencias). El cubo graba todas las sucesiones de 4 palabras, y podría usarse para explorar lo común (o raro) de las frases por autores o período de tiempo.

Este trabajo brinda la posibilidad de crear un DW que ofrezca un buen servicio a sistemas OLAP tipos literarios. Estos cubos permitirán interesantes consultas literarias gracias a la información de sus jerarquías. Hay que señalar que el procesamiento que realizan está orientado solo a la estructura de los textos no a su significado.

### **2.2.2. Estudios sobre data warehousing que utilizan técnicas de descubrimiento de conocimiento para la exploración de textos**

Los trabajos que se referencian a continuación tratan sobre sistemas data warehousing que se relacionan con técnicas de Minería de Datos (Cody et al., 2002) y/o aplican métodos de procesamiento textual para la exploración de documentos textuales (Grimes, 2006), (Tseng y Chou, 2006). También se muestran algunos estudios donde se utilizan técnicas de Recuperación de Información para modelar el conocimiento de documentos textuales en cubos de datos OLAP como por ejemplo en (Ravat et al., 2007), (Pérez et al., 2009) y (Lin et al., 2008), entre otros.

### 2.2.2.1. Propuestas de DW de documentos

En (Tseng y Chou, 2006) se plantea la integración de data warehousing de documentos textuales y minería de datos, para desarrollar el campo de la Inteligencia de Negocio. Normalmente, hay muchos conceptos diversos involucrados en un documento: un documento es multi-dimensional por naturaleza. Los warehousing de documentos, incluyen una extensa información semántica sobre los documentos; las relaciones de rasgos internos que contiene, se agrupan o se dividen para proporcionar más eficiencia en la inteligencia del negocio orientada al texto.

Este estudio propone el concepto de warehousing de documentos, como una importante plataforma para el proceso analítico en línea (OLAP) a nivel de texto y para el análisis interactivo multi-dimensional de los documentos de varias granularidades. El trabajo es esencial para establecer una infraestructura que ayude a combinar texto con tecnologías OLAP.

A continuación se exponen dos de las definiciones fundamentales, la de documento y dimensión.

#### Definición 1. Documento

*Un documento  $T = \{K_1, K_2, \dots, K_i\}$  es una unidad lógica textual caracterizada por un conjunto de palabras claves  $\{K_1, K_2, \dots, K_i\}$*

Para organizar la estructura de un documento se necesita el concepto de dimensión, como se muestra en la Definición 2.

#### Definición 2. Dimensión

*Una dimensión  $D$  es una estructura arbórea con  $m$  niveles,  $m \geq 1$ . Cada uno de los niveles es usado para representar la representación jerárquica del conjunto de palabras. Un nodo de una dimensión es llamado miembro y cada nodo interno contenido en un nodo-hijo se llama miembro resumen, y denotado por \*, los cuales se usan para dar a entender el concepto completo de nodo-hijo de un nodo.*



En la figura 2.5 se describe la arquitectura de un warehousing de documentos. Las fuentes de documentos pueden ser diversas, plantean que pueden ser desde documentos XML, texto plano, email, páginas web, etc. Estas fuentes de datos pasan por un preprocesamiento en los componentes Front-End. El mismo incluye procesos para resumir el texto (Goldstein et al., 1999), (Hahn y Mani, 2000) extracción de características (Feng y Croft, 2001), categorización (Appiani et al., 2001) y técnicas de minería de textos (Lin et al., 1998), (Loh et al., 2000)]. Así se obtienen las características o patrones del texto que son almacenados en Metadatos y los contenidos del texto resumidos se almacenan en los documentos resumidos que se observan en la figura. Por último se muestra que todos esos resultados pueden usarse en procesamientos OLAP, para herramientas de Minería de Datos o Sistemas de Gestión.

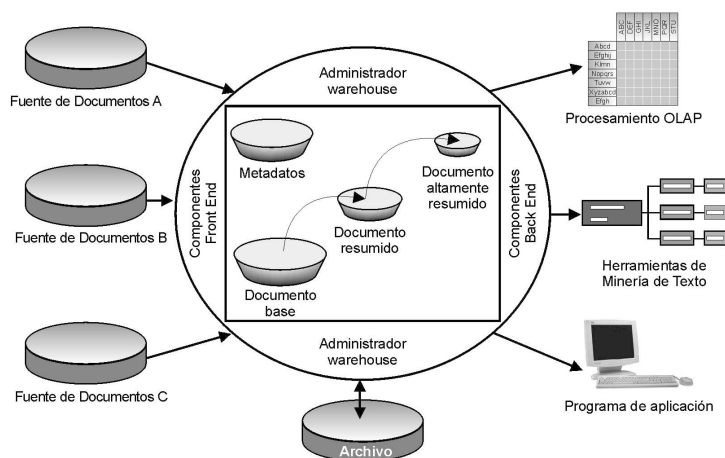


Figura 2.5: Arquitectura de un warehousing de documentos

En este artículo se defiende la importancia de construir un warehouse de documentos para apoyar la inteligencia del negocio vinculada a textos. Muestra mediante ejemplos la implementación de un documento warehousing. Al final plantean como tema de futuras investigaciones una arquitectura paralela para este proceso que mejore los resultados.

El procesamiento textual que realizan en el Front-End de la arquitectura no queda claro, pero por las explicaciones posteriores se puede inferir que éste

se enfoca principalmente en la extracción de contenidos como por ejemplo palabras claves, pero no intentan extraer el conocimiento implícito en los documentos. El modelo que plantean se especializa en el trabajo con documentos textuales como fuente de datos externa, pero no valoran los textos contenidos en atributos de bases de datos. En nuestra propuesta se realizan preprocesamientos textuales de minería para extraer el conocimiento de textos cortos contenidos en dichos atributos textuales. En trabajos futuros podríamos procesar documentos textuales como fuente de datos.

En el trabajo de (Ravat et al., 2007), se introduce un modelo conceptual OLAP multidimensional basado en el concepto de dimensiones que se adapta al análisis multidimensional de un documento. También proporciona un conjunto de operaciones para su manipulación. Las fuentes de datos que usa son documentos RTF, en formato XML, o sea, documentos que contienen en su mayoría textos y tienen una estructura mayormente conocida. Para procesar esos datos, que son no aditivos y no numéricos, se crea una nueva función de agregación *Top\_Keywords* (Ravat et al., 2008). Esta función selecciona las mejores palabras claves del texto con ayuda de técnicas de Recuperación de Información.

Este trabajo desarrolla un modelo multidimensional completo que brinda nuevas formas de análisis de textos y propone el uso de una novedosa función de agregación, especial para datos textuales.

El modelo multidimensional propuesto usa un esquema constelación llamado *Galaxia*, el mismo es definido matemáticamente. Las dimensiones se caracterizan por tener sus atributos organizados jerárquicamente. La jerarquía de una dimensión se representa como una lista ordenada de atributos llamados parámetros. Cada atributo puede estar asociado a atributos débiles que representan una información complementaria de este. A continuación se muestra este modelo multidimensional aplicado a un ejemplo de análisis de publicaciones científicas.

a)	Instituto	Instl		
	Autor	A1	A2	A3
Conferencia				
ER		3	2	1
SSDBM		2	-	-
DaWak		1	1	2

Figura 2.6: El número de veces que un autor es citado en una particular conferencia. .

$$\begin{aligned}
 D^{G1} &= \{D^{Conferencia}, D^{Articulo}, D^{veces}, D^{Autor}, \dots\} \\
 Estrella^{G1} &= \{D^{Conferencia} \longrightarrow (D^{Articulo}, D^{veces}, D^{Autor}), \dots\} \\
 Lk^{G1} &= \{g^{Referencia} : a_{Referencia}^{Articulos} \longrightarrow a_{Articulos}^{Articulos}, \dots\}
 \end{aligned}$$

Dimensión ejemplo:

$$\begin{aligned}
 D^{Conferencia} &= \{A^{Conferencia}, H^{Conferencia}, I^{Conferencia}, I^{estrella}^{Conferencia}\} \\
 A^{Conferencia} &= \{a^{Conferencia}, a^{nombre}, a^{Publicacion}, a^{Rango}\}; H^{Conferencia} = \{HEd, HRa\}; \\
 I^{Conferencia} &= \{i_1^{Conferencia}, \dots, i_q^{Conferencia}\} \\
 I^{estrella}^{Conferencia} &= \{i_k^{Conferencia} \longrightarrow (i_{rk}^{Articulos})^*, (i_{sk}^{veces})^*, (i_{tk}^{Autor})^* \mid \forall k \in [1..q], I_k^{Conferencia} \in \\
 I^{Conferencia} \wedge \exists i_{rk}^{Articulos} \in i^{Articulos} \wedge \exists i_{sk}^{veces} \in i^{veces} \wedge \exists i_{tk}^{Autor} \in i^{Autor}\}
 \end{aligned}$$

Jerarquía ejemplo:

$$\begin{aligned}
 HEd &= \{Param^{HEd}, Weak^{HEd}\} \\
 Param^{HEd} &= \langle a^{Conferencia}, a^{Publicacion} \rangle \text{ and } Weak^{HEd} = \{a^{Conferencia} \longrightarrow \{a^{nombre}\}\}
 \end{aligned}$$

Las figuras 2.6 y 2.7 tratan un ejemplo de análisis con el modelo anterior. En la primera 2.6 se muestran los resultados con un medida clásica y en la segunda 2.7 utilizando la función de agregación Top\_ Keywords.

Es cierto que el análisis que se muestra en la figura 2.7 sería muy complejo hacerlo usando la modelación multidimensional tradicional, debido a que el análisis de documentos textuales implica gran complejidad y no son suficientes para ello las operaciones o funciones de agregación numéricas que propone el modelo. Sería necesario que permitiera que las dimensiones se conviertan en hechos y viceversa, de forma similar a como lo hace este trabajo.

b)	Instituto	Inst1		
	Autor	A1	A2	A3
Conferencia				
ER		XML, Documentos	XML, Data Warehouses	Minería de Datos, Agrupamiento
SSDBM		XML, BD Temporal	-	-
DaWak		Minería de datos	Minería de Datos	Minería de Datos, Agrupamiento

Figura 2.7: El mismo análisis de la figura 2.6 pero con las mejores palabras claves de las publicaciones donde ha sido citado cada autor.

Nuestro trabajo no utiliza estas útiles medidas textuales sino las numéricas, que tradicionalmente han ayudado a la toma de decisiones. Aunque es necesario destacar que nuestras medidas numéricas no son completamente comunes ya que reflejan la combinación de dimensiones clásicas como la fecha, con dimensiones textuales que contienen el conocimiento asociado a su correspondiente información textual. En el capítulo 3 se argumenta sobre esta idea.

En las conclusiones mencionan muy brevemente que se usa XML para el trabajo con los documentos. Este es un detalle importante porque supone el uso de otro lenguaje adicional que es el XML. El prototipo que implementan usa como servidor de bases de datos a Oracle 10g, lo que conlleva a que estén presente todas las desventajas que supone el uso de software propietario (Sánchez, 2004).

En (Pérez et al., 2009) podemos ver otra propuesta de modelo multidimensional extendido, pero para un nuevo data warehouse, llamado data warehouse de contextos de documentos. La novedad de este artículo consiste en brindar una arquitectura de data warehouse capaz de combinar la información que existe en un data warehouse clásico de datos, con un data warehouse de documentos que describe dichos datos, para construir un data warehouse de contextos. Esta arquitectura se sustenta en un nuevo modelo multidimensional que define un R-Cubo (Perez et al., 2007). Este cubo tiene dos dimensiones: la relevancia y el contexto. La relevancia es un valor numérico que mide la importancia de cada hecho en el contexto de análisis establecido y el contexto describe la asociación de cada documento con un valor

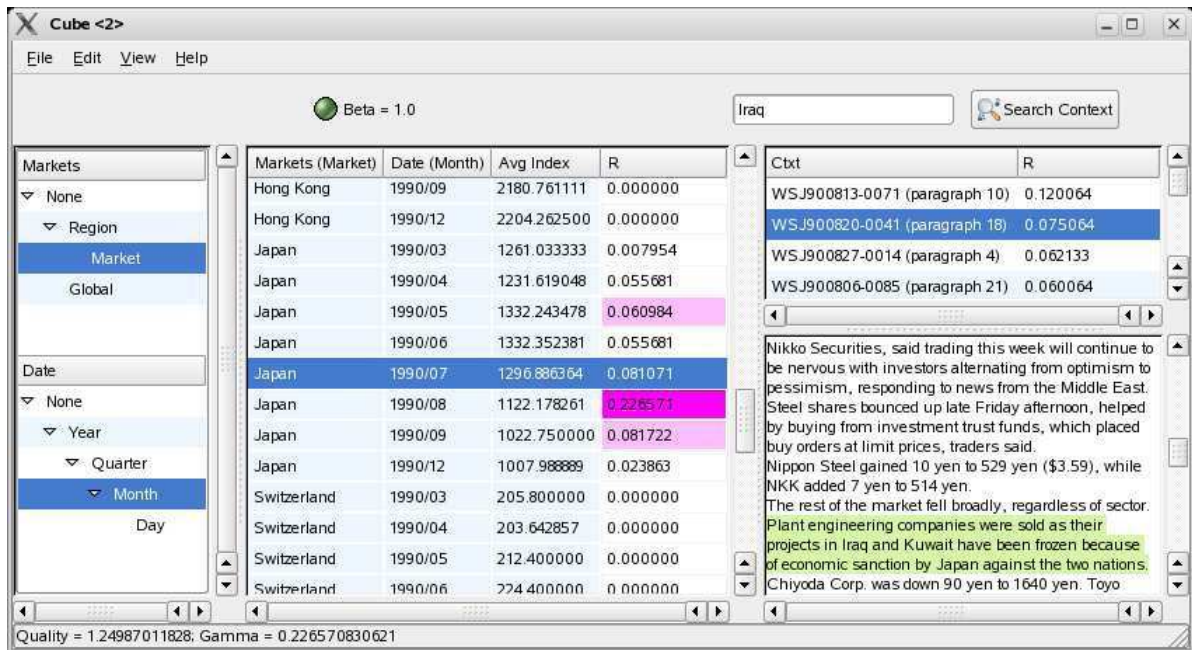


Figura 2.8: Ventana OLAP que muestra un R-Cubo

de relevancia, o sea, cuánto describe cada documento el contexto de análisis seleccionado.

En el R-Cubo se resume la información sobre la integración de datos de diferentes contextos. En la práctica el usuario especifica el contexto del análisis suministrando una secuencia de palabras claves (condiciones de IR) y obtiene un conjunto de documentos y hechos relacionados con dicho contexto, como se puede observar en la figura 2.8. Los resultados reflejan en qué documentos de un data warehouse de documentos se trata de un tema dado (condición IR) y analizar los hechos almacenados en un data warehouse clásico correspondientes a dicho tema.

Este artículo propone una arquitectura que sería de gran utilidad para Sistemas de Apoyo a la Toma de Decisiones (DSS) porque permite al usuario obtener una estrategia de información, combinando fuentes de datos estructurados con documentos no estructurados. Podría decirse que el proceso de data warehousing que realizan se especializa en el trabajo con los contextos, es el aporte fundamental del trabajo, una idea novedosa de analizar docu-

mentos textuales que valdría la pena valorar en trabajos futuros. En (Pérez et al., 2005) los mismos autores hacen un trabajo previo de procesamiento OLAP de documentos con técnicas de IR.

En (McCabe et al., 2000) y (Mothe et al., 2003) ya habían usado técnicas de IR para procesar documentos textuales y modelar los resultados en un esquema OLAP de datos.

A continuación se resumen otras propuestas que usan diversas fuentes de datos, además de documentos textuales, bases de datos con información textual.

### 2.2.2.2. Resumen de otras propuestas que utilizan técnicas de descubrimiento de conocimiento para la exploración de textos

En (Lin et al., 2008) se expone un nuevo modelo de cubo de datos donde llaman a un cubo, Text-Cube. Para ello usan técnicas de Recuperación de Información (RI) y proponen una jerarquía de términos y medidas específicas para datos textuales. Parten del hecho de tener una colección de documentos almacenados en una base de datos multidimensional. Esa jerarquía describe dichos documentos, y ayuda al usuario a navegar por sus diferentes niveles de granularidad. Como medidas de agregación proponen TF (frecuencia de términos) e IV (índice invertido). La primera ofrece la cantidad de veces que los términos descendientes (o términos) aparecen en un documento y la segunda los documentos donde se repite un término descendiente y cuánto se repite. Entre los autores de este artículo se encuentra Jiawei Han, que comparte autoría con otros autores, en el trabajo (Zhang et al., 2009). En esta publicación de forma similar, construyen un nuevo modelo de cubo, llamado, Topic-Cube, desde una base de datos textual multidimensional. Desde el conjunto de datos de documentos almacenados en cada celda de dicha base de datos, utilizan minería de datos para buscar la distribución de cada palabra y así definir los temas o tópicos, y relacionarlas en un árbol jerárquico de niveles de tópicos. Este árbol conforma una jerarquía de tópicos. Los mismos se agregan al cubo de datos base, como una dimensión más, para permitir que el usuario observe los datos seleccionando sus tópicos.

Basándose en los estudios anteriores, en Text-Cube y Topic-Cube, y en un análisis de red de información, tal como RankClus (Sun et al., 2009), en (Yu et al., 2009) construyen el iNextCube, un TexCube para enriquecer una red de información. También en (Yu et al., 2009) realizan una demostración, del uso de un iNextCube en la búsqueda y análisis de bases de datos textuales multidimensionales.

En (Cody et al., 2002) se pretenden unir las técnicas de Inteligencia de Negocio (BI) y Gestión del conocimiento (KM) en el manejo de DW, con técnicas de Minería de Datos, de manera que se puedan analizar datos no textuales y textos de forma unida. Esta unión se nombra BIKM, y está dividida en dos herramientas, eClassifier para realizar la minería de textos y Sapiens para integra datos y textos en un modelo OLAP. El eClassifier puede ser usado para crear la nueva taxonomía del conjunto de documentos seleccionados. Entonces esas nuevas taxonomías se convierten en una nueva jerarquía de dimensiones, sumando así valores a los datos de los documentos existentes en el Sapiens. Estas dimensiones proveen de mayor cantidad de niveles y facilitan la obtención de consultas con mayor nivel de detalles, por ejemplo el costo asociado a un producto de una región en un periodo de tiempo determinado. En este proceso se realiza un exhaustivo análisis de minería de datos sobre documentos textuales, en la herramienta eClassifier; el procesamiento del texto está basado en su taxonomía. El resultado de este análisis es usado para la construcción de las dimensiones del cubo OLAP, logrando así que la información brinde mayor conocimiento. En el propio trabajo se confiesa que no realizan en ningún caso análisis semántico, y plantean que su inclusión supone mejores resultados.

En (Grimes, 2006) se comenta sobre la herramienta Text OLAP como un tercer híbrido dentro del paradigma de OLAP, un artefacto analítico dentro de la colección de PolyAnalyst de Megaputer<sup>2</sup>. Text OLAP ofrece análisis interactivo y estructurado de documentos textuales; de esos documentos extrae términos y conceptos (o entidades) como dimensiones. Las cantidades calculadas ("medidas") se muestran en las celdas de una tabla; indican la fuerza de las relaciones entre las entidades en lugar, por ejemplo de las cantidades de

---

<sup>2</sup>Megaputer es una compañía de software que desarrolla herramientas OLAP con diferentes enfoques, entre ellos, el de minería de texto.

ventas que se podrían obtener con una herramienta de OLAP convencional. El Text OLAP proporciona una matriz para mostrar las entidades con un enlace a sus referencias en el texto. Es una herramienta que realiza procesos de minería típicos y los resultados se insertan en un modelo multidimensional de OLAP, para obtener información sobre el propio procesamiento que hace sobre los documentos textuales. En resumen, Text OLAP brinda información de los procesos de minería que se realizan sobre los documentos textuales.

En (Park et al., 2005) se propone un marco de trabajo para el análisis multidimensional de documentos de XML, llamado XML-OLAP. La base del trabajo consiste en que en los warehouse de XML cada dato de hecho así como los datos de la dimensión se guardan como documentos XML. Éstos pasan con anterioridad por un proceso de limpieza, integración, y actualización de las relaciones entre documentos. Presentan un nuevo lenguaje de consulta para los cubos XML, XML-MDX. Especifican operadores de minería de texto para agregar texto que constituye los datos de la medida. Y por último evalúan a XML-OLAP aplicándolo a un warehouse de XML de patentes.

Para lograr el modelo multidimensional de XML-OLAP recurren a varios trabajos anteriores donde exponen diseños para construir un DW de documentos XML (Nassis et al., 2004), (Golfarelli et al., 2001).

En (Vasilakopoulos et al., 2004) se describe un Editor de Anotaciones (CAFETIERE Anotador Manager), que se inserta dentro un marco de trabajo llamado Parmenides, usado para trabajar con anotaciones de datos textuales. El Sistema Parmenides consiste en la implementación de un Marco de trabajo de Minería de texto que maneja ontologías. El mismo está compuesto por un data warehouse de documentos, herramientas de adquisición y edición de ontologías, un conjunto de módulos para el procesamiento y extracción automática de ontologías, así como por herramientas de minería de datos. El warehouse de documentos (DW) es el módulo responsable de coleccionar los documentos para el posterior análisis. CAFETIERE se encarga de extraer semi-automáticamente y anotar elementos de la semántica básica, hechos, y eventos de textos en lenguaje natural.

Los trabajos resumidos en esta subsección, proponen diferentes técnicas para consultar dentro de los textos de documentos textuales. Se especializan en la



obtención de medidas numéricas específicas para textos (Yu et al., 2009), este hecho es de gran relevancia; y sería aún más si estas medidas se combinaran con las clásicas, como conteos, sumas o promedios.

### **2.2.2.3. Estudios relacionados con fuentes de datos médicos**

En gran parte de los trabajos que se han estudiado hasta aquí, los casos de estudios tratan datos de publicaciones científicas, mercado, entre otros, pero hay otras áreas de conocimiento donde también se cuenta con mucha información textual, como por ejemplo el área de la medicina. Ésta es una de las áreas que demanda de la utilización de tecnologías como OLAP, data warehousing, KDD, entre otras (Hristovski et al., 2000) debido a la necesidad vital de extraer conocimiento de los datos. Las prestaciones analíticas que estas tecnologías ofrecen son de especial aplicación en esta área, donde la toma de decisiones es una tarea diaria.

La creación de un data warehouse y un sistema OLAP para la web se expone en (Fei-Ran Guo y Fang, 2000). Con el objetivo de efectuar la evaluación de los resultados de los servicios de rehabilitación médica. La eficacia de terapias y clínicas, la utilidad esperada de tratamientos y los modelos gráficos son generados por la exploración de datos que se realiza, son resultados de la minería de datos y apoyan la toma de decisiones. Los usuarios pueden recuperar los gráficos y tablas estadísticas sin conocer la estructura de la base de datos o los atributos.

La utilidad esperada es usada para comparar la efectividad entre los tratamientos de grupos específicos de pacientes. Los patrones gráficos son usados para la exploración de los datos. Aunque este trabajo propone, como en (Sun et al., 2004) el uso de minería como apoyo a la toma de decisiones, en ninguno de los dos se valora la exploración de los textos.

En (Sun et al., 2004) proponen un sistema integrado, que nombran RgS-Miner. Éste comprende un sistema de data warehousing de datos biológicos, programas de descubrimientos de patrones, detectores de ocurrencias y asociación de patrones e interfaces de usuario. El sistema está disponible en <http://rgsminer.csie.ncu.edu.tw/>. Realiza un procesamiento significativamente complejo de sitios específicos en el Genoma Humano.

A nuestro criterio, una de las mejores soluciones que aparecen en la literatura es la que se introduce en (Inokuchi y Takeda, 2007). En ella se plantea un método para realizar OLAP sobre datos textuales. Para ello proponen un modelo para la representación de los datos, con sus correspondientes operaciones algebraicas (Selección, Proyección, Diferencia, Agregación, etc.) para la realización de las operaciones OLAP tradicionales. Dicho modelo permite integrar información semántica (en este caso mediante el uso de ontologías) en sistemas OLAP. Esto permite analizar un gran conjunto de documentos textuales con su información semántica subyacente.

El modelo multidimensional se basa en la construcción de las jerarquías de dimensiones utilizando ontologías. Entre las mismas están la Unified Medical Language System (UMLS: <http://umlsinfo.nlm.nih.gov>) y la Gene Ontology (GO: <http://www.geneontology.org>), cada una de ellas necesita ser representadas como un tipo de grafo acíclico orientado.

Cabe destacar que en su solución los autores plantean que dado el gran número de nodos en la jerarquía y la alta complejidad en las relaciones que entre ellos se establecen, no se pueden almacenar dichos datos en un esquema estrella y agregarlos eficientemente con la estructura creada en las jerarquías.

Para dar solución a este problema, se definen un conjunto de funciones que permiten mapear las jerarquías construidas mediante las ontologías y los cubos de datos, cuyas dimensiones normalmente están compuestas por términos relevantes dentro del documento. Dichas funciones por ejemplo retornan un conjunto con las relaciones que posee un término dado en la jerarquía, funciones para calcular agregaciones por palabras claves y categorías, etc. En este caso la medida se corresponde con el número de documentos que contienen los términos que representan las dimensiones.

Hay que señalar que las ontologías antes mencionadas son externas a los datos que se procesan, o sea ya se han obtenido previamente. Esto puede conllevar a que las consultas que realice el usuario relacionadas con ésta, no tengan datos de respuestas, porque no tienen en cuenta, como nosotros, la ontología interna, creada con la información que se tenga en la fuente de datos que se procese.

De forma similar ocurre con el trabajo de (Banek et al., 2006) y en (Torsten

y Günther, 2003). En (Torsten y Günther, 2003) proponen un portal de conocimiento empresarial que integra OLAP y recuperación de información (IR) con una ontología para almacenar los metadatos de los diferentes recursos.

En (Banek et al., 2006) plantean un modelo conceptual de HEWAF está basado en las normas de salud internacionales HL7 [21] para lograr un nivel alto de generalización y portabilidad. Estas normas de HL7 definen el Modelo de Información de Referencia objeto-orientado define (RIM) y la Arquitectura de un Documento Clínico (CDA) así como los documentos XML correspondientes, los cuales especifican la semántica de los documentos clínicos para su intercambio. Usan las anotaciones XML de los datos para facilitar su procesamiento.

Aunque no llegan a realizar data warehousing, es relevante el trabajo que hacen en (Uramoto et al., 2004), aquí también realizan procesamiento textual. Compilan documentos médicos diversos y con la ayuda de técnicas de minería obtienen conocimiento de los mismos.

En (Gibas y Jambeck, 2001) se puede apreciar el significativo desarrollo de la Bioinformática, donde el uso de diversas técnicas de KDD e Inteligencia Artificial de forma general ofrecen ventajas en éste campo (Meyfroidt et al., 2009). Pero todavía según Juan García-Gómez, del grupo de investigación IBIME de la Universidad Politécnica de Valencia ([ibime:www.ibime.upv.es](http://ibime:www.ibime.upv.es)), los CDSS están en fase de evaluación de su valor añadido. Este grupo ha desarrollado trabajos donde, por ejemplo, se puede caracterizar un tumor maligno usando características extraídas de algunas imágenes o predecir la depresión de la madre de un recién nacido.

En (Stolba y Tjoa, 2006) proponen el uso de un data warehouse como una solución conveniente para la integración de fuentes externas de datos médicos basados en evidencias, dentro de un sistema de información clínico. Llevan a cabo técnicas de minerías para encontrar la terapia apropiada para un paciente dado, dada la enfermedad.

En (Bhattacharyya, 2009) tratan sobre el tema de la medicina basada en evidencias. En este trabajo se discute, a un nivel alto, varias metodologías que pueden usarse, junto con la elaboración de varias terminologías asociadas

con data warehousing y el descubrimiento de conocimiento en las bases de datos (KDD), como por ejemplo los Registros Electrónicos Médicos (EMR) y Sistemas Clínicos de Apoyo a la Decisión (CDSS). Exponen la arquitectura de un sistema que integre todos los procesos a seguir para realizar data warehousing y minería de datos sobre los EMR.

## 2.3. Discusión

En el apartado anterior, se estudiaron varios trabajos relacionados con data warehousing y procesamiento textual que proponen técnicas y/o arquitecturas para analizar documentos textuales. Debido a la falta de estructura de éste tipo de datos, su gestión es una tarea compleja, que por lo general involucra el uso de varias disciplinas. Se han expuesto algunas soluciones a este problema, algunas de ellas intentan estructurar el texto, por ejemplo con el uso de XML, para extraer su información y poder gestionarla y otras aplican técnicas de descubrimiento de conocimiento sobre los textos, y así extraen algo más que su información explícita.

Los trabajos de (Keith et al., 2006), (Jensen et al., 2001), (Niemi et al., 2003), (Turkka et al., 2008) y (Park et al., 2005) se limitan a utilizar las ventajas del XML para analizar la información explícita de documentos textuales, por lo que solo consideran como fuentes de datos documentos textuales externos, no valoran la abundante información textual de atributos de bases de datos.

En otros estudios como el de (Tseng y Chou, 2006) y (Ravat et al., 2008) plantean la posibilidad de preprocesar los documentos textuales con ayuda de técnicas de minería y RI, extraer conocimiento de estos y analizar el mismo en modelos multidimensionales específicos. En (Zhang et al., 2009), (Yu et al., 2009) y (Lin et al., 2008) parten de bases de datos textuales multidimensionales y también aplicando minería de datos y/o textos procesan la información textual y proponen nuevos modelos de cubos OLAP.

Otras propuestas combinan herramientas independientes, de minería y OLAP para lograr analizar los resultados de los primeros, como en (Cody et al., 2002), (Grimes, 2006), (Vasilakopoulos et al., 2004).

Se puede apreciar que en estos últimos trabajos se refieren mayormente a la extracción de palabras claves (Ravat et al., 2007), (Pérez et al., 2009)) y no a extraer otro tipo de conocimiento de los textos.

Nuestra propuesta queda enmarcada, en el segundo grupo. Las fuentes de datos que usamos podrían ser fuentes externas textuales, pero, dichas fuentes externas tendrían que ser convertidas a atributos textuales de una base de datos para que puedan ser procesadas. Este proceso forma parte de los trabajos futuros de la presente tesis. El sistema data warehousing que proponemos es capaz de analizar el conocimiento asociado a los atributos textuales de bases de datos y conjuntamente con otros tipos de datos. Nuestro modelo multidimensional brinda soporte a textos libres (atributos textuales) en las dimensiones.

En la tabla 2.4 se resumen los estudios fundamentales que se han analizado en las secciones anteriores. La misma muestra como no son abundantes las propuestas sobre data warehousing y procesamiento textual, donde se consideran las aplicaciones gestión de datos de artículos científicos y de datos médicos. En capítulos posteriores desarrollamos aplicaciones relacionadas con esas fuentes de datos.

Nótese en la tabla que las celdas que tienen los asteriscos están poco referenciadas y son precisamente éstas sobre las que se trata nuestro trabajo. Es decir usamos datos textuales de bases de datos como fuente de datos y las aplicaciones que se implementan se basan en la gestión OLAP de datos médicos y de publicaciones científicas. También es de notar que la mayoría de los trabajos que hacen data warehousing y procesan textos, usan como fuente de datos XML y/o documentos textuales y no los atributos textuales contenidos en bases de datos. Entre las razones por la que ocurre esto es que XML es un lenguaje que brinda muchas facilidades para procesar los textos y por otro lado los sistemas de bases de datos así como los modelos multidimensionales de datos, no ofrecen operadores capaces de procesar con eficiencia la información contenida en los atributos textuales. Debido a esa dificultad de dichos sistemas y modelos hemos usado, como algunos trabajos analizados, técnicas de minería para extraer conocimiento de los atributos textuales y propuesto un nuevo modelo multidimensional para el mismo fin.

## 2.4. Conclusiones

Podemos concluir afirmando que los sistemas OLAP son protagonistas en el tema de procesamiento de textos, debido a que ellos garantizan funciones analíticas, que son útiles en el procesamiento de los mismos. El modelo multidimensional que implementan ofrece ventajas para crear arquitecturas que puedan soportar consultas de un gran nivel de detalle. A pesar de ello, hay que decir que no son muy abundantes los estudios que relacionan el data warehousing con el procesamiento textual, debido a la mencionada dificultad que supone la falta de estructura de los mismos.

Motivado por este planteamiento, hemos analizado varias propuestas que utilizan estrategias para procesar documentos textuales, utilizando formatos XML, o definiendo nuevas arquitecturas sobre los modelos multidimensionales de OLAP.

Algunos de los estudios valorados realizan el procesamiento textual sobre fuentes externas a los sistemas data warehousing; otros aplican técnicas de minería sobre estas fuentes textuales y los resultados se modelan en cubos OLAP y otros utilizan también técnicas de minería y/o Recuperación de Información sobre los textos contenidos en bases de datos textuales multidimensionales para obtener modelos específicos de cubos de datos. De los trabajos consultados, muy pocos de ellos se concentran en el conocimiento implícito de los textos, aunque sí extraen información de relevancia de los mismos, como son las palabras claves. Se destaca el trabajo de Ravat por la medida textual que proponen y el de Inocuchi por el uso de ontologías externas para crear jerarquías en las dimensiones.

Nuestra propuesta introduce un nuevo modelo multidimensional que brinda la posibilidad de aplicar operaciones OLAP clásicas sobre dimensiones textuales, conjuntamente y del mismo modo que el resto de las dimensiones de tipos de datos clásicos. Dichas dimensiones contienen una estructura de conocimiento extraída de los textos cortos almacenados en atributos textuales.

A continuación en el próximo capítulo se desarrolla la propuesta de este modelo. Se exponen todas las formalizaciones matemáticas necesarias y se

ejemplifican las mismas.

ESTRUCTURA DE LA INFORMACIÓN	APLICACIONES			
	Gestión de datos de publicaciones científicas	Gestión de datos médicos	Gestión de datos de ventas de productos	Otros
<i>XML y/o Documentos textuales</i>	(Turkka et al., 2008) (Tseng y Chou, 2006) (Ravat et al., 2007)	(Banek et al., 2006)	(Jensen et al., 2001) (Niemi et al., 2003) (Pérez et al., 2009)	(Keith et al., 2006) (Park et al., 2005) (Nassis et al., 2004) (Golfarelli et al., 2001) (Cody et al., 2002)(Grimes, 2006) (McCabe et al., 2000) (Torsten y Günther, 2003) (Mothe et al., 2003)
<i>Bases de datos con información textual</i>	***** (Yu et al., 2009)	***** (Inokuchi y Takeda, 2007)	(Lin et al., 2008)	(Zhang et al., 2009) (Vasilakopoulos et al., 2004)

Tabla 2.4: Resumen de estudios sobre data warehousing y OLAP con procesamiento textual





## Capítulo 3

# Propuesta de modelo multidimensional con manejo semántico de datos textuales

Retomamos aquí el planteamiento expuesto en el Capítulo 2, sobre la definición de un cubo y sus operaciones, donde se considera que las dimensiones están definidas en dominios de valores discretos y que, si se consideraran otros tipos de dominios, sería necesario extender este concepto. A dicha extensión estará dedicado este capítulo, es decir, extender este concepto considerando atributos cuyo dominio son textos cortos con una semántica concreta y así obtener un nuevo modelo multidimensional que soporte éste tipo de información.

Para trabajar con estos atributos nos proponemos extraer una representación más estructurada que recoja el conocimiento que encierran los mismos, y utilizar ésta para establecer una dimensión que se pueda utilizar en un cubo de datos conjuntamente con otras del tipo "clásico". De esta forma, con este tipo de atributos, se enriquecería el proceso OLAP. Ejemplo de estos atributos pueden ser: descripción de intervenciones médicas, diagnósticos médicos o de otro tipo, títulos de artículos, palabras claves etc., abstracts etc.

En este capítulo nos referimos a definiciones como la de estructura-AP (Martínez, 2008) que son la base del modelo y más adelante describimos nuevas defini-

ciones que enriquecen el trabajo con los atributos textuales, como son la definición de dimensión-AP y la definición de jerarquía de consulta. Además se proponen operaciones de gran utilidad para realizar consultas sobre dicha dimensión-AP.

### 3.1. Antecedentes

Para analizar con éxito tipos de datos textuales como los mencionados anteriormente, proponemos un nuevo modelo multidimensional que facilite el procesamiento de la semántica que se puede obtener de los mismos, con la ayuda de una nueva estructura de almacenamiento. La nueva representación estructurada que se propone para los atributos de textos cortos, pasa por una previa limpieza de los mismos y por un proceso posterior de minería de datos. El resultado de este proceso conduce al concepto de estructura-AP.

Es importante señalar que los conceptos y operaciones pertenecientes a dicha representación estructurada, que se exponen a continuación son referencias de la tesis doctoral del Dr. Sandro Martínez Folgoso.

Los ejemplos ilustrativos sirven de apoyo a la comprensión de las definiciones relacionadas con las estructura-AP. En los mismos se usan atributos textuales de una base de datos correspondientes a un servicio de urgencias médicas. Como se observa en la tabla, se han extraído descripciones textuales referentes al atributo diagnóstico, para algunos registros de la base de datos.

#### 3.1.1. Concepto de estructura-AP y operaciones asociadas (Martínez, 2008)

En esta subsección se comienza la exposición de las definiciones matemáticas bases para la representación formal de los datos. Inicialmente se establece la definición y propiedades de los conjuntos que tienen la propiedad "a priori" (Agrawal y Srikant, 1994) (llamados conjuntos-AP), y luego la definición de la estructura subyacente en los textos que es la de estructura-AP (Martín-Bautista et al., 2006), donde se captura la semántica básica que los textos encierran.

Nu-Enferm.	Tiempo-espera	Ciudad	diagnóstico
1	10	Granada	dolor de pierna
2	5	Gojar	dolor de cabeza y vómitos frecuentes
3	10	Motril	dolor de cabeza y vómitos
4	15	Granada	fractura y vómitos
5	15	Armillá	dolor de cabeza agudo
6	5	Camaguey	dolor agudo de pierna
7	5	Málaga	dolor en la pierna derecha
8	5	Sevilla	dolor de cabeza leve
9	10	Sevilla	dolor de estómago y mareos
10	5	Gojar	fractura de pié
11	10	Granada	fractura de la pierna izquierda
12	5	Santafé	fractura de cabeza
13	5	Madrid	vómitos y acidez de estómago
14	5	Madrid	vómitos y gases en el estómago
15	12	Jaen	dolor agudo de pierna y cadera
16	15	Granada	dolor agudo en la pierna y antebrazo
17	5	Motril	dolor agudo de cabeza y falta de visión
18	10	Motril	dolor agudo en los brazos
19	5	Londres	fractura de pierna
20	15	Madrid	dolor agudo en el estómago y vómitos

Tabla 3.1: Tabla que muestra una porción de una base de datos correspondiente a un servicio de urgencias médicas

### 3.1.2. Definición y propiedades de los conjuntos-AP

#### Definición 3. Conjunto-AP

Sean  $X = \{x_1 \dots x_n\}$  un conjunto referencial de items y  $\mathcal{R} \subseteq \mathcal{P}(X)$  un conjunto de itemsets frecuentes, siendo  $\mathcal{P}(X)$  las partes de  $X$ . Se plantea que  $\mathcal{R}$  es un conjunto-AP sí y sólo sí:

1.  $\forall Z \in \mathcal{R} \Rightarrow \mathcal{P}(Z) \subseteq \mathcal{R}$
2.  $\exists Y \in \mathcal{R}$  tal que :
  - a)  $\text{card}(Y) = \max_{Z \in \mathcal{R}}(\text{card}(Z))$  y no exista  $Y' \in \mathcal{R}$  tal que  $\text{card}(Y') = \text{card}(Y)$
  - b)  $\forall Z \in \mathcal{R}; Z \subseteq Y$

El conjunto  $Y$  de cardinal maximal que caracteriza el conjunto-AP se denomina *conjunto generador de  $\mathcal{R}$* . Notaremos  $\mathcal{R} = g(Y)$ , es decir  $g(Y)$  será

el conjunto-AP con conjunto generador  $Y$ . Llamaremos *Nivel de  $g(Y)$*  al cardinal de  $Y$ . Obviamente, los conjuntos-AP de nivel 1 son los elemento de  $X$ , se considera el conjunto vacío  $\emptyset$  como el conjunto-AP de nivel cero.

### Ejemplo 2

Sea

$X = \{dolor, pierna, cabeza, vómitos, fractura, agudo, estómago\}$  y

$\mathcal{R} = \{\{pierna\}, \{agudo\}, \{dolor\}, \{pierna, agudo\}, \{dolor, agudo, pierna\}\}$ .

Entonces el conjunto generador es  $Y = \{dolor, agudo, pierna\}$ .

Como se observa en el ejemplo anterior y teniendo en cuenta la definición 1, el conjunto generador  $Y = \{dolor, agudo, pierna\}$  se corresponde con el conjunto-AP de mayor cardinalidad y que incluye a su vez, todas las combinaciones presentes en  $\mathcal{R}$ .

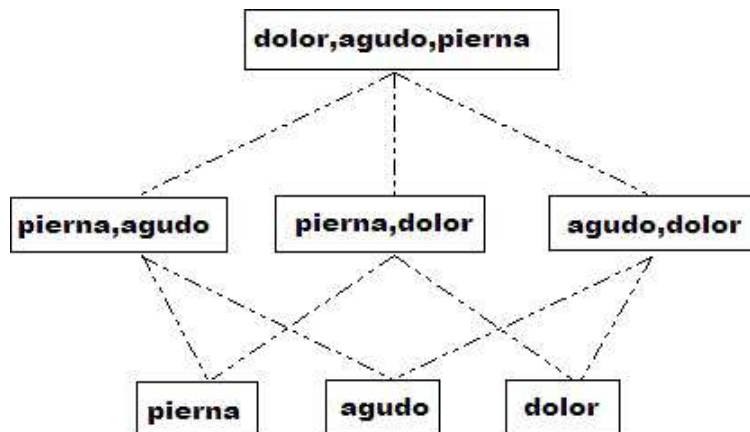


Figura 3.1: Retículo de un Conjunto-AP

A continuación se dan algunas operaciones relacionadas con el conjunto-AP definido. Dichas operaciones serán utilizadas en la definición de otras operaciones como la obtención de la estructura global de conocimiento, que encierra la semántica básica de los datos procesados. Se comienza por la operación que verifica si un conjunto-AP está incluido en otro.

**Definición 4. Inclusión de conjuntos-AP**

Sea  $\mathcal{R} = g(R)$  y  $\mathcal{S} = g(S)$  dos conjuntos-AP con el mismo conjunto referencial de items:

$$\mathcal{R} \subseteq \mathcal{S} \Leftrightarrow R \subseteq S$$

En la definición de inclusión de conjuntos-AP se puede apreciar que, los conjuntos-AP  $\mathcal{R}$  y  $\mathcal{S}$  estarán incluidos unos en otros si alguno de los dos conjuntos generadores  $R$  o  $S$  está contenido uno en otro.

A continuación se introduce una operación importante en el contexto en el que se plantea este modelo, que es el *subconjunto-AP inducido* por un conjunto determinado. Esta operación se encargará de obtener el conjunto-AP particular que se genera, al intersecar el retículo global del conjunto-AP con un conjunto dado.

**Definición 5. Subconjunto-AP inducido**

Sea  $\mathcal{R} = g(R)$  y  $Y \subseteq X$  diremos que  $\mathcal{S}$  es el subconjunto-AP inducido por  $Y$  si y sólo si:

$$\mathcal{S} = g(R \cap Y)$$

Se puede apreciar en la definición que el subconjunto-AP inducido se obtiene de hacer la intersección de los conjuntos generadores de  $\mathcal{R}$  y el conjunto  $Y$ . Ésta intersección nunca será vacía porque el conjunto  $Y$  es un subconjunto del conjunto referencial  $X$ .

**3.1.3. Definición y propiedades de una estructura-AP**

Los conceptos de conjunto-AP establecidos se usan para definir la estructura de información que se construye cuando se calculan itemsets frecuentes. Hay que tener en cuenta que estas estructuras se obtienen de forma constructiva, generando inicialmente itemsets con cardinal igual a 1; luego éstos se combinan para obtener los de cardinal 2, y así sucesivamente hasta obtener itemsets con cardinal maximal, con un soporte mínimo fijado. Por tanto, la estructura final es la de un conjunto de conjuntos-AP, que formalmente se define como estructura-AP (Martín-Bautista et al., 2006).

**Definición 6. Estructura-AP**

Sea  $X = \{x_1...x_n\}$  un conjunto referencial de items y  $S = \{A, B, \dots\} \subseteq \mathcal{P}(X)$  un conjunto de itemsets frecuentes, tal que:

$$\forall A, B \in S; A \not\subseteq B, B \not\subseteq A$$

Llamaremos estructura-AP del generador  $S$ ,  $\mathcal{T} = g(A, B, \dots)$ , al conjunto de conjuntos-AP cuyos conjuntos generadores son  $A, B, \dots$

De la definición anterior queda claro entonces que: la estructura-AP no es más que una colección de conjuntos-AP. Tal como se definió, los conjuntos generadores de la estructura-AP  $A, B, \dots$  no pueden estar contenidos unos en otros, utilizando para esta interpretación la definición de conjunto-AP incluido que se dio anteriormente. Entonces, la estructura-AP quedará constituida por todos los conjuntos generadores que se obtengan de las combinaciones de  $X$  presentes, dentro de todas las posibles ( $\mathcal{P}(X)$ ).

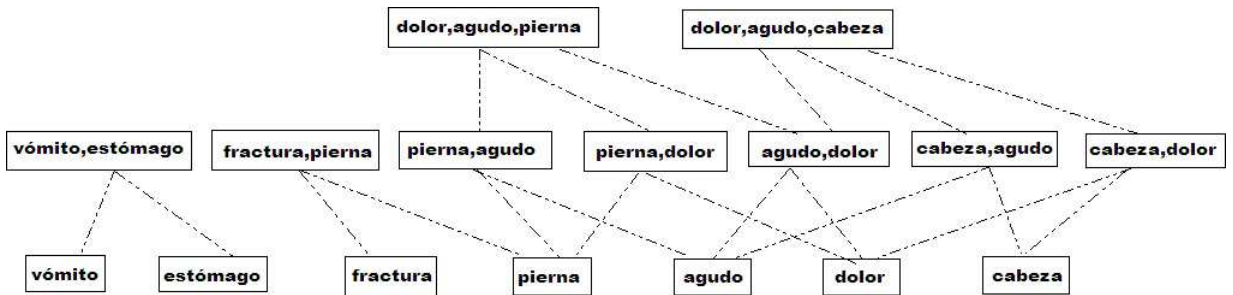


Figura 3.2: Estructura-AP global

Hay que hacer notar que cualquier estructura-AP es un retículo de subconjuntos cuyos extremos superiores son sus conjuntos generadores. La figura 3.2 muestra la estructura-AP global correspondiente a los datos de la tabla 3.1 que se define como  $g(\{vómitos, estómago\}, \{fractura, pierna\}, \{dolor, agudo, pierna\}, \{dolor, agudo, cabeza\})$ . Se brindan a continuación algunas definiciones y propiedades pertenecientes a estas nuevas estructuras.

**Definición 7. Inclusión de estructuras-AP**

Sean  $\mathcal{T}_1, \mathcal{T}_2$ , dos estructuras-AP con el mismo conjunto referencial de items:

$$\mathcal{T}_1 \subseteq \mathcal{T}_2 \Leftrightarrow \forall \mathcal{R} \text{ conjunto-AP de } \mathcal{T}_1 ,$$

$$\exists \mathcal{S} \text{ conjunto-AP de } \mathcal{T}_2 \text{ tal que } \mathcal{R} \subseteq \mathcal{S}$$

De esta definición se puede interpretar que para que una estructura-AP  $\mathcal{T}_1$  esté contenida en otra estructura-AP  $\mathcal{T}_2$ , todos los conjuntos generadores de  $\mathcal{T}_1$  tienen que aparecer incluidos en alguno de los conjuntos generadores de  $\mathcal{T}_2$ .

A continuación se introduce una importante operación sobre esta estructura-AP, la operación subestructura-AP inducida. Ésta no será más que la estructura-AP resultante de intersecar una estructura-AP cualquiera con un conjunto dado.

**Definición 8. Subestructura-AP inducida**

Sea la estructura-AP  $\mathcal{T} = g(A_1, A_2, \dots, A_n)$  con conjunto referencial de items  $X$  y  $Y \subseteq X$ . Definiremos la estructura-AP de  $\mathcal{T}$  inducida por  $Y$  como:

$$\mathcal{T}' = \mathcal{T} \bigwedge Y = g(B_1, B_2, \dots, B_m)$$

donde

$$\forall B_i \in \{B_1, \dots, B_m\} \Rightarrow \exists A_j \in \{A_1, A_2, \dots, A_n\}$$

$$\text{tal que } B_i = A_j \cap Y$$

$$\forall A_j \in \{A_1, \dots, A_n\} \Rightarrow \exists B_i \in \{B_1, B_2, \dots, B_m\}$$

$$\text{tal que } A_j \cap Y \subseteq B_i$$

Está claro que  $\mathcal{T}'$  es la estructura-AP generada por aquellas intersecciones de  $Y$  con los conjuntos generadores  $\mathcal{T}$ , que no están en contradicción con la definición de estructura-AP. El siguiente ejemplo expone estas ideas.

**Ejemplo 3**

Sea  $X = \{\text{dolor, pierna, cabeza, vómitos, fractura, agudo, estómago}\}$ ,



$$\mathcal{T} = g(\{\text{vómitos, estómago}\}, \{\text{fractura, pierna}\}, \{\text{dolor, agudo, pierna}\}, \{\text{dolor, agudo, cabeza}\}),$$

$$Y = \{\text{fractura, pierna, dolor}\}, \text{ entonces se tiene}$$

$$\mathcal{T} \wedge Y = g(\{\text{fractura, pierna}\}, \{\text{pierna, dolor}\}).$$

### 3.1.3.1. Acoplamiento de conjuntos de términos con las estructuras-AP

En esta sección, serán establecidas las definiciones necesarias para consultar la base de datos, donde se cuenta con la estructura-AP como tipo de dato. La idea es que el usuario expresará sus requerimientos como conjuntos de términos, para ser consultados sobre los atributos textuales en la base de datos. Dado que dichos atributos estarán representados por sus estructuras-AP particulares, algunos tipos de acoplamientos tienen que ser dados para satisfacer las consultas sobre dichas estructuras.

Para hacerlo, contamos con dos enfoques distintos, que definimos a continuación.

#### Definición 9. Acoplamiento fuerte

Sea la estructura-AP  $\mathcal{T} = g(A_1, A_2, \dots, A_n)$  con conjunto referencial de items  $X$  y  $Y \subseteq X$ . Se define el acoplamiento fuerte entre  $Y$  y  $\mathcal{T}$  como la operación lógica:

$$Y \odot \mathcal{T} = \begin{cases} \text{verdadero si} & \exists A_i \in \{A_1, A_2, \dots, A_n\} \\ & / Y \subseteq A_i \\ \text{falso} & \text{en otro caso} \end{cases}$$

#### Definición 10. Acoplamiento débil

Sea la estructura-AP  $\mathcal{T} = g(A_1, A_2, \dots, A_n)$  con conjunto referencial de items  $X$  y  $Y \subseteq X$ . Se define el acoplamiento débil entre  $Y$  y  $\mathcal{T}$  como la operación

*lógica:*

$$Y \oplus \mathcal{T} = \begin{cases} \text{verdadero si} & \exists A_i \in \{A_1, A_2, \dots, A_n\} \\ & / Y \cap A_i \neq \emptyset \\ \text{falso} & \text{en otro caso} \end{cases}$$

Estas definiciones se pueden complementar dando alguna medida o índice que cuantifique estos acoplamientos. La idea es considerar que el acoplamiento de un conjunto grande de términos tendrá un índice mayor que uno con un menor número de términos; adicionalmente, si algún conjunto de términos se acopla con más de un conjunto generador, éste tendrá un índice mayor que el de otro, que sólo se acopla con un sólo conjunto. Obviamente, dos índices de acoplamiento pueden ser establecidos, y así se tienen las definiciones siguientes.

### **Definición 11. Índice de acoplamiento fuerte (débil)**

Sea la estructura-AP  $\mathcal{T} = g(A_1, A_2, \dots, A_n)$  con conjunto referencial de items  $X$  y  $Y \subseteq X$ , se define el índice de acoplamiento fuerte (débil) entre  $Y$  y  $\mathcal{T}$  como sigue:

$\forall A_i \in \{A_1, A_2, \dots, A_n\}$  se denota  $m_i(Y) = \text{card}(Y \cap A_i) / \text{card}(A_i)$ ,  $S = \{i \in \{1, \dots, n\} | Y \subseteq A_i\}$ ,  $W = \{i \in \{1, \dots, n\} | Y \cap A_i \neq \emptyset\}$ .

Entonces se define el índice de acoplamiento fuerte y débil entre  $Y$  y  $\mathcal{T}$  como sigue:

$$\text{Índice fuerte} = S(Y|\mathcal{T}) = \sum_{i \in S} m_i(Y) / n$$

$$\text{Índice débil} = W(Y|\mathcal{T}) = \sum_{i \in W} m_i(Y) / n$$

Obviamente:

$$\forall Y \text{ y } \mathcal{T}, S(Y|\mathcal{T}) \in [0, 1], W(Y|\mathcal{T}) \in [0, 1] \text{ y } W(Y|\mathcal{T}) \geq S(Y|\mathcal{T})$$

### 3.1.4. Proceso de transformación de un atributo textual a un atributo-AP

A continuación se describe brevemente el proceso mediante el cual un atributo textual se transforma en un atributo valuado en una estructura-AP, en lo que partir de ahora se denominará *atributo-AP*.

1. Se obtienen los términos frecuentes asociados al atributo textual de partida. Obviamente, esta obtención incluye un proceso de limpieza, eliminación de palabras vacías, manejo de sinónimos, etc. Esto da como resultado el conjunto  $X$  de términos básicos con el que se va a trabajar. En este punto, el valor del atributo textual en cada tupla  $t$  de la base de datos es subconjunto de términos básicos  $T_t$ . Esto permite considerar que las tuplas de la base de datos forman una base de datos transaccional con respecto a dicho atributo textual.
2. Se calculan los itemsets maximales frecuentes asociados a dicha base de datos transaccional (mediante el algoritmo "a priori" por ejemplo). Sean  $\{A_1, \dots, A_n\}$  estos itemsets. De acuerdo a como se han construido, la estructura-AP  $S = g(A_1, \dots, A_n)$  incluye a todos los itemsets frecuentes de la base de datos, y por ello se considera que recoge la semántica básica del atributo textual.
3. Una vez calculada la estructura-AP global, se obtiene el valor del atributo-AP para cada tupla  $t$  de la base de datos de la siguiente manera: Si  $T_t$  es el conjunto de términos asociados a la tupla  $t$ , entonces, el valor del atributo-AP para dicha tupla es:

$$S_t = g(A_1, \dots, A_n) \bigwedge T_t$$

Este proceso permite establecer el dominio de cualquier atributo-AP.

#### Definición 12. Dominio de un atributo-AP

Si se considera una base de datos para la que se ha construido un atributo-AP  $A$  cuya estructura global es  $S(A_1, \dots, A_n)$ , el dominio del atributo  $A$  se define

como:

$$D_A = \{R = g(B_1, ..B_m), /, \forall i \in \{1, .., m\}, \exists, j \in \{1, .., n\} tal que B_i \subseteq A_j\}$$

Es decir  $D_A$  es el conjunto de todas la subestructuras-AP de la estructura-AP global asociada al atributo, ya que estos son los posibles valores que puede tomar el atributo  $A$  cuando se hace la restricción anterior.

### **3.2. Definición de una dimensión asociada a un atributo-AP**

Una vez definido el dominio de un atributo-AP se debe pasar a configurarlo como una dimensión de un modelo multidimensional. Pero antes de definirlo hay que hacer algunas consideraciones.

- Aunque la representación interna de los valores de un atributo-AP son estructuras, la salida y consulta de la información a un usuario se hace en forma de conjuntos de conjuntos de términos ("frases"), que no son más que los conjuntos generadores de las estructuras-AP.
- Esta situación no es diferente cuando se hace OLAP. La entrada a este proceso debe hacerla un usuario estableciendo conjuntos de frases, como valores de la dimensión, aunque definitivamente estos valores serán elementos de un dominio de un atributo-AP.
- De la definición 12 se deduce que se está trabajando con un dominio estructurado y cerrado con respecto a la unión, de forma que un conjunto de elementos del dominio forma parte de dicho dominio. Es decir, que el dominio básico de una dimensión asociada a un atributo-AP, y el dominio donde se valoran las jerarquías son el mismo.

Estas consideraciones llevan a establecer la siguiente definición.

**Definición 13. Partición de estructura-AP asociada a consulta**

Se considera  $C = \{T_1, \dots, T_q\}$  donde  $T_i \subseteq X$  un conjunto de "frases" que un usuario da como posibles valores de la dimensión de un atributo-AP, sea  $S$ , la estructura-AP global asociada a dicho atributo. Se define **partición asociada a  $C$**  como:

$$\mathcal{P} = \{S_1, \dots, S_q, S_{q+1}\}$$

donde

$$S_i = \begin{cases} S \wedge T_i & \text{si } i \in \{1, \dots, q\} \\ S \wedge (X - \bigcup_{i=1}^q T_i) & \text{en otro caso} \end{cases}$$

Con esta definición se puede plantear la organización de una dimensión-AP a partir de una propuesta de usuario. Utilizando la formulación planteada en el epígrafe anterior, un modelo de DW que incluya la dimensión-AP se establece de la siguiente forma:

- $\forall i \in \{1, \dots, q\}$   $f(\dots, S_i, \dots)$  recoge una medida, ( conteo, o agregación numérica de algún tipo etc.) asociada aquellas tuplas que de alguna forma se "acoplan" con  $T_i$ .
- $f(\dots, S_q, \dots)$  recoge la medida asociada a aquellas tuplas que no se acoplan con ninguna frase  $T_i$ , ni con parte de ellas. Es decir, aquellas tuplas que no tienen nada que ver con las frases que ha propuesto el usuario.

Obviamente, el concepto de acoplamiento y las medidas que se consideren han de estar adaptadas a las especiales características de una dimensión-AP. El siguiente ejemplo simple clarificará estos conceptos.

**Ejemplo 4**

Se cuenta con la tabla 3.2, que refleja una simplificación de los datos correspondientes a los ingresos de pacientes en un servicio de urgencias médicas.

1. Número de enfermo

Nu-Enferm.	Tiempo-espera	Ciudad	generadores-atributo-ap
1	10	Granada	(dolor,pierna)
2	5	Gojar	(dolor cabeza), (vómitos)
3	10	Motril	(dolor,cabeza), (vómitos)
4	15	Granada	(fractura) (vómitos)
5	15	Armillá	(dolor,agudo, cabeza)
6	5	Camaguey	(dolor, agudo, pierna)
7	5	Málaga	(dolor, pierna)
8	5	Sevilla	(dolor,cabeza)
9	10	Sevilla	(dolor), (estómago)
10	5	Gojar	(fractura)
11	10	Granada	(fractura pierna)
12	5	Santafé	(fractura) (cabeza)
13	5	Madrid	(vómitos, estómago)
14	5	Madrid	(vómitos, estómago)
15	12	Jaen	(dolor, agudo, pierna)
16	15	Granada	(dolor, agudo, pierna)
17	5	Motril	(dolor, agudo, cabeza)
18	10	Motril	(dolor, agudo)
19	5	Londres	(fractura, pierna)
20	15	Madrid	(dolor, agudo), (vómitos, estómago)

Tabla 3.2: Tabla con atributo-AP correspondiente al Ejemplo 4

2. Tiempo de espera
3. Ciudad donde vive
4. Diagnóstico

Mientras que los atributos 1,2,3 son "atributos clásicos", el atributo 4 es de tipo texto, y a éste se le ha aplicado el proceso descrito en el apartado 3.1.4 , obteniendo la estructura-AP global que se muestra en la figura 3.2 anteriormente expuesta, hasta llegar al atributo generadores-atributo-ap de la tabla 3.2 .

Se supone ahora que se plantea la partición asociada a la consulta:

$$C = \{(dolor, agudo), (vómitos)\}$$

Si se toma como medida de agregación el conteo, y como criterio de acoplamiento el acoplamiento débil dado en la definición 18, se obtiene el cubo unidimensional que aparece en la tabla 3.3.

(dolor agudo)	(vómitos)	Otros	Total
13	6	4	23

Tabla 3.3: Ejemplo de acoplamiento débil

(dolor agudo)	(vómitos)	Otros	Total
7	6	8	21

Tabla 3.4: Ejemplo de acoplamiento fuerte

Si se toma ahora como criterio de acoplamiento, el acoplamiento fuerte, establecido en la definición 17 se obtiene el cubo que aparece en la tabla 3.4.

También se pueden considerar otras dimensiones que no son de tipo "AP". Por ejemplo considerar la dimensión ciudad y agruparla según la jerarquía:

{Prov-Granada, Prov-Malaga, Prov-Jaen, Rest. España, Extranj.}

Tomando nuevamente el conteo como agregación, se obtiene para el acoplamiento débil, los resultados que aparecen en la tabla 3.5.

Asimismo, con el conteo como agregación se obtiene para el acoplamiento fuerte, los resultados que aparecen en la tabla 3.6.

	(dolor agudo)	(vómitos)	Otros	Total
Prov-Granada	7	3	3	13
Prov-Malaga			1	1
Prov-Jaen	1			1
Rest. España	3	3	0	6
Extranj.	1		1	2
Total	13	6	4	23

Tabla 3.5: Ejemplo de cubo bidimensional con acoplamiento débil

### 3.3 - Operaciones sobre una estructura-AP asociada a consulta 79

	(dolor agudo)	(vómitos)	Otros	Total
Prov-Granada	4	3	4	11
Prov-Málaga			1	1
Prov-Jaén	1			1
Rest. España	1	3	2	6
Extranj.	1		1	2
Total	7	6	8	21

Tabla 3.6: Ejemplo de cubo bidimensional con acoplamiento fuerte

	(dolor agudo)	(vómitos)	Otros	Total
Prov-Granada	11,5	10	7,5	9,3
Prov-Málaga			5	5
Prov-Jaén	12			12
Rest. España	15	6,5	7,5	9,6
Extranj.	5		5	5
Total	10,8	8,3	6,6	

Tabla 3.7: Ejemplo de cubo bidimensional con acoplamiento fuerte y agregación tiempo medio

Por último, si se considera como agregación el tiempo medio de espera se obtiene para el acoplamiento fuerte los resultados que aparecen en la tabla 3.7.

Pasamos ahora en el siguiente apartado, a definir las operaciones sobre una estructura-AP asociada a consulta, las mismas se utilizarán en la nueva jerarquía de consulta que debemos definir para el caso de una dimensión-AP.

### 3.3. Operaciones sobre una estructura-AP asociada a consulta

Una vez que se tiene la forma que tienen las particiones de estructuras-AP, cómo se generan y a qué conducen, el paso siguiente es definir las opera-



ciones sobre las mismas. Para ello lo primero será establecer una jerarquía de particiones de una estructura-AP. Obviamente dado que la partición de una estructura-AP está asociada a una consulta, está claro que la jerarquía de particiones estará asociada a una jerarquía de consultas.

### 3.3.1. Definición de jerarquía de consultas

**Definición 14. Jerarquía de una partición de estructura-AP asociada a consulta**

Sean  $C^1 = \{T_1^1, \dots, T_q^1\}$  y  $C^2 = \{T_1^2, \dots, T_n^2\}$  dos conjuntos de "frases" sobre  $X$ ; es decir las posibles consultas sobre la dimensión. Se dice que la consulta  $C^1$  es más fina que la  $C^2$  y notamos por  $C^1 \ll C^2$  si y solo si :

$$C^1 \ll C^2 \Leftrightarrow \forall T_i^1 \in C^1 \quad \exists T_j^2 \in C^2 \quad /T_j^2 \subseteq T_i^1 \quad y \\ \forall T_j^2 \quad \exists T_i^1 /T_j^2 \subseteq T_i^1 \quad y \\ \exists T_i^1 /T_i^1 \subseteq T_j^2$$

Intuitivamente hablando, una consulta es más fina que otra cuando la primera contiene "frases más detalladas" que la segunda. El siguiente ejemplo clarificará este concepto.

#### Ejemplo 5

Sean  $C^1 = \{(dolor, cabeza), (fractura)\}$  y  $C^2 = \{(dolor), (fractura)\}$

Comprobamos entonces que  $C^1 \ll C^2$  ya que se cumple que la definición anterior. O sea podemos ver como cada conjunto perteneciente a  $C^2$  está presente en  $C^1$ .

La siguiente propiedad permite extender la idea de partición "más fina" a las estructuras-AP subyacentes.

**Definición 15. Propiedad 1**

Sean  $C^1 = \{T_1^1, \dots, T_q^1\}$  y  $C^2 = \{T_1^2, \dots, T_h^2\}$  dos consultas sobre  $X$ ; tales que  $C^1 \ll C^2$

### 3.3 - Operaciones sobre una estructura-AP asociada a consulta 81

Sean

$$P^1 = \{S_1^1, \dots, S_q^1, S_{q+1}^1\}$$

$$P^2 = \{S_1^2, \dots, S_h^2, S_{h+1}^2\}$$

Las particiones de subestructuras-AP asociadas a las consultas, se verifica entonces las siguientes propiedades:

1.  $\forall i \in \{1, \dots, q\} \exists j \in \{1, 2..h\}$  tal que  $S_j^2 \subseteq S_i^1$
2.  $\forall j \in \{1, \dots, h\} \exists i \in \{1, 2..q\}$  tal que  $S_j^2 \subseteq S_i^1$
3.  $S_{h+1}^2 \supseteq S_{q+1}^1$

Se irán probando cada una de las partes de la propiedad

1. Propiedad 1.1

Sea  $S_i^1 \in P^1$  de acuerdo con la definición de partición asociada a consulta  $\exists T_i^1 \in C^1$  tal que:

$$S_i^1 = S \wedge T_i^1$$

puesto que  $C^1 \ll C^2$  de la definición 12 se deduce que:

$$\exists T_j^2 \in C^2 / T_i^1 \supseteq T_j^2$$

y para  $T_j^2$  se tiene una subestructura-AP asociada

$$S_j^2 = S \wedge T_j^2$$

realizando el mismo razonamiento que en el caso anterior, se deduce que:

$$S_j^2 \subseteq S_i^1$$

2. Propiedad 1.2

Sea  $S_j^2 \in P^2$  de acuerdo con la definición de partición asociada a consulta

$$\exists T_j^2 \subseteq C^2 \text{ tal que } S_j^2 = S \wedge T_j^2$$

Además puesto que  $C^1 \ll C^2$  se tiene que  $\exists T_i^1 / T_j^2 \subseteq T_i^1$ ; para  $T_i^1$  tiene asociada una subestructura-AP  $S_i^1$  que viene dado por:

$$S_i^1 = S \wedge T_i^1$$

$$\text{dado que } T_j^2 \subseteq T_i^1 \Rightarrow T_i^1 \cap T_j^2 = T_j^2$$

se tiene que:

$S_j^2 = S \wedge T_j^2 = S \wedge (T_i^1 \cap T_j^2)$  y en virtud de las propiedades de las estructuras-AP tenemos que:

$$S \wedge (T_i^1 \cap T_j^2) = (S \wedge T_i^1) \wedge T_j^2 \text{ y por definición}$$

$$(S \wedge T_j^2) = (S \wedge T_i^1) \wedge T_j^2 \subseteq T_i^1$$

$S_i^1 \supseteq S_j^2$  lo que prueba esta parte de la propiedad.

3. Propiedad 1.3

De acuerdo con la definición

$$\begin{aligned} S_{q+1}^1 &= S \wedge (X - \bigcup_{i=1}^q T_i^1) = S \wedge P_1 \\ S_{h+1}^2 &= S \wedge (X - \bigcup_{j=1}^h T_j^2) = S \wedge P_2 \end{aligned}$$

### 3.3 - Operaciones sobre una estructura-AP asociada a consulta 83

---

ahora bien, por la definición de partición más fina tenemos que:

$$P^2 \supseteq P^1$$

En efecto sean:

$$W_1 = \bigcup_{i=1}^q T_i^1 ; W_2 = \bigcup_{j=1}^h T_j^2$$

Se verifica que  $W_2 \subseteq W_1$  ya que de no ser así:

$\exists a \in X$  tal que  $a \in W_2$  y  $a \notin W_1$ . Sea  $T_j^{*2} / a \in T_j^{*2}$  es evidente que como  $a \notin W_1$  no existe  $T_i^1$  que contenga a  $T_j^{*2}$ , lo que contradice la definición de consulta más fina.

Obviamente si  $W_2 \subseteq W_1 \Rightarrow X - W_2 \supseteq X - W_1$  o equivalentemente  $P_2 \supseteq P_1$ .

A partir de aquí se tiene que:

$$S_{h+1}^2 \supseteq S_{q+1}^1$$

La Propiedad 1 tiene importantes consecuencias, ya que, lo que en definitiva dice es que:

Toda jerarquía de consulta definida por la relación "mas fina que"  $\ll$  induce una jerarquía con una jerarquía de secuencias de subestructuras-AP que verifican la relación de inclusión. Si a estas secuencias se le añade la estructura-AP que convierte a esta en una partición; la relación de inclusión se invierte para esta última.

La definición de jerarquía de consulta, da paso a establecer una forma de Roll-Up y Drill-Down sobre la dimensión-AP.

Se considera una consulta  $C^1$  y su partición asociada  $P^1$ .

- Se hace Roll-Up sobre  $C^1$  si se considera una consulta  $C^2$  tal que  $C^1 \ll C^2$  y su partición  $P^2$  asociada; y construimos el cubo de datos. Es decir, si se sube con una consulta donde las " frases son menos precisas".
  
- Se hace Drill-Down si se considera una consulta  $C^3$  tal que  $C^3 \ll C^1$  y su partición  $P^3$  asociada. Es decir si se baja con una consulta donde las frases son más finas que en  $C^1$ .

En resumen, para crear la jerarquía de una consulta  $C^1$  se tienen que crear frases de consulta  $C^2$  que cumplan la condición de que  $C^1 \ll C^2$ , o sea que los elementos de  $C^2$  estén en  $C^1$ , y cumpliendo que  $C^1$  sea la más detallada, la de más elementos. Además se deben crear frases de consulta  $C^3$ , que cumpla la condición  $C^3 \ll C^1$ .

Ahora, ésta sería una jerarquía de tres niveles. El nivel de  $C^1$  que es el nivel más fino o más detallado con respecto al nivel de  $C^2$ . Y el nivel  $C^3$  es el nivel más fino con respecto a  $C^1$  y  $C^2$ . Por tanto para hacer Roll-Up, hay que ir de  $C^1$  a  $C^2$ . El Drill-Down se efectúa cuando se va de  $C^1$  al de  $C^3$ .

Es importante aclarar que las definiciones de las operaciones Roll-Up y Drill-Down para la jerarquía de consulta, no están en contradicción con sus definiciones básicas, a pesar de que se ha invertido el uso de los verbos bajar y subir para cada una. Las siguientes figuras 3.3 y 3.4 ayudan a aclararlo. Nótese que en la figura 3.4, hemos invertido la posición de las frases  $C^2$  y  $C^3$  con respecto a como se encuentran en el retículo de la estructura-AP, para que se pueda apreciar con claridad la equivalencia de esta jerarquía con la clásica de la figura 3.3.

### 3.3 - Operaciones sobre una estructura-AP asociada a consulta 85

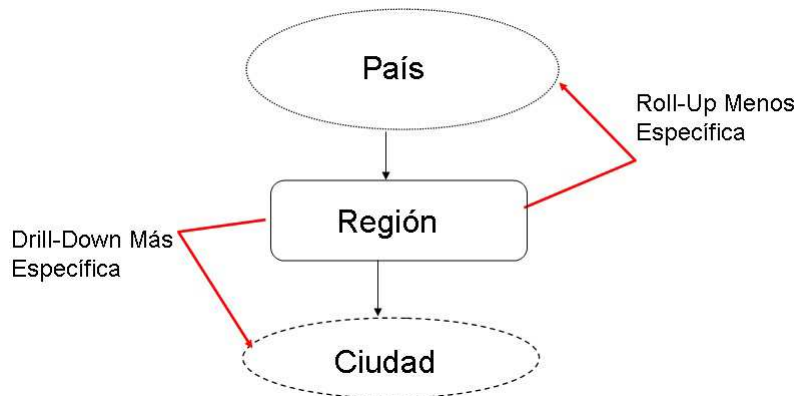


Figura 3.3: Ejemplo básico del uso de las operaciones Roll-Up y Drill-Down

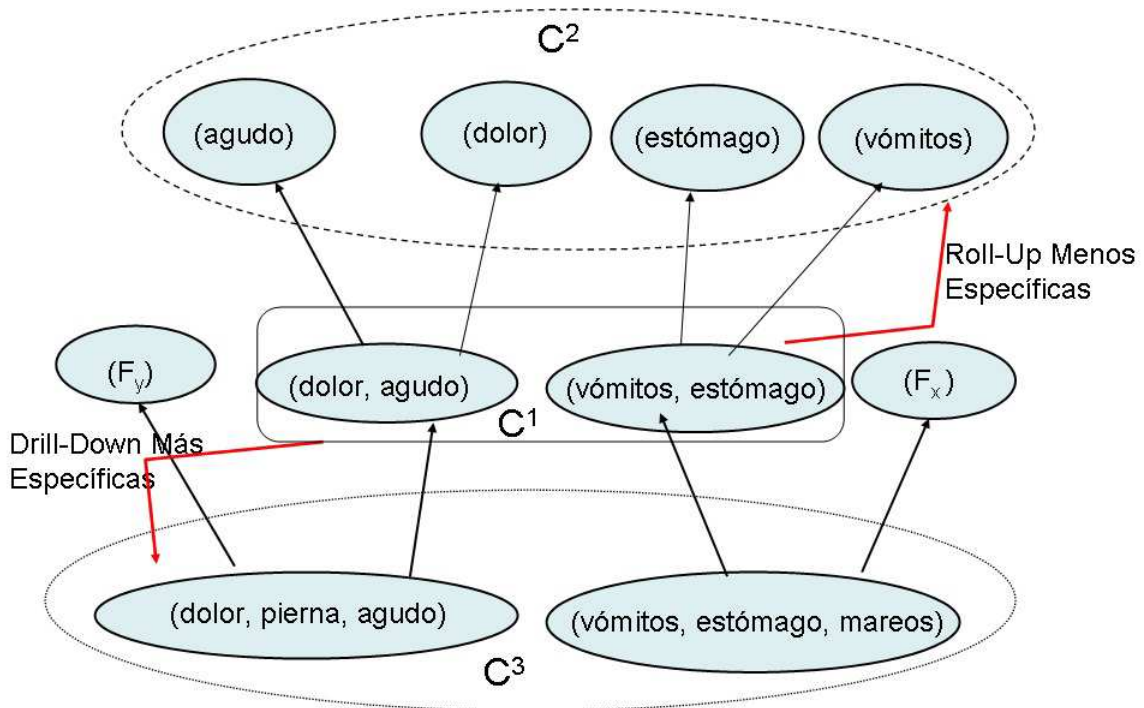


Figura 3.4: Ejemplo del uso de las operaciones Roll-Up y Drill-Down en una jerarquía de consultas

En la gráfica 3.4 se ha mostrado solo una parte de la jerarquía que se pudiera formar, o sea, que se pueden generar muchas más frases de  $C^3$  y de  $C^2$ .

### 3.3.2. Relaciones entre la forma de acoplamiento y las operaciones de Drill-Down y Roll-Up definidas previamente

Sean  $C^1$  y  $C^2$  dos consultas sobre  $X$ ; tales que  $C^1 \ll C^2$

Donde  $C^1 = \{T_1^1, \dots, T_q^1\}$  y  $C^2 = \{T_1^2, \dots, T_h^2\}$

Entonces

$$P^1 = \{S_1^1, \dots, S_q^1, S_{q+1}^1\}$$

$$P^2 = \{S_1^2, \dots, S_h^2, S_{h+1}^2\}$$

son las particiones asociadas a ambas consultas. Se considera entonces una forma de acoplamiento a ( $f$  ó  $d$ ) y una medida  $j$  conteo, asociada a las consultas.

Si se tiene a

$$j_1^1, \dots, j_q^1, j_{q+1}^1$$

$$j_1^2, \dots, j_h^2, j_{h+1}^2$$

como valores asociados a un acoplamiento, donde:

$$\begin{aligned} \forall i \in \{1, \dots, q\} \exists j \in \{1, 2, \dots, h\} \text{ tal que } j_i^1 \leq j_j^2 \\ \forall j \in \{1, \dots, h\} \exists i \in \{1, 2, \dots, q\} \text{ tal que } j_i^1 \leq j_j^2 \end{aligned}$$

### 3.3 - Operaciones sobre una estructura-AP asociada a consulta 87

Es decir habrá valores superiores en  $C^2$  a  $C^1$  cuando se plantea un acoplamiento igual para ambos.

Si se supone ahora a  $jd$  para un nivel de jerarquía y  $jf$  para otro. Entonces lo que ocurre con la relación si se tiene  $jf$  para  $C^2$  y  $jd$  para  $C^1$ , es lo siguiente:

Sea  $i \in \{1, 2..q\}$  y  $j \in \{1, 2..h\}$  tal que:

$$T_i^1 \supseteq T_j^2$$

Es evidente que:

Toda tupla que tenga acoplamiento fuerte con  $T_j^2$  tiene acoplamiento débil con  $T_i^1$ , luego  $jf_j^2 \leq jd_i^1$

Es decir, en definitiva se tiene:

$$\forall i \in \{1, \dots, q\} \exists j \text{ tal que } jf_i^1 \leq jf_j^2 \leq jd_i^1 \leq jd_j^2 \quad (3.1)$$

Como caso particular se puede definir que:

$jf_j^2 \equiv jd_i^1$ ; si  $T_j^2$  es unitérmino ya que en este caso

$$jf = jd$$

y por lo tanto

$$jf_j^2 = jd_i^1 = jd_j^2$$



### 3.4. Operación Dice para una dimensión-AP

**Definición 16. Operación Dice para una dimensión-AP**

*Dada una dimensión-AP con una estructura-AP global  $S = \{(A_1, A_2, \dots, A_n)\}$ . Sea  $T$  un conjunto de términos tal que  $T \subseteq \bigcup_{i=1}^n A_i$ , decimos que hacemos Dice sobre la dimensión-AP asociada a  $T$  cuando se restringe el dominio de la dimensión a la estructura-AP global:*

$$R = S \wedge T$$

*es decir cuando se tiene en cuenta solo a  $R = \{(A_1 \cap T, A_2 \cap T, \dots, A_n \cap T)\}$  como la estructura-AP global básica.*

### 3.5. Ejemplos que ilustran las definiciones anteriores

A partir de los resultados del ejemplo que se define a continuación se obtienen varias demostraciones de casos particulares, que relacionados entre sí, facilitan la comprensión de las definiciones anteriores.

**Ejemplo 6**

Sean  $C^1 = \{(dolor, cabeza), (fractura)\}$  y  $C^2 = \{(dolor), (fractura)\}$

Entonces  $P^1$  y  $P^2$  respectivamente las particiones asociadas a las consultas

$$P^1 = \{(dolor, agudo, cabeza), (fractura, pierna), (otros)\}$$

$$P^2 = \{(dolor, agudo, pierna), (dolor, agudo, cabeza), (fractura, pierna), (otros)\}$$

Estas particiones cumplen las tres propiedades planteadas. Se puede ver que existe al menos un conjunto perteneciente a  $P^2$  que está presente en  $P^1$  y que todos los conjuntos de términos de  $P^1$  están en  $P^2$ .

(dolor,cabeza)	(fractura)	Otros	Total
5	5	10	20

Tabla 3.8: Ejemplo de acoplamiento fuerte con  $C^1$ 

$C_{RU}$	$C_{DD}$
(cabeza)	(dolor,agudo,cabeza)
(dolor)	(fractura,pierna)
(fractura)	-

Tabla 3.9: Ejemplo de jerarquía-AP

El cubo que se genera consultando con  $C^1$  y aplicando acoplamiento fuerte se muestra en la tabla 3.8.

Si se quiere saber la cantidad de diagnósticos que de forma más o menos detalladas se relacionan con las frases anteriores, se puede obtener la jerarquía de consulta asociada a  $C^1$ . Ésta se puede ver en la tabla 3.9, donde se muestran las consultas más y menos finas que  $C^1$ .

Dada la jerarquía de consulta mostrada en la tabla 3.9, el usuario con la ayuda de las frases más y menos detalladas, puede construir una nueva consulta. En el caso de que elija frases más detalladas como (*dolor, agudo, cabeza*) se obtienen los resultados de la tabla 3.10.

En cambio si el usuario elige como consulta final una frase menos detallada como (*cabeza*) o (*dolor*) los resultados serían los que se muestran en las tablas 3.11 y 3.12 respectivamente. La tabla 3.11 indica que la mayoría de

(dolor,agudo,cabeza)	Otros	Total
2	18	20

Tabla 3.10: Ejemplo de consulta más detallada que  $C^1$  con acoplamiento fuerte

(cabeza)	Otros	Total
6	14	20

Tabla 3.11: Ejemplo de consulta menos fina que  $C^1$  con acoplamiento fuerte

(dolor)	Otros	Total
13	7	20

Tabla 3.12: Ejemplo de consulta menos fina que  $C^1$  con acoplamiento fuerte

los pacientes que tienen problemas en la cabeza, ese problema es dolor, son cinco de seis. Y en la tabla 3.12 se refleja que la mayoría de los pacientes tienen dolor.

Tras apreciar el uso de la jerarquía de consulta se pueden comprobar las relaciones entre la forma de acoplamiento y las operaciones Drill-down y Roll-Up que se expusieron en la sección 3.3.2

Si se aplica acoplamiento débil con  $C^1$  los resultados se muestran en la tabla 3.13. Los mismos ayudan a demostrar que:

$$jf_j^2 \leq jd_i^1$$

Si se suman los resultados de esta tabla 3.12 y la anterior 3.11 (6+13) para saber cuánto es  $jf_j^2$ , se obtienen 19 tuplas, pero se deben restar las 5 tuplas que coinciden en ambos resultados, entonces son 14 tuplas. Y en la tabla 3.10 se tiene a  $jd_i^1$ , que es 14. Por lo que así se tiene que  $14 \leq 14$ .

También con estos resultados se demuestra que:

(dolor,cabeza)	(fractura)	Otros	Total
14	5	2	21

Tabla 3.13: Ejemplo de acoplamiento débil con  $C^1$

$$jf_i^1 \leq jf_j^2 \leq jd_i^1 \leq jd_j^2 \text{ (K)}$$

Ya que según la tabla 3.8,  $jf_i^1 = 5$ , entonces  $5 \leq 14 \leq 14 \leq 14$ .

Se cumple así que  $jf_j^2 \equiv jd_i^1$ , y que  $jf_j^2 = jd_i^1 = jd_j^2$ , debido a que  $T_j^2$  es unitérmino, en este caso es (*cabeza*) o (*dolor*).

En el capítulo siguiente se muestra una aplicación práctica del uso de una dimensión-AP de forma conjunta con otras dimensiones clásicas, sobre datos médicos.

### 3.6. Conclusiones

El nuevo modelo de datos multidimensional y las operaciones que sobre él se han definido, ofrecen ventajas analíticas de relevancia. Gracias al uso de una dimensión-AP se puede utilizar la información implícita contenida en los atributos textuales de las bases de datos; ésta constituye una solución al problema planteado de la falta de estructura de estos atributos.

El modelo también garantiza que dichas dimensiones textuales puedan ser combinadas con otras del tipo clásico y sobre ellas se definen las mismas operaciones que tiene definido el modelo clásico sobre las dimensiones clásicas. Por ejemplo el uso de una jerarquía y de operaciones OLAP como el Dice.

La jerarquía asociada a una determinada consulta que se propone usar sobre una dimensión-AP, constituye de gran ayuda al usuario porque mientras que las jerarquías clásicas son agrupaciones propuestas por el cliente, estas nuevas jerarquías deben ser generadas automáticamente por el sistema que implementa el modelo. Como se podrá apreciar en el capítulo siguiente, este sistema ofrecerá al cliente una jerarquía de consultas compuesta por frases de consultas más y menos detalladas con respecto a la consulta inicial.

El análisis de los ejemplos mostrados, demuestra la veracidad de lo planteado hasta aquí, y principalmente la utilidad del modelo; ya que queda claro que se obtiene información relevante sobre atributos textuales usando el nuevo modelo multidimensional propuesto.

En el capítulo que se presenta a continuación se describe la arquitectura del sistema OLAP implementado, Wonder v3.0, junto con otros procesos y herramientas necesarias para realizar una implementación de un data warehousing con el uso de una dimensión-AP. También se realizan implementaciones de data warehousing con datos bibliográficos y médicos que llevan a la práctica el nuevo modelo obtenido.

## Capítulo 4

# Sistema data warehousing con el uso de dimensiones-AP

En este capítulo se presenta el sistema OLAP implementado, Wonder v3.0, junto con otros procesos y herramientas necesarias para realizar una implementación de un data warehousing con el uso de una dimensión-AP. Este sistema forma parte de un marco de trabajo donde se describe además, un resumen del preprocesamiento inicial de los atributos textuales, para obtener el conocimiento asociado a los mismos, necesario en la implementación de dicha dimensión-AP.

Además se exponen dos implementaciones de data warehousing, con datos de publicaciones científicas y datos médicos, respectivamente. En las que se lleva a la práctica el nuevo modelo multidimensional propuesto en el capítulo anterior. Demostrar la factibilidad del mismo en dichos entornos es el objetivo fundamental de este capítulo.

Los ejemplos que se describen muestran consultas sobre cubos de datos que contienen dimensiones-AP y clásicas de forma conjunta. A través de ellos se corrobora el correcto funcionamiento de la herramienta implementada y el cumplimiento del mencionado objetivo.

## 4.1. Descripción del sistema

El sistema desarrollado, Wonder V3.0 (Batista y Tejeda, 2009), es un servidor OLAP de libre disposición. Fue implementado con técnicas y herramientas de software libre y brinda servicios de gestión de cubos OLAP para cualquier base de datos en PostgreSQL. Tiene una particularidad que lo distingue y es que implementa un nuevo modelo multidimensional con soporte a textos libres contenidos en los atributos textuales de base de datos.

Este modelo, como ya sabemos, parte de la premisa de que existe una semántica básica subyacente en determinados atributos textuales de base de datos (Martín-Bautista et al., 2008), debido a que se pueden encontrar conjuntos de términos relacionados semánticamente que aparecen repetidamente en ellos. Una forma de extraer ese conocimiento es, como se ha expuesto, usar una forma intermedia basada en itemsets frecuentes obtenidos vía el algoritmo Apriori. Seguidamente se obtienen todos los conjuntos-AP de la base de datos, y se construye una estructura global intermedia nombrada estructura-AP. Esta estructura refleja el conocimiento contenido en el atributo textual. Bases de datos que contengan esta estructura constituyen la fuente de datos de nuestro sistema.

Se parte de una base de datos inicial que pasa por un proceso de transformación previo. Para ello usamos el sistema de Minería de textos Text Mining Tool (Martínez, 2008). Esta herramienta se encarga de procesar los atributos textuales que se deseen y da como resultado un atributo-AP que contiene la estructura de conocimiento correspondiente a cada atributo textual. De esta manera se obtiene la base de datos modificada que utiliza Wonder como fuente de datos (ver figura 4.1). Seguidamente se pueden definir cubos de datos que cuenten con dimensiones-AP y dimensiones clásicas indistintamente. Sobre esos cubos se podrán realizar operaciones OLAP como por ejemplo el Dice o el uso de una jerarquía de consulta asociada a las dimensiones-AP.

### 4.1.1. Arquitectura general del sistema

La arquitectura general del sistema se muestra en la figura 4.1. A continuación describimos sus componentes.

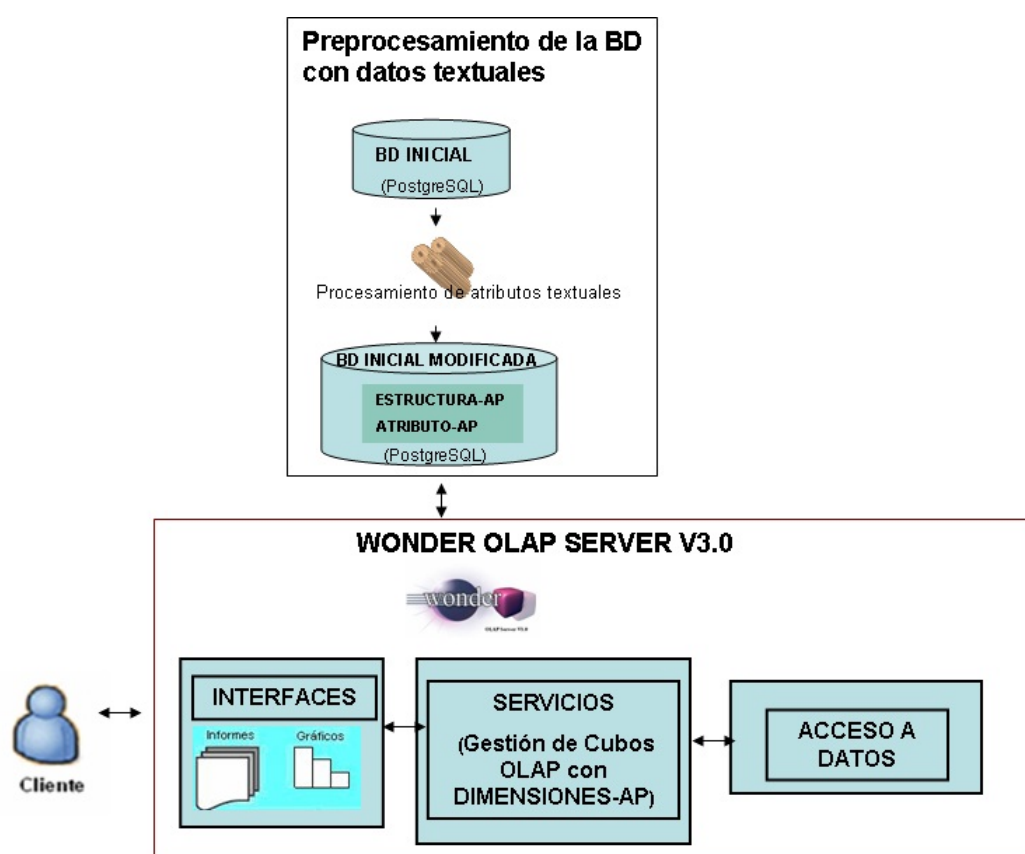


Figura 4.1: Arquitectura general del sistema

- **Procesamiento de la BD con datos textuales:** La base de datos (BD) inicial tiene como característica fundamental que debe estar implementada sobre PostgreSQL, debido a que es el servidor de datos perteneciente a la herramienta OLAP implementada. En esa BD se



almacenan diferentes tipos de datos, incluyendo atributos textuales. Aquellos atributos textuales con relevancia en el proceso de toma de decisiones pueden ser transformados, a fin de realizar consultas útiles. Esta transformación consiste en la realización de un proceso de minería con el fin de obtener una estructura de conocimiento de los mismos. La figura 4.2 lo describe.

La BD Inicial Modificada almacena los resultados obtenidos en dicho proceso. Contiene las nuevas estructuras de conocimiento correspondientes a los atributos textuales de la base de datos inicial, como son la estructura-AP global y las subestructuras-AP inducidas correspondientes a cada tupla del atributo textual original que se procesó; estas subestructuras son los posibles valores de un nuevo atributo, llamado atributo-AP. Sobre la obtención de este atributo-AP se comentará en secciones siguientes.

- **Wonder OLAP Server:** Este sistema tiene como fuente de datos la base de datos modificada anterior; de allí carga la información que necesita para crear sus cubos de datos, y de forma especial, la información contenida en los atributos-AP para la definición de dimensiones-AP.

Está formado por capas, que interactúan entre sí para dar respuesta a las necesidades del cliente. Mediante este sistema, se pueden construir cubos de datos con dimensiones-AP y clásicas indistintamente; de esta forma brinda la posibilidad de convertir en información útil para la toma de decisiones, la información implícita asociada a los atributos textuales contenida en dichas dimensiones-AP. Wonder también permite realizar consultas usando las dimensiones-AP y clásicas de forma conjunta, efectuando las operaciones OLAP usuales del modelo multidimensional (Roll-up, Drill-Down, Slice y Dice) para cualquiera de ellas. Además se destaca el uso de una jerarquía de consulta para las dimensiones-AP que ofrece posibles consultas más y menos específicas que una consulta inicial, usando las funciones de acoplamiento y operaciones que fueron discutidas en el capítulo anterior.

Pasamos a describir con más detalle el preprocesamiento a la base de datos inicial, anteriormente mencionado.

### 4.1.2. Preprocesamiento de la BD Inicial con atributos textuales

Como ya se ha comentado con el uso de la herramienta Text Mining Tool (Martínez, 2008), se puede obtener automáticamente el atributo-AP. Este proceso se realiza sobre los valores del atributo textual deseado, la figura 4.2 describe los pasos que se siguen. A continuación se comentan los componentes de la misma (Martínez, 2008).

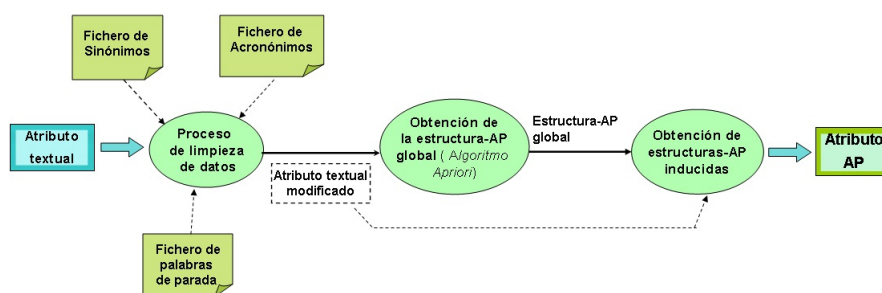


Figura 4.2: Proceso de transformación del atributo textual a la dimensión-AP

1. **Proceso de limpieza de datos:** Aquí se realiza el preprocesamiento del proceso de minería de datos, para realizar la limpieza y homogenización de la muestra inicial. En este caso, se toma como dato de entrada frases de texto corto provenientes de un atributo textual en una base de datos. Utiliza los ficheros de sinónimos y acrónimos para realizar la sustitución de estos en las frases iniciales. También usa el fichero de palabras de paradas para eliminarlas de la frase original. Este proceso da como resultado el fichero de todos los términos diferentes que componen el vocabulario del atributo textual procesado, con su frecuencia de aparición en el total de tuplas procesadas. El producto final que se obtiene, es un nuevo atributo con el texto corto original modificado después de haber realizado la sustitución de sinónimos y acrónimos, y eliminado las palabras de paradas de las frases originales.
2. **Obtención de la estructura-AP global (Algoritmo Apriori):** En este proceso, se parte del texto corto original modificado obtenido anteriormente. Sobre dichos textos se ejecuta el algoritmo Apriori y así se

obtiene el fichero de itemsets frecuentes con su soporte, y los conjuntos-AP detallados con su soporte. Los conjuntos-AP maximales conforman la estructura-AP global que encierra el conocimiento del atributo textual procesado.

3. **Obtención de estructuras-AP inducidas:** Este proceso es el que obtiene las estructuras-AP inducidas para cada tupla de la base de datos, y escribe su representación como un Tipo de Dato Abstracto (TDA)<sup>1</sup> en un nuevo atributo en la tabla desde la que provienen los datos originales. Para obtener el TDA, se realiza la intersección entre la estructura-AP y el texto corto modificado que se obtiene a la salida del proceso de limpieza de datos.
4. **Atributo-AP:** En este atributo ya se tiene la estructura de conocimiento correspondiente a cada tupla del atributo textual original, o sea, el TDA anteriormente mencionado. Como se analizó en el capítulo anterior, este se puede transformar en una dimensión-AP porque se ha comprobado que un atributo-AP cumple con las condiciones de una dimensión: tener un dominio y una partición.

Pasamos a describir con más detalle la arquitectura del cliente Wonder OLAP Server.

### 4.1.3. Arquitectura de Wonder v3.0

La arquitectura de Wonder ha sido implementada tomando como base la arquitectura propuesta por Microsoft para el desarrollo de aplicaciones Cliente-Servidor en varias capas (Microsoft-Co., 2006). Concretamente, esta arquitectura propone un grupo de capas lógicas en las que se puede dividir la aplicación cliente, para un mejor funcionamiento y comunicación entre todos los componentes de un software. En la figura 4.3 se muestra esta arquitectura.

En la figura 4.3 las INTERFACES son las que le permiten a los usuarios ver los cubos, los informes y los gráficos de la aplicación. La capa de los

---

<sup>1</sup>Es un tipo de dato que contiene la estructura de almacenamiento de los datos y los métodos asociados para su manejo.

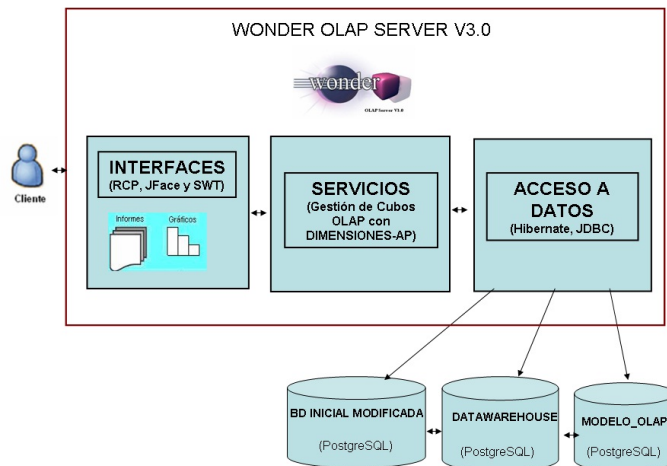


Figura 4.3: Arquitectura en capas de Wonder

SERVICIOS es la contenedora de todos los servicios y la capa ACCESO A DATOS es la encargada del acceso a los metadatos, cubos OLAP y datos fuentes. Esta capa de acceso a datos se comunica con tres bases de datos, la base de datos fuente que es desde donde se extraen los datos para poblar los cubos, la base de datos donde se guardan los modelos y la base de datos donde se guardan los cubos originales o refinados. Las capas pueden describirse en detalle como sigue:

1. INTERFACES: esta capa de la aplicación contiene las interfaces que se muestran al usuario para que interactúe con el software. Permiten la creación de cubos OLAP para después analizar sus datos ya sea a través de gráficos o informes. Las tecnologías que se utilizaron para el diseño de esta capa son: la Plataforma de Cliente Enriquecido de Java (Rich Client Platform (RCP)) (McAffer y Lemieux, 2005), y las librerías JFace y SWT (Bull et al., 2004).

Las interfaces están separadas por vistas, la Vista OLAP que es la vista donde se muestran los cubos creados y las distintas operaciones que se pueden realizar sobre estos; la Vista de Informes que refleja los datos de los cubos y la Vista de Gráficos que es la vista donde se muestran los gráficos obtenidos. La Vista de los Gráficos tiene como limitante que muestra solo los datos bidimensionales, o sea, que si el cubo de datos es de más de dos dimensiones

se le tiene que aplicar la operación Slice.

2. SERVICIOS: es la capa encargada de lograr la sincronización y la organización de las interacciones con el usuario. En ella se encuentran los servicios que están disponibles para cada clase del modelo, además de los servicios que encapsulan toda la lógica del negocio. Esta capa logra la persistencia del modelo de la aplicación a través de la capa de acceso a datos. Aquí son varios los servicios que se implementan. Todos ellos se pueden ver consultando en la ayuda el plugin `cu.jagger.wonder.service.pack.impl`.

3. ACCESO A DATOS: esta capa se encarga de la persistencia de los datos. En ella se encuentran las clases que se encargan de mantener sincronizados los datos que se muestran en la aplicación, con los datos almacenados en las bases de datos. Para cada clase del modelo existe una clase de acceso a datos. Las tecnologías que se utilizaron para su implementación son: Hibernate (Hibernate-org, 2009) para la persistencia de los Metadatos OLAP y JDBC (Conectividad a bases de datos con Java) (Cecchet et al., 2004) para acceder a la fuente de datos y a los cubos OLAP. En esta capa intervienen 3 bases de datos almacenadas en PostgreSQL. Una base de datos fuente (BD Inicial Modificada) que contiene los datos fuentes, una base de datos de modelos (Modelo\_OLAP) donde se almacenan los metadatos correspondientes a los cubos OLAP y otra base de datos (Warehouse), donde se almacenan los cubos de datos construidos, ésta se comunica con la base de modelos para la construcción de sus cubos OLAP y con la base de datos fuente para cargar los datos necesarios para dichos cubos. En Warehouse puede haber cubos que contengan dimensiones clásicas y dimensiones-AP indistintamente.

Como SGBD Wonder V3.0 utiliza PostgreSQL (PostgreSQL-Global-Development-Group, 2009) que además de ser libre es gratuito y se puede descargar libremente de su Sitio Web (<http://www.postgresql.org/>) para multitud de plataformas. La versión oficial actual de PostgreSQL es la 8.4.4, liberada en mayo de 2010. Está considerado como la base de datos de código abierto más avanzada del mundo. PostgreSQL proporciona un gran número de características que normalmente sólo se encontraban en las bases de datos comerciales tales como DB2 u Oracle.

## 4.2. Definición de un cubo de datos en Wonder con el uso de dimensiones-AP

### 4.2.1. Formación del cubo de datos

En esta sección se describe el proceso de creación de un cubo de datos con el uso de Wonder. En la figura 4.4 se muestra este proceso, partiendo de la BD Modificada que se ha comentado anteriormente, como la fuente de datos del sistema.

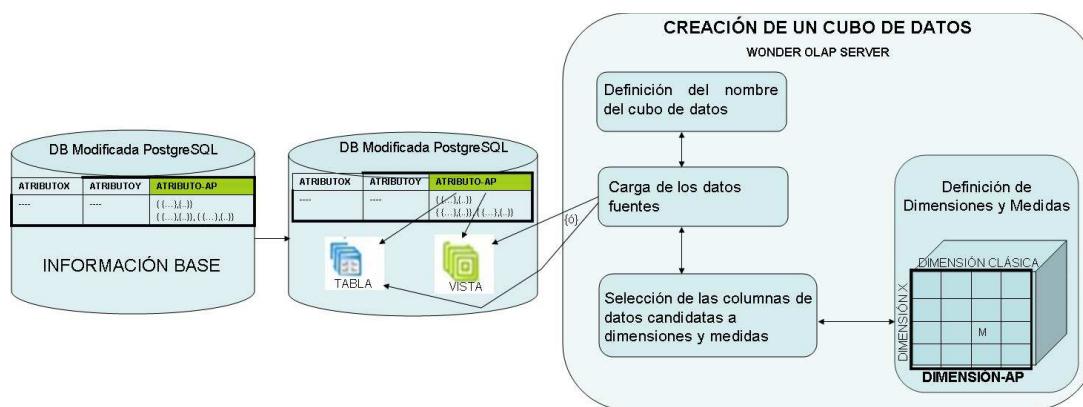


Figura 4.4: Proceso de obtención de cubos de datos con dimensiones-AP

Los pasos que hay que realizar con el fin de implementar un cubo de datos son los siguientes:

1. Lo primero es contar con la INFORMACIÓN BASE. Que es la base de datos (BD Modificada PostgreSQL) que ya hemos comentado y constituye la fuente de datos que utiliza el sistema.
2. Se construyen en la propia base de datos fuente (Base de Datos Modificada PostgreSQL) las tablas o vistas que reúnan toda la información que se desee contengan sus cubos de datos.
3. Se carga la información de un cubo a la vez, desde el Wonder, utilizando el asistente de mapeo de cubos. Los pasos fundamentales en esta tarea son:

- a) Se define el nombre del cubo.
  - b) Se selecciona la tabla o vista desde la que se cargan los datos para dicho cubo OLAP.
  - c) Se seleccionan las columnas de la tabla o vista seleccionada que serán definidas como dimensiones y medidas del cubo.
  - d) A continuación, siguiendo el asistente, se definen cada columna seleccionada como las dimensiones y medidas del cubo.
4. Se finaliza guardando el cubo de datos, que ya estará listo para hacer consultas. Este cubo usa los dominios básicos de todas sus dimensiones, incluyendo las dimensiones-AP.

A continuación se explica cómo a partir de un atributo-AP se obtiene la dimensión-AP del cubo.

#### **4.2.2. Obtención de una dimensión-AP a partir de un atributo-AP**

Como ya se explicó en el capítulo anterior, para definir un atributo-AP como una dimensión de un modelo multidimensional se tuvieron que plantear dos conceptos fundamentales, el de dominio y el de partición.

En la figura 4.5 se puede apreciar la estrecha relación de un atributo-AP con una dimensión-AP. Es importante recordar que aunque en la dimensión-AP se tiene la misma representación interna de los valores de un atributo-AP, o sea estructuras, la salida y consulta de la información a un usuario se hace en forma de conjuntos de conjuntos de términos ("frases"), que no son más que los conjuntos generadores de las estructuras-AP. Es decir en el proceso OLAP de consultas, el usuario establece conjuntos de frases, como valores de la dimensión. En dicho proceso se tiene en cuenta siempre que estos valores serán elementos del dominio definido sobre el atributo-AP, que es básicamente el mismo dominio de la dimensión-AP correspondiente.

En la sección que sigue se ofrecen ejemplos de consultas sobre cubos de datos que contienen dimensiones-AP, para dos aplicaciones diferentes: data warehousing con datos médicos y con datos de artículos científicos.

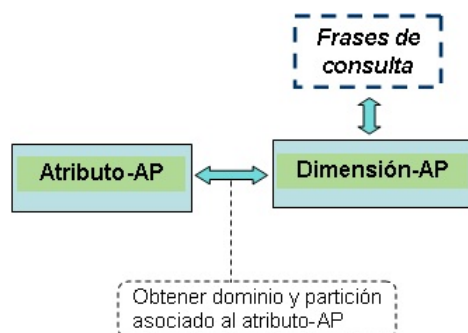


Figura 4.5: Relación de un atributo-AP con una dimensión-AP asociada

### 4.3. Data warehousing de publicaciones científicas

El estudio realizado en el capítulo 2 sobre algunas propuestas de sistemas data Warehousing, respalda la afirmación de que este tipo de sistemas, gracias a su naturaleza analítica, ofrece excepcionales ventajas para el procesamiento de textos. La mayoría de estas referencias proponen la utilización de estrategias para procesar documentos textuales, utilizando formatos XML, UML o definiendo nuevas arquitecturas sobre los modelos multidimensionales de OLAP.

En algunos de esos trabajos, (Tseng y Chou, 2006), (Ravat et al., 2007), (Yu et al., 2009) (Turkka et al., 2008), se utilizan datos de publicaciones científicas como fuente de información de los casos de estudios que implementan, así como para exponer los modelos multidimensionales que proponen. Nos ha parecido oportuno referirnos nuevamente a ellos como un análisis introductorio que nos sirve de motivación a nuestra propia implementación de un data warehousing de artículos científicos.

En Tseng como ya se ha expuesto definen un modelo multidimensional de data warehousing de documentos, como alternativa de modelado multidimensional de textos, facilitando así el procesamiento de documentos textuales. Brindan la ventaja de que la medida clásica cantidad, de sus cubos de datos son además enlaces a los documentos fuentes correspondientes.



En este trabajo se muestran ejemplos de procesamientos OLAP que pueden realizar los usuarios sobre el cubo del documento. Se evidencia en los mismos la utilidad práctica de este tipo de cubo para el caso de publicaciones científicas. El modelo que plantean se especializa en el trabajo con documentos textuales como fuente de datos externa, pero no valora los textos contenidos en atributos de bases de datos, ni la semántica asociada a los mismos.

Este trabajo ofrece la posibilidad de realizar un procesamiento textual sobre información relevante de las publicaciones científicas, como son la fuente, la categoría a la que pertenece y su fecha de publicación, pero, consideramos que no incluye otros datos que sería interesante manejar, como sus autores o palabras claves.

También es importante destacar que el procesamiento textual se basa en la extracción de las partes fundamentales de un documento textual, más allá del contenido explícito, solo obtienen las palabras claves, pero no extraen la semántica asociada a cada uno de los componentes textuales de un documento.

También en (Ravat et al., 2007) relacionan el modelo multidimensional con datos de artículos científicos. Se destaca en este modelo el uso de la función de agregación `Top_KeyWord` para obtener una novedosa medida textual que contiene las palabras claves. Pero no procesan los textos contenidos en las dimensiones. Es importante resaltar que los cubos de datos contienen los enlaces que tienen los artículos entre ellos, como por ejemplo las referencias. A pesar de que nuestro trabajo no valora medidas textuales, si pudiera implementarse en un futuro sin demasiadas complicaciones, porque la estructura textual que definimos como dimensión, podríamos estudiar si se ajusta también como medida.

En las conclusiones mencionan muy brevemente que se usa XML para el trabajo con los documentos. Este es un detalle importante porque supone el uso de otro lenguaje adicional que es el XML. El prototipo que implementan usa como servidor de bases de datos a Oracle 10g, contando así con todas las desventajas que supone el uso de software propietario (Sánchez, 2004).

En (Yu et al., 2009) desarrollan un prototipo de sistema Web llamado `iN-extCube` que integra las dos recientes investigaciones que han desarrollado,

el Text-cube y el Topic-cube. La red de información InextCube se enriquece con esos dos resultados. Las tareas fundamentales que realiza son: la generación de ranking (Yizhou et al., 2009) por autor, conferencias y términos, procesos de clustering con esos ranking, y la construcción de TextCube o TopicCube donde se muestran por autor, conferencia y año, un ranking de los términos frecuentes o tópicos de las publicaciones con su probabilidad de ocurrencia respectivamente. Las funcionalidades del sistema se demuestran en <http://inextcube.cs.uiuc.edu>. Usan como fuente de datos la base de datos DBLP.

La demo que implementa esta propuesta limita al usuario a usar solo las dimensiones conferencias, autor y fecha relativos a los artículos, y el procesamiento textual que brinda con el uso de TextCube y TopicCube solo devuelve como medidas la probabilidad de los términos, el ranking o el score. No se combinan estos procesamientos con las medidas clásicas de OLAP como la suma o el promedio, de forma tal que no solo devolviera resúmenes estadísticos, sino por ejemplo, cantidades orientadas a procesos cotidianos que ayudan a la toma de decisiones. Por ejemplo, sería interesante saber la cantidad de artículos relacionados con un autor, conferencia y fecha que tratan de un tópico dado o cuántas veces contienen un término determinado.

Después de analizar las implementaciones de data warehousing de publicaciones científicas que se plantean en algunos estudios previos, se expone a continuación nuestra propuesta. La misma tiene como objetivos, mostrar soluciones principalmente a dos de las deficiencias anteriormente planteadas, una el número limitado de dimensiones de datos sobre las publicaciones y otra y la más importante es la poca utilidad que le dan a la semántica asociada a las dimensiones textuales y que las dimensiones textuales que proponen por lo general no las tratan conjuntamente con otras del tipo clásicas sin la necesidad de medidas específicas, como en (Yu et al., 2009).

A continuación se expone el diseño de una propuesta similar a la que anteriormente presentamos, de Tseng y Chou. Teniendo como referencia el modelo dimensional que ellos proponen, sugerimos otro con un mayor número de dimensiones, con el objetivo de enriquecer el procesamiento sobre un documento. Se implementan en secciones posteriores la implementación de varios ejemplos de cubos de datos OLAP y consultas. Primero mostramos ejemplos

de consultas clásicas, sin el uso de dimensiones-AP. Por último se podrán apreciar consultas sobre cubos de datos con dimensiones-AP mostrando así la utilidad de las mismas.

### 4.3.1. Origen y descripción del modelo de datos a utilizar

El primer paso para obtener nuestra implementación se refiere a la obtención de los datos. El experimento se realiza, sobre una porción (2235 tuplas referidas a artículos científicos) de datos de la base de datos SICA (Sistema de Información Científica Andalucía) relativos a las TIC, es decir, publicaciones referentes a la rama de la Tecnología de la Información y la Comunicación en Andalucía.

Los datos extraídos pertenecen a varios artículos científicos, las revistas donde estos están publicados, sus autores, las palabras claves definidas por los autores y el año de publicación de los mismos. Es necesario señalar que dichos años de publicación están en el rango de 1993 al 2007. El modelo relacional de tablas se presenta en la figura 4.6.

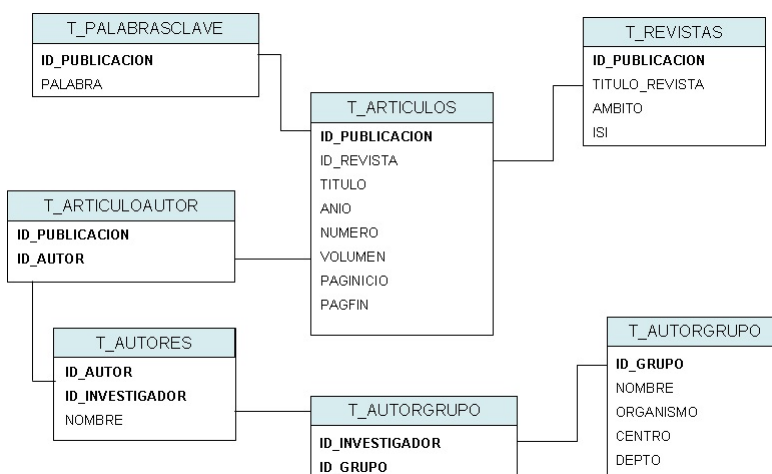


Figura 4.6: Modelo relacional de tablas de la base de datos de publicaciones

Una base de datos relacional como la anteriormente mostrada, almacena entidades en tablas normalizadas relacionadas entre sí. Estas bases de datos tienen como meta evitar la redundancia e inconsistencia de los datos. Su estructura es conveniente para sistemas OLTP ( Procesamiento Transaccional en Línea) pero para complejas consultas multitablas es relativamente lenta. Por razones como éstas surgen otros modelos de datos, como el modelo dimensional.

El modelado dimensional está relacionado estrechamente con OLAP; éste se adapta a entornos como ROLAP y MOLAP; el primero se refiere al procesamiento analítico en línea sobre bases de datos relacionales, y el segundo al procesamiento analítico en línea pero sobre bases de datos multidimensionales. En nuestro caso implementaremos el ROLAP, sobre la base de datos relacional que se ha mostrado.

Antes de pasar al diseño del modelo dimensional, fueron necesarias algunas transformaciones a los datos que ayudaran en el análisis OLAP futuro, como: agregarle a la tabla OLAP\_ARTÍCULOS un campo NUMPAG donde se tenga directamente la cantidad de páginas de cada artículo, limpiar de espacios en blanco los títulos de las revistas y cambiar el tipo de datos de los años de publicación de textual a entero.

Con ayuda del sistema Text Mining Tool, aplicando un soporte 0.01 obtuvimos el atributo-AP correspondiente a los títulos de los artículos. Escogimos este campo textual por la importancia que éste contiene. La mayoría de las búsquedas de información útil sobre publicaciones científicas se realizan sobre los títulos de los mismos. Este atributo-AP se usa para definir la dimensión-AP correspondiente.

### 4.3.2. Transformación del modelo relacional al modelo dimensional

La construcción del modelo dimensional ROLAP, a partir del modelo relacional de la figura 4.6, se realiza de forma similar a (Kara, 2008). En este trabajo, implementan el esquema estrella, que coincide con la técnica tradicional de modelado dimensional sobre bases de datos relacionales. El esque-

ma puede ser en estrella o en copo de nieve (Inmon, 1996). En nuestro caso implementamos un esquema copo de nieve debido a que este esquema está orientado a facilitar mantenimiento de dimensiones grandes y/o complejas. A continuación se muestra todo el proceso.

- Obtención de las dimensiones y jerarquías:

El conjunto de dimensiones proporcionan el contexto para cada medida de la tabla de hechos. Al indicar un valor para cada dimensión, se situará un hecho concreto dentro del espacio del negocio. En pocas palabras, se podría decir que las dimensiones son los descriptores del negocio.

Basándonos en esta definición, y teniendo como base el modelo relacional mostrado en la figura 4.6, obtuvimos las dimensiones: revistas, autores, fecha, palabras claves y título de artículo, entre otras.

Además, se definen como dimensiones las jerarquías derivadas de cada dimensión, formando así el esquema en copo de nieve. La combinación de las dimensiones con los hechos forman el cubo OLAP, sobre el que se realizan las consultas con diferentes niveles de detalle. Estos niveles agrupan a las dimensiones, y en un orden determinado conforman una jerarquía. Cada dimensión puede contener o no jerarquías. En la tabla 4.1 se muestran algunas dimensiones con sus jerarquías.

Dimensión	Jerarquías
Autores	Pertenencia_de_autor (grupo de investigación, organismo y departamento)
Fecha	Últimos 5 años, 2003-1995 y el resto
Revista	Nacionales, Internacionales, ISI, NO-ISI

Tabla 4.1: Dimensiones y sus jerarquías

- Obtención de la tabla de hechos

La tabla de hechos es la tabla primaria del modelo dimensional, con N llaves externas que corresponden a las llaves primarias de las dimensiones y con M columnas de tipo numérico que contienen medidas de un proceso de negocio.

Para obtener la tabla de hechos, es una buena estrategia , analizar todas las tablas del modelo relacional, y considerar las que contienen la mayor cantidad de registros transaccionales, o que constantemente cambian y tienen gran cantidad de filas. También deben de valorarse las tablas que contienen la mayor intersección de tablas del modelo (Kara, 2008). En nuestro caso, es candidata a tabla de hechos, la tabla de artículos. Como se muestra en la figura 4.6 esta tabla contiene la mayoría de la información transaccional y la intercepción de la mayoría de las llaves primarias del resto de las tablas. Como medidas se tienen el número de páginas de los artículos y la cantidad.

Tras obtener los miembros del modelo dimensional: tabla de hechos y dimensiones, finalmente el modelo relacional se ha transformado en el modelo dimensional en esquema copo de nieve que muestra la figura 4.7, sobre este modelo se realizarán los ejemplos que se muestran a continuación.

A continuación se muestran ejemplos de consultas sobre el hipercubo correspondiente al modelo dimensional 4.7. El mismo se construye con la ayuda de la herramienta Wonder OLAP v3.0. En la tabla DIMTITULOARTICULO que se muestra en la figura 4.7 se puede apreciar el campo TDATITULO, que constituye el atributo-AP que mencionamos en la sección anterior. Este campo contiene la semántica asociada a los títulos de los artículos que con Wonder definimos como una dimensión-AP.

### 4.3.3. Ejemplos implementados

Como ya hemos mencionado la experimentación sobre los datos de publicaciones la realizamos sobre PostgreSQL como servidor de bases de datos y

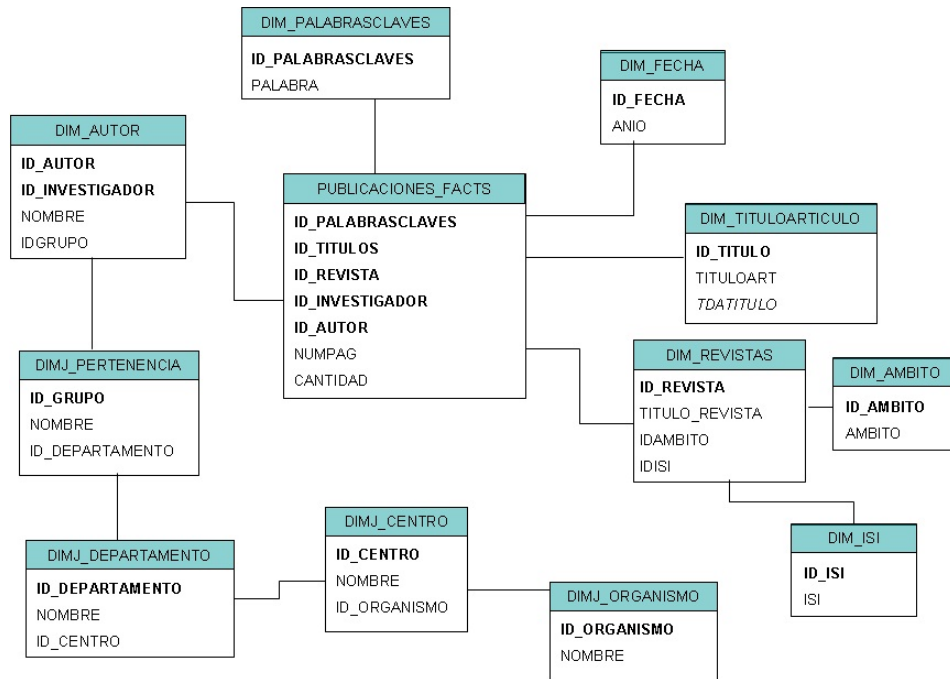


Figura 4.7: Modelo dimensional de los datos de publicaciones

Wonder V3.0 como servidor OLAP de libre disposición (Batista y Tejeda, 2009).

Wonder como servidor OLAP permite que el decisor construya un cubo multidimensional y que posteriormente pueda consultar y refinar esta información almacenada. La misma contiene una estructura definida por él, lo que constituye un paso de avance en la familiarización con la herramienta. En la figura 4.8 podemos apreciar la interfaz del servidor OLAP Wonder V3.0. En el apéndice A se argumenta sobre sus principales características.

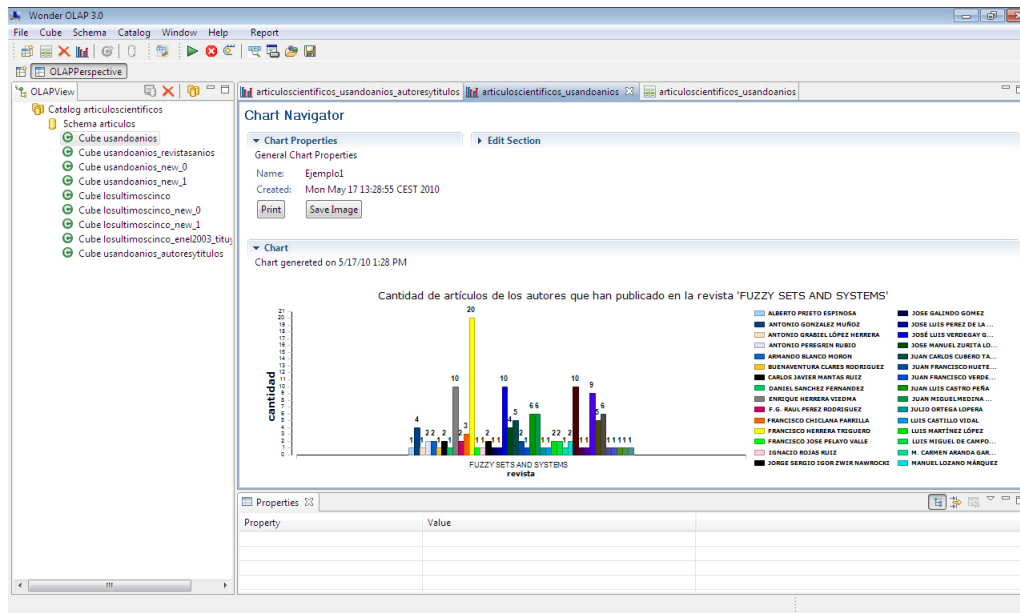


Figura 4.8: Interfaz principal de Wonder

Como se observa en la figura 4.8, la base del trabajo con el servidor OLAP consiste en la definición de cubos de datos sobre consultas a la base de datos de Postgres, o subcubos como resultados de consultas que se realizan sobre los primeros; o sea, son resultado de operaciones (ej. Slice, Dice) sobre cubos de datos ya almacenados en el servidor. El sistema, a partir del árbol de cubos que define en el panel de la izquierda, da la posibilidad de navegar entre ellos y dentro de la estructura de cada uno. La gráfica que se muestra en la parte derecha superior es el resultado de un Dice para la revista "FUZZY SETS AND SYSTEMS" de un cubo que contenía la cantidad total de artículos por revistas y autor.

Las operaciones sobre estos cubos se pueden realizar a partir de diferentes menús contextuales o del menú principal de la aplicación. El panel de la derecha está reservado para la obtención de informes y gráficos como se muestra en la figura. De estos informes se tiene la posibilidad de obtener sus gráficos siempre y cuando el informe sea de dos dimensiones como máximo y un hecho.



## Procedimientos de consultas en Wonder

Antes de pasar a describir los ejemplos es importante conocer los procedimientos que se pueden seguir en Wonder para consultar la información. A continuación se listan los pasos a seguir dentro de la herramienta.

1. Se construyen los hipercubos necesarios, según la estructura dimensional que se tenga. En este caso sería un solo hipercubo con la estructura de la figura 4.7. Teniendo toda la información almacenada físicamente en un cubo de la herramienta, se pueden realizar diferentes operaciones sobre el mismo, en respuesta a las necesidades de información de los usuarios.
2. Para eliminar las dimensiones que no se relacionan con la consulta del usuario, se le realiza al hipercubo una operación *Slice*. Como resultado de ésta operación se almacena un subcubo con las dimensiones deseadas. En Wonder se localiza ésta acción al dar un click derecho sobre el hipercubo y escoger la opción *Create subcube*. Dentro de esta opción se sigue un Wizard donde se pueden combinar las operaciones OLAP: *Slice*, *Dice para* todas las dimensiones y el uso de la *jerarquía asociada a consulta*, esta última para el caso de las dimensiones-AP.
3. Otra opción sería realizar las consultas deseadas directamente sobre la dimensiones del propio hipercubo, con la opción *Cunsulting Cube*. En este caso no se almacena un nuevo cubo (subcubo), sino que se obtienen las gráficas y informes correspondientes directamente.

### Ejemplo 7

Cantidad de artículos por revista y autores: Para dar respuesta a esta consulta se necesita construir un subcubo con la siguiente estructura en el servidor OLAP:

- Dimensiones: revista y autores.
- Medida: cantidad de artículos.
- Función de agregación: SUM

Una vez que se define el subcubo antes mencionado en el ejemplo 7, se puede consultar y ver en un informe o en gráficos sus valores. Sobre él realizamos una operación Dice, que como se explicó anteriormente, es una operación que se define sobre el modelo dimensional y el resultado es que elimina las filas que no cumplen con determinadas condiciones que se definen sobre las dimensiones. Para este ejemplo en concreto la condición fue que la revista tuviera el título "MATHWARE AND SOFT COMPUTING". La figura 4.9 muestra la gráfica obtenida por el servidor OLAP con los datos del subcubo resultante del Dice, con la medida cantidad de trabajos.

Podemos apreciar en la gráfica (figura 4.9) que se genera sobre esa consulta los autores de mayor número de publicaciones en esa revista.

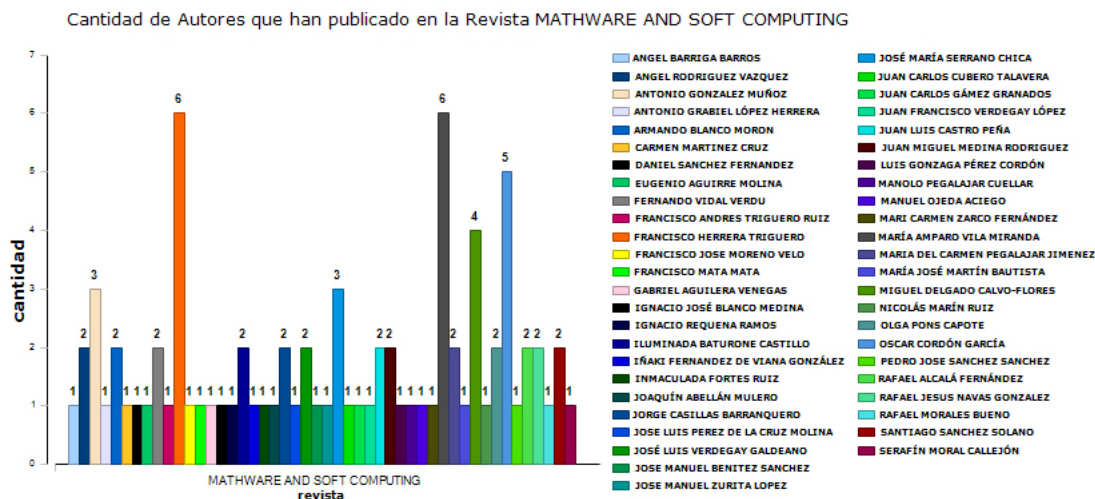


Figura 4.9: Resultado de un Dice por revista al cubo del Ejemplo 7

En los ejemplos de consultas que siguen se utiliza la dimensión-AP *título de artículos*, que combinada con otras dimensiones clásicas conforman subcubos de datos de gran utilidad para la toma de decisiones. De forma tal que se pueden responder interrogantes como: ¿Qué artículos tratan de determinado tema? ¿Qué cantidad de artículos presenta determinada frase en una revista dada? ¿Cuáles son los autores con la mayor cantidad de artículos publicados sobre un tema dado?

### Ejemplo 8

Obtener la cantidad de artículos que contienen completamente en sus títulos (dimensión-AP), la frase  $F = \{(COMPUTATIONAL, MODEL), (DATABASES, DEDUCTIVE, FUNCTIONAL, LOGIC)\}$ , que como vemos está compuesta a su vez por dos frases.

Para obtener los resultados requeridos en el ejemplo 8 se realizó una consulta sobre la dimensión-ap del hipercono, usando como función de acoplamiento el acoplamiento fuerte, como medida la cantidad de artículos y SUM como función de agregación. Podemos ver como en la figura 4.10 se muestran para esta consulta escasos resultados. Solo algunos artículos contienen completamente en sus títulos las frases dadas. Si aplicamos acoplamiento débil en lugar de fuerte, los resultados aumentan, lógicamente, como se observa en la figura 4.11.

Si el usuario lo desea, puede usar la jerarquía de consulta asociada a la consulta realizada. Esta jerarquía ofrece posibles nuevas consultas, más o menos detalladas que la inicial, con las que se garantiza que se encontrará información en la base de datos. De esta forma se pueden obtener nuevos resultados que mantienen relación con los ya obtenidos.



Figura 4.10: Cantidad de artículos que contienen en sus títulos, acoplamiento fuerte con determinadas frases de consultas

En la figura 4.12 se expone la jerarquía obtenida. Como se observa en la ventana correspondiente a esta operación se brindan posibles consultas menos detalladas (Less detail) pero no más detalladas. Eso ocurre porque no existe en la base de datos información que contenga frases más complejas que F.

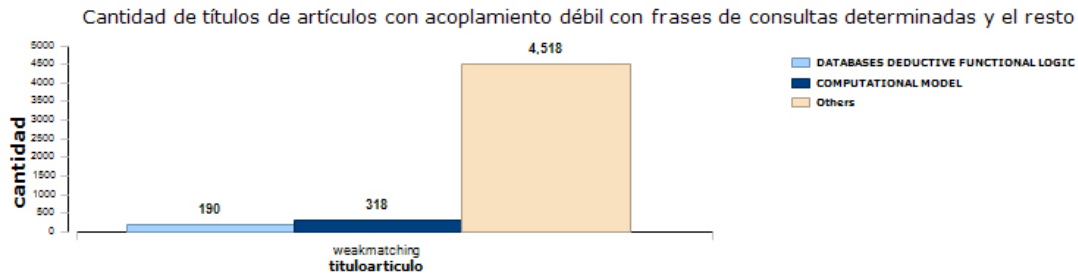


Figura 4.11: Cantidad de artículos que contienen en sus títulos, acoplamiento débil con determinadas frases de consultas

Teniendo la jerarquía se puede construir una nueva consulta (Final Query) con la ayuda de las posibles que se ofrezcan. Seleccionando una o más del grupo de las menos detalladas o de las más detalladas.

Si por ejemplo construimos una consulta final con las frases menos detalladas: {(DATABASES, FUNCTIONAL), (LOGIC), (MODEL)}, aplicando acoplamiento fuerte los resultados que se obtienen se muestran en la figura 4.13. Si aplicamos acoplamiento débil los resultados se observan en la figura 4.14. En esta figura 4.14 podemos apreciar que se han encontrado mayor variedad de resultados, o sea, hemos ofrecido mayor información al usuario sobre el contenido de los datos, que aunque no contienen las frases iniciales, si contienen partes de ellas.

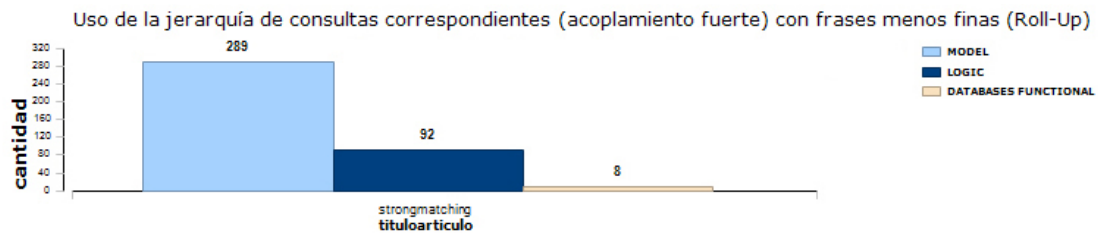


Figura 4.13: Resultados obtenidos con el uso de la Jerarquía de consultas con frases menos finas y acoplamiento fuerte

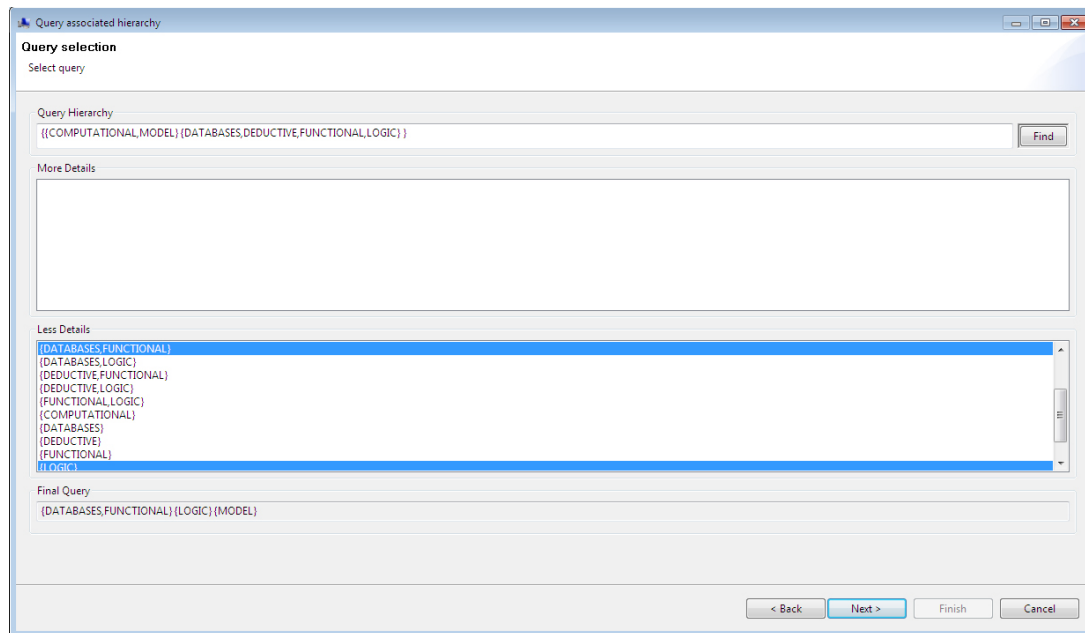


Figura 4.12: Jerarquía correspondiente a la consulta (COMPUTATIONAL, MODEL), (DATABASES, DEDUCTIVE, FUNCTIONAL, LOGIC)

En los ejemplos siguientes se crean subcubos de datos que contienen más de una dimensión. En ellos se pueden ver el uso de una dimensión-AP unidas a dimensiones clásicas como fechas. Primero se combinan dos dimensiones clásicas (revista y fecha) y a partir de los resultados que se obtienen se usa la dimensión-AP (títulos de artículos).

### Ejemplo 9

Obtener la cantidad de artículos publicados en los últimos cinco años en las revistas '*ACM COMPUTING REVIEWS*', '*INTERNATIONAL JOURNAL OR INTELLIGENT SYSTEM*', '*LECTURE NOTES IN COMPUTER SCIENCE*'.

Para dar respuesta a esta consulta se construye un subcubo con la siguiente estructura en el servidor OLAP:

- Dimensiones: revistas y fechas

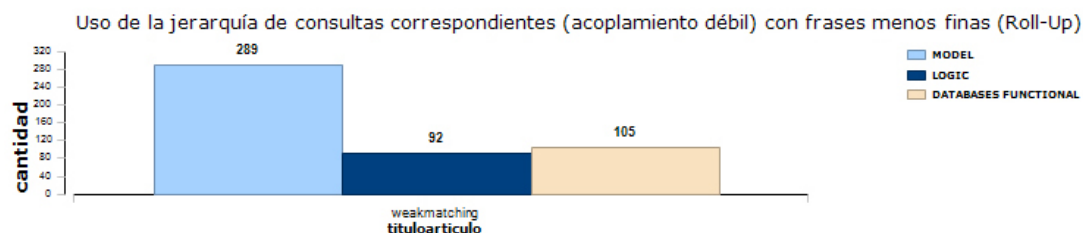


Figura 4.14: Resultados obtenidos con el uso de la Jerarquía de consultas con frases menos finas y acoplamiento débil

- Medida: cantidad de artículos.
- Función de agregación: SUM.

En el caso del ejemplo 9 tras crear el subcubo correspondiente, se aplicó una operación Dice por la dimensión fecha, para quedarnos con los últimos 5 años (teniendo el 2007 como último año en los datos fuentes).

La figura 4.15 muestra los resultados correspondientes a este ejemplo. Nótese que se ha aplicado la operación Dice a la dimensión fecha y se ha evidenciado que en el año 2003 se publicó más que en el resto de los años. Se puede conocer cuántos artículos relacionados con esos temas se han publicado en dichas revistas, en el año 2003. El ejemplo que se plantea a continuación describe tal consulta.

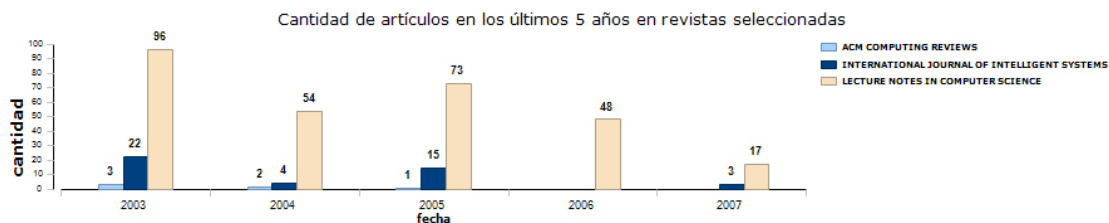


Figura 4.15: Resultados del Ejemplo 9

**Ejemplo 10**

Obtener la cantidad de artículos publicados en el año 2003 en las revistas '*ACM COMPUTING REVIEWS*', '*INTERNATIONAL JOURNAL OR INTELLIGENT SYSTEM*', '*LECTURE NOTES IN COMPUTER SCIENCE*'. Los mismos deben contener completamente en sus títulos (dimensión-AP), la frase  $F = \{(LINGUISTIC, FUZZY), (ANALYSIS, DESIGN)\}$ .

Para dar respuesta a esta consulta se ha decidido construir un subcubo con la siguiente estructura en el servidor OLAP:

- Dimensiones: título de artículo(dimensión-AP), revista, fecha
- Medida: cantidad de artículos.
- Función de agregación: SUM.

En el caso del ejemplo 10 tras crear el subcubo correspondiente, se aplicó una operación Dice por la dimensión fecha, para quedarnos con los publicados en el año 2003.

Ante todo hacer notar que se ha usado el acoplamiento débil, porque este tipo de acoplamiento brinda la posibilidad de crear consultas más abiertas y se querían los artículos relacionados, no solo los que tuvieran en sus títulos exactamente dichas frases, ésto queda a elección del usuario, dependiendo de sus intereses de consulta. No obstante en la figura 4.17 se muestran los resultados usando el acoplamiento fuerte, y como se observa los mismos son escasos. Estos escasos resultados quieren decir que solo en '*LECTURES NOTES IN COMPUTER SCIENCE*' se publicaron en el 2003, títulos que contienen exactamente (acoplamiento fuerte) una de las frases (*ANALYSIS, DESIGN*).

Los resultados de este ejemplo 10, cuando se usa el acoplamiento débil, mostrados en la figura 4.16 son similares a los anteriormente comentados, la mayor cantidad de artículos de ese año relacionados con esas frases, se publicaron en la revista '*LECTURES NOTES AND COMPUTER SCIENCE*'. Esta conclusión puede ser corroborada analizando los resultados del ejemplo anterior (ver figura 4.15). También en la figura 4.16 se aprecia que el uso

de la partición excluyente a las frases de consulta (*others*), es una ayuda para saber la probabilidad de que se puedan encontrar publicaciones sobre esos temas, en esas revistas. Por ejemplo, en la revista *ACM COMPUTING REVIEWS* esa probabilidad es 1 de 3, que es alta (1 trata sobre el tema (ANALYSIS, DESIGN) y 2 sobre otros temas (*others*) que no se relacionan con ninguna de las dos frases de consulta).

Por otro lado el hecho de que no aparezcan publicaciones en esa revista, relacionadas con (LINGUISTIC, FUZZY), puede significar que en la misma se publicaron mayormente artículos que tienen relación con (ANALYSIS, DESIGN) y es más recomendable buscar en la revista *INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS* que es donde aparecen más resultados relacionados con ese tema (4 artículos).

Si en lugar de la cantidad de artículos, el usuario desea conocer el promedio de páginas con que se ha publicado en estas revistas (en el 2003) se realiza una consulta sobre la dimensión-*AP título de artículo* del subcubo de datos creado anteriormente, de forma similar, pero definiendo como medida el número de páginas (*numpag*) y la función de agregación promedio. En la figura 4.18 se pueden ver los resultados.

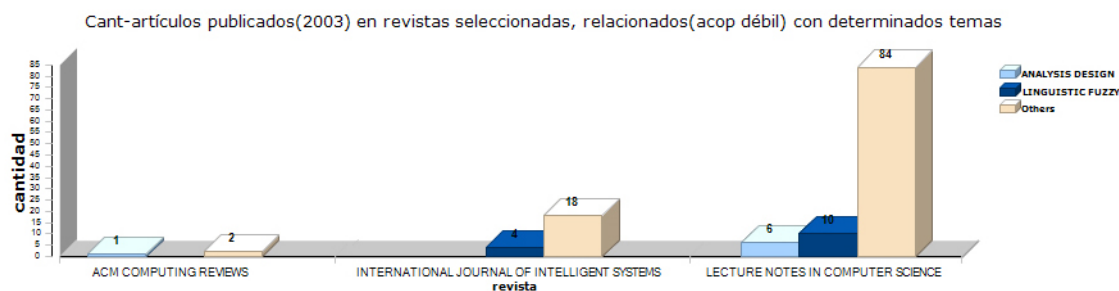


Figura 4.16: Resultados de aplicar acoplamiento débil en el Ejemplo 10

Con la ayuda de la jerarquía de consulta asociada a las frases del ejemplo 10 que se muestra en la figura 4.19 se ha realizado una nueva consulta,



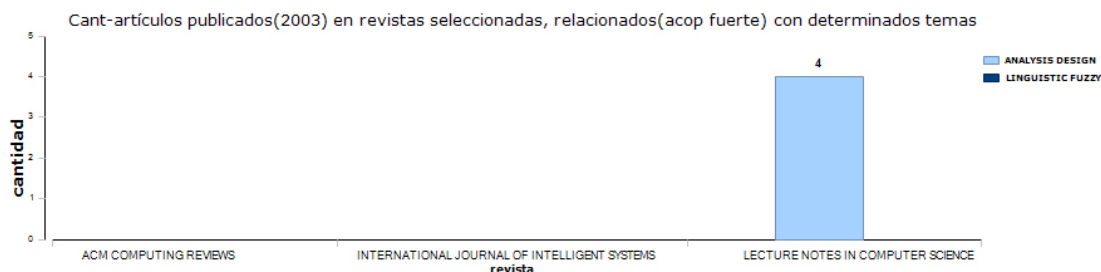


Figura 4.17: Resultados de aplicar acoplamiento fuerte en el Ejemplo 10

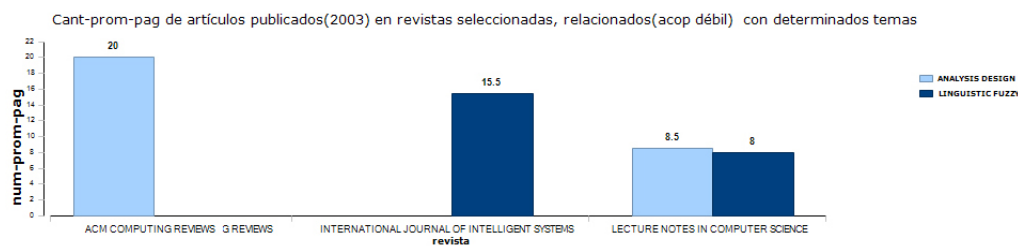


Figura 4.18: Resultados del uso de la cantidad promedio de páginas como medida, en el Ejemplo 10

con la frase  $\{(FUZZY, LEARNING, LINGUISTIC)\}$  perteneciente al grupo de frases más finas o detalladas. Se ha aplicado acoplamiento débil. Los resultados de dicha consulta se observan en la figura 4.20.

Una de las consultas más frecuentes sobre las publicaciones científicas, es la que se realiza para conocer las publicaciones de determinados autores. El ejemplo que se expone a continuación constituye una de esas.

### Ejemplo 11

Obtener la cantidad de artículos con sus autores, que contienen completamente en sus títulos (dimensión-AP), la frase  $F = \{(LEARNING, SYSTEM), (BASED, FUZZY, MODEL)\}$ .

Para dar respuesta a esta consulta se construye un subcubo con la siguiente estructura en el servidor OLAP:

- Dimensiones: título de artículo(dimensión-AP), autor

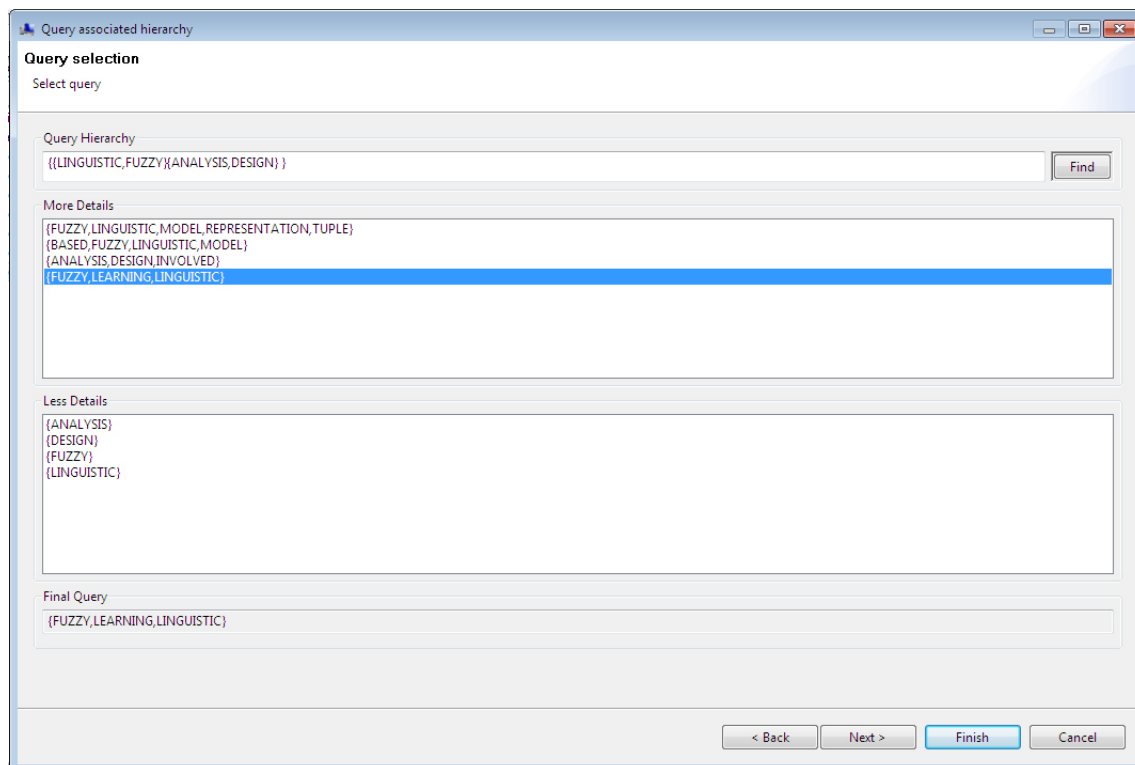


Figura 4.19: Jerarquía de consulta asociada a las frases del Ejemplo 10

- Medida: cantidad de artículos.
- Función de agregación: SUM.

En el ejemplo 11 se realiza una consulta sobre la dimensión-AP título de artículo, usando el acoplamiento fuerte. Los resultados de este ejemplo se muestran en la figura 4.21. Podemos apreciar que según los datos con que se cuentan, los autores que han publicado mayor cantidad de artículos sobre el tema (DATABASES, FUNCTIONAL, RELATIONAL) son Antonio Becerra y Jesús Almendros. Y en el tema (ASSOCIATION, RULES) María Amparo Vila Miranda y Daniel Sánchez Fernández.

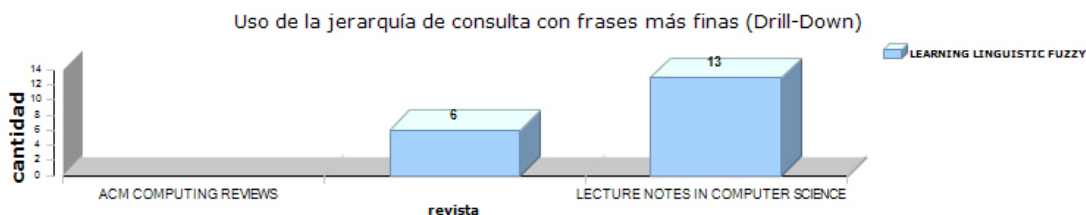


Figura 4.20: Resultados de realizar una consulta más detallada con el uso de la jerarquía de consulta asociada

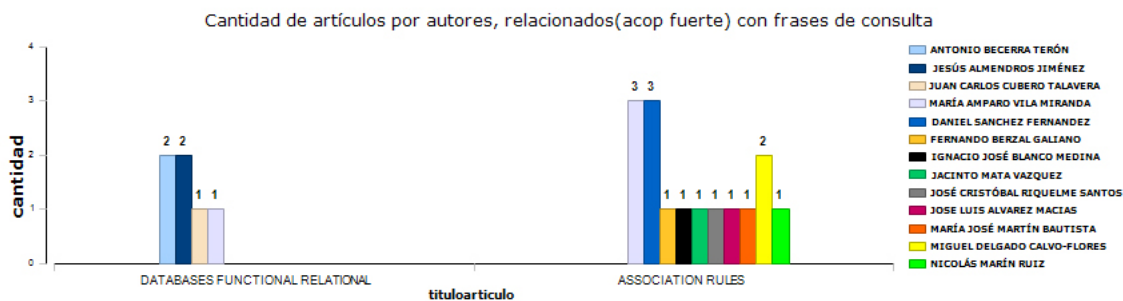


Figura 4.21: Resultados Ejemplo 11

Se expondrá a continuación una implementación similar, pero con el uso de datos médicos. De igual manera se podrá comprobar el buen funcionamiento del sistema y la factibilidad del modelo. Además se muestra como la implementación de un data warehousing de datos médicos, puede considerarse una solución a problemas de procesamientos de datos textuales en el ambiente hospitalario.

#### 4.4. Data warehousing con datos médicos

En (Delgado et al., 2005) se plantea que en el medio hospitalario es frecuente que no se cuente con información suficiente, oportuna y de calidad como un medio para la toma de decisiones. También explican que debido a que los datos que se generan, deben ser plausibles de un análisis estadístico adecuado

y completo, es necesario un proceso de codificación, en el cual se definan y estandaricen cada una de las entidades a medir mediante un vocabulario controlado, como lo es la Décima Revisión de la Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud (CIE-10) de la Organización Panamericana de la Salud.

En el CIE, usan un sistema de codificación o clasificación de diagnósticos médicos con el objetivo de tener una terminología estandarizada, que permita una representación independiente del lenguaje natural, y así lograr que la información esté disponible para propósitos de compatibilidad e integración. Los ejemplos que se mostrarán en esta sección brindan la posibilidad de extraer conocimiento de datos cómo estos y realizar consultas sobre ellos. Se pretende demostrar con los mimos que se puede consultar la información semántica contenida en el tratamiento, el diagnóstico o cualquiera de los textos cortos de interés para la entidad, aunque los mismos no formen parte de codificadores. Es por eso que es de gran importancia contar con una herramienta que de forma eficaz pueda extraer conocimiento de estas informaciones textuales. En este caso dicha herramienta es Wonder OLAP server v3.0.

En el entorno médico, dicho sistema constituye una alternativa a considerar, pues además de dar soporte a consultas semánticas en atributos textuales, ha sido implementado utilizando software libre, algo también importante, en un ámbito donde el costo de todos los recursos es cada día más elevado.

#### **4.4.1. Origen y descripción del modelo de datos a utilizar**

Los conjuntos de datos están compuestos por atributos textuales de una base de datos del Hospital Clínico “San Cecilio” de Granada, España. Se ha efectuado un estudio preliminar de las necesidades de información sobre algunas de las actividades que en este hospital se realizan. Como resultado de ese análisis se ha extraído una porción de los datos, específicamente la información referente a las Intervenciones Quirúrgicas y Urgencias que se encuentran en las tabla TIntervenciones y TUrgencias, con alrededor de 24000 registros

cada una. Además se utilizaron tablas complementarias de codificadores, como las tablas de sexos y anestesia. Se eligieron estos datos con el objetivo de dar respuesta a interrogantes como: ¿Cuántas intervenciones por diagnóstico y sexo se relacionan con una determinada propuesta de tratamiento? ¿Cuántas intervenciones se realizan con el uso de una determinada anestesia?. El modelo relacional de tablas se presenta en la figura 4.22.

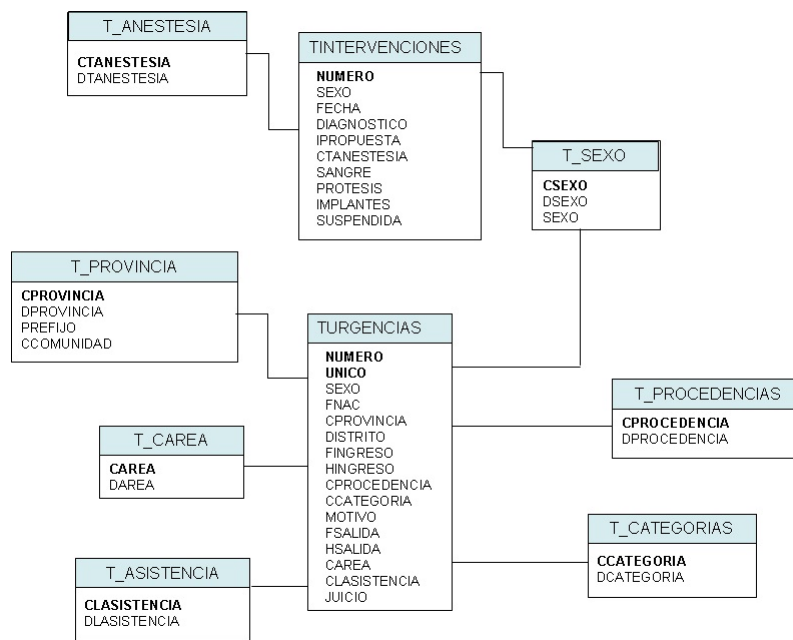


Figura 4.22: Modelo relacional de los datos médicos

Con ayuda del sistema Text Mining Tool, aplicando un soporte 0.01 obtuvimos los atributos-AP correspondientes a los campos propuestas de tratamientos (*ipropuesta*) y los *diagnósticos* de la tabla TInteracciones y al *motivo* de la tabla TURGENCIAS. Estos atributos-AP se usan para definir las dimensiones-AP correspondientes.

### 4.4.2. Transformación del modelo relacional al modelo dimensional

La construcción del modelo dimensional ROLAP, a partir del modelo relacional de la figura 4.22 se realiza con las mismas premisas que planteamos en el data warehousing de publicaciones para obtener la tabla de hechos, pero en este caso se obtiene un esquema estrella.

- Obtención de las dimensiones y jerarquías:

El conjunto de dimensiones proporcionan el contexto para cada medida de la tabla de hechos. En pocas palabras, se podría decir que las dimensiones son los descriptores del negocio.

Basándonos en esta definición y teniendo como base el modelo relacional mostrado en la figura 4.22 se obtuvieron entre otras las dimensiones: tratamiento, diagnóstico, anestesia, sexo, provincia, área y procedencia, entre otras.

- Obtención de la tabla de hechos

Para obtener la tabla de hechos, como ya se ha mencionado se sigue la estrategia de apartados anteriores de analizar todas las tablas del modelo relacional, y considerar las que contienen la mayor cantidad de registros transaccionales, o que constantemente cambian y tienen gran cantidad de filas. En este caso, son candidatas a tablas de hechos, la tablas de Intervenciones y Urgencias.

Tras obtener los miembros del modelo dimensional: tabla de hechos y dimensiones, finalmente el modelo relacional se ha transformado en el modelo dimensional en esquema estrella que muestra la figura 4.23, sobre este modelo se realizarán los ejemplos que se muestran a continuación.

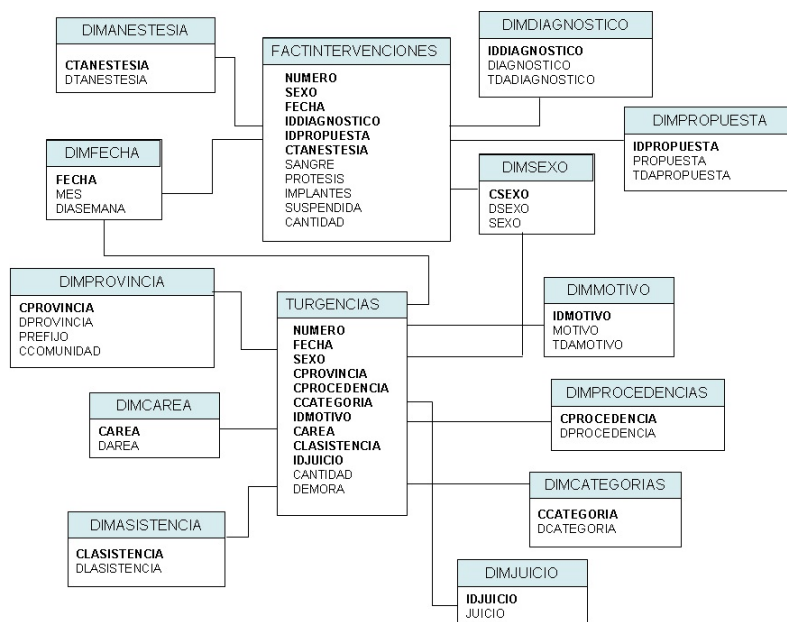


Figura 4.23: Modelo dimensional de los datos

A continuación se muestran ejemplos de consultas sobre el hipercubo correspondiente al modelo dimensional 4.23. El mismo se construye con la ayuda de la herramienta Wonder OLAP v3.0. Nótese que en las tabla DIMDIAGNOSTICO, DIMPROPUESTA, DIMMOTIVO que se muestra en la figura 4.7 se pueden apreciar los respectivos campos **TDA**, que constituyen los atributos-AP que mencionamos en la sección anterior. Estos campos contienen la semántica asociada a las dimensiones textuales que con Wonder definimos como dimensiones-AP.

#### 4.4.3. Ejemplos

De forma similar a como se hizo en el data warehousing de publicaciones científicas, con los datos que se muestran en el esquema dimensional de la figura 4.23 se construye un hipercubo que contendrá todas las dimensiones que se exponen. Sobre éste hipercubo siempre realizamos la operación Slice (quitamos las dimensiones que no se relacionan con la consulta), entre otras operaciones OLAP para obtener los resultados de interés.



Figura 4.24: Resultados del Ejemplo 12

A continuación se muestran ejemplos de subcubos de datos que contienen dimensiones-AP y consultas sobre éstos, útiles en el entorno hospitalario. Es importante destacar que los análisis de las consultas que se realizaron fueron revisados por una especialista en medicina que ha colaborado con esta investigación, la Dra Manuela Zapata Martínez médica general del Servicio Andaluz de Salud.

### Ejemplo 12

Obtener la cantidad de intervenciones, que contienen completamente en sus propuestas de tratamiento (Dimensión-AP), la frase  $F = \{(CIRUGA, ENDOSCOPICA), (BIOPSIA)\}$ .

Para obtener los resultados requeridos en el ejemplo 12 se realizó una consulta sobre la dimensión-ap del hipercubo antes mencionado, usando como función de acoplamiento el acoplamiento fuerte, como medida la cantidad de intervenciones y SUM como función de agregación. Esos resultados se exponen en la figura 4.24.

En la figura 4.24 se observan mayores resultados con la frase (BIOPSIA) y es un resultado muy lógico, porque los tratamientos donde se aplica BIOPSIA son más usados debido a que es menos invasivo. Mientras que la (CIRUGÍA, ENDOSCÓPICA) es un tratamiento para casos más específicos. Los resultados, si se usa el acoplamiento débil se muestran en la figura 4.25.

Si el usuario lo desea, puede usar la jerarquía de consulta asociada a la consulta realizada. Ésta ofrece posibles nuevas consultas, más o menos detalladas que la inicial, con las que se garantiza que se encontrará información



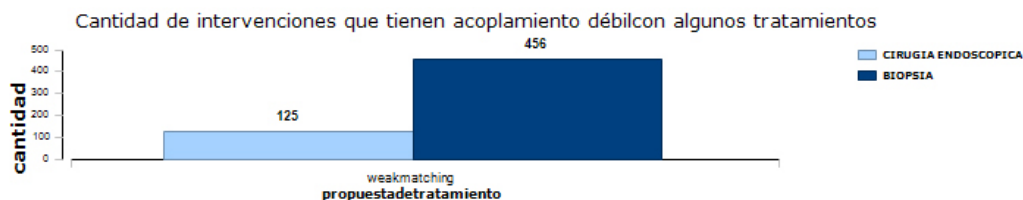


Figura 4.25: Resultados del Ejemplo 12 aplicando acoplamiento débil

en la base de datos. De esta forma se pueden obtener nuevos resultados que mantienen relación con los ya obtenidos. En la figura 4.26 se muestra dicha jerarquía.

Si con la ayuda de la jerarquía de la figura 4.26 construimos la consulta (CIRUGÍA, ENDOSCÓPICA, NASAL), (BIOPSIA, PRÓSTATA), seleccionando frases más detalladas y usando como función de acoplamiento el acoplamiento fuerte los resultados se muestran en la figura 4.27. En ella se puede apreciar que se cuentan con menos resultados porque aumentamos el detalle de la consulta (Drill-Down). En el caso de construir la frase final con las frases menos detalladas  $\{(CIRUGÍA), (ENDOSCÓPICA)\}$  los resultados se muestran en la figura 4.28.

Con los resultados antes obtenidos, mostrados en las figuras 4.24, 4.25, y 4.28, se pueden demostrar las relaciones de las operaciones Roll-Up y Drill-Down con los tipos de acoplamientos, definidas en la sección 3.3.2 del capítulo anterior, en la Ecuación 3.1.

Se considera a la frase  $F = \{(CIRUGÍA, ENDOSCÓPICA), (BIOPSIA)\}$  como  $C^1$  (que es la frase inicial) y a  $\{(CIRUGÍA), (ENDOSCÓPICA)\}$  como los correspondientes valores de  $C^2$  (frases menos detalladas).

Primero puede comprobarse, cómo se cumple que (ver figuras 4.24 y 4.25) los acoplamientos débiles y fuertes devuelven el mismo valor cuando la frase de consulta es unitérmino, como lo es la frase (BIOPSIA). Por lo tanto no fue necesario generar los resultados aplicando acoplamiento débil con el uso de frases menos detalladas (ENDOSCÓPICA y CIRUGÍA), ya que se conocen los resultados gracias a la esa regla de relación.

También es importante notar que después de sumar  $96+111=207$  como el

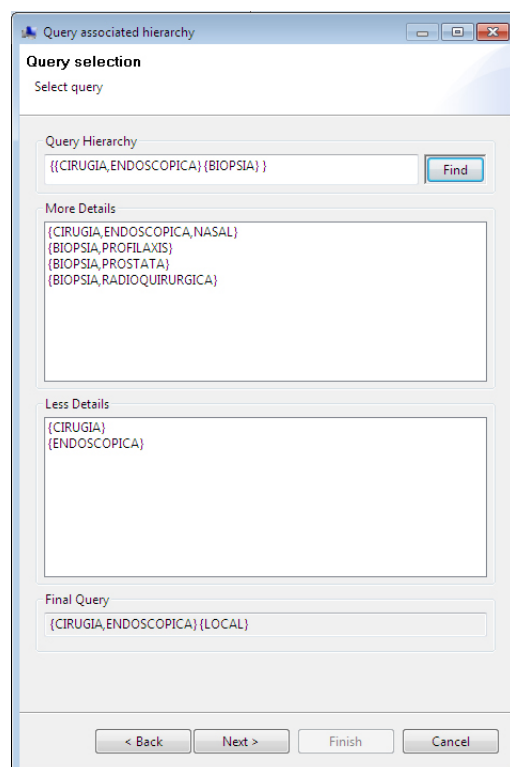


Figura 4.26: Jerarquía asociada a la consulta del Ejemplo 12

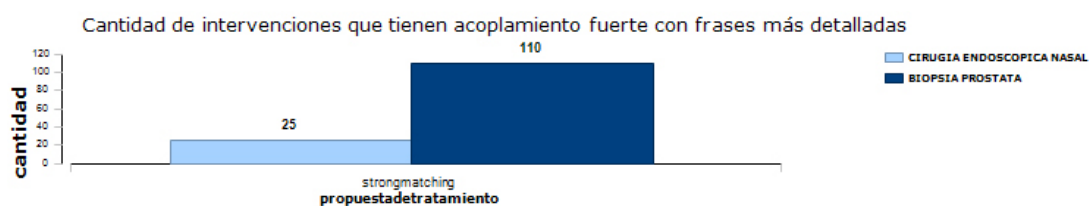


Figura 4.27: Resultados del uso de la jerarquía de consulta del Ejemplo 12

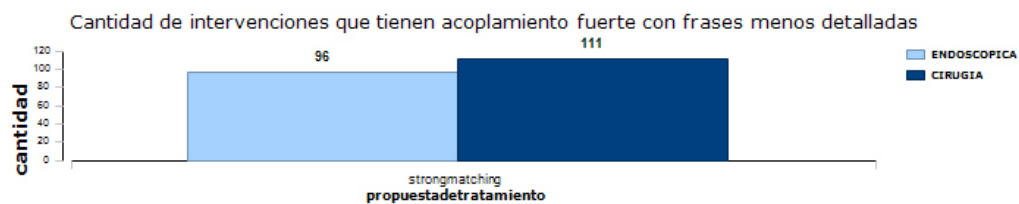


Figura 4.28: Resultados del uso de la jerarquía de consulta del Ejemplo 12 con frases menos detalladas

total que resulta del acoplamiento fuerte y débil con las frases de  $C^2$  (ver figura 4.28) hay que restarle a 207-(los 82 que cumplen con el acoplamiento fuerte con la frase de  $C^1$ : (CIRUGÍA, ENDOSCÓPICA))= 125). Ésto quiere decir que hay 82 propuestas de intervenciones que se acoplan con las dos frases de  $C^2$  a la vez (ver figura 4.24).

Resumiendo, con la ayuda de las figuras 4.24, 4.25, y 4.28, se verifica la Ecuación 3.1, porque efectivamente  $82 \leq 125 \leq 125 \leq 125$

En los siguientes ejemplos se combinan el uso de la dimensión-AP con otras del tipo clásicas.

### **Ejemplo 13**

Obtener la cantidad de intervenciones, para los distintos tipos de anestésicos que se usan, que contienen completamente en sus propuestas de tratamiento (Dimensión-AP), la frase  $F = \{(LAPAROTOMA), (EXTIRPACION, QUISTE), (OSTEOSINTESIS)\}$ . Para dar respuesta a esta consulta se construye un subcubo con la siguiente estructura en el servidor OLAP:

- Dimensiones: propuesta(dimensión-AP), anestesia.
- Medida: cantidad de intervenciones.
- Función de agregación: SUM.

En este ejemplo 13 a la consulta sobre el subcubo de datos creado se le aplica un acoplamiento fuerte de las frases de consulta que se proponen, con las propuestas de intervenciones correspondientes (dimensión-AP). Los resultados se muestran en la figura 4.29.

La figura 4.29 refleja que la anestesia GENERAL se aplica con mayor frecuencia en los tratamientos relacionados con la frase LAPAROTOMÍA y la LOCAL con EXTIRPACIÓN QUISTE. Este resultado es correcto, según la opinión de la doctora que ha colaborado, porque la LAPAROTOMÍA, por su complejidad requiere en mayor medida de la anestesia GENERAL y los tratamientos donde se hace EXTIRPACIÓN QUISTE por ser más comunes y menos complejos requieren de anestesia LOCAL.

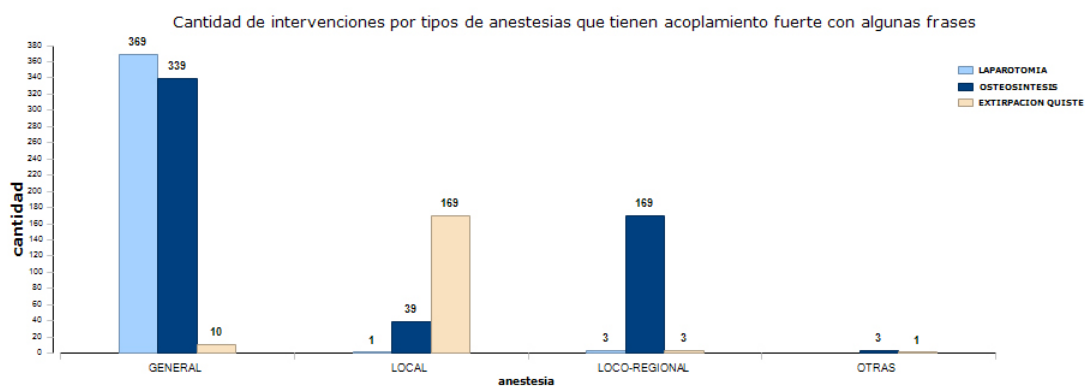


Figura 4.29: Resultados del Ejemplo 13 donde se muestra el acoplamiento con los tipos de anestias

Después de usar la jerarquía asociada a esta consulta, se eligió la frase {CLAVOS ENDER OSTEOSÍNTESIS} como consulta final. En la figura 4.30 se muestran los resultados, estos son escasos y así ratifican el criterio de la especialista de que el uso de CLAVOS ENDER para tratamientos de OSTEOSÍNTESIS es muy específico y que efectivamente cuando se usa, se aplica anestesia GENERAL.

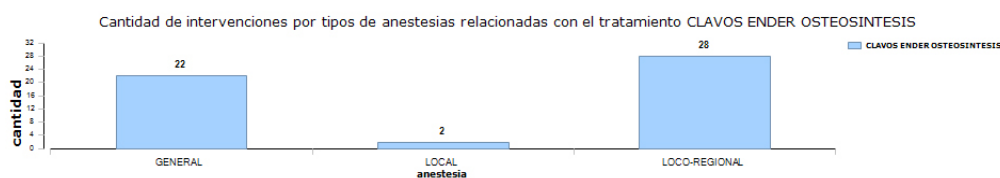


Figura 4.30: Resultados del Ejemplo 13 después de usar la jerarquía de consulta para obtener la frase final CLAVOS ENDER OSTEOSÍNTESIS.

## 4.5. Conclusiones

Al finalizar este capítulo se puede decir que se ha cumplido el objetivo fundamental del mismo: se ha demostrado con varios ejemplos las potencialidades del nuevo modelo multidimensional propuesto en el capítulo anterior, para

la implementación de sistemas data warehousing con especiales prestaciones para dimensiones textuales (dimensiones-AP).

De forma práctica, con datos reales, se han ejemplificado operaciones OLAP sobre dimensiones-AP definidas en el nuevo modelo, como por ejemplo, la operación Dice o el uso de una jerarquía de consulta para ayudar al usuario en la búsqueda de información útil en dichas dimensiones.

También se ha mostrado la arquitectura y principales funcionalidades que posee el sistema OLAP implementado. En este punto cabe destacar, por un lado, que se han usado tecnologías de software libre para su implementación, con todas las ventajas que ésto supone; y por otro, el sistema brinda al usuario la posibilidad de usar múltiples vías para consultar, graficar y hacer resúmenes de cubos de datos, y en especial de cubos con dimensiones-AP.

El siguiente capítulo estará dedicado a la evaluación del modelo propuesto, desde la utilidad práctica que certifiquen expertos en los datos base. O sea se realizan consultas interesantes en los data warehousing implementados, sugeridas y evaluadas por expertos en dichos temas.

# Capítulo 5

## Evaluación del modelo

En este capítulo se realiza una evaluación del nuevo modelo multidimensional de datos propuesto en el capítulo 3. Se mostrarán ejemplos de consultas sobre data warehousing de datos médicos y de datos de publicaciones científicas que demuestren la utilidad del modelo para estos dos tipos de entornos. Para ello era imprescindible contar con la ayuda de expertos en estos temas, de forma tal que ellos validaran la utilidad práctica del mismo, desde los sistemas data warehousing implementados. Para el caso de los datos médicos, tuvimos la valiosa colaboración de la Dra Manuela Zapata Martínez médica general del Servicio Andaluz de Salud, y el Dr Miguel Prados Jefe de los Servicios de Informática del Hospital Clínico San Cecilio. Con relación a los datos de publicaciones científicas contamos con la cooperación del Dr. Víctor Herrero Solana profesor de la facultad de Comunicación y Documentación de la Universidad de Granada, en el área de Bibliometría.

Las consultas que se implementan utilizan los mismos esquemas dimensionales obtenidos en el capítulo anterior (figura 4.23) tienen la particularidad de relacionar más de una dimensión-AP. En este caso se puede consultar la información textual de los tratamientos de intervenciones, de forma combinada con los diagnósticos y con otras dimensiones no textuales.

## 5.1. Utilidad en bases de datos médicas

La utilidad en base de datos médicas se medirá desde dos puntos de vistas diferentes. Primero teniendo en cuenta los criterios técnicos de especialistas como la Dra Zapata, desde su experiencia como médica general. Luego se exponen consultas que resultan de interés desde el punto de vista de la gestión hospitalaria. Éstas han sido resultado de la amable colaboración del Dr Prados que dirige directamente esta actividad desde el punto de vista informático.

Las consultas que se implementan pertenecen a los mismos data warehousing desarrollados en el capítulo anterior, para datos de publicaciones científicas y datos médicos. Los esquemas dimensionales se pueden ver en las figuras 4.7 y 4.23.

### 5.1.1. Criterios técnicos de especialistas

En la primera entrevista realizada a la doctora Zapata comprobamos que las necesidades de información planteadas por ella no pudieron ser satisfechas por el sistema, debido a que solo habíamos preprocesado los datos del atributo *ipropuesta* (*propuesta de intervención o tratamiento*) y también resultaba necesario consultar la información textual del atributo *diagnóstico*.

Después de obtener también la semántica asociada al atributo *diagnóstico*, se pudieron hacer interesantes consultas. Las mismas se muestran a continuación.

#### Ejemplo 14

Para los diagnósticos que se relacionan con el CÁNCER DE MAMA, cuál es el tratamiento quirúrgico más usado, entre la MASTECTOMÍA y la TUMORECTOMÍA.

Para dar respuesta a esta interrogante se crea un subcubo de datos con dos dimensiones textuales (dimensiones-AP): *diagnósticos* y *tratamientos*, cantidad de tratamientos como medida y la suma como función de agregación.

También hay que realizar consultas sobre las dimensiones-AP, usando las funciones de acoplamiento para las frases *mama*, *mastectomía* y *tumorectomía*. En este caso se aplica acoplamiento fuerte.



Figura 5.1: Tratamientos quirúrgicos relacionados con el cáncer de mama

Los resultados del ejemplo 14 se muestran en la gráfica 5.1. En ella se puede apreciar que la mayoría de los tratamientos quirúrgicos que se realizan para el cáncer de mama están relacionados con la mastectomía. Eso indica que todavía hay que trabajar en las medidas de diagnóstico precoz de esa enfermedad, para no tener que llegar a intervenir de forma tan radical.

### Ejemplo 15

¿Cuántas *amputaciones* se realizan para los diagnósticos relacionados con PIÉ DIABÉTICO?.

La respuesta a esta interrogante requiere del uso del mismo subcubo de datos creado en el ejemplo anterior, solo cambiarían las frases de consulta para las dimensiones-AP.

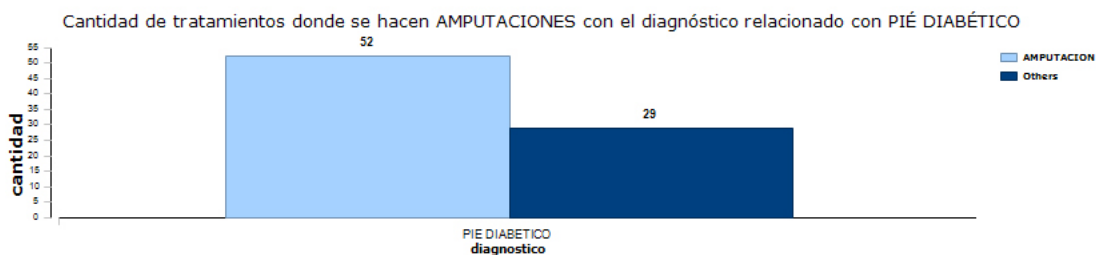


Figura 5.2: Tipos de amputaciones relacionados con el diagnóstico pié diabético



En la figura 5.2 se puede observar que más de la mitad de los diagnósticos relacionados con pie diabético terminan en amputaciones como tratamiento. Ésto indica que se deben reforzar las medidas de prevención para no llegar a tratamientos tan radicales como la amputación.

### Ejemplo 16

¿Cuántos diagnósticos están relacionados con las frases MELANOMA y ESPINOCELULAR?

Los resultados que se presentan en la gráfica 5.3 son el resultado a consultar directamente sobre la dimensión-AP *diagnóstico*, usando acoplamiento débil. Los mismos reflejan que no es tanta la diferencia entre las cantidades de estos diagnósticos. Algo que preocupa porque éste es uno de tipos de cáncer de piel más agresivo y cada vez son más frecuentes.

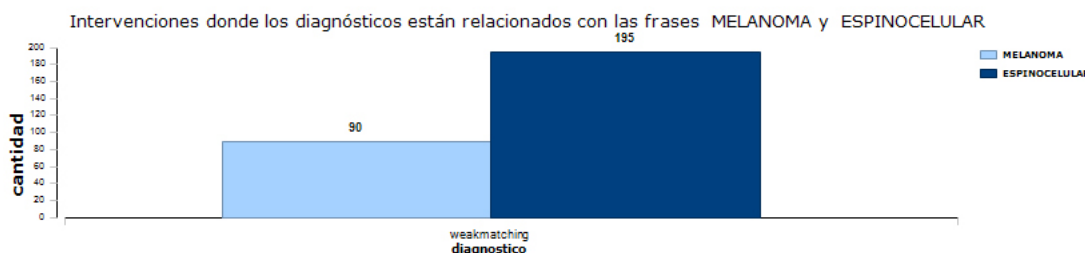


Figura 5.3: Cantidades de intervenciones relacionadas con los diagnósticos melanoma y espinocelular

### Ejemplo 17

¿Es muy significativo el uso de implantes en las cirugías de CATARATA?

En el caso del ejemplo 17 se crea un subcubo de datos con dos dimensiones, una dimensiones-AP *diagnósticos* y otra dimensión clásica *implantes*. Con la cantidad de tratamientos como medida y la suma como función de agregación. También hay que realizar una consulta sobre la dimensión-AP *diagnóstico*, usando la función de acoplamiento fuerte para la frase *catarata*.

Si el usuario lo desea puede realizar una operación Dice al subcubo por la dimensión-AP *diagnóstico* para almacenar solo las intervenciones relacionadas con CATARATA. La figura 5.5 muestra una de las ventanas del

wizard de la opción Create Subcube donde se realiza el Dice. Y la figura 5.6 muestra una de las ventanas del wizard de la opción Consulting Cube donde se observa que después de haber aplicado el Dice, solo se tiene en la dimensión-AP diagnóstico los valores que tienen acoplamiento fuerte (o débil porque si tiene un solo término es igual) con la frase CATARATA.

La figura 5.4 nos dice que en la gran mayoría de las intervenciones quirúrgicas de cataratas se usa implantes. La doctora Zapata como médica de atención primaria plantea que es de gran utilidad en su trabajo conocer cuan agresivas o radicales son las cirugías de diagnósticos muy comunes, como la catarata, porque eso indica que hay que reforzar el trabajo de prevención desde todos los frentes de la vida posibles, no solo desde la atención primaria de salud.

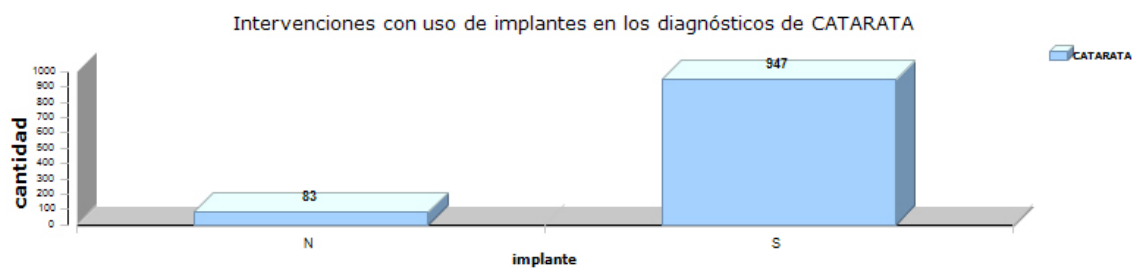


Figura 5.4: Cantidad de implantes en las intervenciones de catarata

### Ejemplo 18

¿Cuántos diagnósticos de algún tipo de *FRACTURA* padecen mujeres y hombres?

Para el caso del diagnóstico fractura las mujeres representan el mayor porcentaje. En la figura 5.7 se puede apreciar. Esto sucede por varias causas, las mujeres tienen una mayor esperanza de vida y también uno de los efectos de la menopausia es que los ovarios dejan de producir estrógenos que son muy necesarios para el buen funcionamiento de los huesos.

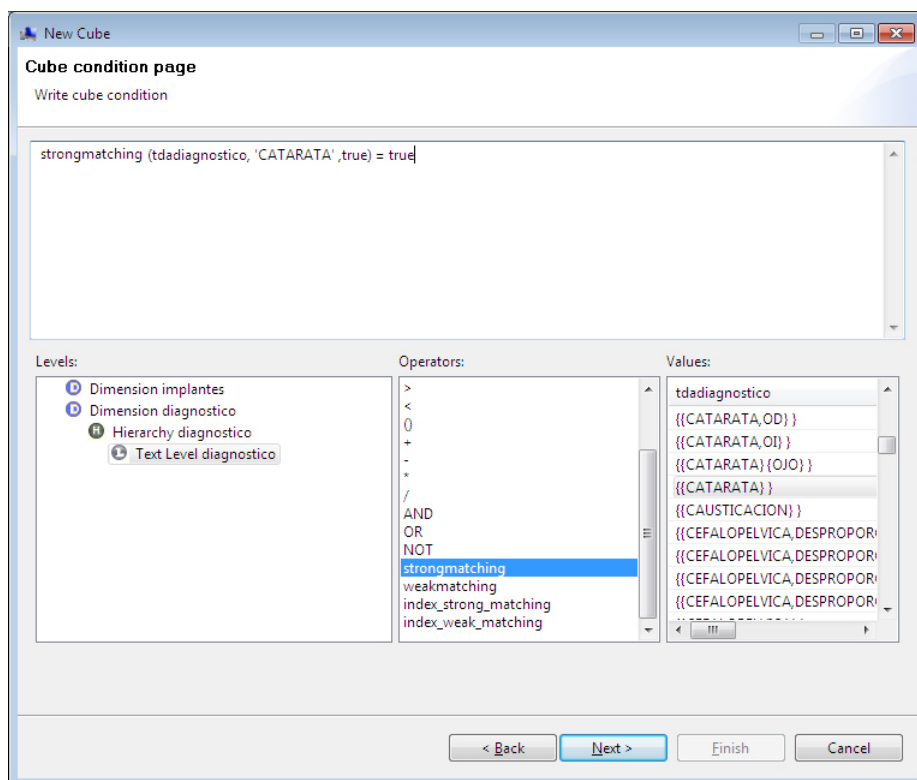


Figura 5.5: Ventana de Wonder donde se realiza la operación Dice

### 5.1.2. Gestores de datos médicos

Las consultas que se muestran a continuación están relacionadas con la gestión hospitalaria. Se relacionan dimensiones como la fecha de las urgencias médicas, que describen el comportamiento de las urgencias en la semana o en los meses y así el funcionamiento o gestión del hospital. También se analizan dimensiones-AP como el motivo de la urgencia, de manera que se pueda saber la relación de éstas con dimensiones como las áreas del hospital. Con la ayuda de los resultados que se muestran a continuación se pueden conocer patrones de urgencias que ayuden a prevenir colapsos en el hospital.

**Ejemplo 19** Se necesita conocer el comportamiento de las urgencias durante la semana.

Para obtener la respuesta requerida por el ejemplo 19 se realiza una consulta

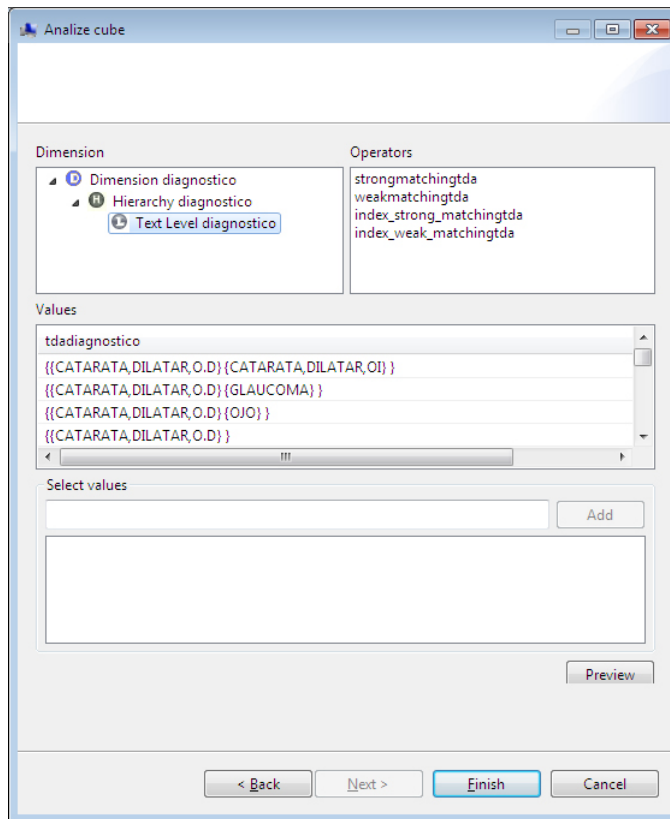


Figura 5.6: Ventana de Wonder de la opción Consulting Cube donde se puede corroborar el Dice

sobre el propio hipercubo base, correspondiente a *TUrgencias*, o sea, con la opción consulting cube se define una consulta sobre las dimensión fecha teniendo en cuenta los días de la semana y los meses de Enero, Febrero y Marzo para tener puntos de comparación. Además se define como medida la cantidad de urgencias con la suma como función de agregación. La figura 5.8 muestra los resultados. En la leyenda 1, 2 y 3 son los tres primeros meses del año y del 0 al 6 son los días de la semana comenzando por el domingo. Las líneas indican que los días pico de urgencia son los domingos y lunes. También reflejan que en el mes de Enero que es un mes con muchos días libres por fiestas, las urgencias se comportan altas durante casi toda la semana, como es lógico porque las personas están en casa y pueden ir al médico.



Figura 5.7: Número de mujeres y hombres con el diagnóstico fractura

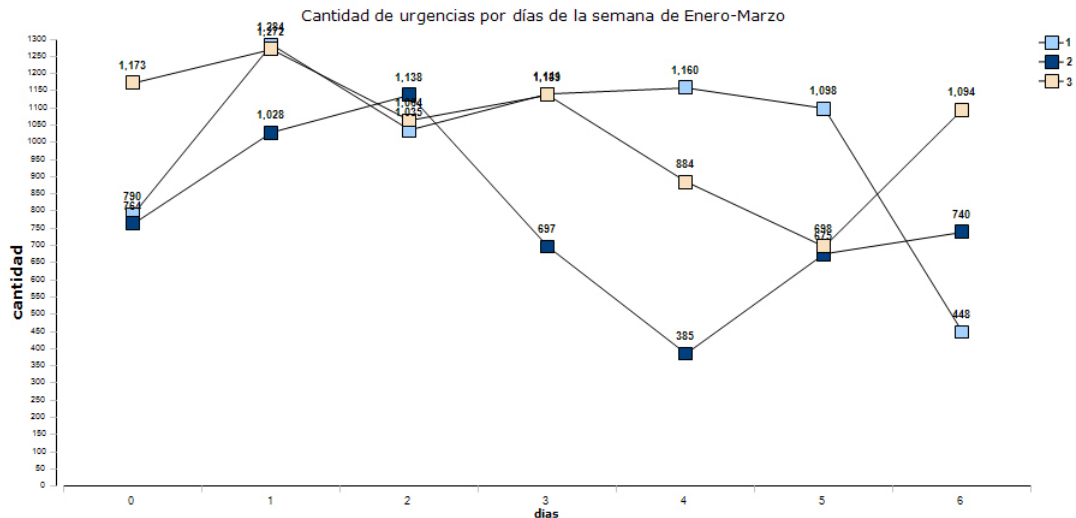


Figura 5.8: Comportamiento de las urgencias en los días de la semana de Enero-Marzo

**Ejemplo 20** Se necesita conocer el comportamiento de las urgencias durante la semana para diferentes causas de urgencias.

Para obtener el comportamiento de las urgencias en la semana y para diferentes causas (ejemplo 20) se realiza una consulta similar al ejemplo anterior, sobre el hipercubo de  $TUrgencias$ , pero además de la dimensión fecha (días de la semana) se tiene en cuenta la dimensión *Categorías* (causas de urgencias). Fueron escogidas tres de las posibles causas de urgencias por ser las más urgentes: Agresión, Accdte Tráfico y Accdte de Trabajo. También para la consulta se define el tiempo de demora de la atención (en minutos) como medida, con el promedio como función de agregación. En la figura 5.9 puede

verse que es relativamente bajo el tiempo promedio de espera de forma general y este resultado es relevante si se trata de las tres causas que requieren de la atención más rápida. Solo se dispara el tiempo de espera el día jueves y nunca llega a una hora. Este resultado puede corroborarse en la gráfica de la figura 5.10 donde se pueden ver los tiempos promedios en los días de la semana y los meses de Enero-Marzo.

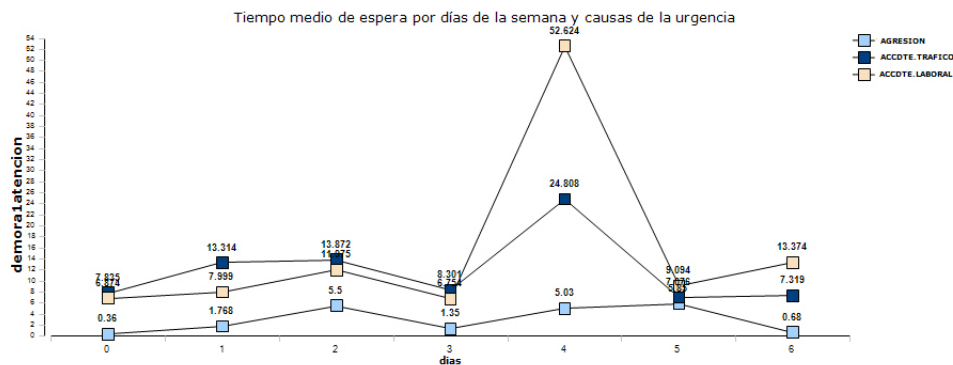


Figura 5.9: Promedio de espera, de atención en urgencias asociado a las causas: Agresión, Accdte Tráfico y Accdte de Trabajo

En los ejemplos siguientes se plantean consultas que valoran dimensiones-AP que relacionadas con otras no AP resultan en interesantes resultados.

**Ejemplo 21** Se necesita conocer la cantidad de las urgencias relacionadas con la frase: (*DOLOR, TORÁCICO*), (*VÓMITOS*), (*FRACTURA*) durante los días de la semana.

En el caso del ejemplo 21 en la consulta que se realiza se combina con los días de la semana, la dimensión-AP *Motivo*. A ésta última se consulta con las frases DOLOR TORÁCICO, VÓMITOS y FRACTURA, aplicando acoplamiento fuerte para garantizar que están presentes las palabras DOLOR y TORÁCICO a la vez.

En la gráfica de la figura 5.11 se puede observar que las urgencias relacionadas con el padecimiento vómitos ocurre más los domingos y lunes, y que es significativamente superior a los otros padecimientos. Las urgencias menores

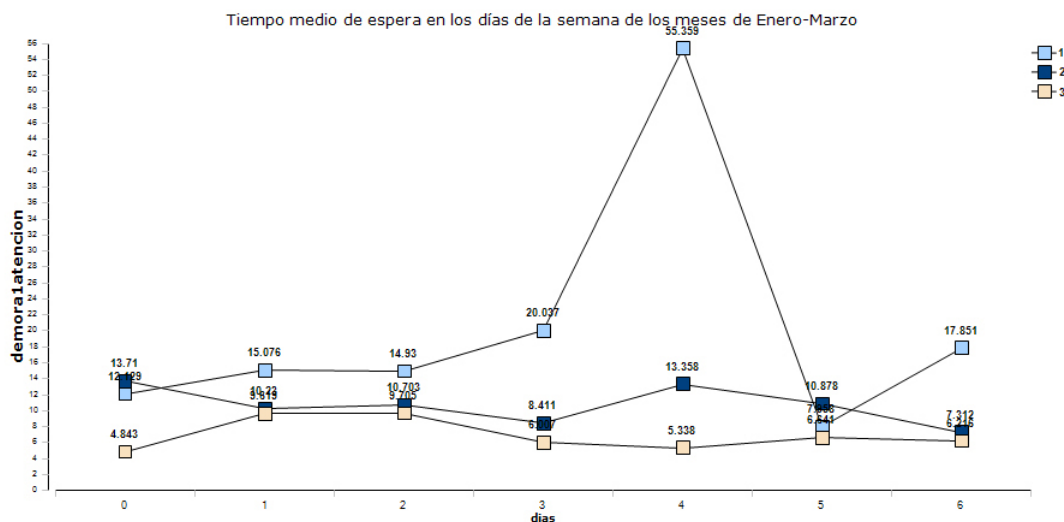


Figura 5.10: Promedio de espera, de atención en urgencias en los días de la semana y el primer trimestre del año

ocurren en el caso de la urgencia fractura, como es lógico porque es más escasa su ocurrencia. La jerarquía asociada a esta consulta puede verse en la figura 5.12. Puede verse que se ofrecen frases de consulta más detalladas, con las que se pudiera conocer por ejemplo, cuando hay vómitos a qué otro padecimiento viene acompañado.

Si después de obtener la jerarquía mostrada en la figura 5.12 se escogen las frases de consulta más detalladas: DOLOR TORÁCICO ATÍPICO Y CEFALEA VÓMITOS los resultados mantienen el patrón de la consulta inicial de que los días de mayor cantidad de urgencias son el domingo y el lunes. La figura 5.13 los muestra.

## 5.2. Utilidad en bases de datos de publicaciones científicas

La evaluación del sistema cuando se usan datos de publicaciones científicas estará orientada al área de la Bibliometría (?). Se considerarán dos ver-

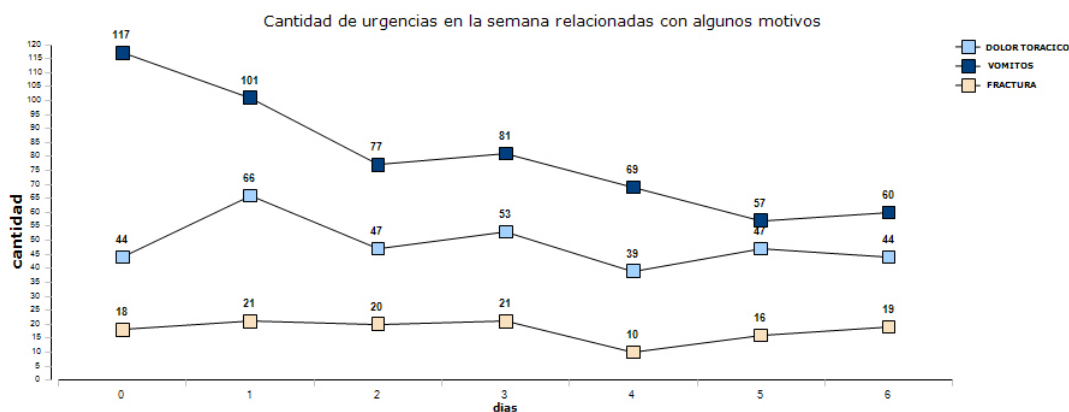


Figura 5.11: Cantidad de urgencias en una semana relacionadas con las urgencias que tienen las frases: DOLOR TORÁCICO, VÓMITOS y FRACTURA

tientes generales de consultas, una referente a la productividad científica y otra enmarcada en un análisis de dominio. Para medir la productividad científica se plantean consultas que ayuden a diagnosticar cómo se comportan las publicaciones por grupo de investigación u organismo, también se muestran consultas que reflejan la composición de los grupos o la productividad media de los autores por temas de investigación. La vertiente del análisis del dominio en este caso refleja el comportamiento científico de las investigaciones. Consultas dentro de dicha vertiente pueden ser, por ejemplo, ¿Cómo se ha comportado en los años registrados las publicaciones de una temática determinada? o ¿Cuáles departamentos son los que trabajan en dicha temática?

### 5.2.1. Bibliometría

Como ya se ha explicado contamos con la gentil colaboración en este tema de Bibliometría del Dr. Víctor Herrero. Después de un estudio conjunto de las posibles necesidades de información en este campo relacionadas con el esquema de datos que se tiene, fueron formuladas las siguientes consultas.

**Consultas que miden la productividad en el ámbito de las publicaciones científicas.**

#### Ejemplo 22



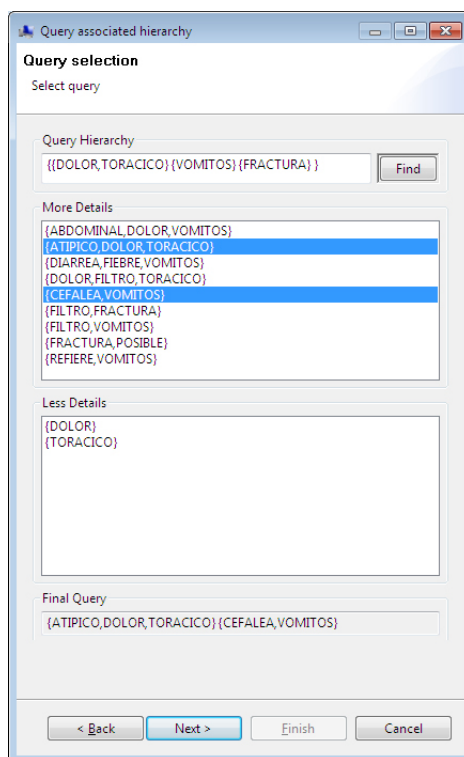


Figura 5.12: Jerarquía de consulta asociada a la consulta del Ejemplo 21

¿Qué grupo de investigación de la Universidad de Granada, en los últimos 5 años almacenados, tiene más publicaciones en revistas ISI?

Para responder la interrogante del ejemplo 22 se construye un subcubo de datos con las dimensiones: *pertenencia*, *autor*, *ISI* y *fecha*. Con la cantidad de artículos como medida y la suma como función de agregación. También se aplica un Dice para el nivel organismo = Universidad de Granada (UGR) y para la fecha los últimos 5 años.

En la figura 5.14 se ve con claridad que el grupo de investigación que más publicó en esos años en la UGR fue Soft Computing y Sistemas de Información. Puede verse a partir de ese resultado, dentro de ese grupo, qué investigadores se destacaron (figura 5.15). Para ello al subcubo resultante anterior se le realiza una operación Dice para ese grupo.

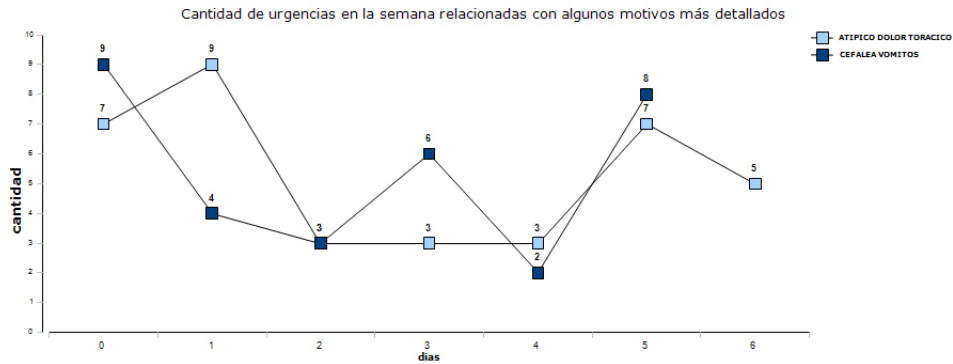


Figura 5.13: Resultados de usar la jerarquía del Ejemplo 21 con frases más detalladas

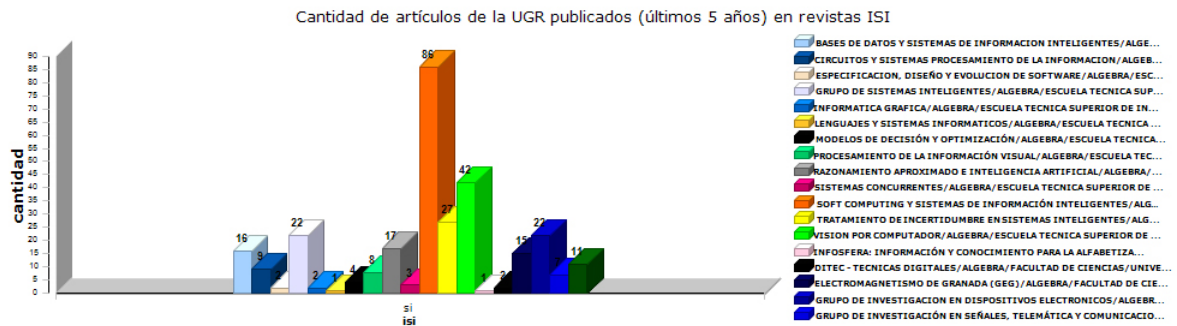


Figura 5.14: Cantidad de artículo publicados en los últimos 5 años del Ejemplo 22

También se podría conocer la producción científica de relevancia (publicaciones en revistas ISI) por organismos de Andalucía. En este caso se puede usar el mismo subcubo pero sin aplicar el Dice por organismos. La figura 5.15 refleja que la Universidad de Granada es el organismo que más publicaciones tiene en revistas ISI, resultado que corrobora los anteriores.

**Consultas que facilitan la realización de un análisis del dominio.**

Los ejemplos de consultas que se muestran a continuación muestran un análisis del dominio que se está estudiando: las investigaciones científicas en el campo de las TIC, en Andalucía.

**Ejemplo 23**



Figura 5.15: Cantidad de publicaciones del grupo Soft Computing y Sistemas de Información Inteligentes

Se necesita saber cuántos artículos se han publicado en los últimos 5 años relacionadas con los temas (*FUZZY*, *DATABASES*), (*PATTERN*, *RECOGNITION*)

Para el caso del ejemplo 23 se construye un subcubo sobre el hipercubo base, con las dimensiones: título artículo (dimensión -AP) y fecha, cantidad de artículos como medida con SUM como función de agregación, y aplicando Dice a la fecha(últimos 5 años). Después de tener ese subcubo, con la opción Consulting Cube se usa el acoplamiento débil para consultar a la dimensión-AP con los temas planteados.

Los resultados de la figura 5.16 reflejan un decremento en las publicaciones científicas en esos 5 años. Además se puede apreciar como ha habido una mayor producción científica en el tema de bases de datos difusas (*FUZZY*, *DATABASES*) por encima del tema de reconocimiento de patrones (*PATTERN*, *RECOGNITION*). El uso de la jerarquía asociada a estas consultas puede ayudar a seguir estudiando el comportamiento de las investigaciones sobre esos temas.

La jerarquía de consulta asociada al ejemplo 23 se muestra en la figura 5.17. De ella se ha escogido una frase más detallada: (*DATABASES*, *FSQL*, *FUZZY*). Se realiza una nueva consulta con dicha frase, pero con el uso de acoplamiento fuerte, con el objetivo de saber cómo se han comportado las publicaciones sobre esa temática exactamente. Los resultados se observan en

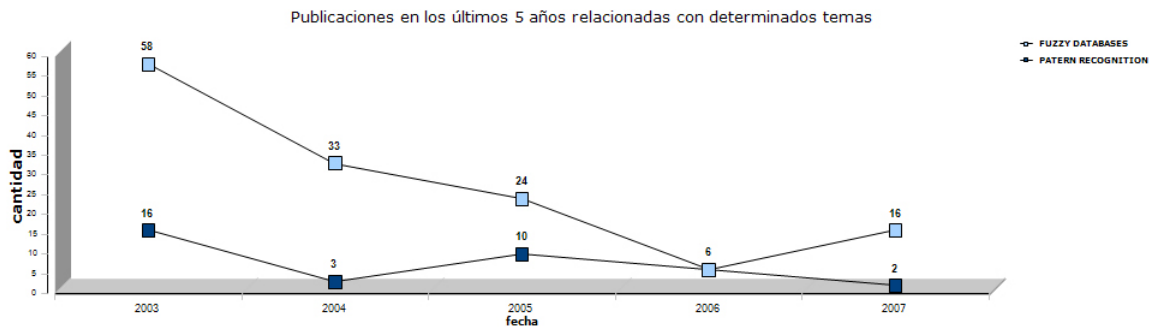


Figura 5.16: Cantidad de artículos de los últimos 5 años que tienen acoplamiento débil con las frases FUZZY DATABASES y PATTERN RECOGNITION

la figura 5.18, los mismos son escasos. Este resultado puede motivar a preguntar qué autores han publicado esos artículos. El ejemplo siguiente trata sobre esta interrogante.

### Ejemplo 24

Se necesita saber qué autores han publicados los artículos que contienen exactamente la frase (*DATABASES, FSQL, FUZZY*).

Los resultados del ejemplo 24 se obtienen consultando directamente las dimensiones *autor* y *título* (dimensión-AP) del hipercubo base. Se define la cantidad de artículos como medida, SUM como función de agregación y se aplica el acoplamiento fuerte de la dimensión-AP con la frase (*DATABASES, FSQL, FUZZY*). La figura 5.19 muestra dichos resultados.

## 5.3. Conclusiones

Durante el desarrollo de este capítulo se ha realizado una evaluación del nuevo modelo propuesto en el capítulo 3. Esa evaluación se ha valorado por concepto de utilidad práctica, valuada por expertos en los datos reales utilizados. Se ha contado con la valiosa colaboración de especialistas en los temas médicos,

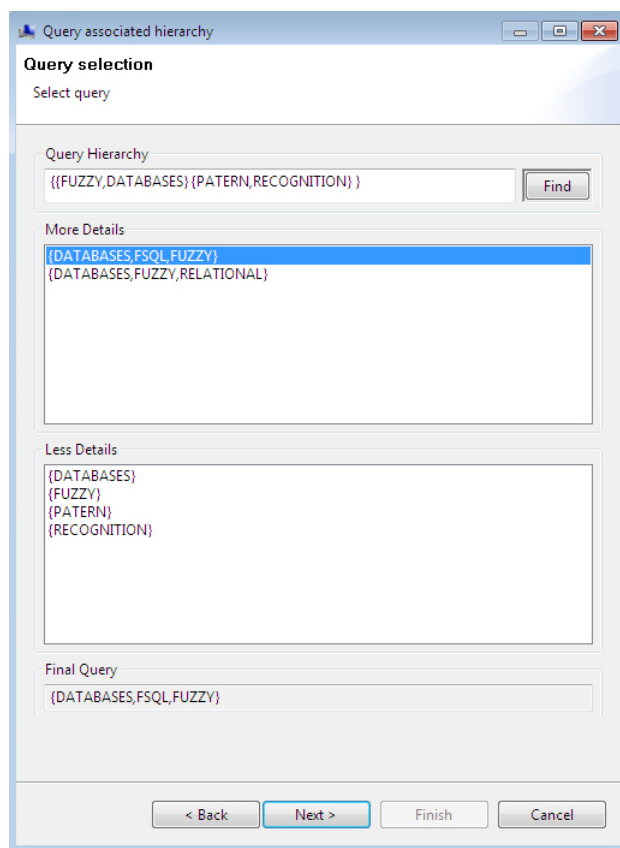


Figura 5.17: Jerarquía asociada al Ejemplo 23

de gestión médica y de bibliometría. La ayuda prestada ha consistido en la propuesta de consultas de interés y en la validación de los resultados.

En el entorno médico se pudieron satisfacer consultas de valor tanto para médicos como para los gestores a nivel hospitalario. La doctora Zapata pudo conocer el comportamiento de las intervenciones relacionadas con determinados diagnósticos, que ella trata en sus consultas médicas. Relacionado con la gestión hospitalaria se pudieron reconocer patrones de ocurrencias de urgencias que influyen en la planificación y funcionamiento de un hospital.

La Bibliometría es una rama que se encarga mayormente de aplicar métodos matemáticos y estadísticos a toda la literatura de carácter científico y a los autores que la producen, con el objetivo de estudiar y analizar la actividad

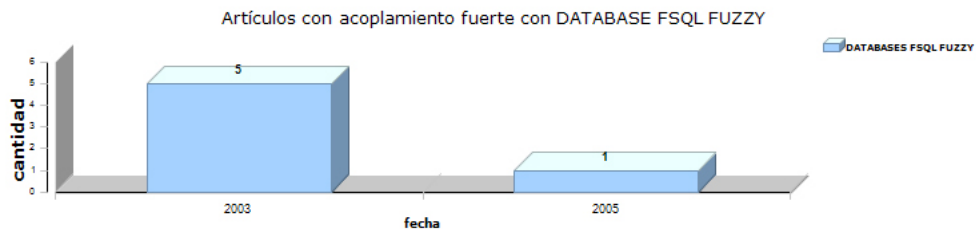


Figura 5.18: Resultado del uso de la jerarquía de consulta del Ejemplo 23

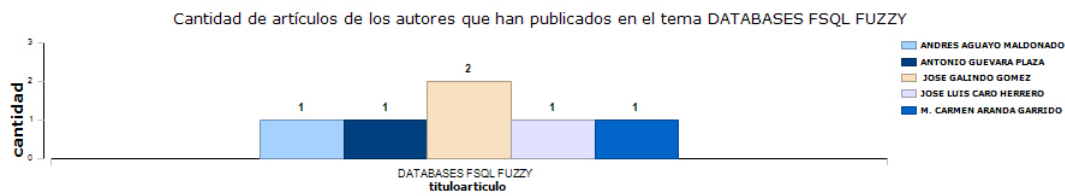


Figura 5.19: Autores que han publicado exactamente sobre el tema (DATABASES, FSQL, FUZZY)

científica. Además de las consultas que contabilizaran la productividad científica de autores, grupos u organismos que se manejan, se pudo analizar el dominio de las publicaciones científicas de Andalucía en las TIC.

El uso de dimensiones-AP y de las operaciones que sobre ellas se implementan, ofrecen ventajas analíticas a las consultas realizadas. La jerarquía de consultas asociada a éstas dimensiones, facilita al usuario la obtención de información útil para la toma de decisiones, porque la misma ofrece todas las posibles consultas que se pueden hacer sobre la información textual analizada. Se ha comprobado que esas dimensiones se pueden combinar con otras de tipos clásicos, explotando así la utilidad de la semántica básica que ellas contienen.



# Capítulo 6

## Conclusiones y trabajos futuros

El presente capítulo resume los objetivos que se han logrado y que han sido descritos en la presente memoria. A continuación de éstos, se exponen ciertas líneas futuras que extenderán los resultados obtenidos aquí presentados.

### 6.1. Conclusiones

Al comienzo de este trabajo, se planteó como objetivo fundamental la definición e implementación de un nuevo modelo multidimensional con soporte en sus dimensiones a atributos textuales en bases de datos. Dicho objetivo fue planteado con la finalidad de poder manejar los mencionados atributos textuales como el resto de los atributos de la base de datos, en un modelo multidimensional en ambiente OLAP.

De manera general, para conseguir estos objetivos, se ha profundizado en el estudio de las técnicas y herramientas necesarias para la definición del modelo propuesto y su implementación. Se ha realizado la formulación matemática de las estructuras y operaciones que componen dicho modelo, teniendo como centro el concepto de dimensión-AP. También se ha logrado su implementación mediante el desarrollo de una herramienta de software llamada Wonder OLAP Server v3.0. Dicha herramienta OLAP es pionera en dar soporte al manejo, conjuntamente y del mismo modo, a atributos textuales



como al resto de los atributos de la base de datos en procesos OLAP.

Es de destacar que, la solución obtenida, a pesar de haber sido utilizada en el contexto de bases de datos médicas y de artículos científicos, puede ser aplicada en otros contextos y ayudar a resolver problemas diversos sobre atributos textuales. Esto permitirá incrementar sustancialmente la información para la toma de decisiones, si se tiene en cuenta que hoy día más del 80 por ciento de la información de las organizaciones se encuentra en el formato texto.

De forma más detallada, la labor realizada queda descrita a continuación:

1. *Se ha formalizado matemáticamente el nuevo modelo multidimensional, que garantiza el procesamiento del conocimiento implícito de atributos textuales de bases de datos obtenido de forma previa, en entorno OLAP.*
  - La base de dicho modelo es la estructura definida como dimensión-AP. Dicha estructura es obtenida mediante la definición de una partición y un dominio sobre un atributo-AP, que son el requisito fundamental para poder definir cualquier dimensión.
  - Se definieron las operaciones OLAP Dice, Roll-Up y Drill-Down sobre una dimensión-AP. Para el caso de los agrupamientos, se definen sobre una nueva jerarquía, llamada jerarquía asociada a consultas, que brinda la posibilidad de subir y bajar el nivel de detalles de las consultas a realizar sobre una dimensión-AP. También se demostraron reglas que se cumplen en las relaciones entre las formas de acoplamiento que se usan en dichas consultas, que ayudan al usuario en tal proceso.
  - El nuevo modelo de datos multidimensional y las operaciones que sobre él se han definido, ofrecen ventajas analíticas de relevancia. Gracias al uso de una dimensión-AP se puede utilizar el conocimiento extraído de los atributos textuales de las bases de datos; ésta constituye una solución al problema planteado de la falta de estructura de estos atributos.
  - El modelo también garantiza que dichas dimensiones textuales puedan ser combinadas con otras del tipo clásico y sobre ellas

se definen las mismas operaciones que tiene definido el modelo clásico sobre las dimensiones clásicas. Por ejemplo, el uso de una jerarquía y de operaciones OLAP como el Roll-Up, Drill-Down y Dice.

- La jerarquía asociada a una determinada consulta que se propone usar sobre una dimensión-AP, constituye de gran ayuda al usuario porque, mientras que las jerarquías clásicas son agrupaciones propuestas por el cliente, éstas nuevas jerarquías son generadas automáticamente por el sistema que implementa el modelo gracias a la estructura de la dimensión-AP.

2. *Se ha implementado una herramienta OLAP capaz de modelar y consultar cubos multidimensionales, con soporte a atributos textuales en sus dimensiones.*

- La implementación que realiza Wonder del nuevo modelo multidimensional propuesto, permite las funcionalidades requeridas en sistemas OLAP, incluyendo soporte a datos textuales. Las gráficas e informes presentadas en los ejemplos implementados demuestran la utilidad de dicha herramienta para la toma informada de decisiones.
- Wonder construye automáticamente las jerarquías de consultas sobre las dimensiones textuales para las consultas que son de interés del usuario. Dicha jerarquía está compuesta por términos más y menos específicos con respecto a la consulta inicial.
- Para analizar con éxito dichos tipos de datos textuales, se han mostrado ejemplos donde se usan dimensiones-AP. En ellos se refleja que se facilita el procesamiento de la información útil contenida en estos tipos de datos.
- Se pudo comprobar la viabilidad del uso de una dimensión-AP en sistemas data warehousing. La información que esta nueva dimensión encierra es de gran valor para la toma de decisiones en cualquier entorno, como por ejemplo el hospitalario. Las consultas que se mostraron en los ejemplos implementados así lo corroboran.

- Las ventajas analíticas que normalmente ofrecen los sistemas data warehousing clásicos, se acrecientan con la implementación de este modelo. La posibilidad que brinda Wonder, de consultar el conocimiento obtenido asociado a los atributos textuales con el uso de las operaciones OLAP representa una ventaja analítica considerable.
  - La implementación que se realizó de un data warehousing de publicaciones brindó soluciones a algunos de los problemas encontrados en los estudios previos relacionados, como lo es la imposibilidad de procesar en cubos de datos, la estructura de conocimiento obtenida correspondiente a atributos textuales de bases de datos, conjuntamente y del mismo modo que el resto de los datos.
3. *Se ha evaluado y refinado el sistema y el modelo obtenido con datos reales, utilizando el criterio de expertos.*
- La evaluación se ha realizado por concepto de utilidad práctica, según las necesidades reales de información de los expertos consultados. Dicha información se ha obtenido consultando atributos textuales y su combinación con atributos de tipos clásicos, almacenada en bases de datos médicas y de bibliometría. Gracias a la propuesta realizada en esta memoria, dichos atributos textuales han podido ser tenidos en cuenta para la toma de decisiones, cuando antes no era posible debido a su falta de estructura.
  - El uso de dimensiones-AP y de las operaciones que sobre ellas se implementan, ofrecen ventajas analíticas a las consultas realizadas. La jerarquía de consultas asociada a estas dimensiones, facilita al usuario la obtención de información útil para la toma de decisiones, porque la misma le sugiere todas las posibles consultas que se pueden realizar sobre la información textual analizada. Se ha comprobado que esas dimensiones se pueden combinar con otras de tipos clásicos, explotando así la utilidad de la información implícita que ellas contienen.
  - En el entorno médico se pudieron satisfacer consultas de valor tanto para médicos como para los gestores a nivel hospitalario. Se

podieron dar respuestas a interrogantes como el comportamiento de las intervenciones relacionadas con determinados diagnósticos. Por ejemplo, para los diagnósticos que se relacionan con *cáncer de mama*, cuál es el tratamiento quirúrgico más utilizado.

- Relacionado con la gestión hospitalaria, se pudieron reconocer patrones de ocurrencias de urgencias que influyen en la mejor planificación y funcionamiento de un hospital. Gracias al análisis de la información textual se pudo conocer por ejemplo, la distribución de algunos motivos de urgencias de especial interés por día de la semana. También se logró precisar el promedio de espera para recibir la primera atención, asociado a diferentes causas como *agresión, accidente laboral y accidente de tráfico*.
- En el entorno de la Bibliometría, se realizaron consultas específicas con el objetivo de estudiar y analizar la actividad científica, basadas en la literatura de carácter científico y en los autores que la producen. Además de las consultas que contabilizarán la productividad científica de autores, grupos u organismos que se manejan, se pudo analizar el dominio de las publicaciones científicas de Andalucía en las TIC. Entre los hallazgos encontrados, y corroborados por el especialista, está la cantidad de artículos producidos por los grupos de investigación en temáticas como *Bases de Datos Difusas y Reconocimiento de Patrones*. A partir de dicha información, se le dio seguimiento a estos temas (utilizando la jerarquía de la dimensión-AP) para encontrar otros que de ellos se derivan, y resultan también de interés.

## 6.2. Trabajos futuros

El trabajo descrito en esta memoria abre todo un conjunto de ideas sobre las que continuar la investigación. Algunas de ellas no han sido abordadas por estar fuera de los objetivos inicialmente planteados, y otras han aparecido como consecuencia de las aportaciones aquí realizadas. A continuación, presentamos algunas de las líneas que hemos considerado más interesantes

de cara a futuras aportaciones.

1. *Desde el punto de vista de extensión del modelo abstracto:*

- Introducción de incertidumbre y soporte en las estructuras del modelo y sus operaciones. Permitiría plantear el modelo de forma difusa y utilizar el soporte de los itemsets en la definición de las operaciones.
- Estudiar la posibilidad de modelar un atributo-AP como medida de un cubo de datos.

2. *Profundizar en el estudio del rendimiento del sistema propuesto:*

- Considerar otro tipo de fuente de datos, como por ejemplo documentos externos, datos provenientes de otras bases de datos más allá del entorno hospitalario y revistas científicas (resúmenes de artículos, correos electrónicos, etc ).

# Apéndice A

## Manual de usuario de Wonder 3.0

Este apéndice estará dedicado a la presentación de un manual de usuario básico y actualizado de la herramienta Wonder OLAP Server v3.0, teniendo en cuenta que ya se ha descrito su arquitectura y principales funcionalidades en los capítulos 4 y 5 de esta memoria.

En este manual se describen las capacidades y funcionalidades del sistema. Se ha desarrollado con el objetivo de brindar todos los detalles que puedan ser de utilidad al usuario en el manejo del mismo. Con la ayuda de las imágenes de sus principales ventanas se exponen los pasos necesarios para explotar al máximo sus prestaciones. Es importante señalar que la mayoría de las opciones fueron implementadas a partir del uso de wizard (asistentes). Esto facilita el trabajo con el sistema, pues dichos asistentes garantizan que se completen las operaciones. También es necesario que los usuarios sepan que éste mismo manual puede hallarlo en forma de ayuda dentro de la propia aplicación, a través de un navegador sencillo con el árbol típico de libros de contenidos.

### A.1. Características de Wonder 3.0

Wonder V.3.0 es una herramienta OLAP, capaz de analizar grandes cantidades de datos con la creación de cubos de datos multidimensionales. Ha sido

implementado usando Java como lenguaje de programación, y como servidor de datos PostgreSQL, ambas tecnologías punteras dentro del software libre. Como novedad, esta versión permite analizar el conocimiento contenido en atributos textuales de bases de datos, definiendo el mismo como una dimensión más de un cubo de datos, llamada dimensión-AP. El objetivo principal de la herramienta es agilizar el proceso de toma de decisiones en distintos ambientes donde sea utilizado.

### **A.1.1. Requerimientos para su ejecución**

Como requerimientos para su ejecución Wonder necesita:

- De un procesador Pentium o superior con memoria RAM de 512 Mb mínimo y velocidad de 500 MHz mínima.
- El espacio libre en disco duro debe ser de 100 Mb, la resolución del monitor debe ser de 800 x 600 px o mayor.
- La fuente de datos debe estar implementada en un servidor PostgreSQL y la computadora donde se encuentre el mismo sea una Pentium III o superior con memoria RAM 256 Mb mínimo, velocidad 500 MHz mínimo y espacio libre en disco duro de 200 Mb.

## **A.2. Guía de explotación**

En la barra de herramientas del software se pueden encontrar distintas operaciones a realizar, en esta sección explicaremos cada una de ellas detalladamente.

### **A.2.1. Conexión al servidor de datos**

La conexión con el servidor de datos de Postgres se realiza a través de un botón situado en la vista OLAP, señalado en la figura A.1. Al hacer click en ese botón se muestra la ventana de la figura A.2.

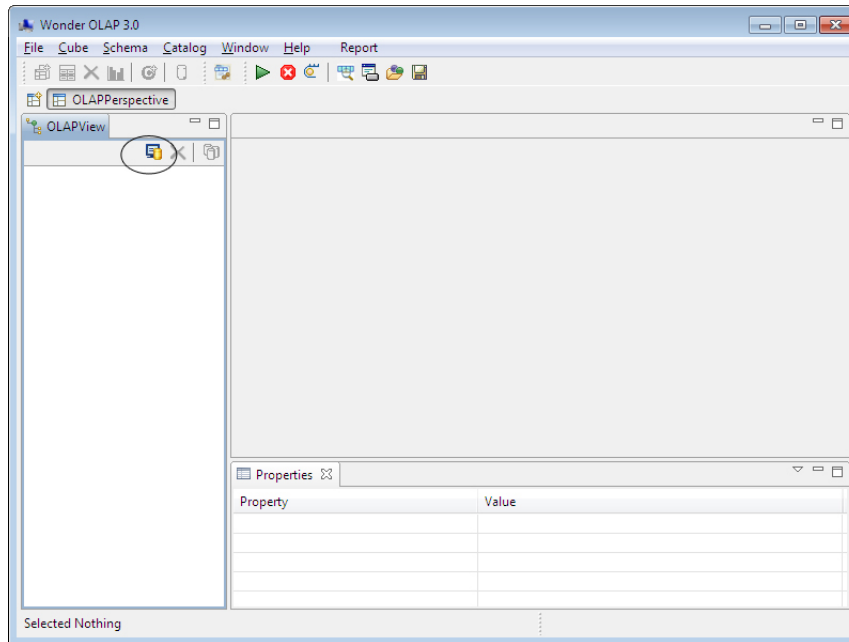


Figura A.1: Icono para realizar la conexión con el servidor de datos

Como se puede observar en la figura A.2 la ventana de conexión cuenta con varios parámetros de entrada necesarios para establecer la conexión con el servidor de bases de datos. A continuación serán explicados cada uno de ellos.

- Host: servidor de base de datos.
- Port: puerto por el que se realizará la conexión a la base de datos.
- DataBase Server: nombre del gestor de base de datos.
- User: usuario para la conexión al servidor.
- Password: clave para establecer la conexión.

Si todos estos parámetros son entrados de la forma correcta se accede a la vista principal de la aplicación. La misma se presenta en la figura A.3. Ya establecida la conexión se puede comenzar a trabajar con el sistema. Primero se crean los catálogos desde un icono (botón de acceso directo)



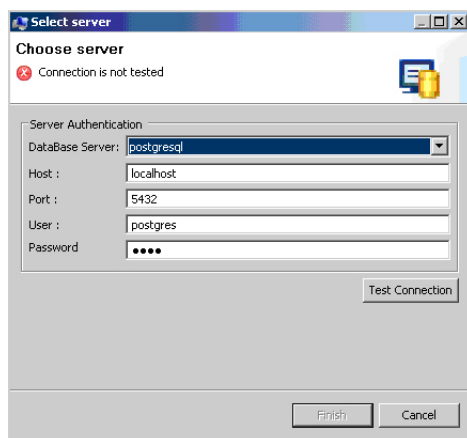


Figura A.2: Ventana de conexión con el servidor de datos

situado seguidamente después del eliminar conexión. Con click derecho sobre el catálogo se crean los esquemas donde estarán los cubos de datos o usando el menú principal de la aplicación.

### A.2.2. Vistas de la aplicación

La vista más usada es la Vista OLAP. Como vista principal de la aplicación contiene todos los cubos que han sido mapeados en el servidor OLAP. Para lograr que la información de los cubos y el manejo con los mismos sea de forma clara y organizada, esta vista ha sido implementada como un árbol jerárquico donde el nivel más externo corresponde a los catálogos y dentro de éstos los esquemas. Dentro de los esquemas se encuentran los cubos, los cuales contienen las dimensiones y las medidas correspondientes. Dentro de las dimensiones se encuentran las jerarquías y dentro de estas los distintos niveles que la conforman.

Esta vista proporciona brinda un conjunto de opciones a través de un menú contextual en dependencia del elemento seleccionado en la vista. La siguiente figura A.5 es el menú que se muestra cuando está seleccionado (click derecho) un cubo. Las opciones de dicho menú se pueden ver con claridad. Es importante destacar que las opciones *Perform Dice* y *Perform Slice* también

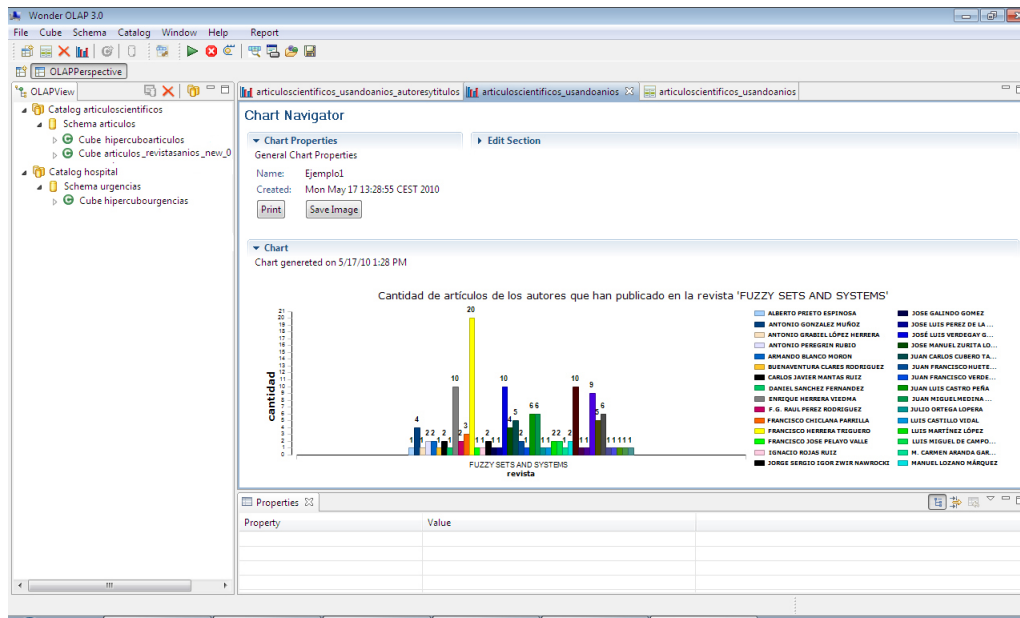


Figura A.3: Interfaz principal de Wonder

se pueden realizar de forma combinada dentro de *Create Subcube*, y que las opciones *Show report* y *Show chart* se pueden obtener como resultado del *Consulting Cube*. Éstas relaciones entre las opciones se han implementado con el objetivo de brindar la mayor flexibilidad y reutilización de funcionalidades.

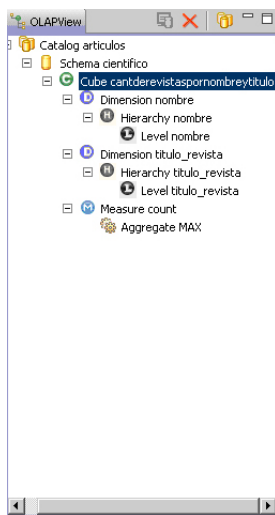


Figura A.4: Vista OLAP de la aplicación

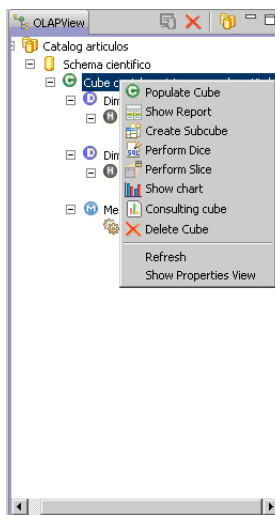
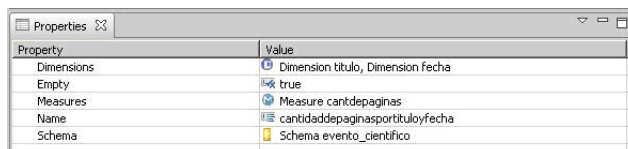


Figura A.5: Vista OLAP con su menú contextual

Otra de las vistas es la Vista de propiedades. Es la que contiene las propiedades fundamentales del objeto que esté seleccionado en ese momento en la vista

OLAP. En la figura A.6 se muestran las propiedades de un cubo, que son el nombre , el esquema al que pertenece, sus dimensiones, así como las medidas. También muestra si no está vacío.



Property	Value
Dimensions	Dimension titulo, Dimension fecha
Empty	true
Measures	Measure cantdepaginas
Name	cantidaddepaginasportituloylecha
Schema	Schema evento_cientifico

Figura A.6: Vista de propiedades

### A.2.3. Crear un cubo

Las siguientes 5 figuras (desde la figura A.7 hasta la figura A.11), muestran el asistente para crear un cubo. Se ha implementado esta y la mayoría de las funcionalidades del sistema, a través de un asistente para que el usuario logre hacer sus operaciones de la forma más cómoda posible y pueda deshacer operaciones realizadas previamente sin necesidad de cancelar su operación.

La figura A.7 muestra la página inicial del asistente, para conectarse con la base de datos desde la que desea importar los datos del cubo. El usuario selecciona la fuente de datos donde se encuentran los datos que desea analizar y prueba la conexión. Si la conexión fue satisfactoria, se le permite continuar a la siguiente página.

En la figura A.8 se observa la posibilidad que brinda el sistema de seleccionar las columnas, a partir de las cuales se crearán las dimensiones y medidas del nuevo cubo. Se llenan por defecto los campos Schema y Catalog a que corresponde el cubo. En esta página el usuario debe entrar el nombre del cubo, si desea poblarlo en ese momento, así como seleccionar la tabla o vista donde se encuentran las columnas que serán dimensiones o medidas.

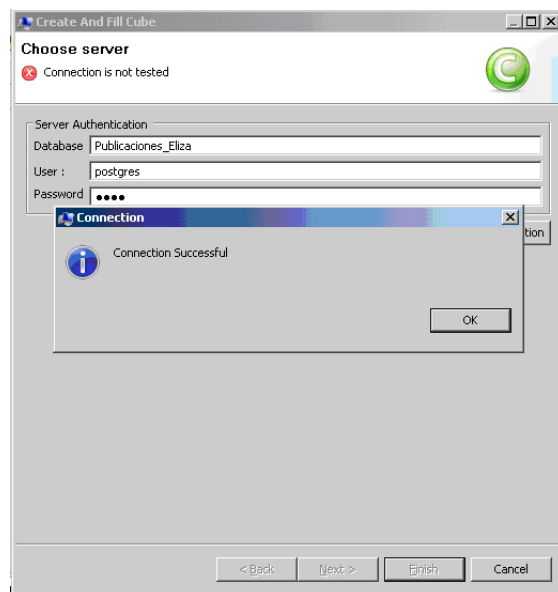


Figura A.7: Primera ventana para crear un cubo: Establecer conexión con la fuente de datos

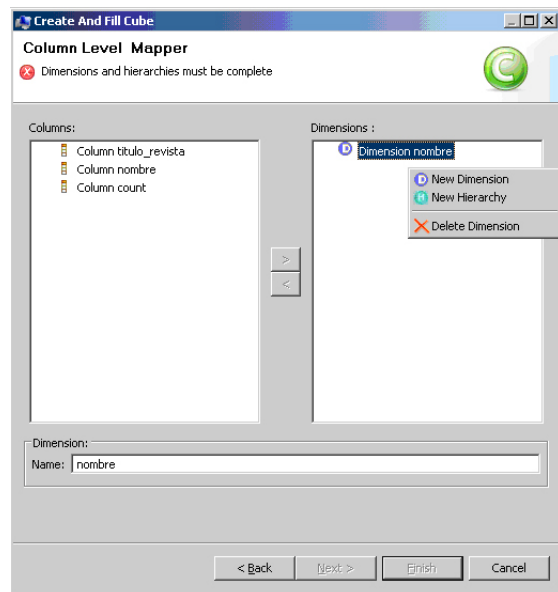


Figura A.9: Definición y mapeo de las dimensiones

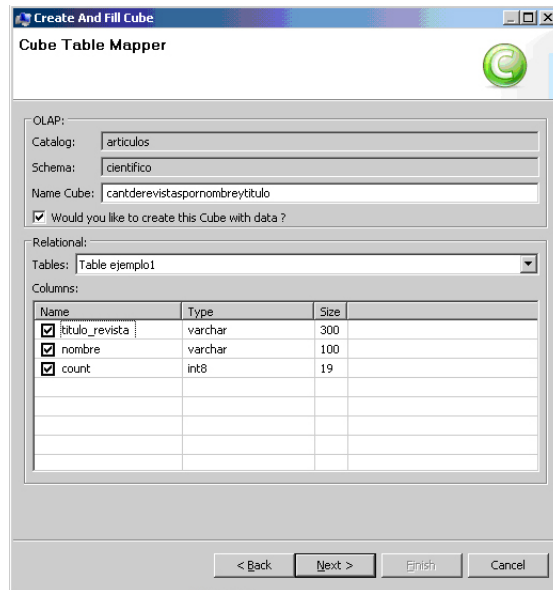


Figura A.8: Definición del nombre del cubo y mapeo con la fuente de datos

La página que permite crear las dimensiones, jerarquías y niveles del nuevo cubo se muestra en la figura A.9. Mediante un menú contextual el usuario puede crear dimensiones y dentro de estas jerarquías. Luego puede seleccionar la columna que desea mapear y con el botón ">" adicionarle esta columna a la jerarquía como nivel. Es importante aclarar que, para que este botón se active debe estar seleccionada una columna y una jerarquía.

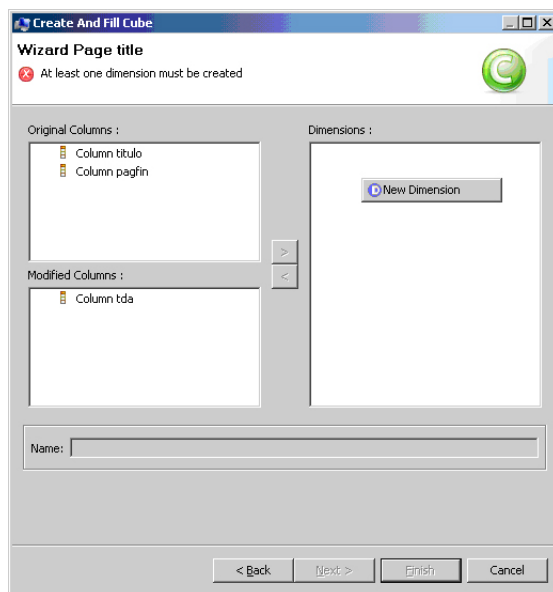


Figura A.10: Definición de una dimensión-AP

En la figura A.10 se muestra la página que al igual que la anterior permite crear las dimensiones, jerarquías y niveles pero de tipo texto libre (dimensión-AP). Mediante un menú contextual el usuario puede crear dimensiones y dentro de estas jerarquías. Luego puede seleccionar la columna de tipo texto y otra columna de tipo tda que desea mapear y con el botón ">" adicionarle este nivel a la jerarquía (para que este botón se active debe estar seleccionada una columna tipo texto, una del tipo tda y una jerarquía).

Como se observa en la figura A.11 en la página que permite seleccionar las medidas, el usuario debe seleccionar la columna que desea mapear como medida, luego seleccionar la función agregada que le desea aplicar y con el botón '>' adicionar esta medida al cubo (para que este botón se active debe estar seleccionada una columna y una función de agregación).

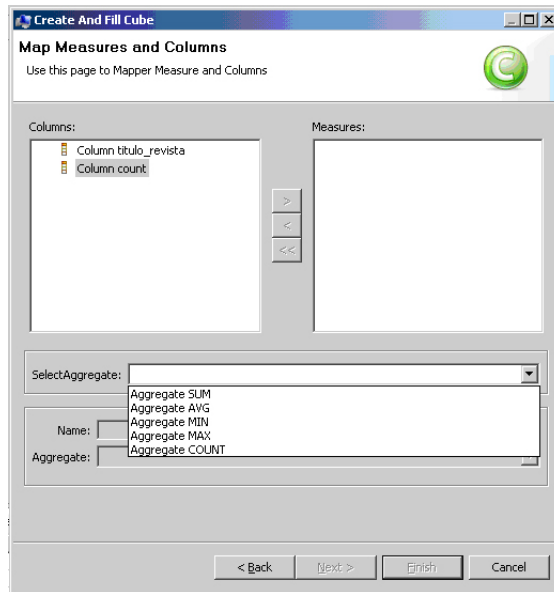


Figura A.11: Definición de las medidas del cubo

#### A.2.4. Mostrar un gráfico

Las siguientes 2 figuras (figura A.12 y A.13) muestran el asistente creado para mostrar el gráfico de un cubo, con el click derecho sobre el mismo. En figura A.12 se observa la página que permite configurar el título del gráfico y el tipo (Bar, Line, Scatter). En la figura A.13 se permite configurar la manera en que se van a analizar los datos del cubo en el gráfico. Se permite seleccionar la medida, que dimensión se mostrará en la leyenda y cuál como categoría. Para no repetir las mismas imágenes se ha reservado mostrar un gráfico generado para el final de la opción *Consulting Cube* que de forma similar lo obtiene.



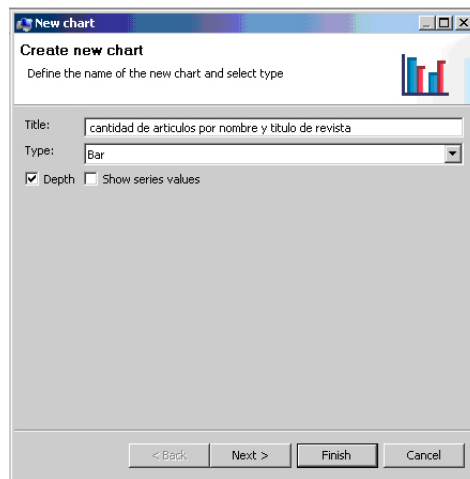


Figura A.12: Definición del título y tipo de un gráfico

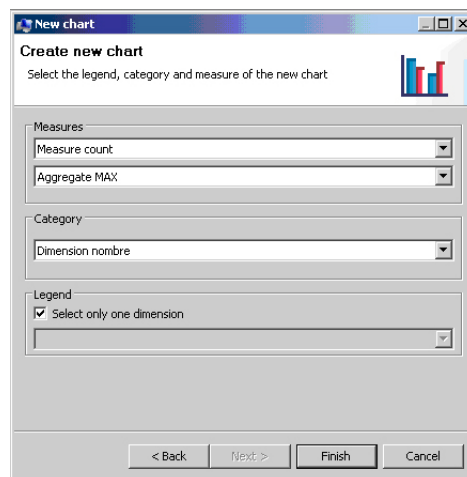


Figura A.13: Configuración de los parámetros de un gráfico

### A.2.5. Mostrar un informe

Las siguientes 4 figuras (desde la figura A.14 hasta la A.17), muestran el asistente creado para mostrar el informe de un cubo.

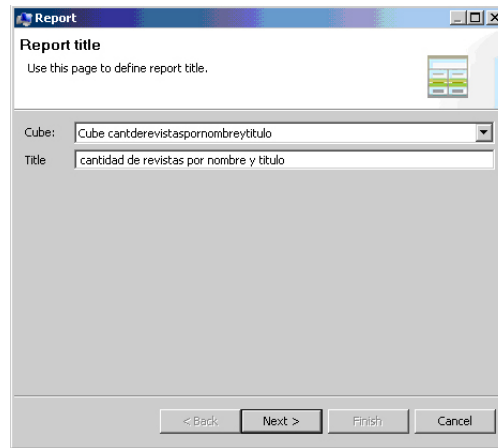


Figura A.14: Definición inicial de un informe

La página inicial del asistente para crear un informe se expone en la figura A.14. Teniendo en cuenta que esta funcionalidad debe ser invocada teniendo seleccionado un cubo, por defecto en los campos Title y Cube aparecen los valores correspondientes al cubo seleccionado, pero se brinda la posibilidad de seleccionar otro cubo.

En la figura A.15 se observa la posibilidad que brinda el sistema de seleccionar las dimensiones que se van a mostrar en el informe. En la lista de la izquierda aparecen todas las dimensiones del cubo original, y se pueden ir adicionando a la lista de la derecha, las que se utilizarán para crear el informe. Una vez que una dimensión está en la lista de salida (derecha) es posible continuar con la siguiente página.

En la figura A.16 puede verse que el sistema permite seleccionar la medida del cubo a partir de la cual se creará el informe y en la figura A.17 se puede ver un ejemplo de informe generado. Nótese que el mismo se puede salvar como pdf o imprimir directamente.

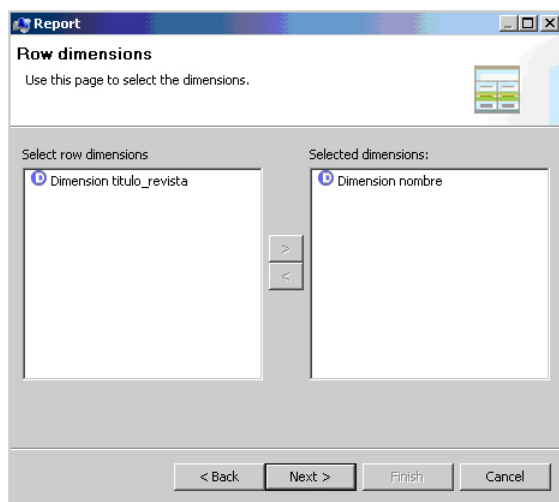


Figura A.15: Configuración de las dimensiones por las que estará compuesto el informe

## A.2.6. Consultar un cubo

En las 3 figuras siguientes (desde figura A.18 hasta la A.21), se expone el asistente correspondiente a la opción *Consulting Cube*, que como ya se ha mencionado se puede acceder a ella cuando se da click derecho sobre un cubo. Esta opción permite hacer consultas sobre las dimensiones del cubo seleccionado, generando al final del asistente un gráfico (por defecto) y un informe (opcional).

La figura A.18 muestra la página inicial del asistente para consultar un cubo. El usuario tiene la facilidad de introducir el nombre del gráfico que se obtendrá, que será el mismo del informe.

Durante la configuración del resto de los parámetros para consultar un cubo, puede verse en la figura A.19 que el usuario debe seleccionar la medida con la función agregada que le desea aplicar, la dimensión que será la categoría en el gráfico y la que será leyenda. Además puede decidir si quiere analizar los valores que no cumplen con la consulta aplicada y si desea mostrar el informe.

Si el usuario seleccionó una sola dimensión ésta será la última página del

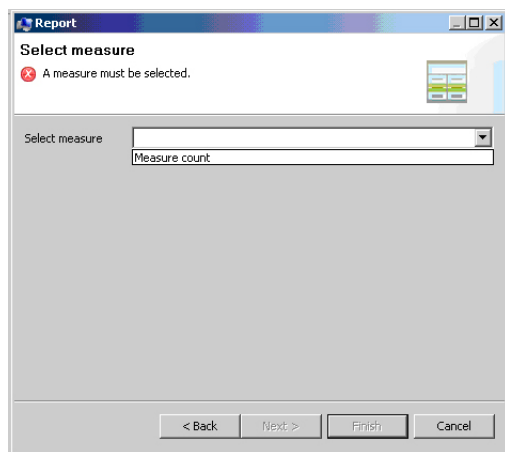


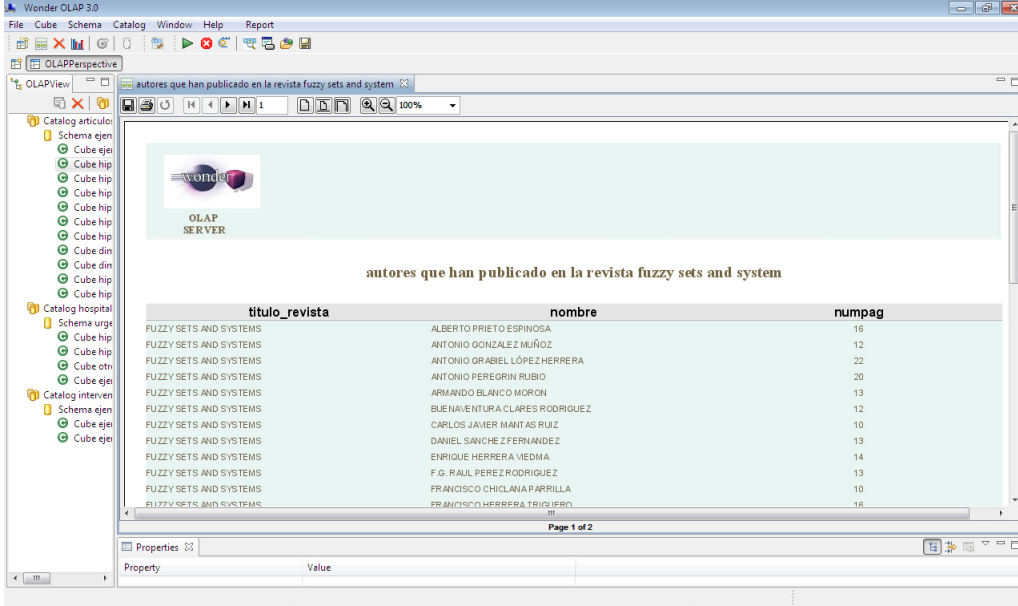
Figura A.16: Selección de la medida que se mostrará en el informe

asistente de lo contrario se mostrará otra similar con la otra dimensión seleccionada. El usuario debe seleccionar el operador que desea aplicar y los valores para consultar puede escribirlos directamente y con el botón Add agregarlos o seleccionarlos de la lista de valores mostrada. Los resultados que se obtienen, un gráfico y/o un informe pueden verse en la figura A.21.

### A.2.7. Operaciones Roll-up y Drill-Down sobre una dimensión

Las operaciones de Roll-Up y Drill-Down dentro de la jerarquía de una dimensión clásica (no dimensión-AP) puede verse en la figura . Cuando el nivel seleccionado es el primero, con click derecho sobre la jerarquía, aparece solo la opción Drill-Down, una vez que se está en cualquier otra aparecen las dos operaciones.

En el caso de las dimensiones-AP la jerarquía que se define es diferente, es una jerarquía de consulta con algunas especificidades. Las siguientes 3 figuras (desde figura A.23 hasta la A.25), muestran el asistente creado para realizar las operaciones Roll-up y Drill-down sobre una jerarquía de consulta asociada a una dimensión-AP de un cubo.



autores que han publicado en la revista fuzzy sets and system

titulo_revista	nombre	numpag
FUZZY SETS AND SYSTEMS	ALBERTO PRIETO ESPINOSA	16
FUZZY SETS AND SYSTEMS	ANTONIO GONZALEZ MUÑOZ	12
FUZZY SETS AND SYSTEMS	ANTONIO GRABEL LÓPEZ HERRERA	22
FUZZY SETS AND SYSTEMS	ANTONIO PEREGRIN RUBIO	20
FUZZY SETS AND SYSTEMS	ARMANDO BLANCO MORON	13
FUZZY SETS AND SYSTEMS	BUENAVENTURA CLARES RODRIGUEZ	12
FUZZY SETS AND SYSTEMS	CARLOS JAVIER MANTAS RUIZ	10
FUZZY SETS AND SYSTEMS	DANIEL SANCHEZ FERNANDEZ	13
FUZZY SETS AND SYSTEMS	ENRIQUE HERRERA VIEDMA	14
FUZZY SETS AND SYSTEMS	F.G. RAUL PEREZ RODRIGUEZ	13
FUZZY SETS AND SYSTEMS	FRANCISCO CHICLANA PARRILLA	10
FUZZY SETS AND SYSTEMS	FRANCISCO HERRERA TRIGUERO	16

Page 1 of 2

Figura A.17: Ejemplo de informe

En la figura A.23 se muestra la página para seleccionar el valor del tda mejor contenga las frases de consulta que desee. Luego se pasa a la figura A.24 donde el usuario puede editar el tda seleccionado hasta obtener exactamente la frase de consulta deseada. Seguidamente presionado el botón Find el sistema le devuelve las posibles frases de consulta más o menos detallada que la formulada y el usuario selecciona de dichos grupos las que desee como consulta final.

En la figura A.25 se muestra la última página del asistente y es aquí donde el usuario selecciona la función de acoplamiento que quiere aplicar, así como los valores de configuración de dicha función.

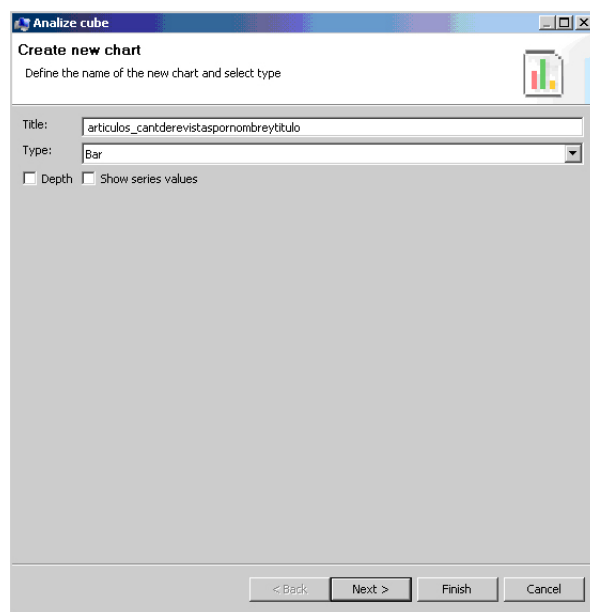


Figura A.18: Definición inicial de la opción Consulting Cube

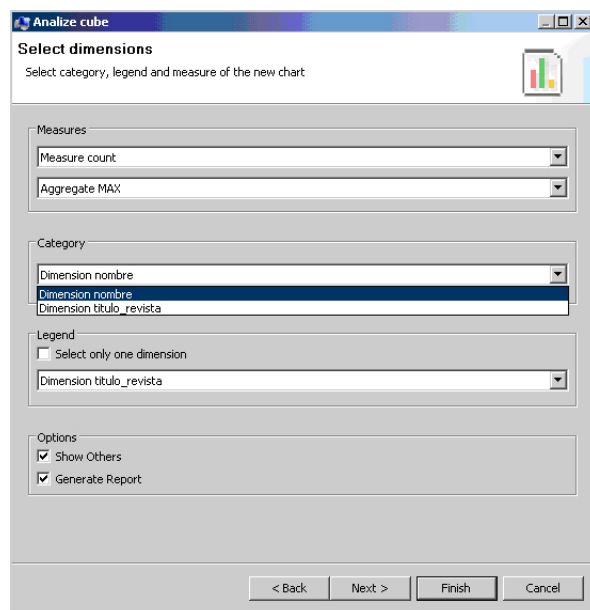


Figura A.19: Configuración de los parámetros necesarios para consultar un cubo

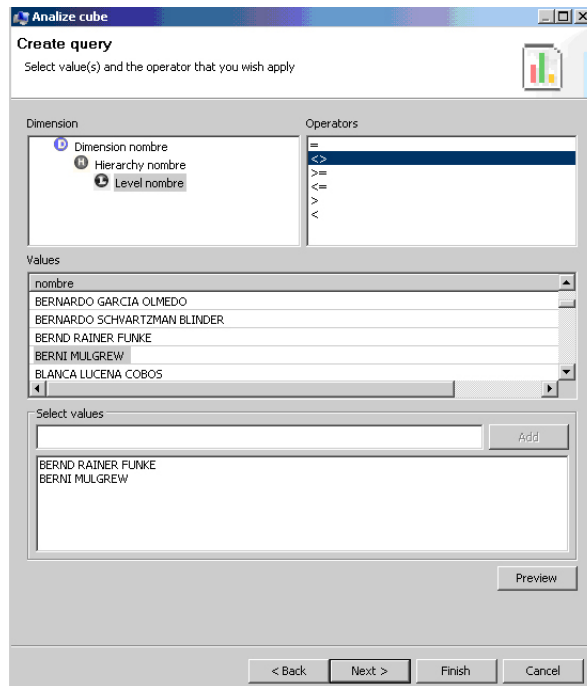


Figura A.20: Consultando las dimensiones del cubo

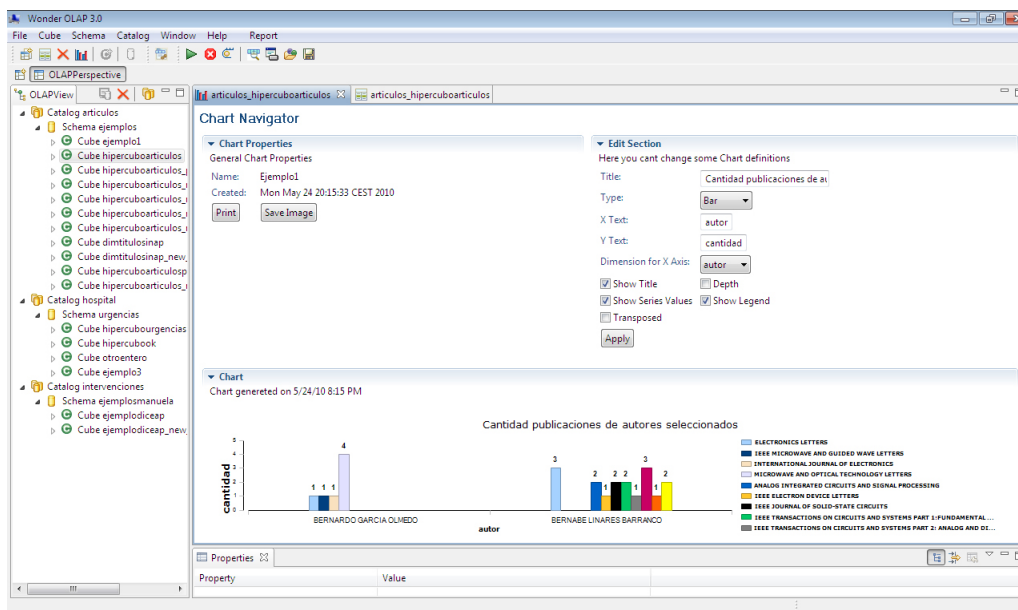


Figura A.21: Resultados que se obtienen de consultar un cubo

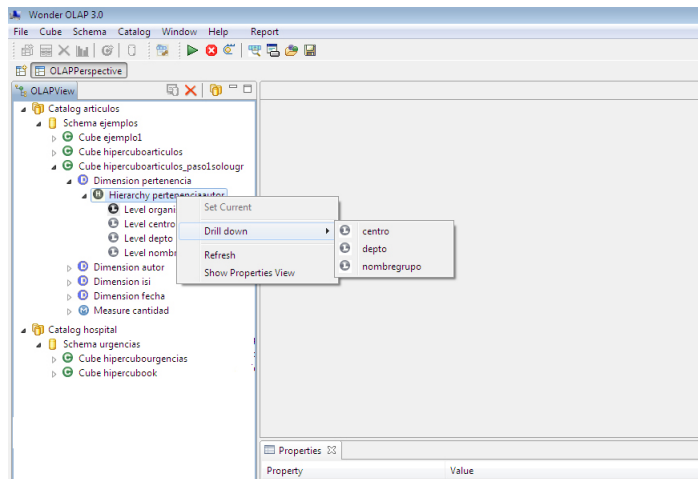


Figura A.22: Uso de una jerarquía clásica

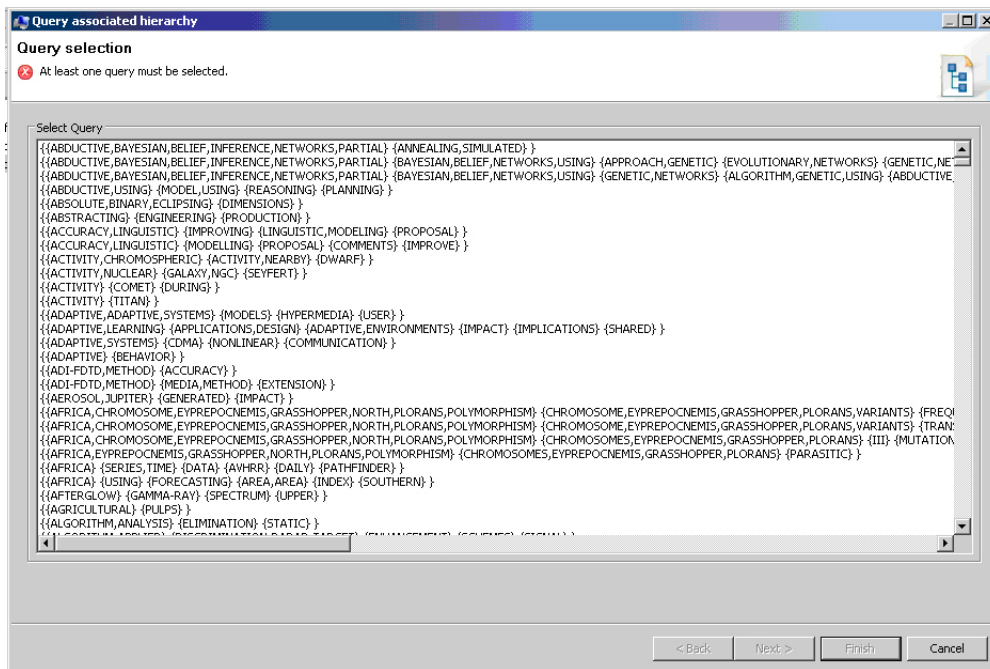


Figura A.23: Uso de una jerarquía de consulta



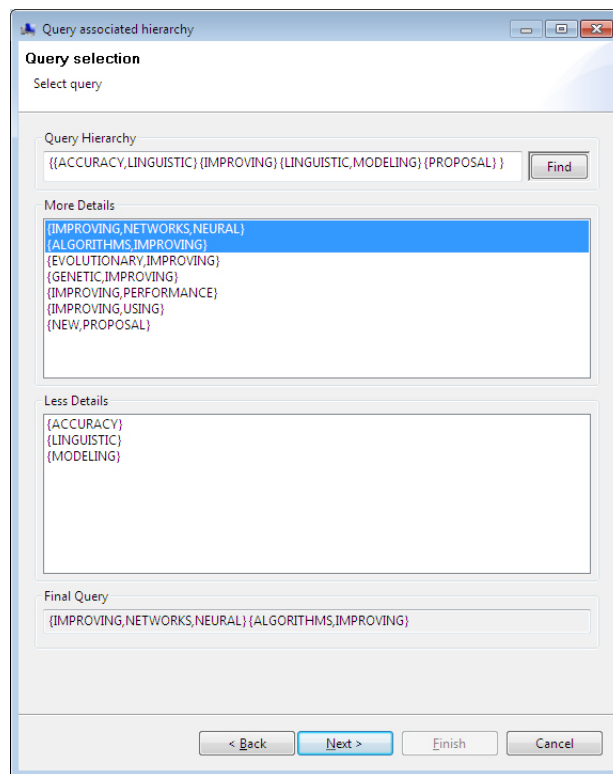


Figura A.24: Obtención de los niveles de la jerarquía de consulta

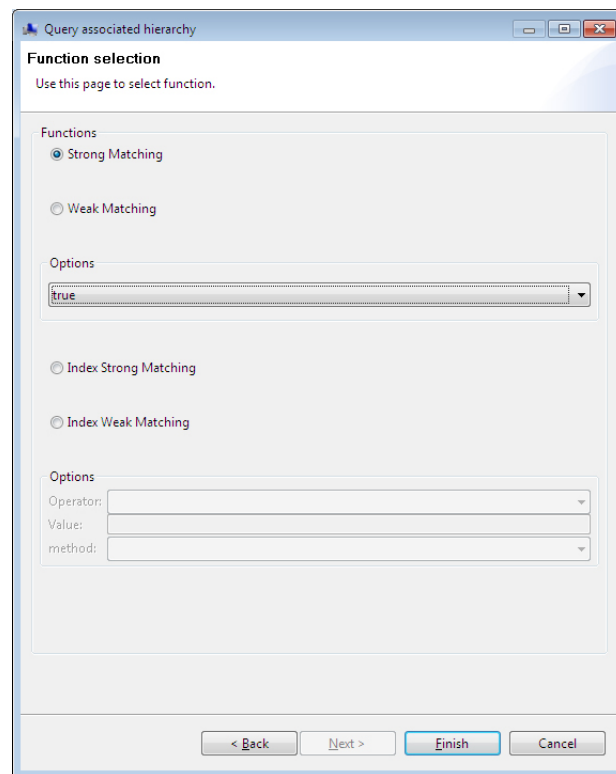


Figura A.25: Elección del tipo de acoplamiento a aplicar



# Bibliografía

Agrawal, R., Srikant, R., Sept 1994. Fast algorithms for mining association rules. En: In Proceedings of International Conference on Very Large Data Bases( VLDB'94). Santiago, Chile.

Ahonen, H., Heinonen, O., Klemettinen, M., Verkamo, A. I., 1998. Applying data mining techniques for descriptive phrase extraction in digital document collections. En: In Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries( ADL). pp. 2–11.

Appiani, E., Cesarini, F., Colla, A. M., Diligenti, M., Gori, M., Marinai, S., Soda, G., 2001. Automatic document classification and indexing in high-volume applications. International Journal on Document Analysis and Recognition (IJ DAR) 4 (2), pp. 69–83.

Banek, M., Tjoa, A. M., Stolba, N., September 2006. Integrating different grain levels in a medical data warehouse federation. En: Data Warehousing and Knowledge Discovery (DaWaK). Krakow, Poland, pp. pp. 185–194.

Batista, K. G., Tejada, E., 2009. Wonder v 3.0: Servidor olap de libre disposición con soporte a textos libres. Monografía (Universidad de Camaguey).

Berstel, J., Boasson, L., 2000. Xml grammars. En: Mathematical Foundations of Computer Science. Lecture Notes in Computer Science. Springer, Bratislava, Slovakia, pp. 182–191.

Bhattacharyya, S. B., 2009. Data warehousing and data mining: an overview. <http://www.sbbhattacharyya.info/>.

- Bull, R. I., Best, C., Storey, M.-A., 2004. Advanced widgets for eclipse. En: Eclipse '04: Proceedings of the 2004, Conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA) workshop on eclipse technology eXchange. ACM, New York, NY, USA, pp. 6–11.
- Cecchet, E., Marguerite, J., Zwaenepole, W., 2004. C-jdbc: flexible database clustering middleware. En: Proceedings of the annual conference on USENIX Annual Technical Conference (ATEC '04:). USENIX Association, Berkeley, CA, USA, pp. 26–26.
- Cody, W. F., Kreulen, J. T., Krishna, V., Spangler, W. S., 2002. The integration of business intelligence and knowledge management. *IBM Systems Journal* 41 (4), pp. 697–713.
- Cosijn, E., Keskustalo, H., Pirkola, A., De-Wet, K., Jarvelin, K., Oct 2004. Afrikaans - english cross-language information retrieval. En: In Proceedings of the 3rd biennial Development of Information Science in South Africa (DISSAnet) Conference. Pretoria, South Africa, pp. 97–110.
- Delgado, M., in Bautista, M. M., Sánchez, D., Vila, M., September 2002. Mining text data: Special features and patterns. En: Proceedings of Pattern Detection and Discovery: European Science Foundation (ESF) Exploratory Workshop. Springer-Verlag Heidelberg.
- Delgado, R. P., Zavalaga, L. F. L., Morales, E. A. C., 2005. Concordancia entre el diagnóstico médico y la codificación de informática, considerando el cie-10, en la consulta externa de pediatría en el hospital nacional cayetano heredia, lima-perú. *Revista Médica Herediana*. [online] vol.16 (4), p.239–245.
- E.F. Codd, S. C., Salley, C., 1993. Providing olap to user-analysts: An it mandate. Tech. rep., E.F. Codd Associates.
- Fei-Ran Guo, Bambang Parmanto, J. J. I. J. W., Fang, H., 2000. The implementation of data warehouse and olap for rehabilitation outcome evaluation: Redwine system. American Medical Informatics Association (AMIA), 1023.

- Feldman, R., Sanger, J., 2007. The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data. Cambridge University Press, NY, USA.
- Feng, F., Croft, W. B., 2001. Probabilistic techniques for phrase extraction. *Information Processing & Management* 37 (2), pp. 199 – 220.  
URL <http://www.sciencedirect.com/science/article/B6VC8-4292GM8-2/2/222aeade7b79cf0e299f5811eb4f9eac>
- Fuhr, N., 2003. Xml information retrieval and information extraction. En: *Text Mining. Theoretical Aspects and Applications*. Physica Verlag, pp. pp. 21–32.
- Galhardas, H., Florescu, D., Shasha, D., Simon, E., Saita, C., 2001. Improving data cleaning quality using a data lineage facility. En: *Proceedings of Design and Management of Data Warehouses( DMDW)*. Interlaken, Switzerland, p. 3.
- Gibas, C., Jambeck, P., 2001. *Developing Bioinformatics Computer Skills*. O'Reilly Media, Inc.
- Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999. Summarizing text documents: sentence selection and evaluation metrics. En: *SIGIR '99: Proceedings of the 22nd annual international ACM Special Interest Group on Information Retrieval(SIGIR) conference on Research and development in information retrieval*. Association Computing Machinery( ACM) Press, New York, NY, USA, pp. 121–128.
- Golfarelli, M., Rizzi, S., Vrdoljak, B., 2001. Data warehouse design from xml sources. En: *Proceedings of the fourth Association Computing Machinery( ACM) international workshop on Data warehousing and OLAP*. Association Computing Machinery( ACM) Press.
- Grimes, S., Mar 2006. New directions for olap.  
URL <http://www.intelligententerprise.com/showArticle.jhtml;jsessionid=ICPGHXWQEFX2SQSNDL0SKHSCJUNN2JVN?articleID=181401812>

- Hahn, U., Mani, I., 2000. The challenges of automatic summarization. Institute of Electrical and Electronics Engineers( IEEE) Computer 33 (11), 29–36.
- Han, J., Kamber, M., 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.
- Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., Jarvelin, K., 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Information Retrieval 4, pp. 209–230.
- Hibernate-org, Enero 2009. Relational persistence for java and .net.  
URL <http://www.hibernate.org/>
- Hristovski, D., Rogac, M., Markota, M., 2000. Using data warehousing and olap in public health care. Proceedings of American Medical Informatics Association ( AMIA) Symposium, pp. 369–373.
- Inmon, W. H., Mar 1996. Building the Data Warehouse, third edition Edición. Wiley Computer Publishing.
- Inokuchi, A., Takeda, K., 2007. A method for online analytical processing of text data. En: Proceedings of the sixteenth Association Computing Machinery( ACM) conference on Conference on information and knowledge Management( CIKM '07). Association Computing Machinery( ACM), New York, NY, USA, pp. 455–464.
- Jensen, M. R., Moller, T. H., Pedersen, T. B., 2001. Specifying olap cubes on xml data. Journal Of Intelligent Information Systems 17, 200–1.
- Justicia, C., 2004. Formas intermedias de representacion en mineria de texto. Memoria para el Diploma de Estudios Avanzados, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Granada, España.
- Kara, K. T., marzo 2008. Relational model to dimensional model.  
URL <http://kubilaykara.blogspot.com/2008/03/relational-model-to-dimensional-model.html>

- Keith, S., Kaser, O., Lemire, D., 2006. Analyzing large collections of electronic text using olap. The Computing Research Repository (CoRR) abs/cs/0605127.
- Lin, C. X., Ding, B., Han, J., Zhu, F., Zhao, B., December 2008. Text cube: Computing ir measures for multidimensional text database analysis. En: International Conference on Data Mining( ICDM'2008). Pisa, Italy, pp. pp. 905–910.
- Lin, S.-H., Shih, C.-S., Chen, M. C., Ho, J.-M., Ko, M.-T., Huang, Y.-M., 1998. Extracting classification knowledge of internet documents with mining term associations: A semantic approach. En: Special Interest Group on Information Retrieval( SIGIR'98). Association Computing Machinery( ACM) Press, Melbourne, Australia, pp. pp. 241–249.
- Llavori, R. B., Cabo, M. J. A., García, S., Sanz, I., Noviembre 1999. Chronology: Una aproximación al almacenamiento y recuperación de información de actualidad. En: Jornadas de Ingeniería del Software y Bases de Datos (JISBD). Cáceres, Spain, pp. 111–122.
- Loh, S., Wives, L. K., de Oliveira., J. P. M., 2000. Concept-based knowledge discovery in texts extracted from the web. ACM Special Interest Group on Knowledge Discovery and Data Mining( SIGKDD) Explorations 2 (1), pp. 29–39.
- Martín-Bautista, M., Martínez, S., Vila, M., September 2008. A new semantic representation for short texts. En: To appear in Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2008), LNCS Springer-Verlag. Turin, Italy.
- Martín-Bautista, M., Prados, M., Vila, M., Martínez, S., July 2006. A knowledge representation for short texts based on frequent itemsets. En: Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU). Paris, France.
- Martínez, S., Julio 2008. Una solución semántica al tratamiento de atributos textuales en un modelo relacional orientado a objetos: Implementación



- en software libre. Tesis Doctoral, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Granada, España.
- Martínez-Méndez, F., Rodríguez, J. V., January 2003. Síntesis y crítica de las evaluaciones de la efectividad de los motores de búsqueda en la web. *Information Research* 8 (2).
- McAffer, J., Lemieux, J.-M., October 2005. *Eclipse Rich Client Platform: Designing, Coding, and Packaging Java(TM) Applications*, 1st Edición. Addison-Wesley Professional.
- McCabe, M. C., Lee, J., Chowdhury, A., Grossman, D., Frieder, O., 2000. On the design and evaluation of a multi-dimensional approach to information retrieval (poster session). En: *Special Interest Group on Information Retrieval (SIGIR): Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. Association Computing Machinery (ACM), New York, NY, USA, pp. 363–365.
- Meyfroidt, G., Guiza, F., Ramon, J., Bruynooghe, M., 2009. Machine learning techniques to examine large patient databases. *Baillire's Best Practice & Research. Clinical Anaesthesiology* 23, pp. 127–143.
- Microsoft-Co., Jun 2006. Directivas de seguridad, administración operativa y comunicaciones. *Microsoft Developer Network (MSDN), Patterns & Practices*.  
URL <http://msdn.microsoft.com/es-es/library/ms978348.aspx>
- Mothe, J., Chrisment, C., Dousset, B., Alaux, J., 2003. Doccube: multi-dimensional visualisation and exploration of large document sets. *Journal of the American Society for Information Science and Technology( JASIST)*, Special 54, 650659.
- Nassis, V., Rajugan, R., Dillon, T. S., Rahayu, J. W., September 2004. Conceptual design of xml document warehouses. En: *Proceedings of the 6th International Conference Data Warehousing and Knowledge Discovery (DaWaK 2004)*. Springer, Zaragoza, Spain, pp. 1–14.

- Nemati, H. R., Steiger, D. M., Iyer, L. S., Herschel, R. T., June 2002. Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems* 33 (2), pp. 143–161.
- Niemi, T., Ninimäki, M., Nummenmaa, J., Thanisch, P., Sep 2003. Applying grid technologies to xml based olap cube construction. En: In *Proceedings of Design and Management of Data Warehouses( DMDW' 03) Workshop*. Berlin, Germany, pp. 2003–004.
- Park, B.-K., Han, H., Song, I.-Y., August 2005. Xml-olap: A multidimensional analysis framework for xml warehouses. En: *DaWaK*. Springer, Copenhagen, Denmark, pp. 32–42.
- Pedersen, T. B., Jensen, C. S., 2001. Multidimensional database technology. *IEEE Computer* 34 (12), pp. 40–46.
- Perez, J. M., Berlanga, R., Aramburu, M. J., Pedersen, T. B., 2007. R-cubes: Olap cubes contextualized with documents. *International Conference on Data Engineering*, pp. 1477–1478.
- Pérez, J. M., Llavori, R. B., Cabo, M. J. A., 2009. A relevance model for a data warehouse contextualized with documents. *Information Processing & Management* 45 (3), pp. 356–367.
- Pérez, J. M., Pedersen, T. B., Llavori, R. B., Aramburu, M. J., 2005. Ir and olap in xml document warehouses. En: *European Conference on IR Research (ECIR)*. pp. 536–539.
- PostgreSQL-Global-Devel-Group, Enero 2009. PostgreSQL.  
URL <http://www.postgresql.org/>
- Raghavan, S., García-Molina, H., 2001. Integrating diverse information management systems: a brief survey. *Institute of Electrical and Electronics Engineers( IEEE) Data Engineering Bulletin* 24 (4), pp. 44–52.
- Ravat, F., Teste, O., Tournier, R., Zurfluh, G., 2007. A conceptual model for multidimensional analysis of documents. En: *Conceptual Modeling - ER 2007*. Vol. 4801/2007. Springer Berlin / Heidelberg.

- Ravat, F., Teste, O., Tournier, R., Zurfluh, G., 2008. Top\_keyword: An aggregation function for textual document olap. En: Data Warehousing and Knowledge Discovery (DaWaK). pp. pp. 55–64.
- Rusty, H. E., 2001. XML Bible, second edition Edición. Hungry Minds, Inc.
- Salton, G., McGill, M., 1983. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., NY, USA.
- Sánchez, R. G., 2004. Software libre vs. software propietario: Programando nuestro futuro.  
URL <http://www.historia-actual.org/Publicaciones/index.php/haol/index>
- Shanmugasundaram, J., Tufte, K., He, G., Zhang, C., Dewitt, D., Naughton, J., September 7-10 1999. Relational databases for querying xml documents: Limitations and opportunities. En: Proceedings of 25th International Conference on Very Large Data Bases( VLDB'99). Edinburgh, Scotland, UK, pp. 302–314.
- Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., Carlsson, C., 2002. Past, present, and future of decision support technology. Decision Support Systems 33 (2), pp. 111–126.
- Stolba, N., Tjoa, A. M., 2006. Tjoa a m.: An approach towards the fulfilment of security requirements for decision support systems. En: in the Field of Evidence-Based Healthcare, Knowledge Rights - Legal, Societal and Related Technological Aspects (KnowRight'2006). pp. 51–59.
- Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T., Mar 2009. Rankclus: Integrating clustering with ranking for heterogeneous information network analysis. En: In Proceedings of International Conference on Extending Database Technology (EDBT'09). Saint-Petersburg, Russia.
- Sun, Y.-M., Huang, H.-D., Horng, J.-T., Huang, S.-L., Tsou, A.-P., 30 Aug - 3 Sep 2004. Rgs-miner: A biological data warehousing, analyzing and mining system for identifying transcriptional regulatory sites in human genome. En: In Proceedings of 15th International Workshop on Database

- and Expert Systems Applications (DEXA'2004). Zaragoza, Spain, pp. 751–760.
- Thomsen, E., 2002. *Olap Solutions: Building Multidimensional Information Systems*. John Wiley & Sons, Inc., New York, NY, USA.
- Torsten, P., Günther, P., 2003. Ontology-based integration of olap and information retrieval. En: *DEXA '03: Proceedings of the 14th International Workshop on Database and Expert Systems Applications*. IEEE Computer Society, Washington, DC, USA, p. 610.
- Trujillo, J., Song, I.-Y., Apr 2008. New trends in data warehousing and olap. *Decision Support Systems* 45 (1), pp. 1–3.
- Tseng, F. S., Chou, A. Y., 2006. The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. *Decision Support Systems* 42 (2), pp. 727–744.
- Turkka, N., Kalervo, J., Niemi.Timo, 2008. A tool for data cube construction from structurally heterogeneous xml documents. *Journal of the American Society for Information Science and Technology( JASIST)* 59 (3), PP. 435–449.
- Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H., Takeda, K., 2004. A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal* 43 (3), PP. 516–533.
- Vasilakopoulos, A., Bersani, M., Black, W. J., May 2004. A suite of tools for marking up textual data for temporal text mining scenarios. En: *In Proceedings 4th International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon, Portugal, pp. 24–30.
- Vassiliadis, P., Sellis, T., 1999. A survey of logical models for olap databases. *Special Interest Group on Management of Data( SIGMOD) Record* 28 (4), pp. 64–69.
- Weiss, S., Indurkha, N., Zhang, T., Damerou, F., Oct 2004. *Text Mining: Predictive Methods for Analyzing Unstructured Information*, first edition Edición. Springer.

Wolff, C. G., Ago 2002. La tecnología datawarehousing. Ingeniería Informática: La revista electrónica del DIICC.

URL <http://www.inf.udec.cl/revista/ediciones/edicion3/cwolff.PDF>

Yizhou, S., Yintao, Y., Jiawei, H., 2009. Ranking-based clustering of heterogeneous information networks with star network schema. En: In Proceedings of the 15th ACM Special Group on Knowledge Discovery and Data Mining (SIGKDD) international conference on Knowledge discovery and data mining. Association Computing Machinery (ACM) Press, New York, NY, USA, pp. 797–806.

Yu, Y., Lin, C. X., Sun, Y., Chen, C., Han, J., Liao, B., Wu, T., Zhai, C., Zhang, D., Zhao, B., 2009. inextcube: information network-enhanced text cube. Proceedings of the Very Large Data Bases (VLDB) Endowment 2 (2), 1622–1625.

Zhang, D., Zhai, C., Han, J., 2009. Topic cube: Topic modeling for olap on multidimensional text databases. En: SDM. pp. 1123–1134.