Instituto Politécnico Nacional

Centro de Investigación en Computación

Secretaría de Investigación y Posgrado

VISUALIZACIÓN DE LA INFORMACIÓN POR JERARQUÍAS

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A

EL ING. CRUZ ALVAREZ ALFREDO CÉSAR

DIRECTOR DE TESIS:
Dr. GILBERTO MARTINEZ LUNA
Dr. ADOLFO GUZMÁN ARENAS

MÉXICO, D.F. **2011**

SIP-14

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 12:00 horas del día 29 del mes de Junio de 2011 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"VISUALIZACIÓN DE LA INFORMACIÓN POR JERARQUÍAS"

Presentada por el alumno:

CRUZ	**ÁLVAREZ**	**ALFREDO CÉSAR**
Apellido paterno	Apellido materno	Nombre(s)

Con registro: **B 0 9 1 6 4 6**

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de Tesis

Dr. Gilberto Lorenzo Martínez Luna

Dr. Adolfo Guzmán Arenas

Dr. Marco Antonio Moreno Ibarra

M. en C. Sandra Dinora Orantes Jiménez

Dra. Hortensia Gómez Víquez

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México el día 27 del mes Junio del año 2011, el (la) que suscribe Alfredo César Cruz Álvarez alumno (a) del Programa de Maestría en Ciencias de la Computación con número de registro B091646, adscrito a Centro de Investigación en Computación, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de Dr. Gilberto Lorenzo Martínez Luna, Dr. Adolfo Guzmán Arenas y cede los derechos del trabajo intitulado Visualización de la información por jerarquías, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección fredcess23@gmail.com, lluna@cic.ipn.mx . Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Alfredo César Cruz Alvarez

Agradecimientos

"El principio de la sabiduría es el temor de Jehová;
Buen entendimiento tienen todos los que practican
sus mandamientos; Su loor permanece para siempre"
Salmos 111.10

Agradezco a dios por sus bendiciones, por la vida, la salud y por permitirme compartir momentos tan especiales con mi familia y seres queridos.

A mi familia que me ha apoyado en todo momento, por el amor y cuidado que siempre han tenido hacia mí.

Al Instituto Politécnico Nacional y al Centro de Investigación en Computación que me dieron la oportunidad de estudiar la Maestría en Ciencias de la Computación, lo cual fue siempre mi anhelo.

A mis directores de tesis por la dirección de este trabajo, por su constante apoyo y por compartir sus conocimientos.

Al comité tutorial por sus observaciones y aportaciones, las cuales hicieron posible la culminación del presente trabajo.

Resumen

En la actualidad el análisis en las bases de datos se dificulta por diversas razones, ya sea por tratar con los grandes volúmenes que se almacenan o por comprender las estructuras internas definidas en los datos, esto provoca que la búsqueda de anomalías o situaciones de interés en un conjunto de datos sea compleja. Supóngase que en una empresa de ventas de productos *"Se desea saber en qué nivel de la clasificación (jerarquía) de productos se tienen altos niveles de ventas, digamos arriba del 80% con respecto al año anterior"*. Esta consulta de negocio llamada "tendencia con niveles jerárquicos" consiste en localizar aquellos elementos en una jerarquía establecida en los datos que presenten una situación de interés para el analista, esto significa encontrar los elementos relevantes y sus subelementos necesarios para alcanzar los puntos de interés.

En este trabajo se muestra un análisis, diseño e implementación de una herramienta para resolver el tipo de pregunta de negocio planteada, lo cual consiste en la búsqueda de los elementos de interés dentro del árbol de jerarquías de una dimensión de un cubo de datos y posteriormente presentar los resultados en representaciones visuales recomendadas por los expertos de la visualización de la información. Se presentan 3 tipos de representaciones lo cuales son: Mapas de nodos, Mapas de calor y Mapas Pastel Multi-Nivel. Además de tableros de control para analizar las anomalías o puntos de interés en otros niveles de la jerarquía permitiendo así una navegación sobre los distintos niveles de granularidad y sobre otras dimensiones involucradas en el análisis.

Palabras clave: Visualización de la información, Jerarquías, Cubo de datos, Tableros de control.

Clasificación ACM:

H. Information Systems / H.2 Database Management / H.2.8 Database Applications
H. Information Systems / H.5 Information Interfaces and presentation / H.5.2 User Interfaces

Abstract

At present analysis in databases is difficult due to various reasons: either because of large volumes stored, or internal structures in the data set. This causes the search for anomalies or situations of interest in a data set to be complex. Suppose in a product sales company this situation occurs: They want to know what level of classification (hierarchy) of products with high sales levels, say above 80% over the previous year. This consulting business called "trend in hierarchy" is locating those items in a hierarchy in the data submitted by a situation of interest to be analyzed. This means finding the relevant elements and the elements necessary to achieve the landmarks.

This paper presents an analysis, design and implementation of a tool to solve the kind of business question posed. This consists in finding items of interest within the hierarchy tree of data cube dimension and then presents the results in visual representations recommended by information display scientists. There are 3 types of representations which are nodes maps, heat and Pastel Multi-Level maps. In addition to dashboards analyzing anomalies or points of interest in other levels of the hierarchy, which allows navigation on different granularity levels and other dimensions involved in the analysis.

Keywords: Information visualization, hierarchies, data cube, dashboard.

ACM Classification:

H. Information Systems / H.2 Database Management / H.2.8 Database Applications
H. Information Systems / H.5 Information Interfaces and presentation / H.5.2 User Interfaces

Contenido

Índice de figuras

Índice de tablas

Índice de gráficas

1

INTRODUCCIÓN

1 Introducción

1.1 Antecedentes

En la actualidad existen bases de datos de grandes volúmenes como son las bases de datos comerciales, científicas, bancarias, entre otras, que son generadas por múltiples fuentes a partir de sistemas transaccionales, archivos planos, sistemas multimedia e Internet [Feng, 2002], [Han, 2006].

Las organizaciones necesitan herramientas que les permitan explotar el sistema visual humano para extraer información, visualizar tendencias, patrones y relaciones entre los datos; para así poder comprenderlos y llevar a cabo posibles tomas de decisiones, una de las formas de hacer lo anterior, es lo que se conoce como visualización de la información [Chen, 2010], [Hanrahan, 2009], [Tegarden, 1999].

A través de la visualización de la información, no solo se busca crear gráficas de estructuras de información complejas, esta incluye actividades de conocimiento, sociales y colaborativas para la comprensión clara y ágil de esta, proveniente de grandes cantidades de datos [Chen, 2006].

Los avances tecnológicos en hardware y software permiten ahora explotar las habilidades humanas visuales-espaciales para resolver problemas analizando los grandes volúmenes de datos de las organizaciones. Además de que la visualización de la información ha influenciado a los actuales estándares de los navegadores Web.

1.2 Planteamiento del problema

La gran cantidad de volúmenes de datos, sus estructuras internas y las definidas por el analista de negocio dificultan la revisión o análisis de los elementos de interés como son ventas, unidades vendidas, pacientes atendidos, ingreso y egreso de estudiantes, entre otros indicadores también de interés. Una forma de declarar o plantear los análisis deseados a resolver, es por medio de consultas o preguntas de negocio, como por ejemplo, “En una empresa de venta de productos se desea saber en qué nivel de la jerarquía de productos se tienen bajos niveles de ventas, digamos entre un 20% y 30% con respecto al año anterior”. Este tipo consulta de negocio ha sido planteada en artículos y tesis relacionados al análisis de datos como son [Guzmán, 2008], [Agrawal, 1998], [Martínez, 2007].

Esta consulta puede presentarse en cualquier dominio de datos como son comercial, científico, bancario, entre otros. Por ejemplo en un dominio científico es necesario contar con herramientas que permitan revisar continuamente el estatus de la producción científica para poder realizar una evaluación. Evaluación en cuanto a los elementos que participan y ayudan a su generación, como los investigadores, áreas de conocimiento, artículos, conferencias, revistas, tesis, entre otros. Un ejemplo de una consulta de negocio en un dominio científico puede ser “saber en qué nivel de la clasificación ACM de tesis de una institución académica existe un incremento en número en 2 años determinados”.

Como resultado al tipo de pregunta planteada (en cualquier dominio), se desea que los datos sean presentados en forma visual, por medio de mapas de situaciones de interés o tableros de control que permitan evaluar el desempeño y así poder decidir sobre las políticas de apoyo según el dominio de los datos.

1.3 Objetivos

1.3.1 Objetivo general

- Construir un prototipo Visualizador de la información, usando mapas de anomalías por medio de jerarquías, el cual permita visualizar y analizar situaciones de interés en los datos.

1.3.2 Objetivos particulares

- Diseñar y definir los indicadores y la forma de presentación en un tablero, utilizando la jerarquía de la información.

- Demostrar que es posible ampliar el sistema de visualización de la información a otras aéreas.

- Analizar y plantear un proceso de solución través de Dashboards u otras herramientas de visualización.

1.4 Justificación

El análisis de grandes volúmenes de datos dificulta el proceso del descubrimiento del conocimiento. En la actualidad existen herramientas del tipo OLAP que optimizan el acceso a grandes volúmenes de datos, herramientas de visualización de información que buscan agilizar la detección del conocimiento y algoritmos que permitan navegar en las estructuras virtuales internas en el dominio de los datos como son las jerarquías.

El presente trabajo muestra una forma de cómo aprovechar las herramientas OLAP y de visualización, junto con algoritmos de minería de datos para demostrar que el descubrimiento del conocimiento se vuelve una tarea ágil e intuitiva, tal y como se demuestra en [Han, 2006], [Chen, 2010].

1.5 Beneficios esperados

Lograr un nivel de automatización en el proceso del descubrimiento del conocimiento de situaciones de interés planteadas, a través de herramientas OLAP, herramientas de visualización y algoritmos de minería de datos que faciliten la presentación de la información, visualizaciones como son:

- ✓ Mapas de situaciones de interés

  Es la visualización de anomalías o puntos de interés dentro de una estructura jerárquica, indicando la ruta a recorrer de cada situación de interés. Se hace uso de formas, colores, texturas y posiciones aprovechando el sistema de percepción humano.

- ✓ Dashboard (tableros de control)

  Permiten navegar de manera detallada dentro de los puntos de interés mostrando un análisis a partir de las dimensiones definidas en el cubo de datos.

- ✓ Indicadores

  Permiten visualizar por medio de formas, colores y texturas, los hechos o medidas definidos en el cubo de datos que se este analizando.

1.6 Alcances y límites

El proyecto tiene como alcance resolver los escenarios del tipo de pregunta de negocio "Tendencias en niveles jerárquicos" la cual ha sido planteada en artículos y tesis relacionados al análisis de datos como son [Guzmán, 2008], [Agrawal, 1998], [Martínez, 2007]. Presentando los resultados por medio de visualizaciones recomendadas en artículos y congresos como InfoVis [Chignell, 2005], [Schulz, 2006]. Además de lograr un nivel de generalización del prototipo visualizador de situaciones de interés por medio de jerarquías, lo que significa tener la capacidad de trabajar en cualquier dominio de datos, sea comercial, científico u otros.

2

Marco teórico y Estado del arte

2 Marco Teórico y Estado del arte

Con la inundación de datos producidos por los sistemas de información de hoy en día, algo debe ser hecho para poder extraer conocimiento y/o tomar decisiones, extrayendo el contenido de los datos [Rozeva, 2007]. Los recientes avances en tecnologías de la visualización proveen la capacidad de usar las habilidades visuales humanas para resolver los problemas abstractos. Si un problema que involucra más de 2 variables o dimensiones puede ser visualizado con una representación apropiada, entonces puede ser posible usar las habilidades visuales que nos permitan tomar decisiones [Hanrahan, 2009]. Es posible aprovechar los beneficios de la visualización sobre almacenes de datos, constituidos por cubos de datos, combinando operaciones de navegación OLAP y algoritmos de minería de datos.
Por esta razón se presentan algunos conceptos necesarios del campo de minería de datos en el presente capitulo. Además de los conceptos básicos de la visualización de la información como son la historia, principios y características.

2.1 Visualización

De acuerdo al diccionario de gráficos y realidad virtual, la visualización es el proceso de representar datos de una manera visual [Tegarden, 1999], [Chen, 2006].

Los datos pueden representar objetos concretos, tales como cuartos o carros, o los datos pueden también representar objetos abstractos, tales como ventas o costos. Si el dato es abstracto, entonces una visualización análoga debe ser creada. Una visualización análoga típica es una gráfica de pastel o una gráfica de barras.

La visualización permite:

- Explotar el sistema visual humano para extraer información de los datos.
- Provee una visión general de la complejidad de los conjuntos de datos.
- Identifica estructuras, patrones, tendencias, anomalías y relaciones entre datos.
- Ayuda a identificar las áreas de interés.

En otras palabras, la visualización permite realizar la toma de decisiones usando las habilidades visuales naturales que poseemos.

Las tecnologías de visualización se pueden clasificar en 3 clases generales: Visualización científica, visualización de la información/datos y realidad virtual [Chen, 2010].

Visualización científica, como el nombre indica, es la transformación de los datos producidos a través de cálculos científicos, de ingeniería o experimentos con imágenes, ejemplo: Checkerboard que presenta las diferencias en las medidas de los cuboides de un cubo [Chen, 2009].

Visualización de la información/datos se refiere a la transformación de datos no-espaciales dentro de imágenes visuales que representan una analogía, ejemplo: Uso de deslizadores dinámicos en mapas geográficos [Bao, 2003].

En el contexto de visualización de la información de negocio, la **realidad virtual (VR)** es simplemente un 3D, cálculos generados simulando el ambiente que se crea en tiempo real de acuerdo al comportamiento del usuario. VR también ha sido referida como realidad artificial, ciber-espacial y ambiente virtual, ejemplo: "LandMarks" en la Web [Chen, 2006].

Cabe mencionar que la visualización tiene también una relación con la representación de mapas geográficos, los cuales representan la información físico natural y que se rigen por métodos cartográficos. En la actualidad han surgido los sistemas de información geográfica (SIG) que dan inicio al estudio de la información espacial.

2.1.1 Historia de la visualización

La visualización no es nueva. Por ejemplo hace más de 20 000 años, en Francia fueron realizados dibujos en cavernas, además de que los chinos crearon el primer mapa conocido en el siglo XII. Sin embargo la primera representación multidimensional no apareció hasta el siglo XIX.

Dos de los mejores ejemplos fueron creados por el Dr. John Snow y Charles Joseph Minard, en 1854 cuando el Dr. Snow realizó un diagrama ubicando las muertes por cólera en el centro de Londres (Figura 2.1), observó que el cólera ocurrió en medio de esos que vivían cerca de "The Broad Street". Cada muerte fue mostrada como un punto sobre el mapa de London y fue la concentración de puntos la que reveló la conexión oculta entre las muertes y los pozos contaminados. Basado en esta observación, el Dr. Snow tenía el indicador y decidió cerrar los pozos contaminados, con lo cual terminó con la epidemia del cólera [Chen, 2010].

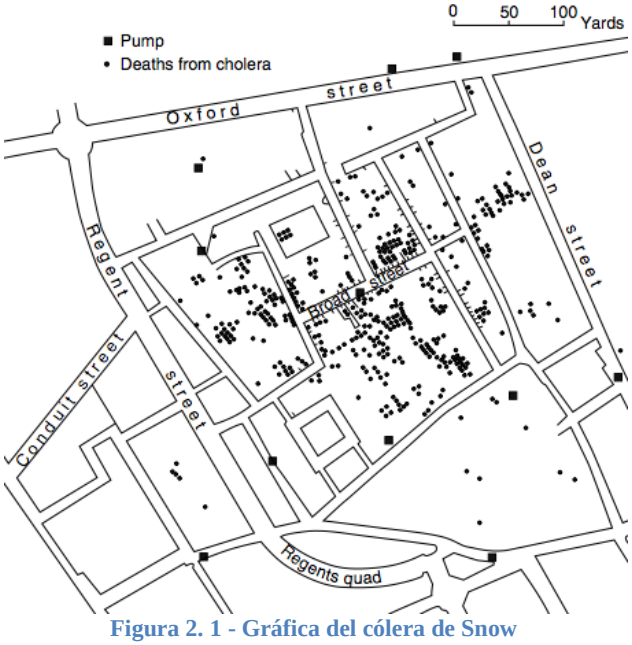


Figura 2. 1 - Gráfica del cólera de Snow

En 1861, Minard creó posiblemente la primera gráfica estadística alguna vez dibujada. Se trata de un ejemplo clásico que muestra un mapa que revela claramente las pérdidas de la armada de Napoleón en 1812 durante la invasión a Moscú. El tamaño de la armada es mostrado como el ancho de la banda en el mapa, comenzando en las orillas Ruso-Polacas con 422,000 hombres. En el momento en que ellos llegaron a "Moscú" en septiembre, el tamaño de la armada se redujo a 100,000. Eventualmente, solo una pequeña fracción de la armada original de Napoleón sobrevivió [Chen, 2010].

El mapa ayuda a revelar las causas que provocaron la disminución del ejército, por ejemplo la línea superior de color café representa la invasión, la banda obscura la retirada y la línea inferior roja representa la temperatura. Se concluye que la constante caída de temperatura fue la mayor causa de la disminución de la armada, así es como se determina la relación "temperatura – tamaño" de la armada, lo cual significa que mientras temperaturas mas bajas se presentaban, el tamaño de la armada "disminuía mas" [Chen, 2010], [Tegarden, 1999].

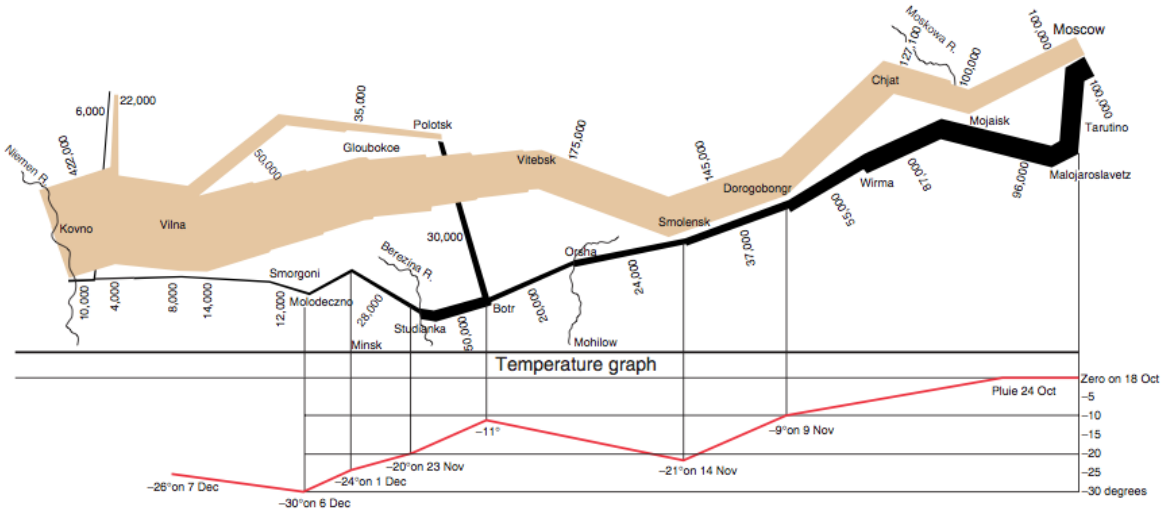


Figura 2. 2 - Pérdidas sufridas de Napoleón durante su invasión a Rusia en 1812

2.2 Visualización de la información

El término de visualización de la información se refiere al diseño, desarrollo y aplicación de cálculos, y representaciones gráficas interactivas de información.

El campo de la visualización de la información involucra a la comunidad científica de investigadores quienes están o han contribuido al campo de estudio. La visualización de la información a menudo implica tratar con datos abstractos o espaciales. La transformación de esos datos a representaciones gráficas intuitivas y significativas es fundamental en el campo. La transformación es por lo tanto un proceso creativo en el cual los diseñadores asignan nuevos significados a los elementos gráficos [Chen, 2010], [Tegarden, 1999].

Es posible modelar el proceso de la visualización de la información en términos de la transformación de los datos, transformación de la visualización y transformación del mapeo visual.
La transformación de los datos convierte los datos en bruto a formas matemáticas. La transformación de la visualización establece un modelo visual-espacial de los datos.
La transformación del mapeo visual determinar la apariencia del modelo visual-espacial al usuario.

El arte y la funcionalidad son parte integral de la visualización de la información. Los investigadores y artistas han intentado derivar criterios que nos digan cuando la visualización de la información es arte, cuando no lo es y cuando es ambas. Las discusiones de estética son inevitables en el diseño de estudios y búsqueda de fundamentos teóricos de la visualización de la información.

Como arte, la visualización de la información pretende comunicar ideas complejas a su audiencia e inspirar a sus usuarios nuevas conexiones. Como ciencia, la visualización de la información debe presentar información y patrones asociados rigurosamente y con exactitud.

La conexión entre aspectos científicos y artísticos de la visualización de la información es discutida en términos de: *visualización de la información funcional y visualización de la información estética.* El principal rol de la visualización de la información funcional es comunicar un mensaje al usuario, mientras que la meta de la visualización de la información estética es presentar una impresión subjetiva de un conjunto de datos provocando una respuesta emotiva del usuario.

2.2.1 Análisis visual

El análisis visual es un campo emergente que se origina de la visualización de la información, tiene como objetivo apoyar el razonamiento analítico y toma de decisiones a través del uso de la visualización de la información, análisis estadístico, minería de datos y otros campos [Hanrahan, 2009].

Ha llegado a ser la manera más rápida en que las personas exploran y comprenden grandes volúmenes de datos. Las grandes compañías reconocen la necesidad de incrementar los estándares visuales. La visualización de la información está siendo adoptada por grandes compañías, universidades y agencias de gobierno.

Compañías como son: Apple, Pfizer, Microsoft, Coca Cola, Google.
Universidades: Georgetown University, MIT, Harvard University por mencionar las más importantes.

La gente ahora reconoce que el análisis visual de los datos acelera el análisis de negocio y la principal característica de las aplicaciones de análisis visual es que las aplicaciones unifican los pasos de consultas a las bases de datos, explorando y visualizando al mismo tiempo [Hanrahan, 2009].

Suponiendo que el usuario no tiene una pregunta específica hacia el sistema visualizador, pero navegando sobre este, se logra enterar de aspectos relevantes de su negocio. Esto significa que la visualización en una aplicación de análisis visual permite a las personas detenerse y analizar la información presentada, de esta manera una aplicación de análisis visual ayuda a las personas a realizar un análisis y evaluación.

2.2.2 Puntos de vista en la visualización

En años anteriores de la visualización de la información, era creído que la habilidad de ver la totalidad de un conjunto de datos era importante para descubrir interesantes conexiones ocultas y otros patrones. Más recientemente se descubrió que con el aumento del **análisis visual**, la visualización de la información hace énfasis también al proceso de búsqueda de puntos de vista (insights) [Chen, 2010].

Los investigadores de la visualización de la información han discutido en *¿Cómo medir el grado de interés en los puntos de vista?* A diferencia de los investigadores de minería de datos sobre el grado de interés que pretende desarrollar métricas y algoritmos para determinar el grado de interés en un conjunto de datos dado, pocas métricas han sido desarrolladas en el campo de la visualización.

2.2.3 Evaluación de puntos de vista (Insights)

La definición de puntos de vista en la visualización de la información en la literatura ha sido vaga y ambigua, pocas conexiones han sido establecidas entre el estudio de los puntos de vista en otras disciplinas y el campo de la visualización de la información.

En la comunidad de la visualización se ha planteado *¿Cómo establecer la efectividad de interacción con las interfaces de visualización de la información?*

Notables esfuerzos sobre la caracterización y medición de los puntos de vista (insights) incluyen enfoques exploratorios, por ejemplo un interesante framework de evaluación de visualizaciones interactivas ha sido propuesto recientemente nombrado *"Don Norman`s Seven Stages of Action"* [Norman, 1990] del cual 2 escenarios con interfaces de computadora son particularmente problemáticas: ejecución y evaluación.

La ejecución debe ser reducida de modo que los usuarios puedan lograr sus tareas sin problemas, mientras que la evaluación debe ser reducida a que los usuarios puedan juzgar su progreso con precisión.

Teniendo en cuenta los recientes fundamentos teóricos en el campo, se espera que este campo sea un tema importante de investigación.

2.2.4 Frameworks teóricos

Recientes estudios reportan que la visualización de la información en la actualidad carece de suficientes fundamentos teóricos. En el año del 2007 un seminario tuvo lugar en

Dagstuhl, Alemania cuyo objetivo fue desarrollar nuevas teorías de visualización. Esta ausencia de teorías ha provocado un creciente interés en la comunidad y la búsqueda de estos fundamentos ha introducido y adoptado teorías y conceptos de otros campos y disciplinas.

Muchas visualizaciones de la información carecen de medidas cuantitativas que puedan indicar el resumen de calidad, incertidumbre, novedad y otras métricas. Sin embargo la ejecución y evaluación han tenido un potencial progreso en esta dirección.

2.2.5 Principios de Gestalt

Un espacio de información implica una definición de métricas que midan la distancia en el espacio abstracto. La noción de un espacio abstracto aprovecha la psicología Gestalt la cual da principios de nuestra tendencia a ver patrones de elementos. Los principios de Gestalt son proximidad, similaridad, continuidad, figura-fondo y simetría [Chen, 2010].

Proximidad

El principio de proximidad dice que tendemos a ver agrupaciones de elementos en una estructura visual basada en la proximidad entre esos elementos. Los elementos que están relativamente cerca tienden a darnos un sentido de similaridad. Este principio ha sido adaptado por la comunidad de la visualización de la información [Chen, 2006].

Similaridad

Por otro lado el principio de similaridad de la psicología de Gestalt dice que los atributos visuales tales como: la forma, el color y textura son señales para nosotros de cómo agrupar elementos [Chen, 2006].

Continuidad

El principio de continuidad menciona que los detalles que mantienen un patrón o dirección tienden a agruparse juntos, como parte de un modelo. Es decir, percibir elementos continuos aunque interrumpidos entre si [Chen, 2006].

Figura - fondo

Establece el hecho de que el cerebro no puede interpretar un objeto como figura o fondo al mismo tiempo. Depende de la percepción del objeto será la imagen a observar [Chen, 2006].

Simetría

El principio de simetría dice que las imágenes simétricas son percibidas como un solo elemento, en la distancia [Chen, 2006].

2.3 OLAP (On-Line Analytical Processing)

El termino OLAP se refiere al procesamiento analítico en línea (On-Line Analytical Processing) y tiene como objetivo consultar de manera rápida y eficaz grandes cantidades de datos [MansmannSvetlana, 2006]. Forma parte de las soluciones de Business Intelligence.

OLAP hace uso de cubos de datos para manipular y visualizar la información. Esta tecnología involucra herramientas de reporteo y graficado que permiten desplegar datos agregados, permitiendo detectar variables de interés y de esta manera navegar en los niveles de las dimensiones, el conjunto de niveles forma una estructura llamada jerarquía en la dimensión, esta jerarquía permite consultar los datos de forma general o especializada [Han, 2006].

2.3.1 Separación de sistemas OLAP y OLTP

La razón por la cual los almacenes de datos o sistemas OLAP suelen separase de las bases de datos operacionales OLTP (On-Line Transactional Processing), es para mejorar el desempeño de ambos sistemas.

Una base de datos operacional está diseñada en base a indexamiento y mapeos usando llaves primarias, buscando registros particulares. De otra manera las consultas en los almacenes de datos son complejas, involucrando el cálculo de grandes cantidades de datos en distintos niveles. Lo cual indica que si ejecutamos consultas OLAP en bases de datos operacionales podría degradar el desempeño de las tareas operacionales [Han, 2006].

Los dos sistemas ofrecen distintas funcionalidades y requieren diferentes tipos de datos, es por esto que es necesario mantenerlos separados.

Sin embargo, muchos vendedores de SGBD (sistemas de gestión de base de datos) están empezando a optimizar tales sistemas para soportar consultas OLAP, si esta tendencia continua, la separación entre sistemas OLAP y OLTP decrecerá.

2.4 Modelos de datos multidimensionales

Los almacenes de datos y las herramientas OLAP están basados en un modelo de datos multidimensional. Este modelo visualiza los datos en forma de un cubo.

Un cubo de datos permiten a los datos ser modelados y vistos en múltiples dimensiones y está definido por dimensiones y hechos [Han, 2006].

En términos generales, las **dimensiones** son las perspectivas o entidades que una organización quiere mantener registradas, por ejemplo: tiempo, producto, sucursal y localización. Estas dimensiones permiten almacenar los registros de cosas como: ventas mensuales de productos, las sucursales y localizaciones en las cuales los productos fueron

vendidos. Cada dimensión podría tener una tabla asociada, llamada **tabla de dimensión**, en la cual se describe la dimensión [Han, 2006], [Rozeva, 2007], [Vassiliadis, 1998].

Los modelos de datos multidimensionales son organizados alrededor de un tema central, como por ejemplo: ventas. Este tema es representado por una tabla de hechos. Los **hechos** son medidas numéricas, pensamos en ellas como las cantidades por las cuales queremos analizar las relaciones entre las dimensiones.

Un ejemplo de hechos podría ser: ventas o unidades vendidas. La **tabla de hechos** contiene los nombres de los hechos o medidas. Para comprender el modelo multidimensional, comenzamos viendo un cubo de 2 dimensiones, que de hecho es una tabla o una hoja de cálculo de las ventas de una tienda de electrónica (Tabla 2.1).

ubicación = "Vancouver"				
	producto			
tiempo	hogar	computación	telefonía	seguridad
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	80	580

Tabla 2. 1 - Venta de productos por trimestre en la sucursal Vancouver

La tabla 2.1 presenta las ventas de productos por trimestre en la ciudad de Vancouver. Las ventas son mostradas con respecto a la dimensión tiempo y la dimensión producto (organizada de acuerdo al producto vendido). El hecho o medida es desplegado en dólares (miles).

Ahora suponemos que deseamos ver las ventas en 3 dimensiones. Queremos ver los datos de acuerdo al tiempo, producto y ubicación para las ciudades de Chicago, Nueva York, Toronto y Vancouver. Los datos en 3 dimensiones de la tabla 2.2 representan series de tablas en 2 dimensiones. Podemos conceptualizar esta tabla en forma de un cubo de 3 dimensiones (Figura 2.3).

	Ubicación : "Chicago"				Ubicación: "Nueva York"				Ubicación: "Toronto"				Ubicación: "Vancouver"			
	Producto				producto				producto				producto			
tiempo	**hogar**	**comp**	**tel**	**seg**	**hogar**	**comp**	**tel**	**seg**	**hogar**	**comp**	**tel**	**seg**	**hogar**	**comp**	**tel**	**seg**
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Tabla 2. 2 - Venta de productos por trimestre en 4 sucursales

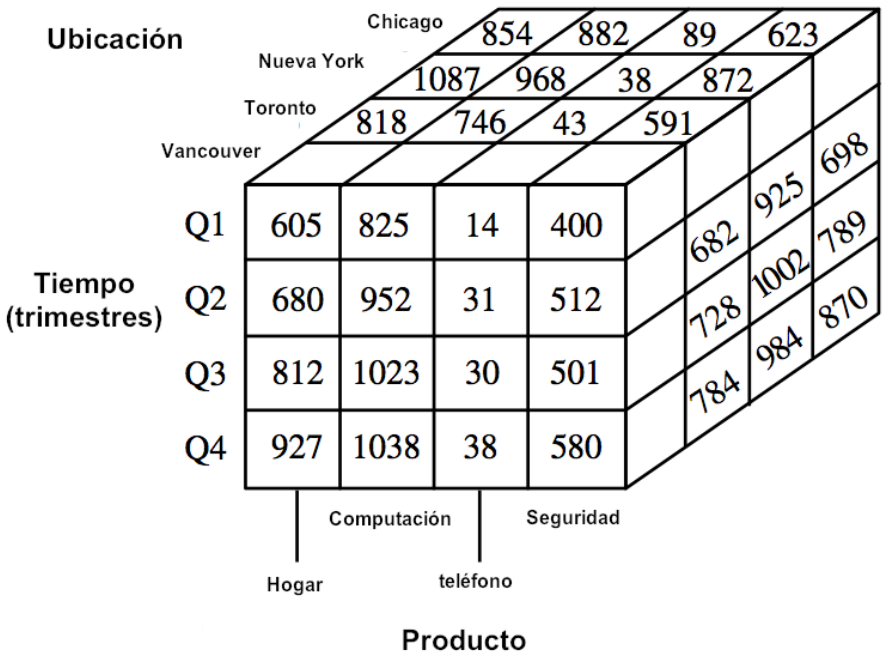


Figura 2. 3 - Cubo de datos con 3 dimensiones

Los cubos de datos presentados son llamados también **cuboides**. Dado un conjunto de dimensiones, podemos generar un cuboide para cada uno de los posibles subconjuntos de las dimensiones dadas. El resultado podría formar una lattice de cuboides, cada una mostrando los datos en diferentes niveles de resumen o group by.
De modo que la lattice de cuboides se refiere a un cubo de datos.
La figura 2.4 muestra una lattice de cuboides formando un cubo de datos para las dimensiones: tiempo, producto, ubicación y proveedor.

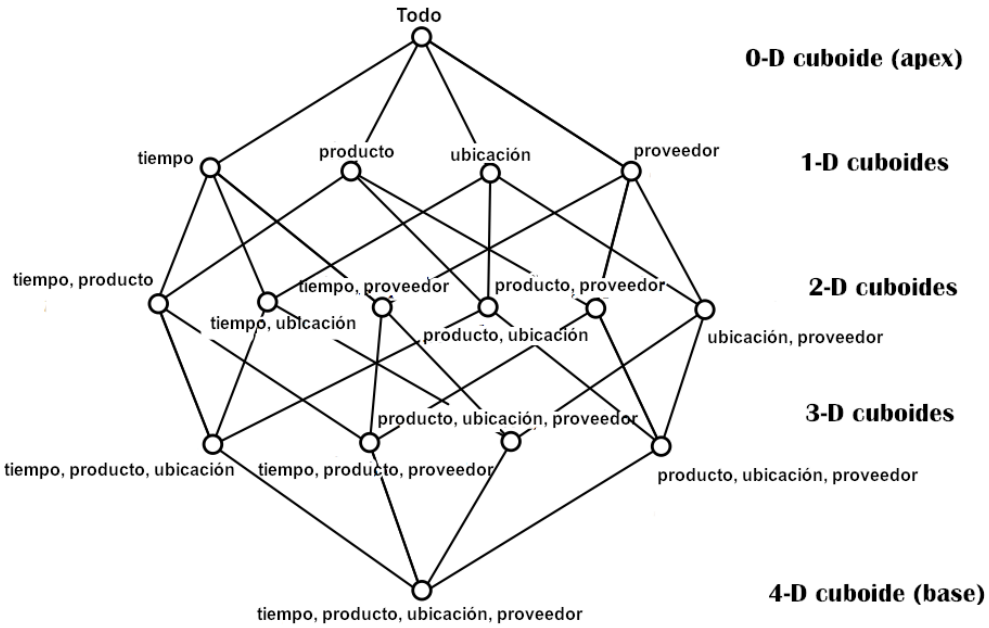


Figura 2. 4 - Lattice de cuboides compuesta por 4 dimensiones

El cuboide que mantiene el nivel más bajo de resumen, es llamado “base cuboid”, mientras que el cuboide de cero dimensiones, que mantiene el nivel más alto de resumen es llamado el “apex cuboid” [Han, 2006].

2.5 Esquemas para bases de datos multidimensionales: estrella, copo de nieve y constelación de hechos

Un almacén de datos requiere un esquema conciso y orientado a temas que facilite el análisis de los datos en línea.

El modelo de datos más popular para un almacén de datos es el modelo multidimensional. Tal modelo puede existir en la forma de: esquema estrella, esquema copo de nieve y esquema constelación de hechos [Han, 2006].

Esquema estrella.- Es el modelo más común en el cual el almacén de datos contiene:
Una gran tabla central **(tabla de hechos)** que contiene la mayoría de los datos, sin redundancia, un conjunto de pequeñas tablas asistiendo a la tabla de hechos **(tablas de dimensiones)**. Una para cada dimensión [Rozeva, 2007], [Vassiliadis, 1998].

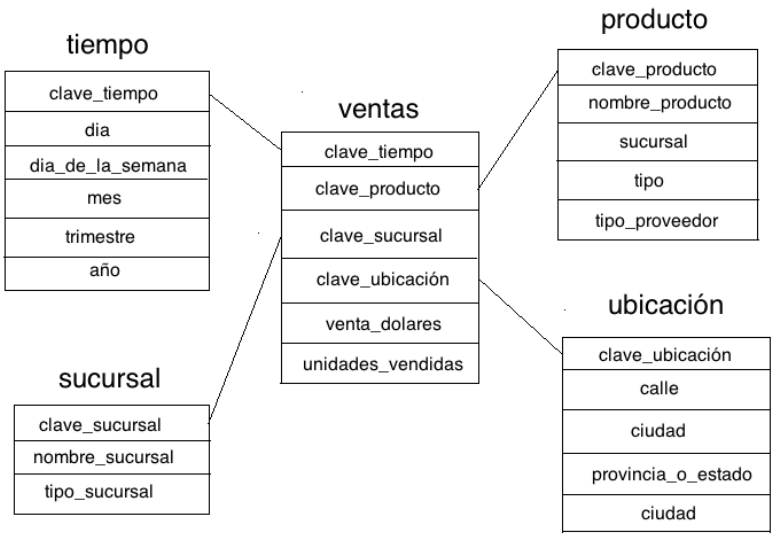


Figura 2. 5 - Esquema estrella

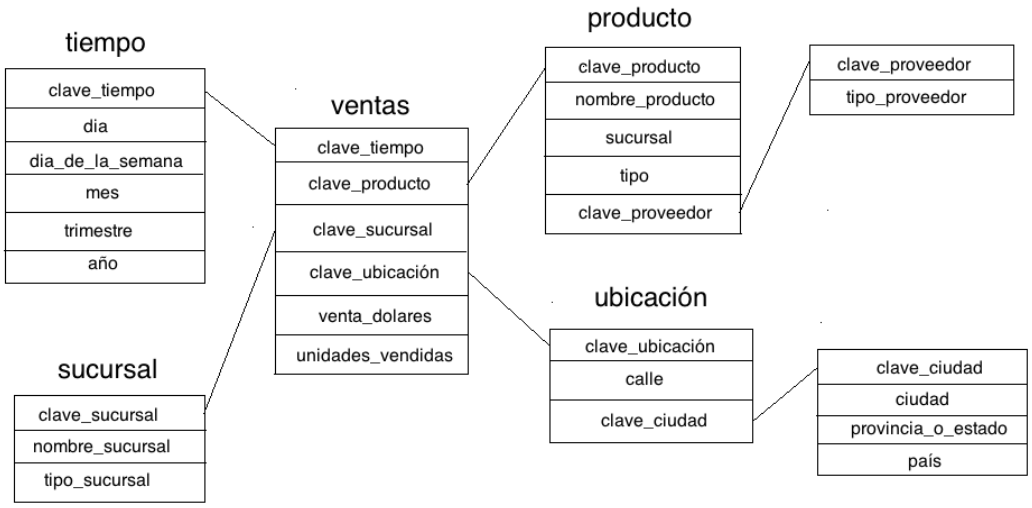


Figura 2. 6 - Esquema copo de nieve (snowflake)

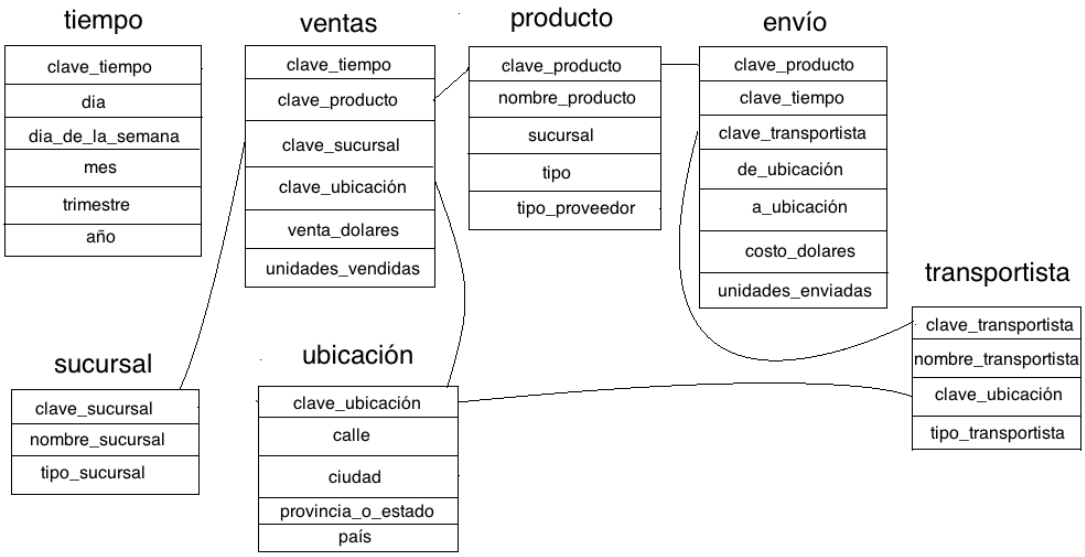


Figura 2. 7 - Esquema Constelación de hechos

2.6 Lenguaje de consulta en minería de datos

Así como el lenguaje de consultas relacional SQL (Structured Query Language) es usado para especificar consultas relacionales. Un lenguaje de consulta en minería de datos DMQL (Data Mining Query Language) puede ser especificado para tareas de minería de datos. Los almacenes de datos (Data Warehouse) y los Data Marts pueden ser definidos usando 2 lenguajes, uno para la definición de cubos y otro para la definición de dimensiones [Han, 2006].

La definición de cubos tiene la siguiente sintaxis:

define cube <nombre_cubo> [<lista_dimensiones>]: <lista_medidas>

La definición de las dimensiones tiene la siguiente sintaxis:

define dimensión <nombre_dimension> as (<lista_ atributos>)

2.7 Conceptos de jerarquía

El concepto de jerarquía define una secuencia de mapeos desde un conjunto de conceptos de bajo nivel hasta un nivel de concepto más alto (más general) [Chignell, 2005], [Malinowski, 2006].

Si consideramos el concepto de jerarquía en la dimensión "ubicación", los valores para ciudad incluyen: Vancouver, Toronto, Nueva York o Chicago. Estos datos pueden ser mapeados a provincia o estado al cual pertenecen. Por ejemplo Vancouver puede ser mapeado a Colombia Británica y Chicago a Illinois. Mientras que la provincia o estado pueden ser mapeados al país al cual pertenecen, Canadá y EUA.

Estos mapeos forman una jerarquía conceptual para la dimensión "ubicación" (Figura 2.8).

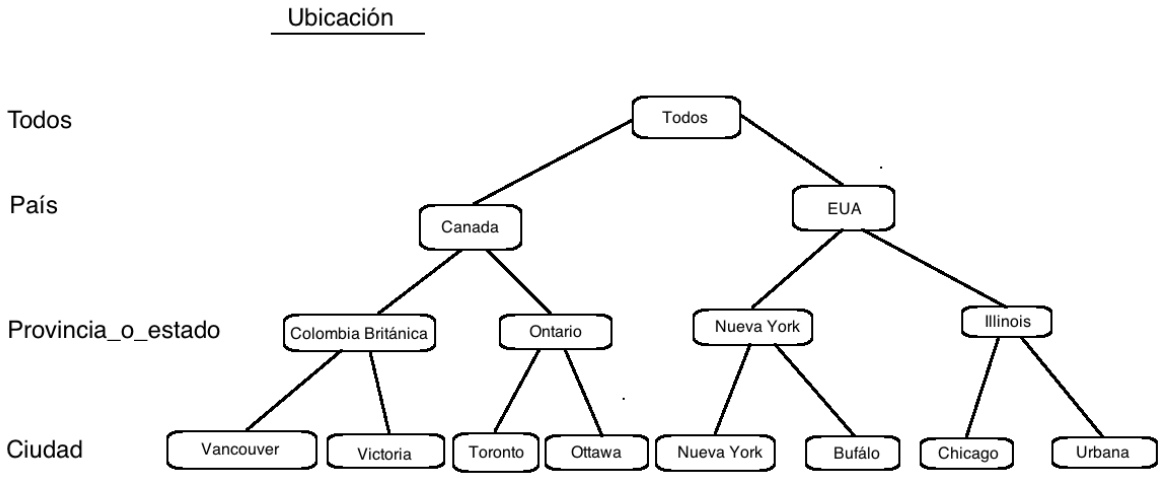


Figura 2. 8 - Jerarquía de la dimensión "ubicación"

Muchas jerarquías están implícitas en los esquemas de base de datos. Por ejemplo si suponemos que la dimensión "ubicación" esta descrita por los atributos: calle, ciudad, provincia_o_estado y país. El orden de la jerarquía es:
"calle < ciudad < provincia_o_estado < país".

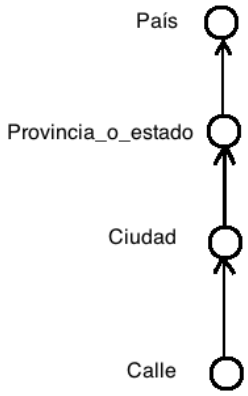


Figura 2. 9 - Niveles de la jerarquía de la dimensión "ubicación"

Alternativamente los atributos de una dimensión podrían estar organizados en un orden parcial, formando una lattice. Un ejemplo de un orden parcial para la dimensión tiempo es: día, semana, mes, trimestre y año, día< {mes<trimestre; semana} < año".

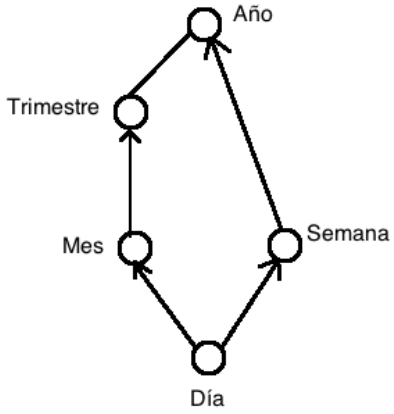


Figura 2. 10 - Lattice de la dimensión "tiempo"

La jerarquía que es de orden parcial o total entre los atributos en el esquema de la base de datos es llamado **esquema de jerarquía**.

Los sistemas de minería de datos deben proveer al usuario la flexibilidad de predefinir jerarquías de acuerdo a sus necesidades particulares.

La jerarquía también puede definirse agrupando o discretizando valores de una dimensión o atributo determinado, lo que resulta en una **jerarquía de agrupación**. Un ejemplo de jerarquía de agrupación es mostrada en la figura 2.11 para la dimensión precio, donde un intervalo ($X…$Y] denota un rango.

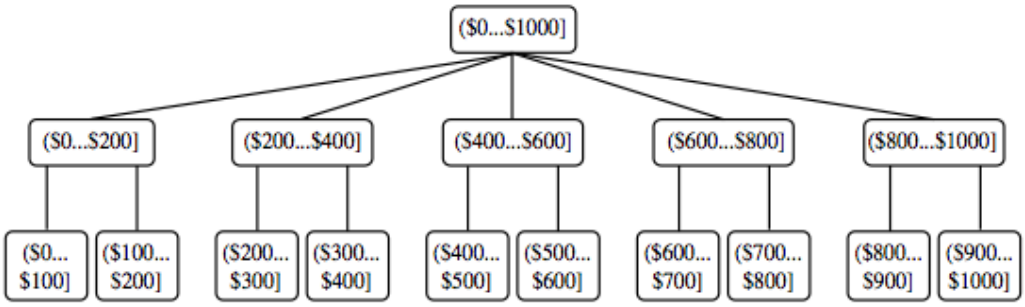


Figura 2. 11 - Jerarquía para el atributo precio

Podría haber más de una jerarquía para un atributo o dimensión, basado en diferentes puntos de vista. Más de un concepto de jerarquía puede ser definido para el mismo atributo de acuerdo a las necesidades de distintos usuarios.

Las jerarquías pueden ser proporcionadas de forma manual por los usuarios del sistema, los expertos del dominio, o ingenieros del conocimiento, o pueden ser generar automáticamente basadas en un análisis estadístico de la distribución de los datos.

2.7.1 Discretización de datos y generación de jerarquías

Las técnicas de discretización de datos pueden ser usadas para reducir el número de valores para un atributo continuo dado, dividiendo el rango del atributo dentro de intervalos. La etiqueta del intervalo puede ser usada para remplazar valores de datos. Remplazando numerosos valores de un atributo continuo por un número pequeño de etiqueta de intervalos, de tal modo se reduce y simplifica los datos originales.
Así la representación de conocimiento se vuelve concisa y fácil de comprender.

Las técnicas de discretización se clasifican en como la discretización es ejecutada, por ejemplo, si se utiliza la información de clase o en qué dirección se procede (top-down, bottom-up). Si el proceso de discretización usa la información de clase, entonces se dice que es una discretización supervisada. De otra forma es no supervisada. Si el proceso comienza encontrando uno o pocos puntos para dividir el rango del atributo (splits points), y repite esto recursivamente sobre el intervalo de resultados es llamado **discretización top-down o splitting**.

La discretización bottom-up o merging comienza considerando todos los valores continuos como splits-points, uniendo los valores vecinos para formar intervalos, este proceso es aplicado recursivamente a los intervalos resultantes.

La discretizacion puede ser ejecutada recursivamente sobre un atributo para proporcionar una jerarquía o una partición multi-resolución de los valores de los atributos.

Las jerarquías son útiles para minar en varios niveles de abstracción. Son usadas para reducir los datos, remplazando conceptos de bajo nivel (valores numéricos como edad) con conceptos de alto nivel (tales como joven, mediana edad, mayor).

Además, la minería en un conjunto reducido de datos requiere menos operaciones de entrada - salida y es más eficiente que la minería en un conjunto mayor.

Debido a estos beneficios, las técnicas de discretización y conceptos de jerarquías son aplicados antes de la minería de datos, como un pre procesamiento, en lugar de ser aplicado en la etapa de minado.

La definición manual de jerarquías puede ser una tarea tediosa y tardada para los usuarios. Afortunadamente varios métodos de discretización pueden ser usados para generar automáticamente jerarquías en atributos numéricos. Además muchas jerarquías para atributos categóricos están implícitas dentro del esquema de la base de datos y pueden ser automáticamente definidas en el nivel de definición del esquema.

2.8 Herramientas OLAP

Las herramientas OLAP (Procesamiento analítico en línea) se dividen en 2 tipos: servidores OLAP (motores) y clientes OLAP. Los servidores OLAP ofrecen los servicios de creación y administración del modelo de datos multidimensional, además de permitir el acceso e iteración a los mismos, mientras que el cliente OLAP es una interfaz que permite visualizar grandes cantidades de datos y navegar sobre ellos de forma interactiva y fácil.

Mondrian

Mondrian es un motor ROLAP (Relational On-Line Analytical Processing) desarrollado en Java, que permite analizar grandes conjuntos de datos que se encuentran almacenados en un almacén de datos. Mondrian se considera un motor porque que se encarga de recibir consultas dimensionales en lenguaje MDX (MultiDimensional eXpressions) y entregar los datos del cubo que correspondan a la consulta. El cubo se representa como un conjunto de metadatos que definen cómo se han de mapear estas consultas dimensionales a sentencias SQL para obtener de la base de datos la información necesaria para satisfacer la consulta dimensional. Utiliza una memoria caché para almacenar los resultados de las consultas que se acceden múltiples veces.

Mondrian es usado para:

- ✓Alto desempeño, análisis interactivo de grandes o pequeños volúmenes de información.
- ✓Exploración dimensional de los datos, por ejemplo analizando ventas por marcas de productos, región o periodo de tiempo.
- ✓Análisis de expresiones en lenguaje MDX a expresiones en SQL para recuperar respuestas a consultas dimensionales.
- ✓Cálculos avanzados utilizando las expresiones de cálculo del lenguaje MDX

Jedox PALO

Jedox Palo es un servidor de bases de datos multidimensional capaz de centralizar y administrar casi un número infinito de hojas de cálculo. El sistema opera en tiempo real, soporta la consolidación de jerarquías así como numerosas funciones de inteligencia empresarial y es un servidor de código abierto. Palo es un servidor de datos multidimensional MOLAP (Multidimensional On-Line Analytical Processing) orientado a celdas, específicamente desarrollado para almacenamiento y análisis de datos en hojas de cálculo.

Comparación entre servidores OLAP

La siguiente tabla compara las características principales de 3 servidores OLAP (Online analytical processing).

Servidor OLAP	Compañía	Ultima version	Licencia
Mondrian OLAP	Pentaho	3.2	Eclipse Public License
Palo	Jedox	3.1	General Public License
Oracle Database OLAP option	Oracle	11g R2	Propiedad

Tabla 2. 3 - Información general de servidores OLAP

Servidor OLAP	MOLAP	ROLAP	HOLAP
Mondrian OLAP	NO	SI	NO
Palo	SI	NO	NO
Oracle Database OLAP option	SI	SI	SI

Tabla 2. 4 - Modo de almacenamiento de datos en servidores OLAP

Servidor OLAP	XML for Analysis	OLE DB for OLAP	MDX	SQL
Mondrian OLAP	SI	SI	SI	NO
Palo	SI	SI	SI	NO
Oracle Database OLAP option	NO	SI	SI	SI

Tabla 2. 5 - API y lenguaje de consulta de servidores OLAP

Servidor OLAP	Windows	Linux	OS
Mondrian OLAP	SI	SI	SI
Palo	SI	SI	NO
Oracle Database OLAP option	SI	SI	SI

Tabla 2. 6 - Sistemas operativos compatibles con servidores OLAP

2.9 Estado del Arte

En este tema se exponen algunos trabajos de investigación previos relacionados a la exploración de cubos de datos, búsqueda de anomalías, búsqueda de diferencias entre cuboides y representaciones visuales de los resultados. Estas investigaciones han sido desarrolladas y publicadas por la comunidad científica en el área de minería de datos y de visualización. Su estudio es conveniente debido a que se tratan de herramientas y técnicas previas a las que se pretenden desarrollar en el presente trabajo de tesis.

2.9.1 Exploración de cubos OLAP usando un descubrimiento impulsado

Los analistas usan OLAP para identificar regiones de anomalías que representan problemas en áreas o bien nuevas oportunidades. Usando las operaciones OLAP los analistas navegan a través de un enorme espacio de búsqueda buscando excepciones. En esta investigación se propone un nuevo paradigma de exploración del descubrimiento que mine los datos para cada excepción y resuma la excepción en niveles apropiados por adelantado. Entonces se usan las excepciones para llevar al analista a regiones interesantes del cubo durante la navegación a través de un enfoque estadístico y haciendo el proceso eficiente sobre bases de datos multidimensionales.

Este trabajo presenta técnicas de cálculo que hacen el proceso de encontrar excepciones en grandes conjuntos de datos la cual usa operaciones OLAP y habilita una rutina de pre cálculos de agregados para encontrar las excepciones. Esta técnica reconoce que los datos podrían ser muy grandes para almacenarlos en memoria y los resultados inmediatos podrían ser escritos en disco requiriendo una optimización cuidadosa.

Exploración dirigida

Para la búsqueda de anomalías, la exploración típica comienza en el nivel más alto de la jerarquía de la dimensión del cubo, usando secuencias de drill-down. Desde el nivel más alto de la jerarquía el analista excava a niveles más bajos de la jerarquía buscando valores y visualmente identificando los valores de interés.
Si una exploración a lo largo de un camino (path) no lleva a un resultado interesante entonces es necesario realizar una operación roll-up al path y comenzar investigando otra rama.
Esta exploración conducida tiene varias deficiencias una de ellas es el espacio de búsqueda cuando es muy grande.

Exploración descubrimiento dirigido

La búsqueda de anomalías es guiada por indicadores precalculados de excepciones en varios niveles de detalle. Esto incrementa las oportunidades de notar patrones anormales en los datos en cualquier nivel de agregación. Se considera un valor en una celda del cubo como una excepción si es significativamente diferente del valor anticipado basado en el modelo estadístico.

Por ejemplo un gran incremento en ventas en diciembre (Dec) podría parecer excepcional cuando miremos en la dimensión tiempo (time), pero cuando miremos en otras dimensiones como producto (item) este incremento no parecería excepcional si otros productos también tienen incrementos similares (Tabla 2.7).

Sum of sales	*Month*											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Total		1%	−1%	0%	1%	3%	−1%	−9%	−1%	2%	−4%	3%

Tabla 2. 7 – Ventas mensuales totales de productos.

Los cálculos de los indicadores de excepción son llevados a cabo junto con la construcción del cubo de datos. Tres medidas son usadas como indicadores de excepción para ayudar a identificar anomalías en los datos. Esas medidas indican el grado de sorpresa que la cantidad en la celda mantiene, el valor de sorpresa indica cuan anómalo es una cantidad en una celda con respecto a otras celdas y está compuesto de 3 valores.

1.- SelfExp: Representa la sorpresa de una celda relativa a otra celda en el mismo nivel de agregación.
2.- InExp: Representa el grado de sorpresa en algún lugar bajo esta celda si se excava desde una celda.
3.- PathExp: Representa el grado de sorpresa para cada ruta desde una celda.

Ejemplo:

Supongamos que se desea analizar las ventas mensuales de una empresa como diferencias en porcentajes de meses previos. Las dimensiones involucradas son “producto (item)” y “región (region)”. Los indicadores de excepción se traducen como señales visuales, el color de fondo en cada celda esta basado en su SelfExp, el color y espesor del rectángulo alrededor de cada celda esta en función de su valor InExp.
Por ejemplo los indicadores en las celdas de julio (Jul), agosto (Aug) y septiembre (Sep) señalan al usuario explorar a niveles de agregación más bajos por medio de operaciones Drill-Down.
Por otro lado para saber que dimensión presenta más excepciones se colorea cada dimensión basada en su valor PathExp.

La tabla 2.8 muestra las ventas en el tiempo (time) para cada producto (item). Considerando la diferencia en ventas del 41% para “Sony b/w printers” en septiembre (Sep). Esta celda tiene un fondo obscuro, indicando un alto valor de SelfExp lo que significa que la celda es una excepción. Considerando ahora la diferencia de ventas del -15% para “Sony b/w printers” en noviembre (Nov) y del -11% en diciembre (Dec). El -11% para diciembre es marcado como una excepción mientras que el -15% no lo es, a pesar de que -15% es una desviación más grande que -11%. Esto es debido a que la mayoría de los productos en diciembre tienen un valor positivo mientras que las ventas en noviembre no.

Avg. sales	*Month*											
Item	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Sony b/w printer		9%	−8%	2%	−5%	14%	−4%	0%	41%	−13%	−15%	−11%
Sony color printer		0%	0%	3%	2%	4%	−10%	−13%	0%	4%	−6%	4%
HP b/w printer		−2%	1%	2%	3%	8%	0%	−12%	−9%	3%	−3%	6%
HP color printer		0%	0%	−2%	1%	0%	−1%	−7%	−2%	1%	−4%	1%
IBM desktop computer		1%	−2%	−1%	−1%	3%	3%	−10%	4%	1%	−4%	−1%
IBM laptop computer		0%	0%	−1%	3%	4%	2%	−10%	−2%	0%	−9%	3%
Toshiba desktop computer		−2%	−5%	1%	1%	−1%	1%	5%	−3%	−5%	−1%	−1%
Toshiba laptop computer		1%	0%	3%	0%	−2%	−2%	−5%	3%	2%	−1%	0%
Logitech mouse		3%	−2%	−1%	0%	4%	6%	−11%	2%	1%	−4%	0%
Ergo-way mouse		0%	0%	2%	3%	1%	−2%	−2%	−5%	0%	−5%	8%

Tabla 2. 8 - Ventas mensuales para cada producto

Un análisis mas profundo de este trabajo se presenta en [Agrawal, 1998].

2.9.2 Exploración y visualización de cubos OLAP con pruebas estadísticas

En este trabajo se propone la combinación de la lattice transversal de dimensiones OLAP y las pruebas estadísticas para descubrir diferencias significativas entre grupos altamente similares. Las pruebas estadísticas permiten comparar pares de celdas vecinas en cuboides. La visualización de los resultados se presenta en un "tablero" parecido a un "Mapa de Karnaugh" que permite explorar interactivamente el cubo y así poder comprender las diferencias entre 2 cuboides que difieren en una dimensión. De manera que las diferencias entre las celdas del tablero representan celdas de interés al usuario.

La figura 2.12 presenta un ejemplo de un cubo con 3 dimensiones D_1, D_2, D_3. Cada cara representa un cuboide de 2 dimensiones. Como se puede ver en la figura existen 2 conjuntos de celdas pares con un cuboide que difiere en exactamente una dimensión. La diferencia en los patrones de relleno indica que hay una diferencia significante en las medidas o hechos.

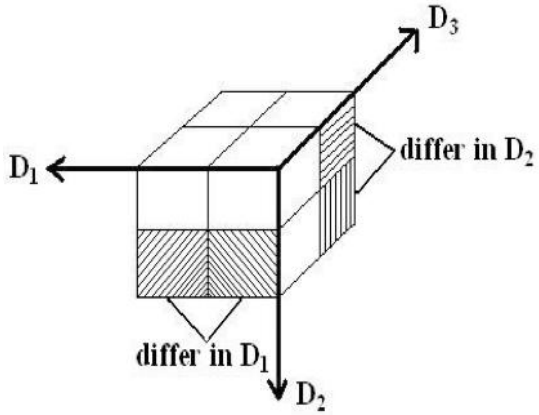


Figura 2. 12 - Vista de cubo con 3 dimensiones

El enfoque presentado en este trabajo tiene como características.

- 2 grupos con número de elementos diferentes pueden ser comparados (conjuntos grandes y pequeños).
- La comparación de medias toma en cuenta las varianzas.

Las pruebas estadísticas tienen dos metas:

1) Encontrar diferencias significantes entre dos grupos en un cuboide.
2) Cuando existe una diferencia significante, el usuario se concentra en los grupos que difieren en una dimensión.

Se presenta un algoritmo que integra la exploración de un cubo, las pruebas estadísticas y la visualización. Este algoritmo tiene las siguientes metas:

- Explorar todos los cuboides.
- Ejecutar las pruebas estadísticas para cada par.
- Seleccionar los pares que presenten diferencias significantes.
- Explorar visualmente de forma interactiva el cubo junto con los resultados de la pruebas.

Los parámetros de entrada y salida del algoritmo son los siguientes:

- Parámetros de entrada: umbral máximo del número de dimensiones diferentes (generalmente umbral = 1).
- Una tabla C que contiene todos los pares de celdas que difieren en (umbral) dimensiones.

Las fases del algoritmo son:

- Precalcular el cubo con d dimensiones.
- Calcular las estadísticas para cada grupo en la lattice de dimensiones.
- Crear pares de grupos que difieren en al menos (umbral) dimensiones.
- Calcular parámetros de subpoblación.
- Calcular las pruebas estadísticas para cada par de celdas en el mismo nivel de agregación.
- Seleccionar los pares que tiene diferencias significantes y categorizar los resultados dentro de capas teniendo 1, 2 y 3 dimensiones diferentes.

Después de estas fases el usuario puede explorar y visualizar los resultados en 2 dimensiones. La navegación del cubo permite al usuario saltar de una celda a otra dentro del tablero (Figura 2.13).

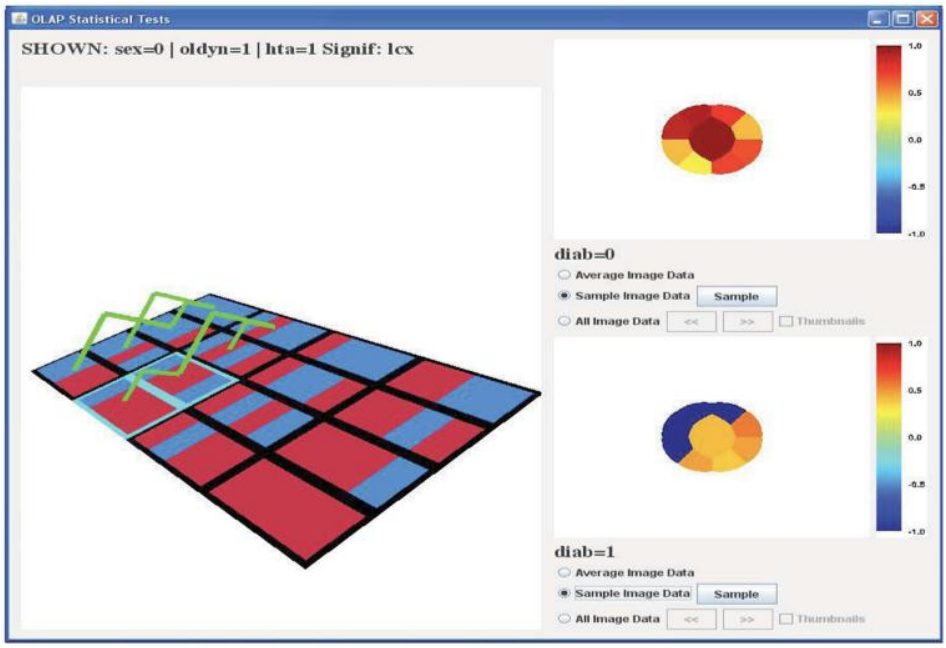


Figura 2. 13 - Vista del tablero de resultados

Cada celda que difiere en una dimensión es visualmente vinculada. Cada celda tiene patrón de relleno determinado el cual especifica el tipo de agregación que este representa.

Las pruebas de ente enfoque se realizan sobre un conjunto de datos médicos almacenados en una base de datos relacional. El prototipo que implementa este enfoque es desarrollado en Java y se conecta por medio de JDBC (Java Database Connectivity) al administrador de bases de datos.

Finalmente la herramienta desarrollada muestra que las pruebas estadísticas son una técnica prometedora para explorar cubos de datos. Estas pruebas estadísticas producen resultados confiables tanto con grandes conjuntos de datos como con pequeños, a diferencia de las técnicas de minería de datos o aprendizaje automático que requieren de una gran cantidad de datos.

Más información del modelo estadístico propuesto en este trabajo puede ser estudiado en [Chen, 2009].

2.9.3 Comparación empírica de deslizadores e histogramas

Las consultas dinámicas facilitan la exploración de la información en tiempo real, visualizando la formulación de la consulta y los resultados. Los deslizadores de consultas dinámicas son usados para el filtrado de datos y una alternativa a las consultas dinámicas es usar varias visualizaciones simples tales como histogramas.

En este trabajo se compara los 2 enfoques en un experimento empírico sobre “DynaMaps”, una herramienta de visualización de datos geográficos y de esta manera se determina las ventajas y desventajas de cada una.

En la visualización de la información, las consultas dinámicas permiten a los usuarios formular rápidamente consultas con widgets gráficos, tales como deslizadores (sliders) para manipular directamente la base de datos. Sin embargo un problema con el diseño inicial ocurre cuando los datos no están distribuidos uniformemente, pequeños ajustes a los deslizadores pueden repentinamente filtrar la mayor parte de los datos de la pantalla, provocando la desorientación del usuario. Por esta razón se puede hacer uso de histogramas brushing ideales para datos no distribuidos.

Para demostrar el uso de los deslizadores e histogramas se realizan pruebas sobre datos con varios atributos representando estadísticas de censo en cada uno de los 50 estados de EUA (Figura 2.14).

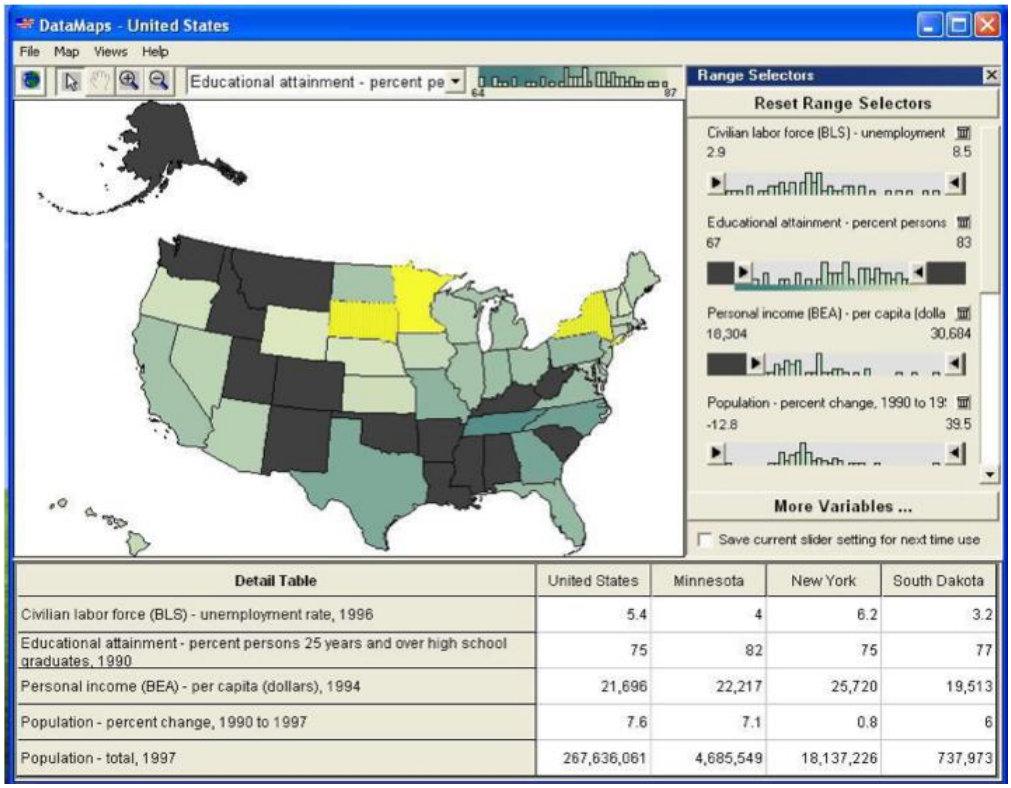


Figura 2. 14 - Interfaz de consultas dinámicas

Cada dynamic query (DQ) slider es un slider doble representando un atributo y un filtro de los valores del atributo. Los filtros son coloreados de gris obscuro en el mapa ya cada DQ slider le es agregado un histograma estático mostrando la distribución de los datos en el atributo (Figura 2.15).

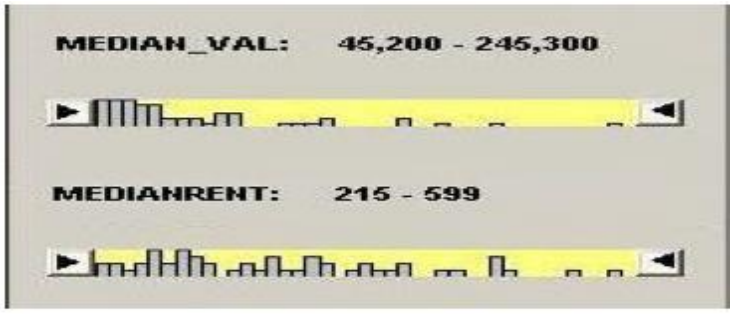


Figura 2. 15 - DQ slider

Una versión alterna al uso de DQ sliders son los histogramas brushing donde los usuarios pueden directamente seleccionar las barras en el histograma para resaltar los estados correspondientes en el mapa, además de resaltar también la barra seleccionada. Así el usuario resalta los estados de interés en lugar de filtrar los estados no deseados, teniendo la ventaja de que el usuario puede seleccionar múltiples rangos discontinuos en el histograma (Figura 2.16).

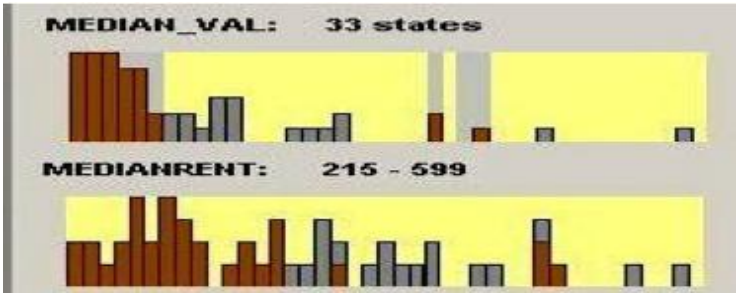


Figura 2. 16 - Histogramas brushing

Se realiza un estudio en [Bao, 2003] y se tiene como resultado que existen tareas específicas ideales para trabajar con deslizadores e histogramas.

Tarea 1 y 2.- Contar estados con rangos de valores simple.
Tarea 3.- Encontrar estados con múltiples criterios.
Tarea 4.- Comparación de estados.
Tarea 5.- Descubrimiento de patrones entre múltiples atributos.

Y se concluye que los histogramas son mejores para tareas complejas de descubrimiento y mejor valorados por los usuarios para la identificación de relaciones. Por otra parte, los DQ sliders son superiores para tareas simples de especificación de rangos y funcionan mejor como un control auxiliar para otras visualizaciones.

2.10 Resumen del capitulo

Se han descrito conceptos necesarios para la solución del problema planteado en 1.2, tales como: Jerarquías, Cubos de datos, Visualización de la información, entre otros. Se han estudiado las herramientas de apoyo actuales, que nos permiten y ayudan a resolver el tipo de pregunta planteada. Además de analizar las características principales de las herramientas y soluciones que trabajan sobre tipos de preguntas específicos. Estas herramientas han sido desarrolladas por la comunidad científica y han sido publicadas en diversos congresos internacionales.

3

Análisis y diseño de la aplicación VisJ

3 Análisis y diseño de la aplicación VisJ

Para el diseño y análisis de la solución planteada, este capítulo se divide en 5 secciones.

Sección 1.- Se presenta el planteamiento actual del problema a resolver en este trabajo.
Sección 2.- Se esboza el tipo de pregunta de negocio que se desea resolver, dando la definición y ejemplos de esta.
Sección 3.- Se modelan de manera formal las jerarquías en las dimensiones de un cubo de datos, cuyo estudio es relevante para comprender etapas del diseño.
Sección 4.- Se describen los elementos necesarios para que nuestra aplicación pueda ser considerada como un sistema visual analítico, además de analizar el espacio de representación visual que el sistema presentará.
Sección 5.- Se presenta el análisis del espacio de representación de visualizaciones posibles. El cual consiste de la visualización de resultados que el sistema es capaz de mostrar.

3.1 Planteamiento de pregunta de Negocio

Una forma de declarar o plantear un análisis sobre las bases de datos, es por medio de consultas o preguntas de negocio, como por ejemplo, ***"se desea saber en qué nivel de la clasificación (jerarquía) de productos se tienen bajos niveles de ventas, digamos abajo del 20% con respecto al año anterior"***.

Esta pregunta puede ser resuelta de varias formas, ya sea por una combinación de consultas SQL (Structured Query Language) cuando se tienen los datos en una base de datos relacional o por un análisis dirigido por el usuario.

En este capítulo se presenta un tipo de pregunta o consulta de negocio que ha sido planteada en [Agrawal, 1998], [Martínez, 2007], [Guzmán, 2008], la manera de resolverla es: combinando el análisis de los datos con OLAP junto con herramientas de visualización y técnicas de minería de datos.

3.1.1 Tendencia con niveles jerárquicos

Cuando las dimensiones de interés presentan una estructura interna llamada también jerarquía que describe la granularidad de los datos, es necesario plantear una consulta que involucre ambos conceptos: Tendencia-Jerarquía. De modo que sea posible encontrar el comportamiento de los elementos de interés en cualquier nivel de la jerarquía en un periodo de tiempo.

Nuestra definición de tendencia con niveles jerárquicos, hace uso de la definición simple de tendencia presentada en 3.2.2 y la definición de jerarquía en las dimensiones presentada en 3.3, combinando las características de ambos conceptos planteamos la siguiente definición.

Definición

La tendencia con niveles jerárquicos se refiere a localizar un conjunto de elementos (hechos) que presenten un comportamiento creciente, decreciente o constante a través del tiempo. La localización se realiza en cualquier nivel de la jerarquía de una dimensión, en un periodo de tiempo determinado.

El crecimiento o decremento es determinado comparando los hechos de interés entre 2 cuboides de datos, correspondiente al mismo nivel de la misma dimensión de interés y al mismo periodo de tiempo.

Ejemplo

Un ejemplo de esta pregunta es:
En una empresa de venta de productos se desea saber en qué nivel de la clasificación (jerarquía) de productos se tienen bajos niveles de ventas, digamos abajo del 20% con respecto al año anterior.

La figura 3.1 da un aspecto visual de los 2 cuboides que se comparan, correspondientes a distintos años de ventas.

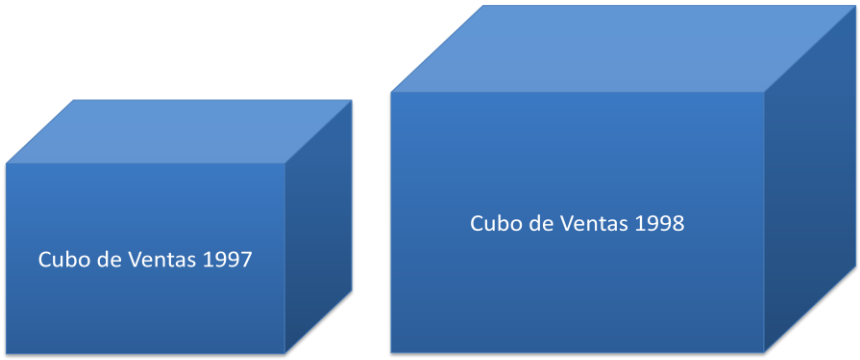


Figura 3. 1 - Comparación de 2 cubos de datos

El porcentaje de crecimiento o decremento se puede definir matemáticamente como:

$$Crecimiento/Decremento = 100 \times \frac{(Cubo\ 2 - Cubo\ 1)}{Cubo\ 1} \qquad (ecuación\ 1)$$

Donde:

Cubo 1.- El cubo de referencia en la comparación
Cubo 2.- El cubo en el cual se desea saber la situación de interés.

Escenarios de tendencia con niveles jerárquicos

La pregunta de tendencia con niveles jerárquicos puede ser planteada de diferentes maneras, se puede consultar por crecimientos o decrementos (tipo de tendencia) en los hechos de los elementos de la dimensión de interés, indicando la cantidad de elementos, un rango o un porcentaje, además de indicar la unidad de tiempo (mes, trimestre, año).

Una revisión o modelado de las posibilidades de la pregunta de tendencia con niveles jerárquicos se observa en la figura 3.2.

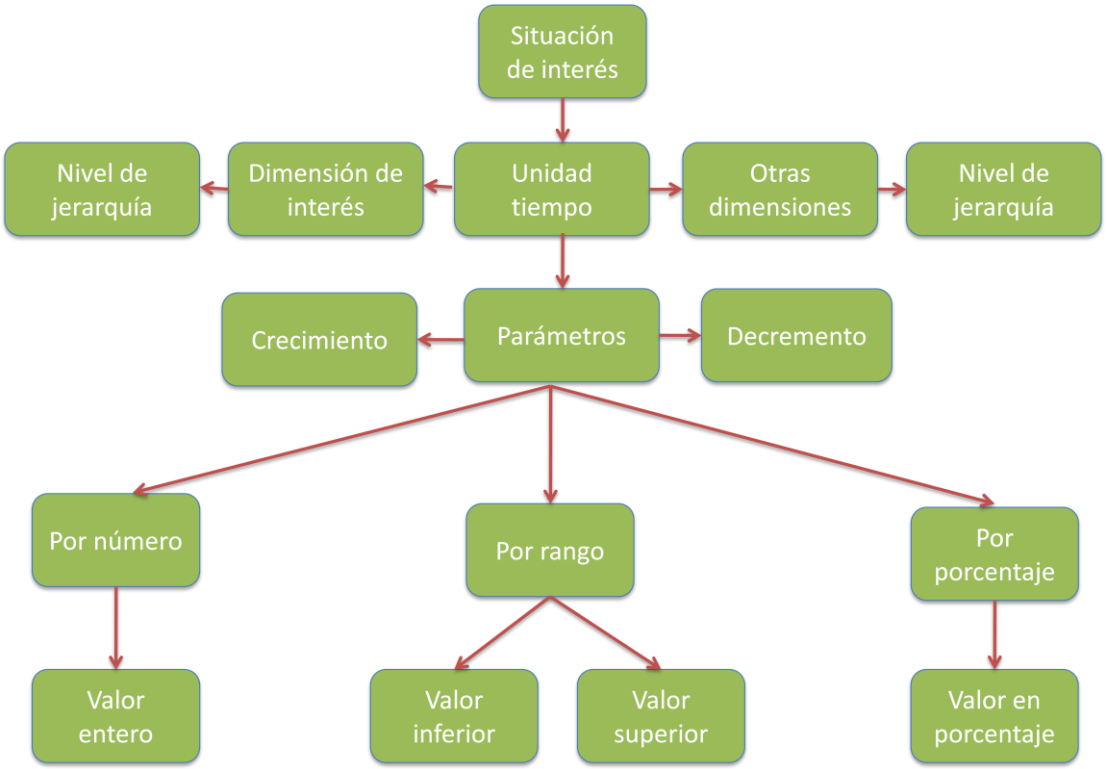


Figura 3. 2 - Escenarios de la consulta

Para indicar los parámetros de la pregunta se inicia con la selección de nodos en la parte superior, hasta llegar a los nodos en el último nivel, de esta forma se especializa la pregunta, eligiendo una ruta que la describa.

Esto es, si se desea plantear la consulta:

En una empresa de venta de productos se desea saber en qué nivel de la clasificación (jerarquía) de productos se tienen bajos niveles de ventas, digamos abajo del 20% con respecto al año anterior

La forma de describirla es:

1.- Seleccionar el nodo “Decremento”.
2.- Seleccionar el nodo “Por porcentaje”.
3.- Seleccionar el nodo “Valor en porcentaje”

Es posible especificar el espacio de búsqueda de los elementos de interés, indicando un valor y nivel especifico de la jerarquía en la dimensión de interés y dimensiones alternas, además de seleccionar la unidad de tiempo (día, mes, año). De esta manera una variante a la pregunta de eficiencia seria:

¿En qué departamentos de la Familia de productos "Comida" se obtuvo un incremento en ventas (50%), en las sucursales de California, en el primer trimestre del año 1998 con respecto al primer trimestre del año 1997?

3.1.2 Tendencia

Las consultas de tendencia simple tienen como objetivo localizar comportamientos en los elementos, ya sea de tipo "creciente", "decreciente" o "constante", indicando el momento en el tiempo en que se mantiene esta tendencia [Martínez, 2007].

Definición

En base a los hechos, buscar en varios periodos de tiempo, si existen productos o servicios (dimensión a analizar) que mantienen una tendencia (comportamiento) en un número determinado de lapsos de tiempo [Guzmán, 2008], [Martínez, 2007].

Ejemplo

En una empresa de venta de productos se desea saber el comportamiento en la venta de productos en los últimos 2 años.

La tendencia suele graficarse como se muestra en la figura 3.3, donde el eje "x" representa la dimensión tiempo y el eje "y" representa los valores de los hechos de los elementos de interés, por ejemplo venta de productos.

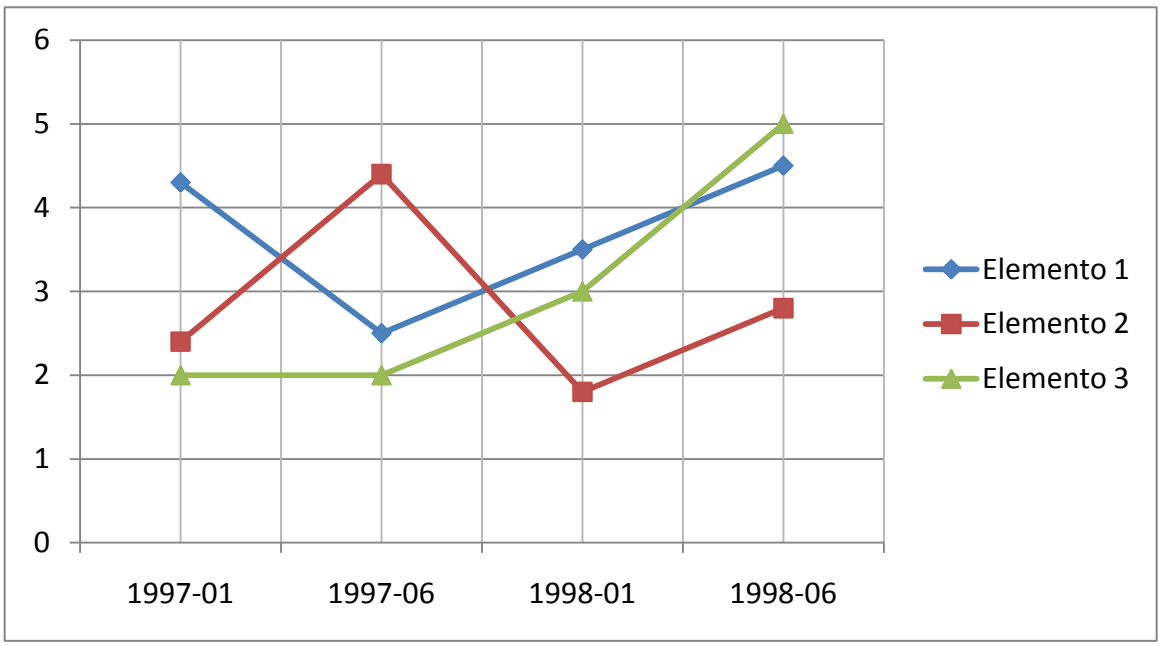


Figura 3. 3 - Tendencia en los años 1997 y 1998

3.2 Jerarquías en la dimensión

El propósito de las jerarquías es proveer una estructura de navegabilidad en una dimensión, de modo que las medidas en diferentes niveles de agregación puedan ser obtenidas por medio de operaciones OLAP como son drill-down o roll-up [Rozeva, 2008].

A continuación se presentan algunos conceptos necesarios para el diseño de las jerarquías en una dimensión y su definición formal, la cual será utilizada a lo largo del capítulo 4.

3.2.1 Diseño de jerarquías en el modelo lógico

Una de las fases más importantes es el diseño de las jerarquías en las dimensiones dentro del modelo lógico de datos, debido a que las jerarquías permiten obtener vistas de los datos con diferente granularidad, esto es resumir o detallar a través de operaciones roll-up y drill-down respectivamente [Malinowski, 2006].

Estas jerarquías pueden ser modeladas en niveles lógicos ordenados para asegurar una estructura consistente y coherente. Las jerarquías tienen una estructura árbol generada por las relaciones one-to-many (padre-hijo).

Existen 2 tipos de jerarquías que pueden presentarse en una dimensión, jerarquías simétricas y asimétricas.

Una jerarquía es simétrica si existe una única ruta (path) desde los miembros de nivel bajo a los miembros del nivel alto y todos los niveles son obligatorios [Rozeva, 2007] (Figura 3.4).

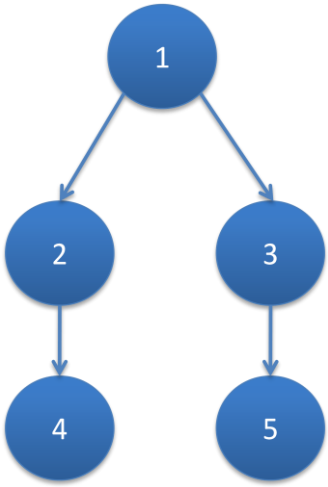


Figura 3. 4 - Jerarquía simétrica

Por otro lado una jerarquía es asimétrica cuando:

- No todos los niveles son obligatorios, esto es que pueden existir rutas que no cubren todos los niveles de la jerarquía [Rozeva, 2007].
- Hay niveles padre sin hijos.

Tales jerarquías no son completamente resumibles (Figura 3.5).

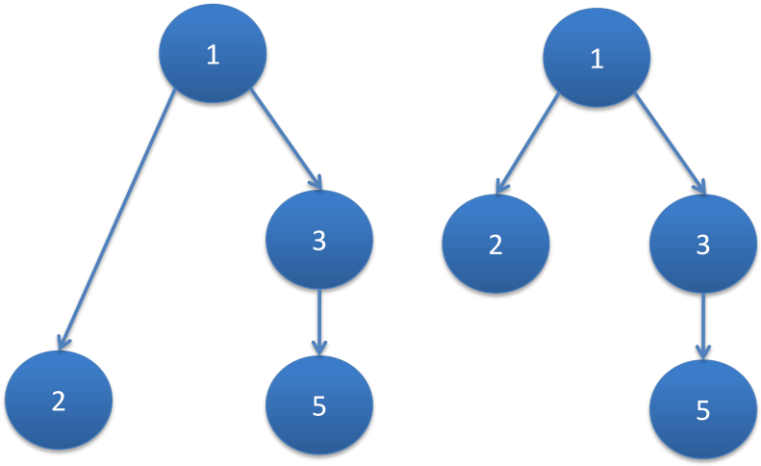


Figura 3. 5 - Jerarquía asimétrica

En un esquema lógico una dimensión representa una relación sobre un conjunto de atributos y consecuentemente una jerarquía involucra relaciones padre-hijo entre 2 columnas de la relación de tablas.

Para definir una jerarquía que cumpla con las condiciones de sumarización, existen 4 dependencias sobre la relación que aseguran la exactitud de los agregados entre los niveles.

- Dependencia transitiva anti-cierre
- Dependencia funcional
- Dependencia sin confusión
- Dependencia de equilibrio

Dependencia transitiva anti-cierre

Existe la posibilidad que una dimensión presente múltiples jerarquías (Figura 3.6). La dependencia transitiva anti-cierre restringe múltiples rutas de un nivel padre a un nivel hijo, previniendo una ruta roll-up por nodos intermedios. De esta manera si hay una ruta más larga entre 2 nodos, la ruta directa entre ellos no está permitida [Rozeva, 2007].
Cuando el grafo de la figura 3.6 es forzado con esta dependencia, la ruta directa entre los nodos 1 y 4 es excluida.

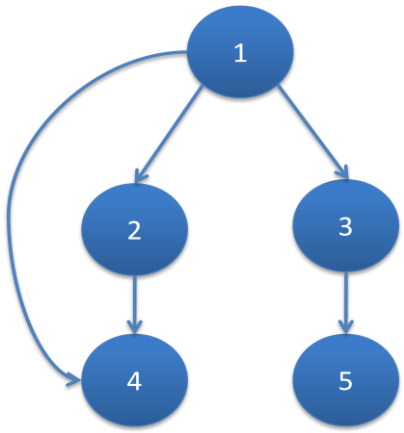


Figura 3. 6 - Múltiples jerarquías

Dependencia funcional

La dependencia funcional fuerza una jerarquía a una estructura árbol, lo cual significa que cada nodo hijo tiene un único nodo padre [Rozeva, 2007].

$$C \rightarrow P$$

Donde:

C es un nodo hijo
P es un nodo padre.

Dependencia sin confusión

Esta dependencia se asegura que exista un nodo en cada nivel de la jerarquía en cada ruta roll-up [Rozeva, 2007]. En otras palabras todos los hijos de un padre están en el mismo nivel.

$$CP \rightarrow L$$

Donde:

C es un nodo hijo.
P es un nodo padre.
L es un identificador de nivel.

Dependencia de equilibrio

Si la jerarquía esta desbalanceada, la dependencia de equilibrio o balance, obliga que todos los nodos hoja estén en el mismo nivel [Rozeva, 2007].

3.2.2 Descripción formal de la jerarquía en la dimensión

Una vez presentadas las dependencias en las jerarquías, a continuación se da una definición formal de las jerarquías presentes en las dimensiones.
Logrando así, formalizar el concepto de jerarquía en una dimensión dentro de un cubo de datos.

Una dimensión $\boldsymbol{D}$ contiene un conjunto de valores $\boldsymbol{V} = \{v_1, v_2, \ldots, v_n\}$.

Una jerarquía de profundidad $\boldsymbol{h}$ sobre $\boldsymbol{D}$ es un conjunto ordenado de $\boldsymbol{h+1}$ niveles, por ejemplo $\boldsymbol{H} = \{L_0, \ldots, L_h\}$.

Cada nivel de la jerarquía $\boldsymbol{i}$ de $\boldsymbol{H}$ sobre $\boldsymbol{D}$ es un conjunto de conjuntos [Bayer, 1999].

$$\boldsymbol{L_i} = \{m^i_1, \ldots, m^i_j\} \text{ con } m^i_k \subseteq \boldsymbol{V} \text{ para } k = 1, \ldots j.$$

Cada $\boldsymbol{m} \in \boldsymbol{L_i}$ es un conjunto miembro de la jerarquía del nivel **i** conteniendo todos los elementos de una categoría.

A un miembro **m** le es asignado un nombre de etiqueta (***label(m)***), por ejemplo *"Drink"* para $\boldsymbol{m^1_2}$. (Figura 3.7)

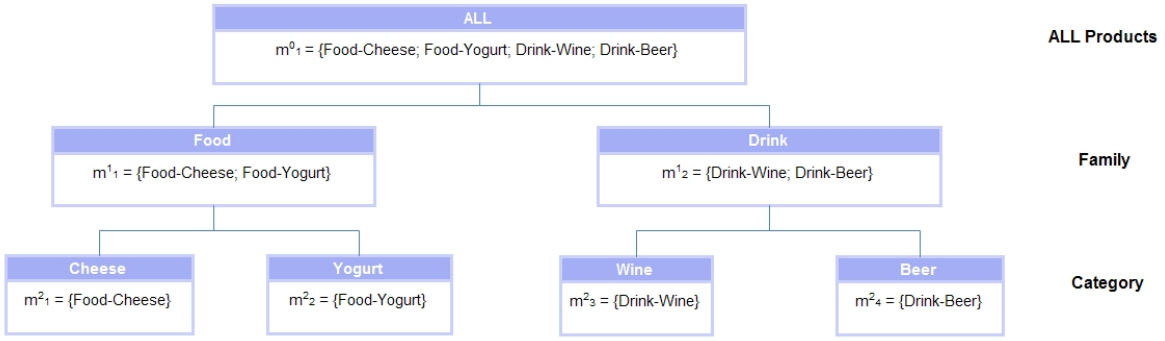


Figura 3. 7 - Jerarquía en la dimensión Producto

La relación $\subseteq$ entre los miembros de 2 niveles vecinos L_i y L_{i+1} define una relación jerárquica (orden parcial) [Bayer, 1999] entre los niveles (por ejemplo: Food-Cheese está en la familia Food).

Incrementando el nivel de la jerarquía, incrementa la granularidad de la categorización, esto es, el dato es clasificado de acuerdo a categorías más finas.

Una jerarquía ***H*** construye un árbol jerárquico donde los nodos de ***H*** son los miembros jerárquicos (etiqueta de miembro) conectados por aristas las cuales están definidas por la relación subconjunto entre 2 niveles vecinos [Bayer, 1999].

Los hijos de un miembro m^i_k del nivel i son todos los miembros de m^{i+1}_l del nivel más bajo $i+1$ que son subconjuntos de m^i_k, esto es:

$$Children(m^i_k) = \{ m^{i+1}_l \in L_{i+1} \mid m^{i+1}_l \subseteq m^i_k \}$$

Esto es, el conjunto {{Food-Cheese},{Food-Yogurt}} es el conjunto hijo de "Food".

El padre del miembro m^i_k de nivel i es el miembro m^{i-1}_l del nivel superior $i-1$, m^{i-1}_k es el superconjunto de $\mathbf{m^i_k}$.

$$\mathbf{parent(m^i_k) = \{ m^{i-1}_l \in L_{i-1} \mid m^{i-1}_l \supseteq m^i_k \}}$$

Esto es, el conjunto *Food* es el padre de *{Food-Cheese}*.

3.3 Análisis de la visualización de la información

Como se mencionó en el capítulo 2, la meta de la visualización es ayudar a comprender un conjunto de datos aprovechando el sistema visual humano el cual permite la habilidad de ver patrones, tendencias e identificar anomalías.

La creación de visualizaciones requiere identificar y seleccionar la codificación visual más efectiva para mapear un conjunto de datos a características gráficas tales como: posición, tamaño, forma, color.

Sabiendo que el espacio de posibles diseños de visualizaciones es amplio, investigadores en computación, psicólogos y estadísticos han estudiado como diferentes codificaciones facilitan la comprensión de los datos tales como números, categorías o redes. No sin olvidar que debe existir un balance entre la interacción del diseño y la estética.

En este proyecto se realiza un análisis de los espacios de visualización empleados para representar las situaciones de interés encontradas, en base a la pregunta de negocio descrita en 3.2, mostrando las técnicas para visualizar e interactuar con diversos conjuntos de datos.

Las visualizaciones propuestas son creadas usando PowerChart [RE07], una herramienta Web de visualización, la cual hace uso del lenguaje XML para la construcción de las visualizaciones.

3.3.1 Análisis visual del sistema

El análisis visual es una nueva área de la tecnología y añade características especiales a los sistemas de análisis de datos. Es el proceso de la facilidad de razonamiento analítico interactuando con interfaces visuales, además de ser un medio de exploración y comprensión de los datos [Hanrahan, 2009].

Es posible hacer una pregunta, conseguir la respuesta y seguir preguntando, todo esto dentro de las interfaces visuales.

En resumen, el análisis visual permite ir en cualquier dirección con los pensamientos mientras se aprovecha el sistema de percepción visual humano para guiar hacia las rutas interesantes y útiles.

El análisis visual es una manera rápida para las personas de explorar y comprender los datos de cualquier tamaño. El sistema desarrollado debe de cumplir con al menos 6 elementos esenciales los cuales se mencionan a continuación y que son claves para una verdadera aplicación visual analítica [Hanrahan, 2009].

- Exploración visual
- Aumento de la percepción humana
- Expresividad visual
- Visualización automática
- Cambio de perspectivas visuales
- Enlace de perspectivas visuales
- Visualización colaborativa

Además también que el sistema desarrollado en este trabajo incluye características como tableros de control.

Exploración visual

Esta característica es considerada la más importante; es donde la aplicación unifica los pasos de consulta, exploración y visualización de datos dentro de un único proceso.
Los usuarios pueden aun no tener una pregunta específica, pero conforme ellos se van moviendo a través de la visualización de los datos, ellos notan algo y hacen la pregunta. Esto significa que las visualizaciones en una aplicación visual analítica permiten a los usuarios detenerse y mirar más de cerca.
Los filtros, agrupamiento, ordenamiento y operaciones OLAP tienen lugar dentro la visualización. Un usuario podría comenzar con una pregunta básica y basado en las señales visuales o puntos de vista profundizar la investigación.

Aumento de la percepción humana

Las aplicaciones visuales analíticas fomentan el pensamiento visual aprovechando los poderes de la percepción humana. El cerebro humano posee una capacidad asombrosa de procesar gráficos mucho más rápido que procesar tablas de números. Desafortunadamente la mayoría de los sistemas de "Inteligencia de negocio" y "hojas de cálculo" no aprovechan estas capacidades de percepción del cerebro humano.
Las representaciones del sistema visual analítico generan un preciso uso del tamaño, color, forma y texto para hacer las diferencias y cuando son usados de forma adecuada ayudan a la interpretación.

Expresividad visual

La expresividad visual es especialmente importante cuando se necesita mirar más de 2 ó 3 dimensiones de un problema simultáneamente. Las aplicaciones visuales analíticas permiten a las personas visualizar múltiples dimensiones de un problema sin esfuerzos, en formatos que sean fáciles de comprender. Lo que significa que estas aplicaciones muestran problemas complejos con simplicidad y elegancia.

La expresividad multidimensional es particularmente importante cuando se involucra la dimensión tiempo, manejar apropiadamente la dimensión tiempo no es tan simple como añadir líneas de tendencia. Debe tenerse habilidad de mostrar visualmente datos y tiempos en múltiples niveles de detalle simultáneamente (Ventas por año, por mes, por día).

Visualización automática

La visualización automática incluye la sugerencia automática a visualizaciones efectivas para un problema específico. Esto también ayuda a las personas a aprender y pensar visualmente.

Cambio de perspectivas visuales

No existe una visualización única que ofrezca el mejor resumen de los resultados. Típicamente las personas necesitan mirar una variedad de visualizaciones, dependiendo de las tareas que se quieran lograr.
Las aplicaciones visuales analíticas efectivas sugieren una serie de alternativas de visualizaciones.
Este cambio de perspectivas sobre un problema es una gran manera de generar nuevas preguntas, produciendo curiosidad acerca de lo que esta actualmente pasando en los datos.

Enlace de perspectivas visuales

En enlace de perspectivas es una adición a la característica de cambio de perspectivas.
Por ejemplo una visualización podría mostrar un conjunto de anomalías y el usuario puede seleccionar una anomalía e instantáneamente ver otra visualización que despliega el detalle de los datos.
Otro ejemplo es, los usuarios pueden seleccionar una línea de tendencia en una primera vista y ver las entidades geográficas relacionadas a esa línea en una segunda vista. Esto significa una correlación de la información, en resumen una visualización lleva a otra.

Visualización colaborativa

La visualización colaborativa es la habilidad de crear interactivamente útiles visualizaciones de la información en equipo. Las personas publican resultados de forma segura interactivamente y disponibles en la red.

Una vez planteados los 7 elementos necesarios en el diseño del sistema analítico a desarrollar, veamos el espacio de visualización propuesto.

3.3.2 Análisis del espacio de Visualizaciones

El objetivo es presentar un mapa de situaciones de interés que ayude a identificar las estructuras internas de los datos (jerarquía), reconociendo las rutas (paths) de las situaciones de interés encontradas durante la etapa de análisis de la pregunta de negocio.

Se presentan 3 formas de representación visual de los resultados, comenzando con una de las más populares “Mapas de Nodos”, continuando con “Mapas de calor” y completando con “Mapas Pastel Multi-Nivel”. Todas estas visualizaciones ideales para la representación de jerarquías en los datos [Chignell, 2005], [Schulz, 2006], de acuerdo a las investigaciones presentadas en congresos de visualización de la información como es InfoVis.

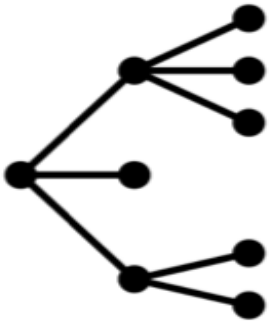


Figura 3. 9 - Mapa de Nodos

Figura 3. 8 - Mapa de Calor

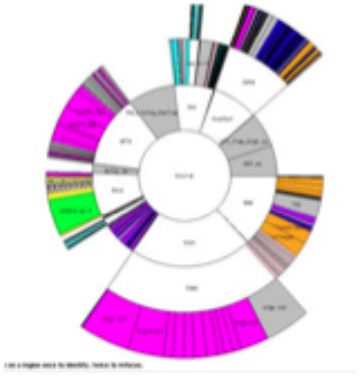


Figura 3. 10 - Mapa Pastel Multi-Nivel

Mapa de Nodos.

Este tipo de mapa representa cada elemento de la jerarquía de una dimensión por medio de nodos, de esta manera permite visualizar las entidades relacionadas de manera jerárquica, reconociendo la ruta de una situación de interés. Es un mapa intuitivo y fácil de comprender [Chen, 2006].

Sus características principales son:

- Cada elemento es representado por un nodo esfera.
- Los nodos están relacionados por conectores (aristas).
- Cada nodo es representado por un color que identifica el nivel de jerarquía.

Mapa de Calor

Otra forma de presentar las situaciones de interés es por medio de un mapa de calor, el cual hace uso del color para representar cada ruta de un elemento de interés dentro de una tabla, cada fila representa la ruta de un objeto de interés encontrado y las columnas corresponden al nivel de jerarquía. Los mapas de calor son ideales para representar una gran cantidad de

datos, por ejemplo cuando se plantee una consulta de negocio que necesite presentar muchos elementos de interés [Chen, 2006].

Mapa Pastel Multi-Nivel

El mapa Pastel Multi-Nivel también llamado "Sunburst" muestra una estructura árbol en forma de pastel, la ventaja de presentar los datos en forma de pastel en vez de una simple visualización de árbol, es la posibilidad de observar una vista instantánea de los datos, trazando la ruta (path) de los nodos hijos al nodo padre [Chen, 2006], [Bostock, 2010].

Cada nivel corresponde a un sub-pastel, el número de sub-pasteles es igual al número de niveles de la jerarquía de la dimensión.

El radio del pastel de cada nivel es dividido en segmentos, dependiendo el número de hijos del segmento padre.

Esta diversidad de mapas indica que el sistema visual analítico propuesto permite el "Cambio de perspectivas visuales". La iniciativa de mezclar interacciones es motivada por la observación que se experimenta como diseñadores, al no conocer cómo construir visualizaciones efectivas debido a la gran diversidad de características, por lo cual es efectivo explorar el mismo conjunto de datos desde diferentes perspectivas a través de diferentes diseños de visualización.

3.3.3 Ventajas y desventajas de los Mapas de visualización

Una vez propuesto el espacio de representación de visualizaciones, se analizan las ventajas y desventajas de cada una de las visualizaciones de forma detallada [Chignell, 2005].

	Mapa de Nodos	**Mapa de Calor**	**Mapa Pastel Multi-Nivel**
Ventajas	- Familiar - Muestra la estructura jerárquica y sus elementos	- Escalables, uso eficiente del espacio - Facilita la comparación de elementos	- Escalables, uso eficiente del espacio - Posibilidad de trazar la ruta de forma dinámica
Desventajas	- Difícil de escalar (50 nodos)	- Menos familiar - Difícil de ver la estructura jerárquica	- Menos familiar

Tabla 3. 1 - Ventajas y desventajas de mapas

3.4 Análisis de la solución a pregunta de negocio

En este tema se propone un proceso de solución automática a la pregunta planteada en 3.2, aprovechando las herramientas OLAP y de visualización, junto con técnicas de minería de

datos, para obtener un **Mapa de situaciones de interés**, en el cual se pueda demostrar que el descubrimiento del conocimiento se vuelve una tarea ágil e intuitiva.

Como primer paso se plantea una modelación formal de la consulta de negocio, indicando los parámetros requeridos para resolverla.

3.4.1 Modelado de tendencia con niveles jerárquicos

Se plantea la definición formal de los elementos necesarios para el análisis de la consulta de tendencia con niveles jerárquicos.

Sea $C_{target}=<D_1, D_2,\ldots,D_k; H_1, H_2,\ldots,H_n>$. El cubo de datos a analizar.

Donde:

D_k son las dimensiones del cubo.
H_n son los hechos definidos en el cubo.

Se define también elementos generales para definir el espacio de búsqueda y donde realizar la creación dinámica de cubos:

G_1. Es la dimensión de interés, D_i

G_2. Es el hecho o medida de interés, H_i

G_3. Rangos de otras dimensiones: $R_j=[v_{ji}, v_{jf}]$, donde $R_j \neq D_i$ y v_{ji}, v_{jf} son los valores que definen el espacio de búsqueda.

G_4. Lapsos de tiempo.

Al tratarse de la tendencia en niveles jerárquicos, es necesario especificar las características de los objetos de interés, esto es:

C_1 = {“crecimiento”, “decremento”}
C_2 = {“número de objetos”, “porcentaje”, “rango”}
C_3 = {valor de número de objetos, valor de porcentaje, valores que definen el rango}

Finalmente se define las características de visualización, por las cuales se desea obtener los resultados.

$\mathbf{V_1}$= {Mapa de Nodos, Mapa de Calor, Mapa Pastel Multi-Nivel}

Donde:

$\mathbf{V_1}$=Es el espacio de opciones de visualización.

Reuniendo los elementos anteriores es posible definir la consulta de tendencia en niveles jerárquicos.

3.4.2 Algoritmo de tendencia con niveles jerárquicos

Los pasos necesarios para resolver la consulta de tendencia con niveles jerárquicos en distintos niveles de la jerarquía son los siguientes:

P.1.- Definición y carga del cubo de datos.
P.2.- Cálculo de los agregados por jerarquía de la dimensión de interés d_i.
P.3.- Calculo del crecimiento/decremento.
P.4.- Ordenamiento del crecimiento/decremento.
P.5.- Selección de los parámetros de los objetos de interés.
P.6.- Visualización de los resultados.

P.1.- Definición y carga del cubo de datos.

Consiste en la descripción del modelo lógico multidimensional en el esquema XML y su carga al motor OLAP.

P.2.- Cálculo de los agregados por jerarquía de la dimensión de interés d_i.

El cálculo de los agregados se realiza junto con la definición del esquema XML, en el se declaran las medidas o hechos de cada cubo de datos. (Ver tema 3.4).

P.3.- Calculo del crecimiento/decremento.

El crecimiento/decremento de cada par de cuboides, se obtiene aplicando la ecuación:

$$\text{Crecimiento/decremento} = 100 \times \frac{(\text{Cubo 2} - \text{Cubo 1})}{\text{Cubo 1}} \qquad \textit{(ecuación 1)}$$

P.4.-Ordenamiento del crecimiento/decremento.

Se ordenan los porcentajes calculados en cada elemento del nivel de la jerarquía, nivel que es especificado por el usuario. Se obtiene una lista ordenada de acuerdo a los parámetros del usuario (crecimiento, decremento).

P.5.- Selección de los parámetros de los objetos de interés.

Se selecciona las características que describen a la consulta de negocio, son los escenarios de tendencia con niveles jerárquicos.

P.6.- Visualización de los resultados.

Finalmente de especifica el tipo de visualización por la cual se desea obtener los resultados.

3.5 Resumen del capitulo

En este capitulo se estudio el tipo de pregunta de negocio que se desea resolver, se analizaron conceptos como es la jerarquía en las dimensiones, que forman parte del cubo de datos. Se presentó también un análisis de los elementos que definen a una aplicación visual analítica y se plantearon tipos de vistas disponibles para representar estructuras jerárquicas, estas vistas han sido publicadas en artículos y conferencias de visualización de la información y son recomendadas por la comunidad científica.

4

Desarrollo e implementación de la aplicación VisJ

4 Desarrollo e implementación de la aplicación VisJ

Se analiza en 5 secciones el desarrollo e implementación del sistema visualizador de situaciones de interés, estas secciones son:

Sección 1.- Se presenta la descripción de las bases de datos del dominio comercial y científico.
Sección 2.- Se describe la creación de cubos OLAP (On-Line Analytical Processing) a través del motor ROLAP (Relational On-Line Analytical Processing) Mondrian.
Sección 3.- Se diseñan las dimensiones del cubo de datos, sus niveles y miembros, presentando ejemplos reales de modelación.
Sección 4.- Se describe el proceso de solución manual que se realiza para poder resolver el tipo de consulta que se describe en 3.2. Se hace uso de un visor OLAP para realizar la exploración.
Sección 5.- Se presenta el análisis de los procesos necesarios que el sistema visualizador de anomalías implementa para resolver de manera automática la consulta planteada en 3.2. Además también se diseña el tipo de arquitectura del sistema y la modelación en diagramas UML (Unified Model Language).

4.1 Descripción de las bases de datos (Modelo físico)

Uno de los objetivos del sistema visualizador de situaciones de interés o anomalías es la posibilidad de trabajar en varios dominios de datos, siempre y cuando los dominios presenten un modelo multidimensional definido con dimensiones y hechos, cuyas dimensiones presenten una estructura interna o también llamada jerarquía.

Se hace uso de dos dominios: comercial y científico, presentando los modelos multidimensionales y el diseño de los cubos de datos, junto con sus dimensiones y hechos.

4.1.1 Conjunto de datos comerciales

Se usa una base de datos diseñada de forma multidimensional llamada “FoodMart” cuyo dominio es información de ventas de un supermercado. Esta base de datos está disponible en la página oficial de Mondrian, como sentencias DML (Data Manipulation Language). La base de datos contiene 37 tablas cuya información es productos, sucursales, clientes, promociones, ventas de almacén, costos de almacén, unidades vendidas, entre otros.

Esta base de datos ha sido diseñada cuidadosamente para su uso con ROLAP, esto es se tienen tablas que contienen agregados (tabla de hechos) los cuales resumen combinaciones de distintas dimensiones (productos, sucursal, cliente) y estas tablas están relacionadas con tablas que almacenan información específica de cada dimensión (tabla de dimensión).

Los elementos dentro de la jerarquía en la dimensión del cubo de datos son organizados de una forma bien definida, garantizando así la correcta clasificación de los datos.

Las tablas de hechos son: sales_fact, sales_fact_dec_.

Las tablas de dimensiones son: product, employee, store, promotion, category, region, days.

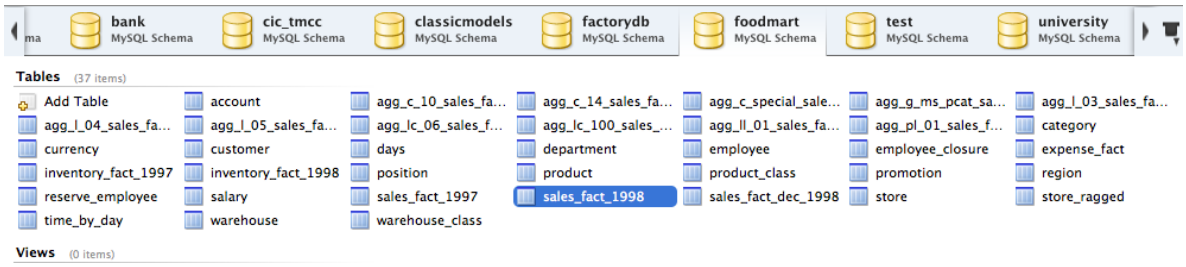


Figura 4. 1- Vista de las tablas en la base de datos FoodMart

La figura 4.1 muestra el número de tablas que componen a la base de datos foodmart. La figura 4.2 presenta una tabla de hechos llamada: sales_fact_ (central), 5 tablas de dimensión relacionadas a la tabla de hechos por medio de una llave foránea y una tabla de detalle de la dimensión producto llamada: product_class. Esta estructura tiene como nombre esquema de copo de nieve. A partir de este diseño se modela un cubo OLAP usando el motor Mondrian.

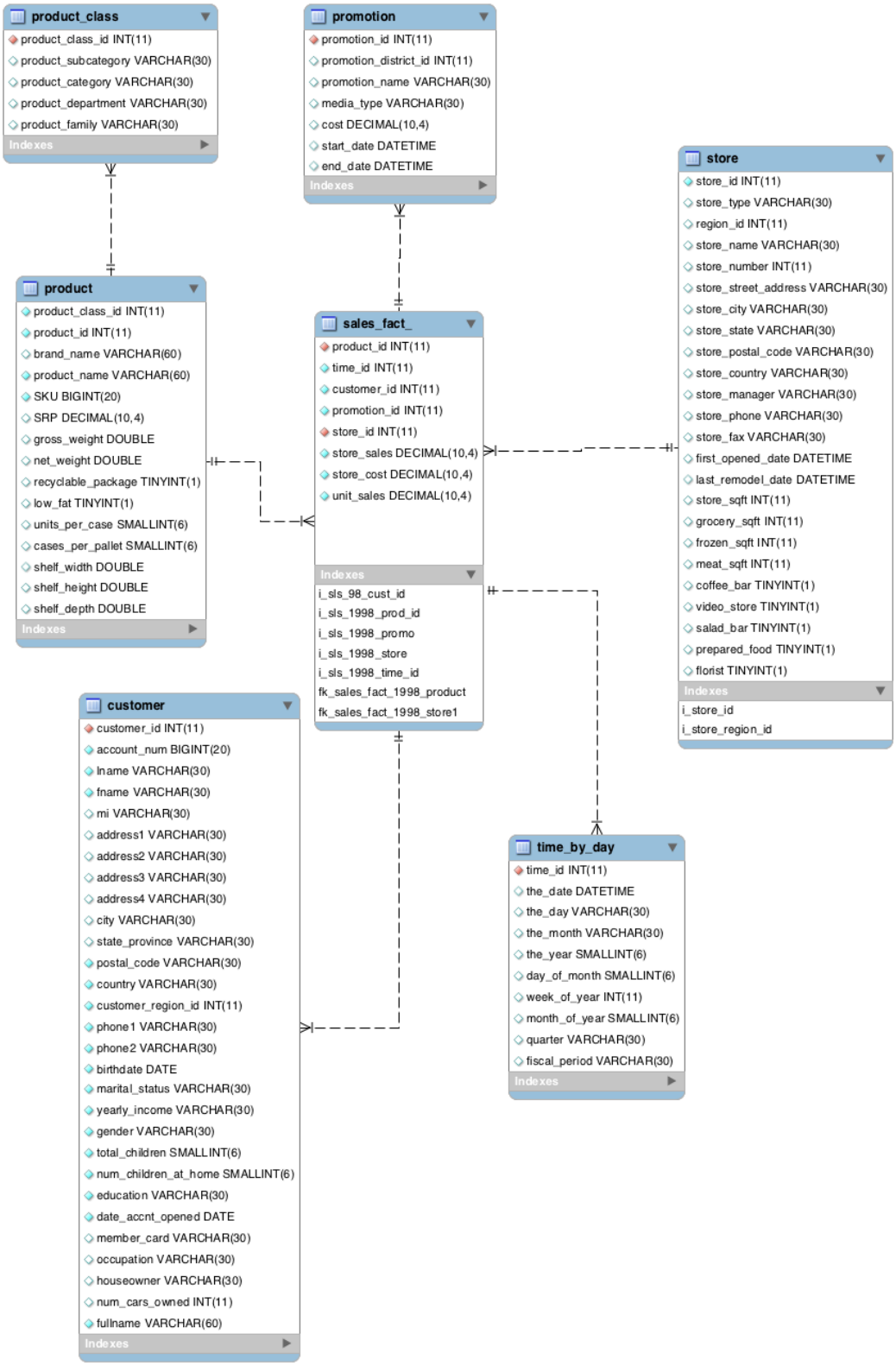


Figura 4. 2 - Esquema copo de nieve del dominio comercial

Esta base de datos es almacenada en MySQL 5 aunque pudo haberse usado cualquier otro administrador como es Oracle, MSAccess o SQL Server.

4.1.2 Conjunto de datos científicos

La base de datos multidimensional del dominio científico almacena información relacionada a un conjunto de tesis y su clasificación ACM (Association for Computing Machinery). Estas tesis pertenecen al Centro de investigación en computación del IPN. La base de datos está conformada por una tabla de hechos llamada: tesis_cic y una tabla de dimensión: tiempo. La tabla de hechos contiene la clasificación ACM de cada tesis, de manera que en realidad existen 2 dimensiones dentro del dominio científico: clasificación y tiempo. (Ver figura 4.3).

De manera similar a la base de datos comercial, los elementos dentro de la jerarquía en la dimensión del cubo de datos científico son organizados de una forma bien definida, garantizando así la correcta clasificación de los datos.

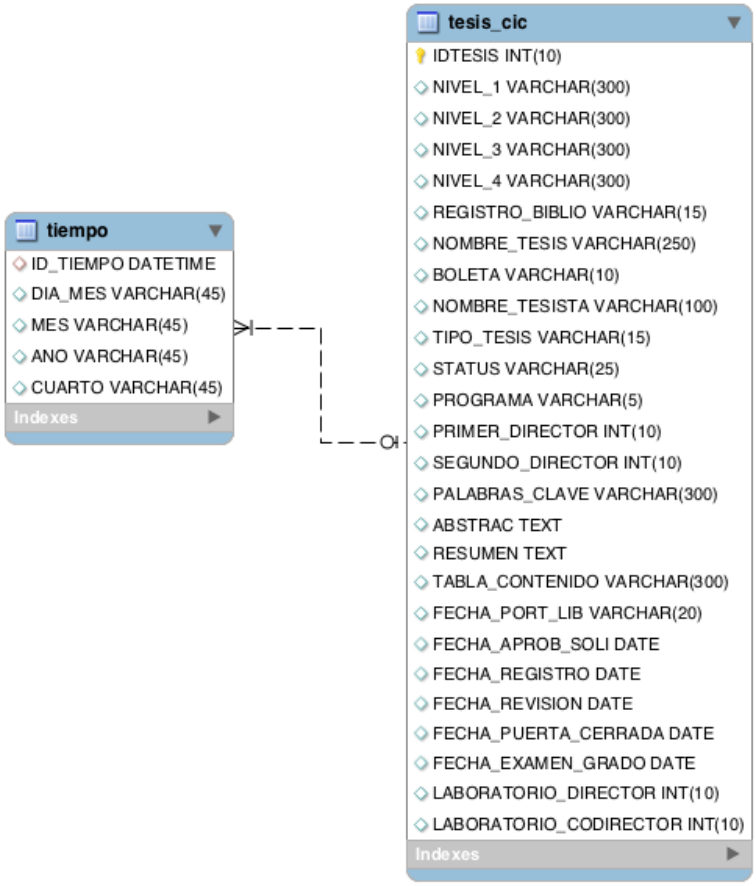


Figura 4. 3 - Esquema Estrella del dominio científico

De la misma manera que el dominio comercial, esta base de datos fue almacenada en MySQL 5.

4.2 Modelo lógico

Para el desarrollo del sistema utilizaremos un servidor de tipo ROLAP llamado Mondrian, esto significa que almacena los datos en una base de datos relacional.

Este servidor ROLAP aprovecha el administrador de almacenamiento de un RDBMS. La idea general de usar Mondrian es delegar a la base de datos lo que es de la base de datos y añadir características tipo OLAP como son el diseño de cubos al sistema.

4.2.1 Diseño de cubos OLAP en el dominio comercial

Para crear el modelo lógico se hará uso de un esquema XML, en el cual se define la base de datos multidimensional que consiste de cubos, jerarquías, miembros y un mapeo del modelo lógico al modelo físico. Las siguientes líneas describen la definición del cubo de datos Sales (Ventas).

```
<Schema>
<Cube name="Sales" defaultMeasure="Unit Sales">

<Table name="sales_fact_">
</Table>

<Dimension name="Store">
<Hierarchy hasAll="true" primaryKey="store_id">
<Table name="store"/>
<Level name="Store Country" column="store_country" uniqueMembers="true"/>
<Level name="Store State" column="store_state" uniqueMembers="true"/>
<Level name="Store City" column="store_city" uniqueMembers="false"/>
<Level name="Store Name" column="store_name" uniqueMembers="true"/>
</Hierarchy>
</Dimension>
Otras dimensiones …

<Measure name="Store Cost" column="store_cost" aggregator="sum" formatString="#,###.00"/>
<Measure name="Unit Sales" column="unit_sales" aggregator="sum"  formatString="Standard"/>
Otras medidas…
</Cube>
</Schema>
```

El modelo lógico se compone de los objetos utilizados para escribir consultas multidimensionales, estos objetos son cubos, dimensiones, jerarquías, niveles y miembros. Mientras que el modelo físico es la fuente de datos, datos que son presentados a través del modelo lógico.

En el esquema XML anterior se define un cubo de datos de nombre: “Sales”. En el cual se indica el nombre de la tabla de hechos “sales_fact_”.

```
<Table name="sales_fact_">
</Table>
```

Se define una dimensión “Store” con 4 niveles de jerarquía: Store Country, Store State, Store City, Store Name, estos niveles pertenece a la tabla de dimensión store.

```
<Table name="store"/>…
```

Y cada nivel es mapeado a una columna de la tabla de dimensión.

```
<Level name="Store Country" column="store_country" uniqueMembers="true"/>
```

Una vez declaras las dimensiones del cubo, se realiza la definición de los hechos o medidas, como ejemplo se tiene 2 medidas: Store Cost, Unit Sales.

```
<Measure name="Store Cost" column="store_cost" aggregator="sum" formatString="#,###.00"/>
<Measure name="Unit Sales" column="unit_sales" aggregator="sum" formatString="Standard"/>
```

Estas medidas son mapeadas ahora a una columna de la tabla de hechos sales_fact_, indicando el tipo de agregado y formato del número.

4.2.2 Diseño de cubos OLAP en el dominio científico

La definición del cubo de datos del dominio científicoestá formado por 2 dimensiones: “clasificación”, ”tiempo” y una medida: “Número de tesis” con un agregado tipo: count.

```
<Cube name="Tesis" defaultMeasure="Número Tesis">
<Table name="tesis_cic"/>

<DimensionUsage name="Clasificacion" source="Clasificacion" foreignKey="IDTESIS"/>
<DimensionUsage name="tiempo" source="tiempo" foreignKey="FECHA_EXAMEN_GRADO"/>

<Measure name="Número Tesis" column="IDTESIS" aggregator="count" formatString="#,###"/>

</Cube>
```

La dimensión “Clasificación” tiene una jerarquía de 4 niveles que describen la clasificación ACM de tesis, cada nivel es mapeado a una columna de la tabla de dimensión.

```
<Dimension name="Clasificacion">
<Hierarchy hasAll="true" primaryKey="IDTESIS">
<Table name="tesis_cic"/>
<Level name="Nivel 1" column="NIVEL_1" uniqueMembers="false"/>
<Level name="Nivel 2" column="NIVEL_2" uniqueMembers="false"/>
<Level name="Nivel 3" column="NIVEL_3" uniqueMembers="false"/>
<Level name="Nivel 4" column="NIVEL_4" uniqueMembers="false"/>
</Hierarchy>
</Dimension>
```

La dimensión “Tiempo” tiene una jerarquía de 3 niveles que describen la fecha de terminación de cada tesis, de la misma manera cada nivel es mapeado a una columna de la tabla de dimensión.

```
<Dimension name="tiempo">
<Hierarchy hasAll="true" primaryKey="ID_TIEMPO">
<Table name="tiempo"/>
<Level name="ano" column="ANO" uniqueMembers="false"/>
<Level name="mes" column="MES" uniqueMembers="false"/>
<Level name="dia" column="DIA_MES" uniqueMembers="false"/>
</Hierarchy>
</Dimension>
```

4.2.3 Arquitectura del motor OLAP

Una vez que se define el cubo de datos dentro del archivo XML (eXtensible Markup Language). El paso siguiente es cargarlo al motor OLAP.

El motor OLAP consiste de 4 capas (Figura 4.4)

- Capa de presentación
- Capa dimensional
- Capa estrella
- Capa de almacenamiento

A continuación se describe la función de cada capa en el servidor OLAP.

Capa de presentación

La capa de presentación representa una interfaz de usuario, la cual interactúa con el sistema OLAP, realizando preguntas y obteniendo respuesta del servidor OLAP.

Es en esta capa en donde nos enfocaremos a desarrollar el sistema visualizador de anomalías, por medio de las preguntas planteadas.

Capa dimensional

La capa dimensional analiza, valida y ejecuta las consultas MDX (MultiDimensional eXpressions). Esta capa envía las celdas requeridas a la capa de agregación (estrella) en lotes. Además de contener el metadatos que describe el modelo dimensional y el mapeo entre este y el modelo relacional.

Capa estrella

La capa estrella es responsable de mantener en memoria un cache de agregados.
Como se menciono, la capa dimensional envía peticiones de conjuntos de celdas requeridas, si la celda requerida no está en la cache, el administrador de agregación envía una petición a la capa de almacenamiento.

Capa de almacenamiento

La capa de almacenamiento es un sistema administrador de bases de datos relacional RDBMS (Relational Database Management System), el cual es responsable de proporcionar los agregados y miembros de la tabla de hechos y de dimensiones respectivamente.

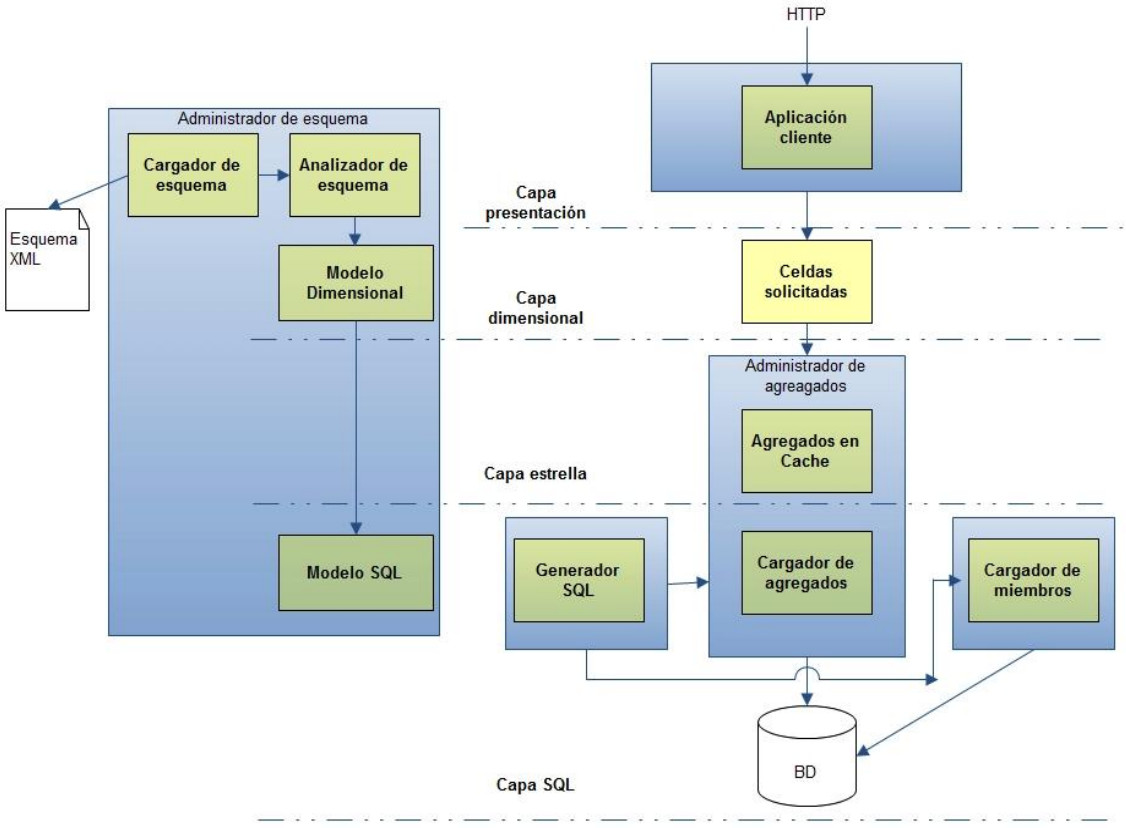


Figura 4. 4 - Arquitectura del motor OLAP

El módulo de memoria cache es la parte más importante porque mantiene los agregados pre- calculados en memoria, entonces consultas subsecuentes pueden acceder a celdas sin ir a consultar al disco.
Si la cache mantiene un conjunto de datos de un nivel más bajo de agregación, es posible calcular el conjunto de datos requeridos por medio de una operación roll up. En otras palabras si la cache contiene los agregados para todos los hijos de un miembro, entonces es posible calcular los agregados para el miembro padre por la operación roll-up.

4.3 Diseño de la jerarquía en las dimensiones

Como se menciono anteriormente, cada cubo de datos tiene definido las dimensiones que lo componen, estas dimensiones tienen una estructura interna llamada jerarquía.

Las dimensiones Producto (Product), Almacén (Store), Tiempo (Time), Cliente (Customers) conformar el cubo de datos con el cual trabajemos (cubo Ventas).

```
<Schema>
<Cube name="Sales" defaultMeasure="Unit Sales">

<Table name="sales_fact_1998">
</Table>

<Dimension name="Product"> … </Dimension>
<Dimension name="Store"> … </Dimension>
<Dimension name="Time"> … </Dimension>
<Dimension name="Customers"> … </Dimension>

…

</Cube>
</Schema>
```

En la dimensión Producto se tiene 6 niveles de jerarquía: Familia, Departamento, Categoría, Subcategoría, Marca y Nombre de Producto (Figura 4.5 y 4.6).

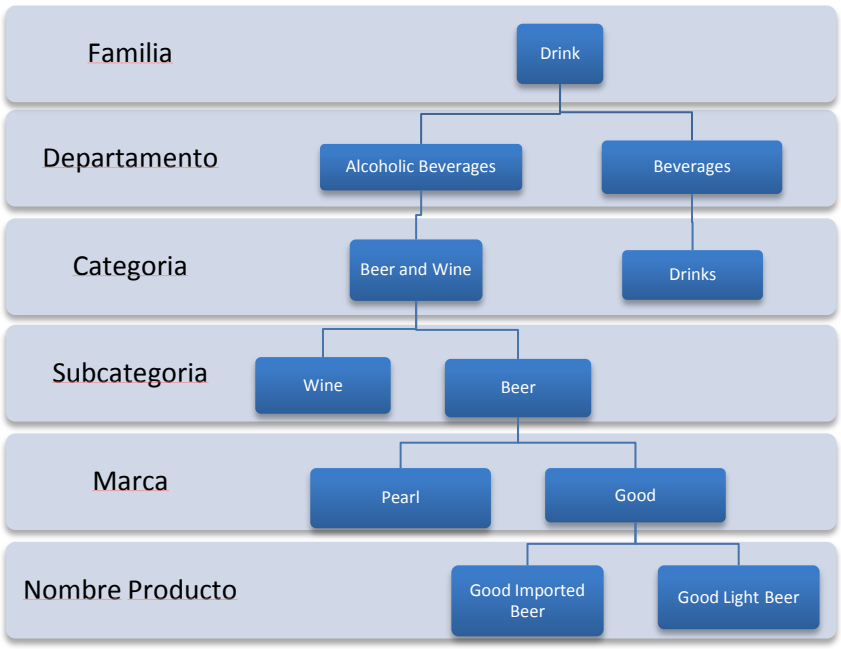


Figura 4. 5 - Jerarquía en la dimensión Producto - Drink

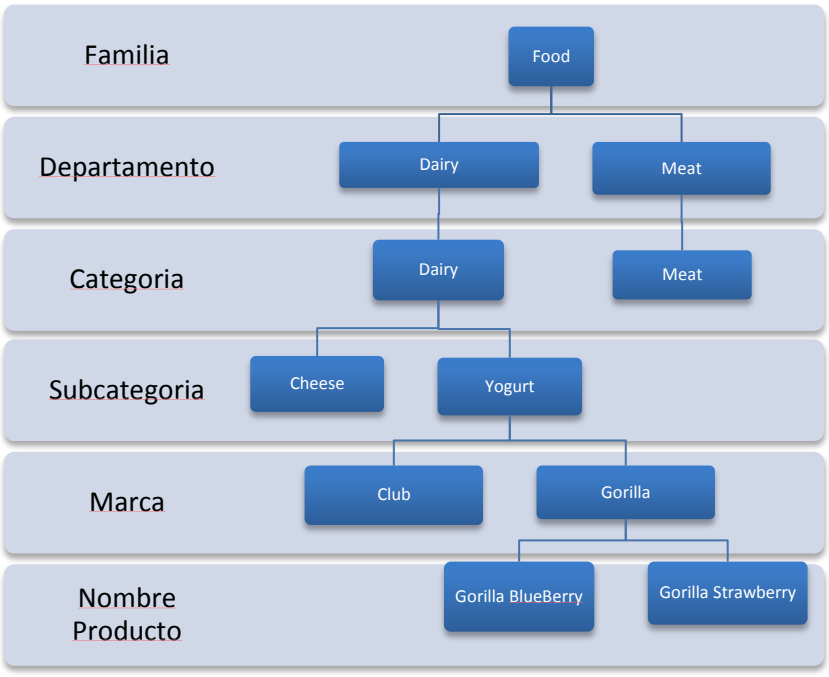


Figura 4. 6 - Jerarquía en la dimensión Producto - Food

Cada nivel de la Jerarquía tiene un número de elementos, en el caso específico de la Jerarquía Producto se tiene:

Nivel de la Jerarquía	**Número de elementos**
Familia	3 elementos
Departamento	23 elementos
Categoría	55 elementos
Subcategoría	102 elementos
Marca	512 elementos
Nombre de Producto	1560 elementos

Tabla 4. 1 - Elementos en la dimensión Producto

La tabla 4.1 indica que conforme se recorre los niveles de arriba hacia abajo, el número de elementos en cada nivel aumenta, ya que se está realizando una especialización de los datos, como consecuencia el dominio de datos crece.

Ejemplo:

Food → Dairy → Dairy → Yogurt →Gorilla →Gorilla BlueBerry

Esta ruta (path) tomada de la figura 4.6, indica que el producto: Gorilla BlueBerry pertenece a la marca: Gorilla, esta marca pertenece a la subcategoría: Yogurt, que a su vez

pertenece a la categoría: Dairy, esta pertenece al departamento: Dairy y finalmente este departamento forma parte de la familia: Food.

De este ejemplo es posible modelar la definición siguiente:

$$E_{nk} \in E_{nk-1} \in E_{nk-2} \in E_{nk-3} \ldots \in E_{nk-n}$$

Donde:

E_{nk}= Un elemento en un nivel k de la jerarquía de una dimensión.

La figura 4.5 y 4.6 muestran la jerarquía de la dimensión Producto como una estructura árbol, como primer nivel se tiene Drink y Food como nodos padre, a partir de ellos se derivan nodos hijo de segundo, tercer, cuarto, quinto y sexto nivel, dependiendo el número de niveles de cada jerarquía, conforme se especializan los datos el número de nodos hijo aumenta. Finalmente cuando se llega al último nivel de la jerarquía se encuentran los nodos finales, a los cuales se les llamarán nodos hoja.

Una vez comentada la estructura interna de los datos, veamos cómo definir la jerarquía en el modelo lógico.

Dentro de la definición de la dimensión Producto (Product) en el esquema XML, se indica la estructura siguiente:

```
<Dimension name="Product">
<Hierarchy hasAll="true" primaryKey="product_id" primaryKeyTable="product">

<Join leftKey="product_class_id" rightKey="product_class_id">
<Table name="product"/>
<Table name="product_class"/>
</Join>

<Level name="Product Family" table="product_class" column="product_family" uniqueMembers="true"/>

<Level name="Product Department" table="product_class" column="product_department"
uniqueMembers="false"/>

<Level name="Product Category" table="product_class" column="product_category" uniqueMembers="false"/>

<Level name="Product Subcategory" table="product_class" column="product_subcategory"
uniqueMembers="false"/>

<Level name="Brand Name" table="product" column="brand_name" uniqueMembers="false"/>

<Level name="Product Name" table="product" column="product_name" uniqueMembers="true"/>

</Hierarchy>
</Dimension>
```

La etiqueta <Hierarchy> envuelve la definición de los niveles de la jerarquía, en ella se indica la llave primaria de la tabla dimensión.

```
<Hierarchy hasAll="true" primaryKey="product_id" primaryKeyTable="product">
```

Como se mencionó al principio del capítulo, se está trabajando con un modelo copo de nieve, por lo tanto se tienen tablas con información detallada de una tabla de dimensión.

La etiqueta <Join> es necesaria para poder unir una tabla de dimensión con su tabla de detalle, a través de llaves foráneas.

```
<Join leftKey="product_class_id" rightKey="product_class_id">
<Table name="product"/>
<Table name="product_class"/>
</Join>
```

En las líneas anteriores se une la tabla de dimensión “product” con la tabla de detalle “product_class”.

Una vez realizada la unión de tablas, se define los niveles, por medio de la etiqueta <Level>, en la cual se da el nombre de la dimensión, el nombre de la tabla y columna del nivel que contiene los datos.

```
<Level name="Product Family" table="product_class" column="product_family" uniqueMembers="true"/>
.
.
.
<Level name="Product Name" table="product" column="product_name" uniqueMembers="true"/>
```

4.4 Proceso de solución manual a la pregunta de negocio

Anteriormente se mencionó que las formas de resolver la consulta de negocio planteada en el tema 3.2, era por medio de consultas SQL o por un análisis dirigido por el usuario.

En este tema se muestra la forma de realizar un análisis dirigido por el usuario, por el cual intentamos encontrar las situaciones de interés, que satisfagan la pregunta:

En qué nivel de la clasificación (jerarquía) de productos se tiene un decremento del 20% en ventas en los últimos dos años.

Que corresponde a la tendencia con niveles jerárquicos. De esta forma se demuestra que el proceso manual de búsqueda de las situaciones de interés en una base de datos multidimensional se vuelve una tarea laboriosa y costosa.

Hacemos uso de los datos del cubo definido en el esquema descrito en el tema 4.1, junto con un visor OLAP para poder explorar el modelo multidimensional, el visor OLAP elegido es JPivot, el cual es parte de la distribución de Mondrian y trabaja en ambiente Web por medio de Servlets y paginas JSP.
Como primer paso seleccionamos la dimensión de interés y el hecho de interés que se desea analizar, estos son “Producto”y “Store Sales” respectivamente (Figura 4.7 y 4.8).

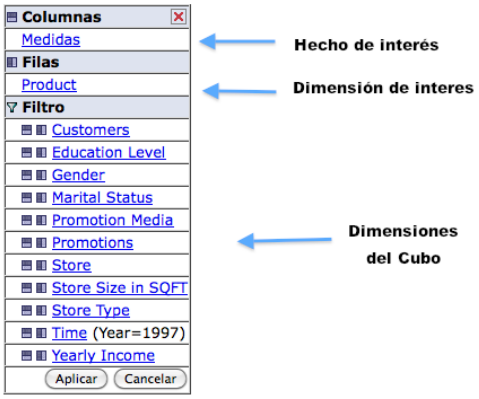


Figura 4. 7 - Selección de dimensión en el Visor OLAP

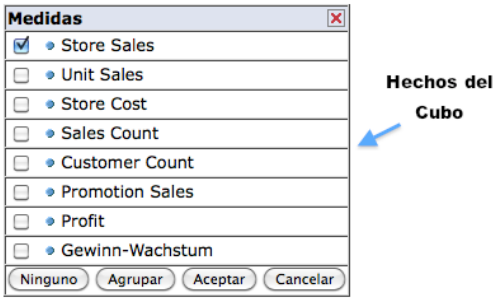


Figura 4. 8 - Selección de hechos en el visor OLAP

Una vez seleccionados, se muestra una tabla con los datos de interés. Esta se divide en filas y columnas. Las filas son las dimensiones y las columnas son los hechos (Tabla 4.2).
Las filas contienen los niveles de la Jerarquía de la dimensión Producto, como primer nivel se tiene "All Products" al dar click sobre el icono (+) de este elemento, comienza la exploración en el cubo, ya que se expanden y visualizan los nodos hijo de "All Product", los cuales son Drink, Food y Non-Consumable, que corresponden al segundo nivel de la jerarquía, a su vez cada uno de ellos contiene sus propios nodos hijo.

En este segundo nivel de la jerarquía se observa la cantidad en las celdas, que representan el cruce del elemento de la dimensión de interés y el hecho, para los años 1997 y 1998.

	Medidas	Medidas
Product	Store Sales	Store Sales
−All Products	565.238,13	1.079.147,47
+Drink	48.836,21	93.742,16
+Food	409.035,59	778.135,80
+Non-Consumable	107.366,33	207.269,51

Slicer: [Year=1997] Slicer: [Year=1998]

Tabla 4. 2 - Vista del segundo nivel de la jerarquía producto

Aplicando la ecuación de eficiencia grupal (ecuación1), se obtiene las siguientes cantidades para este nivel de la jerarquía (Tabla 4.3).

Producto (Familia)	Eficiencia
All Products	90.91 %
Drink	91.95 %
Food	90.23 %
Non - Consumable	93.04 %

Tabla 4. 3 - Calculo de eficiencias en el segundo nivel de la jerarquía

Se observa que es este segundo nivel (nivel familia), todos los elementos presentan eficiencias arriba del 90%. Por consiguiente es necesario seguir el análisis en niveles inferiores hasta encontrar algún elemento que tenga una eficiencia menor al 20%.

Se elige el nodo "Drink" para continuar con el análisis y se observa que este nodo tiene 3 elementos como hijos, que son: "Alcoholic Beverages","Beverages" y"Dairy". Cada uno con su hecho para el año 1997 y 1998.

	Medidas	Medidas
Product	Store Sales	Store Sales
-All Products	565.238,13	1.079.147,47
-Drink	48.836,21	93.742,16
+Alcoholic Beverages	14.029,08	27.107,99
+Beverages	27.748,53	52.403,74
+Dairy	7.058,60	14.230,43
+Food	409.035,59	778.135,80
+Non-Consumable	107.366,33	207.269,51

Slicer: [Year=1997] Slicer: [Year=1998]

Tabla 4. 4 - Vista del tercer nivel de la jerarquía producto

Nuevamente aplicando la ecuación de eficiencia grupal, se resume que para el tercer nivel (nivel departamento), las eficiencias superan el 90% (Tabla 4.5), de manera que seguimos explorando en un nivel inferior.

Producto (Departamento)	Eficiencia
Alcoholic Beverages	93.22 %
Beverages	88.85 %
Dairy	101.61 %

Tabla 4. 5 - Calculo de eficiencias en el tercer nivel de la jerarquía

Continuamos el análisis con la selección del nodo "Beverages", expandimos el nodo y encontramos que este elemento contiene 4 nodos hijo: "Carbonated Beverages", "Drinks", "Hot Beverages" y "Pure Juice Beverages" (Tabla 4.6).

	Medidas	Medidas
Product	Store Sales	Store Sales
−All Products	565.238,13	1.079.147,47
−Drink	48.836,21	93.742,16
−Alcoholic Beverages	14.029,08	27.107,99
+Beer and Wine	14.029,08	27.107,99
−Beverages	27.748,53	52.403,74
+Carbonated Beverages	6.236,35	11.518,33
+Drinks	5.642,29	11.386,09
+Hot Beverages	9.261,74	16.965,72
+Pure Juice Beverages	6.608,15	12.533,60
+Dairy	7.058,60	14.230,43
+Food	409.035,59	778.135,80
+Non-Consumable	107.366,33	207.269,51

Slicer: [Year=1997] Slicer: [Year=1998]

Tabla 4. 6 - Vista del cuarto nivel de la jerarquía producto

Se calcula la eficiencia grupal de cada elemento del nivel Categoría, se obtiene resultados que sobre pasan el 80%. Se observa que en este nivel el umbral de porcentaje se vió reducido con respecto al porcentaje en niveles superiores (Tabla 4.7). Sin embargo aún no se ha encontrado ninguna situación de interés.

Producto (Categoria)	Eficiencia
Carbonated Beverages	84.70 %
Drinks	101.80 %
Hot Beverages	83.18 %
Pure Juice Beverages	89.66 %

Tabla 4. 7 - Calculo de eficiencias en el cuarto nivel de la jerarquía

Se continúa analizando el siguiente nivel de la jerarquía, se elige el elemento “Drinks” y se observa que contiene 1 subelemento “Flavored Drinks” el cual contiene 5 elementos hijos, los cuales son: “Excellent”, ”Fabulous”, ”Skinner”, ”Token” y ”Washington” (Tabla 4.8).

	Medidas	Medidas
Product	Store Sales	Store Sales
−All Products	565.238,13	1.079.147,47
−Drink	48.836,21	93.742,16
+Alcoholic Beverages	14.029,08	27.107,99
−Beverages	27.748,53	52.403,74
+Carbonated Beverages	6.236,35	11.518,33
−Drinks	5.642,29	11.386,09
−Flavored Drinks	5.642,29	11.386,09
+Excellent	1.230,51	2.585,32
+Fabulous	1.483,46	3.367,67
+Skinner	1.127,48	2.438,75
+Token	713,59	1.604,82
−Washington	1.087,25	1.389,53
Washington Apple Drink	835,38	965,25
Washington Mango Drink	126,54	203,50
Washington Strawberry Drink	125,33	220,78
+Hot Beverages	9.261,74	16.965,72
+Pure Juice Beverages	6.608,15	12.533,60
+Dairy	7.058,60	14.230,43
+Food	409.035,59	778.135,80
+Non-Consumable	107.366,33	207.269,51

Slicer: [Year=1997] Slicer: [Year=1998]

Tabla 4. 8 - Vista del quinto nivel de la jerarquía producto

Se calculan las eficiencias en este nivel y se observa que existe un elemento cuyo valor está por debajo del 30%, el cual es la marca “Washington” (Tabla 4.9).

Producto (Marca)	Eficiencia
Excelent	110 %
Fabulous	127 %
Skinner	116.3 %
Token	124.9 %
Washington	27.78 %

Tabla 4. 9 - Calculo de eficiencias en el quinto nivel de la jerarquía

Por lo tanto resulta interesante seguir la exploración en el elemento “Washington”. Se encuentra que contiene 3 nodos hijo o también llamados nodos hoja (Por ser los últimos nodos de la jerarquía), calculamos sus eficiencias y encontramos una situación de interés.

El producto Washington Apple Drink presenta una eficiencia debajo del 20% con respecto al año anterior (Tabla 4.10), este resultado satisface los parámetros de la consulta planteada al principio del análisis.

Producto (Nombre)	Eficiencia
Washington Apple Drink	15.56 %
Washington Mango Drink	61.11 %
Washington Strawberry Drink	43.18 %

Tabla 4. 10 - Calculo de eficiencias en el sexto nivel de la jerarquía

De este resultado puede deducirse la siguiente expresión:

Una eficiencia "alta" en un nivel determinado de la jerarquía no implica que todos los elementos del nivel inferior también tengan una eficiencia alta (y viceversa).

Esto significa que el elemento "Drinks" presentó la mayor eficiencia a nivel Categoría, sin embargo la marca Washington no tiene una eficiencia alta.

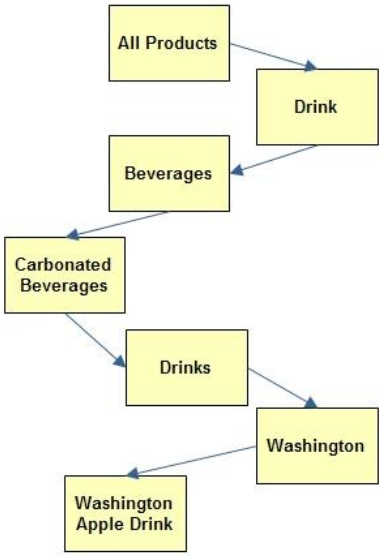


Figura 4. 9 - Ruta o path del punto de interés

Concluido el análisis de la ruta o path elegida (Figura 4.9), hemos solo encontrado una situación de interés, de todas las posibles rutas desde el nodo padre (All Products), por lo tanto se vuelve necesario analizar cada ruta o rama del árbol.

Claramente se ve, que el proceso de búsqueda requiere tiempo de análisis y es costoso. Esto significa que no es viable resolver preguntas de negocio de forma manual usando un visor OLAP.

4.5 Diseño del sistema visualizador de situaciones de interés con jerarquías

Como parte del desarrollo de una solución automática al proceso de respuesta a la consulta de negocio planteada en 3.2, se propone el desarrollo de un "sistema visualizador analítico de situaciones de interés con jerarquías", el cual muestre los elementos que cumplan con los parámetros de la consulta de negocio descrita, aprovechando los tipos de visualización estudiados en 3.4.2, que permiten realizar un análisis visual para explorar jerárquicamente los elementos de interés.

El sistema tiene como objetivo lograr aprovechar las herramientas actuales de minería de datos, herramientas de visualización y herramientas Web.

A continuación se describen los módulos y la arquitectura del sistema visualizador, además también del diseño del sistema haciendo uso de UML (Lenguaje Unificado de Modelado).

4.5.1 Módulos y requerimientos en la solución

El sistema propuesto consiste de tres módulos principales: "Interfaz de usuario", "Exploración y búsqueda en cubo de datos" y el "proceso de transformación de los datos". Como etapa previa es necesario describir de manera adecuada los cubos de datos, sus dimensiones, jerarquías y hechos dentro del esquema XML, además de contar con la base de datos que almacene el modelo multidimensional. Como resultado a la consulta de negocio ejecutada en el sistema, se obtiene el conjunto de visualizaciones descrito en 3.4.2.

La figura 4.10 presenta un diagrama de los módulos del sistema y los requerimientos necesarios para que este trabaje de forma adecuada.

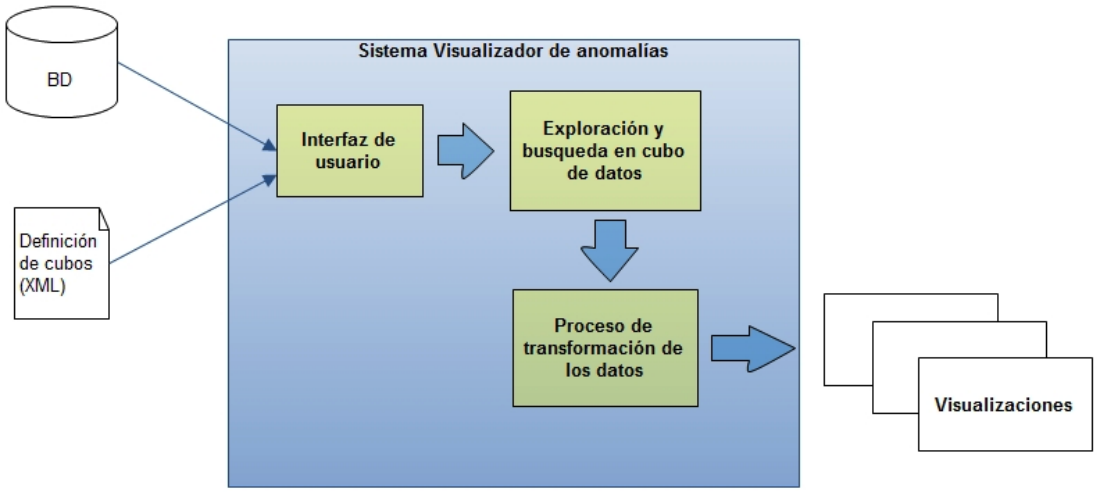


Figura 4. 10 - Módulos generales del sistema Visualizador

.

Requerimientos previos

Los requerimientos necesarios para el correcto funcionamiento del sistema son:

Base de datos

Es necesaria una base de datos que almacena la tabla de hechos y dimensiones. En este caso se hace uso de una base de datos en MySQL 5. Esta base de datos se describe en 4.1

Definición de cubos XML

Se trata del esquema XML en el que se define y modela los cubos de datos, dimensiones, jerarquías y hechos, tal como se describió en 4.2.1.

En resumen el sistema solo necesita que el usuario cuente con una base de datos que almacena el modelo multidimensional y un documento XML que defina los objetos multidimensionales.

Módulos en el sistema visualizador de anomalías

Interfaz de usuario

La interfaz de usuario tiene como objetivo mostrar los parámetros de la consulta de negocio, la carga de los cubos de datos, permitir la selección del espacio de búsqueda y finalmente enviar estos parámetros al módulo de exploración y búsqueda.

Exploración y búsqueda en cubos de datos

En este módulo se lleva a cabo la exploración en los cuboides y en la estructura jerárquica de la dimensión de interés que forman parte del espacio de búsqueda. Se selecciona y filtra aquellos elementos que cumplan con los parámetros definidos en la etapa previa y se entrega como resultado la lista de elementos que cumplieron al módulo de "transformación de los datos" (ver figura 4.10). Cada elemento de la lista esta descrito por la jerarquía de la dimensión de interés de la siguiente manera: *Nivel 1* . *Nivel 2* . *Nivel 3* . *Nivel n.*
La razón por la cual se recupera esta estructura es debido al API OLAP4J (OLAP for Java) utilizada para la consulta de cubos de datos.

Proceso de transformación de los datos

La transformación de los datos consiste en convertir la lista de elementos generada por el módulo de exploración a representaciones gráficas intuitivas y significativas como son las visualizaciones presentadas en 3.4.2. La transformación es un proceso creativo en el cual se asignan significados a los elementos gráficos, este proceso forma parte de la visualización de la información y tiene como finalidad establecer el modelo visual-espacial de los datos.

Este módulo convierte la lista entregada por el módulo de "exploración y búsqueda en cubos de datos" a una estructura de datos como es un árbol. Esta estructura árbol forma parte de modelo visual-espacial y su creación es necesaria para poder llevar a cabo las representaciones visuales (ver figura 4.11).

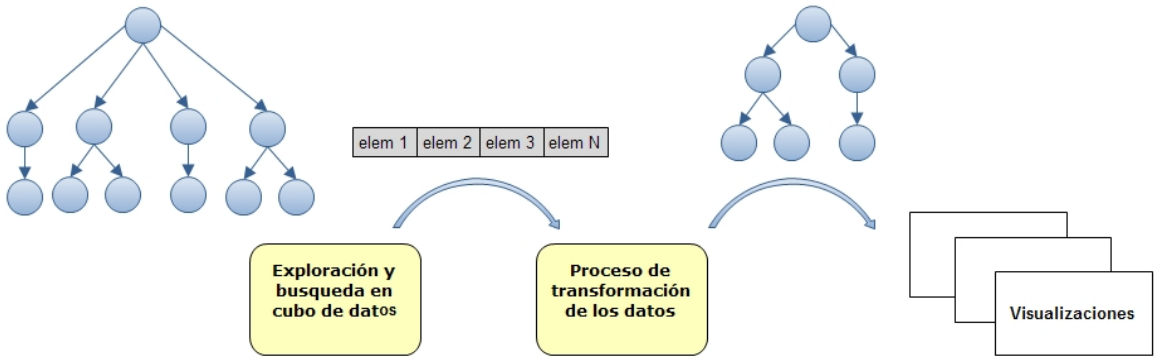


Figura 4. 11 - Proceso de transformación lista-Árbol

La razón de generar una estructura árbol en memoria es debido a que es necesario presentar a la herramienta de visualización (PowerChart) un documento XML cuyas etiquetas están estructuradas como padre-hijo. De manera que la forma considerada para presentar los datos es generando un árbol en memoria.

Como ejemplo del proceso de exploración y transformación supongamos que se recupera la siguiente lista de elementos que cumplieron con los parámetros definidos en la pregunta de negocio.

Washington Apple Drink
Hermanos Honey
Hermanos Broccoli
Nationeel Beef Jerky

Cada elemento a su vez esta descrito por la jerarquía completa a la que pertenece.

All Product . Drink.Beverages. Carbonated Beverages. Drinks. Washington. **Washington Apple Drink**
Food.Produce.Fruit.Fresh Fruit.Hermanos.**Hermanos Honey**
Food.Produce.Vegetables.Fresh Vegetables.Hermanos.**Hermanos Broccoli**
Food.Snack Foods.Dried Meat.Dried Meat.Nationeel.**Nationeel Beef Jerky**

Es en este punto donde el módulo "Proceso de transformación de los datos" crea la estructura árbol en memoria correspondiente a partir de la lista descrita.

4.5.2 Arquitectura del sistema Visualizador

El sistema será desarrollado como una aplicación Web, en la cual proponemos el uso de una arquitectura MVC (Modelo Vista Controlador), cuyas capas se describen a continuación:

Capa Vista: Es la capa encargada de generar las respuestas dinámicas que deben ser entregadas al usuario.

Capa Controlador: Todas las peticiones del usuario son dirigidas a los controladores.

Capa Modelo: La lógica de negocio de la aplicación incluyendo el acceso al cubo de datos y su manipulación esta encapsulada en esta capa.

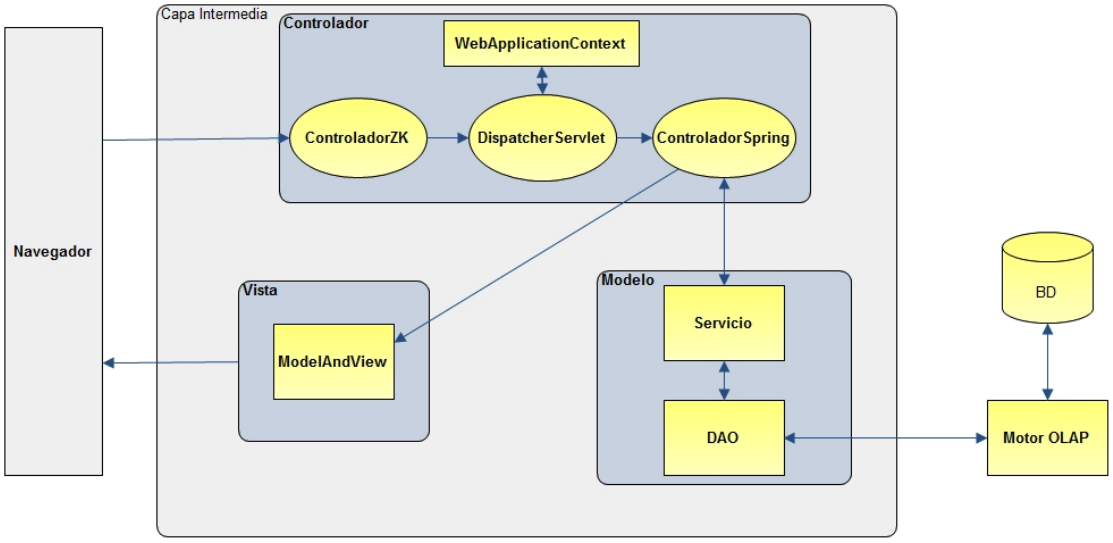


Figura 4. 12 - Arquitectura del sistema Visualizador

La figura 4.12 describe la arquitectura MVC planteada, la cual consiste de los siguientes módulos:

Capa Controlador

El módulo ControladorZK se encarga de generar la vista inicial del sistema, que será mostrada al usuario y de capturar los parámetros de la consulta de negocio.
El módulo ControladorSpring es el cerebro de la aplicación, recibe todas las peticiones y parámetros enviados del ControladorZK y determina las acciones a realizar para cada una de las peticiones. En resumen es el coordinador de todo el proceso.

Capa Modelo

El módulo Servicio es el módulo que se encargará de la lógica de negocio del sistema, esto es realizará tareas como son cálculos, ordenamientos, comparaciones entre otras.

El módulo DAO (Data Access Object) tiene la tarea de consultar los cubos de datos definidos en el motor OLAP.

Capa Vista

El módulo ModelAndView es el encargado de entregar una vista al usuario, dependiendo de la petición enviada al módulo ControladorSpring.

El siguiente paso es el diseño de sistema, usando UML. A continuación de muestra el diagrama de casos de uso, el diagrama de clases, el diagrama de secuencia y el diagrama de despliegue diseñados para el desarrollo del sistema Visualizador.

4.5.3 Diagrama de casos de uso - Consulta mapa de situaciones de interés

Se diseña el caso de uso consulta de mapa de situaciones de interés, el cual contiene 7 casos de uso y un solo actor, este se presenta en la figura 4.13 y su descripción es presentada en la tabla 4.11.

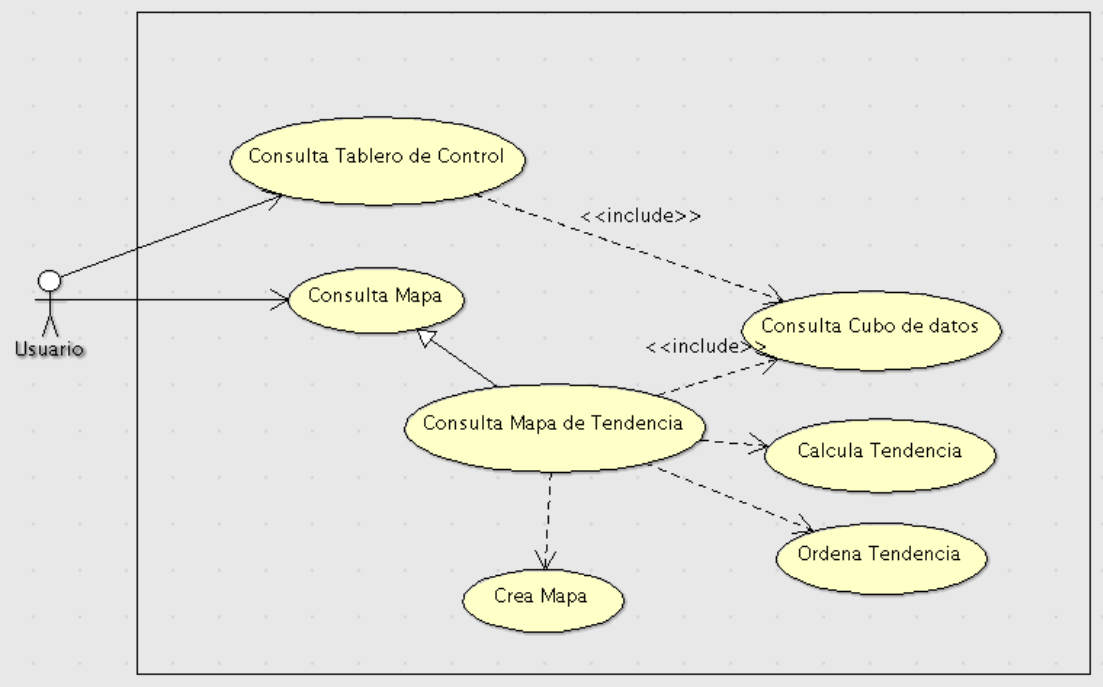


Figura 4. 13 - Diagrama de casos de uso consulta mapa

Nombre de caso de uso	Consulta Mapa de Situaciones de Interés (Tendencia con niveles jerárquicos)
Actores participantes	Usuario del sistema
Condiciones iníciales	Ingresar los parámetros de la consulta de negocio.
Flujo de eventos	El usuario ingresa y envía los parámetros de la consulta del cubo de datos, se calculan y ordenan los valores porcentuales de tendencias de los cuboides, finalmente se construye el mapa solicitado.
Condiciones de salida	El usuario recibe como respuesta a su consulta un mapa de situaciones de interés.
Requerimientos especiales	

Tabla 4. 11 - Descripción de caso de uso: consulta mapa

4.5.4 *Diagrama de Clases - Consulta mapa de situaciones de interés (Tendencia con niveles jerárquicos)*

El diagrama de clases del sistema se presenta en la figura 4.14, describe una organización MVC.

Consiste básicamente de clases controlador, interfaces de servicio, clases de servicio, clases DAO y clases de creación de documentos XML.

Capa Vista: Clase ModelAndView

Capa Controlador: ControladorZK, ControladorSpring

Capa Modelo: EficienciaServiceImpl, TableroServiceImpl, EficienciaDAO, TableroDAO

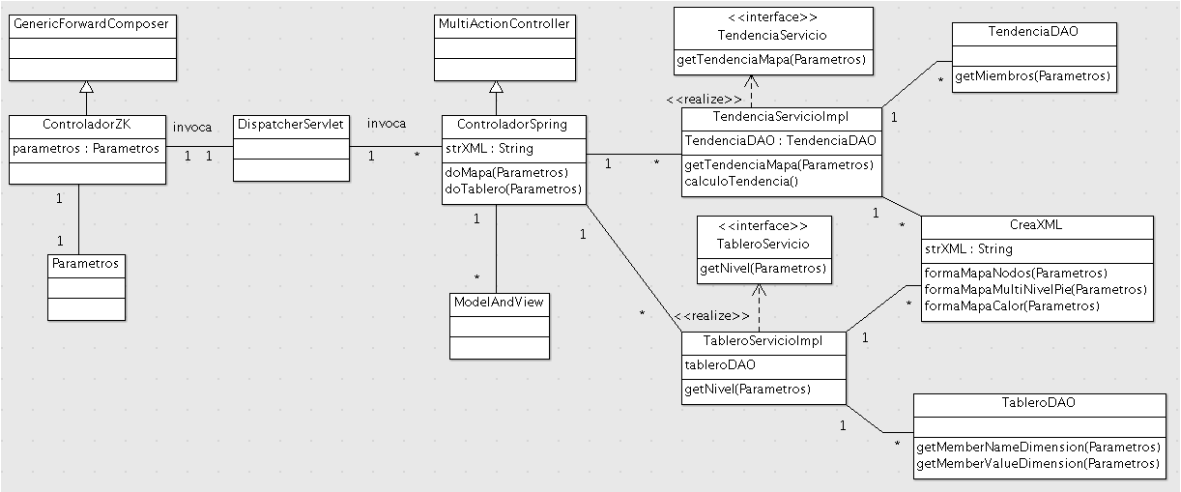


Figura 4. 14 - Diagrama de clases de la consulta mapa de situaciones de interés

4.5.5 *Diagrama de Secuencia - Consulta mapa de situaciones de interés*

El diagrama se secuencia de la figura 4.15 presenta la forma en que interactúan los objetos involucrados en la creación del mapa de situaciones de interés, a través de los mensajes en cada flecha.

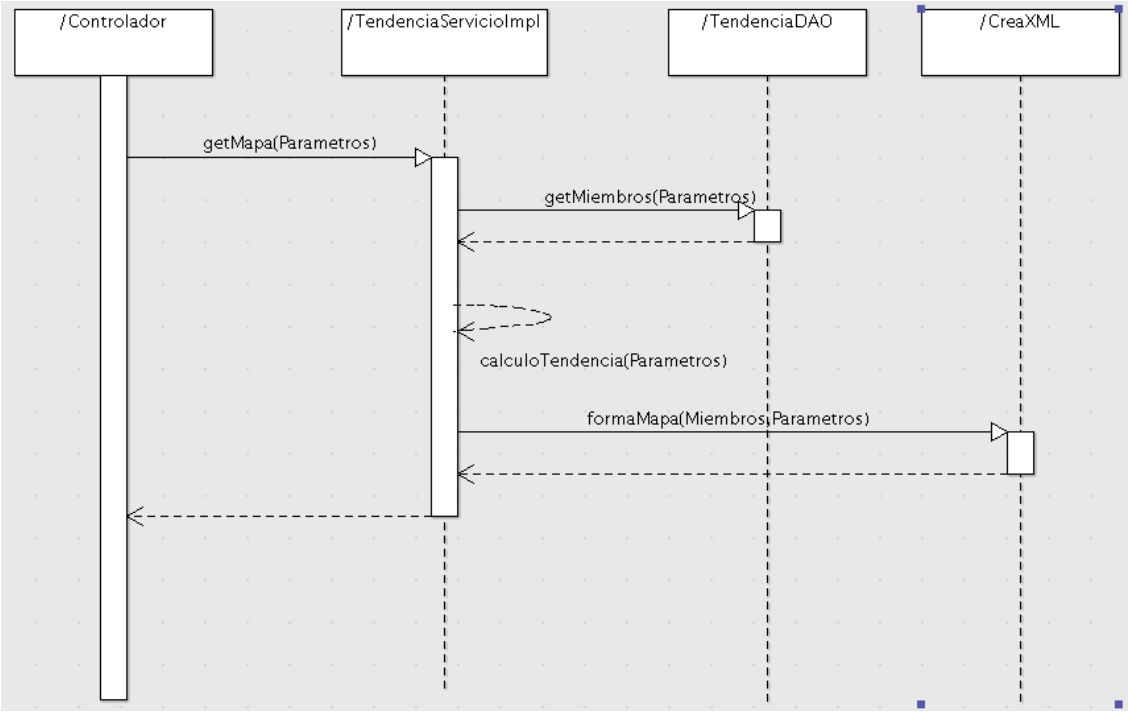


Figura 4. 15 - Diagrama de secuencia de la consulta mapa de situaciones de interés

4.5.6 Diagrama de despliegue del sistema visualizador (VisJ)

El diagrama de despliegue de la figura 4.16, muestra los componentes en tiempo de ejecución que conforman del sistema visualizador, se compone de:

Un navegador Web que accede a un servidor Tomcat. El servidor Tomcat a su vez accede a un Servidor OLAP que define el modelo lógico (Cubos) del sistema. EL servidor OLAP consulta al servidor de bases de datos MySQL que almacena el modelo físico de los datos.

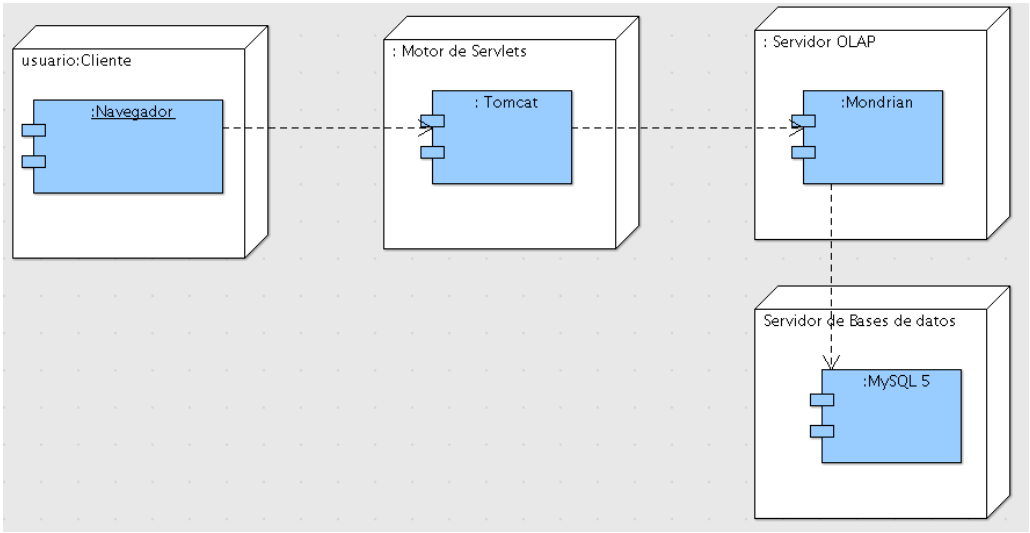


Figura 4. 16 - Diagrama de despliegue de la consulta mapa de situaciones de interés

4.5.7 Interfaz de usuario para la definición de la consulta de negocio

Una vez que los cubos de datos han sido construidos con N dimensiones, el sistema desarrollado permite la definición de la pregunta de negocio planteada en 3.2, visualizando los resultados de la exploración en los cubos de datos por medio de un "Mapa de situaciones de interés" y de "tableros de control". El sistema consiste de 4 módulos necesarios para la definición de la consulta de negocio, además también de un módulo de conexión a los cubos de datos que definen el esquema multidimensional.
A continuación se muestra los módulos del sistema necesarios para plantear la consulta de negocio.

Módulo de conexión a cubos de datos

Antes de definir una consulta de negocio, es necesario especificar el dominio en el cual el sistema trabajara en este caso en un dominio comercial o científico. El módulo de conexión de datos permite la comunicación con los datos definidos en los cubos diseñados.

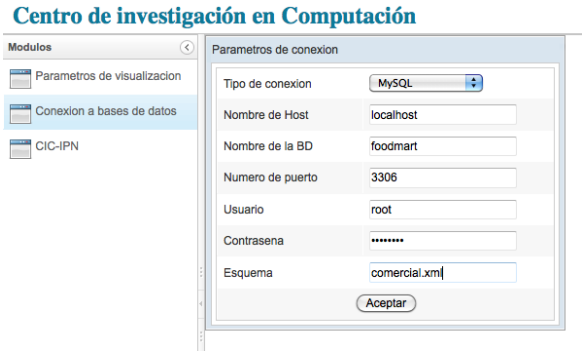


Figura 4. 17 - Módulo de conexión a cubos de datos

Módulo de cubos de datos

En este módulo se visualiza en una estructura árbol los niveles dentro de la jerarquía de las dimensiones definidas en el cubo de datos seleccionado. Por ejemplo el cubo "Sales" tiene como dimensiones "Store", "Time", "Product" entre otros, y cada dimensión presenta niveles de la jerarquía.

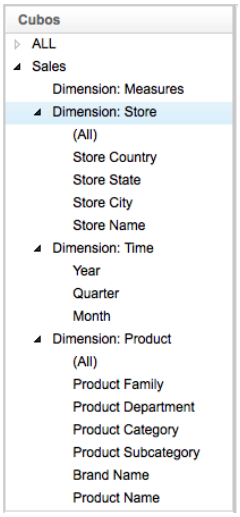


Figura 4. 18 - Módulo de cubos de datos

Módulo de características de elementos de interés

Es el módulo en el cual se define el escenario de la consulta, escenarios descritos en 3.2.1.

Figura 4. 19 - Módulo de características de elementos de interés

Módulo de parámetros de cubo de datos

En el módulo de parámetros de cubos de datos se especifica el espacio de búsqueda en el cubo de datos. Seleccionando la dimensión de interés y el dominio de las demás dimensiones.

Figura 4. 20 - Módulo de parámetros de cubo de datos

Módulo de parámetros de visualización

A través de este módulo de define el tipo de mapa deseado para visualizar los puntos de interés o anomalías.

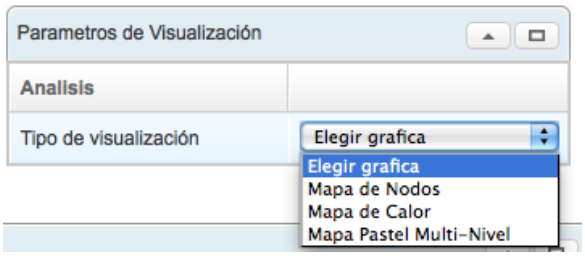


Figura 4. 21 - Módulo de parámetros de visualización

4.6 Resumen del capitulo

Se realizó la descripción del conjunto de datos comercial y científico necesarios para probar la aplicación, se analizó detalladamente el diseño y creación de los cubos de datos por medio del motor Mondrian. Se presentó un ejercicio de demostración que resuelve el tipo de pregunta planteada en 3.2 de forma manual y se presenta también el proceso de solución automática. Finalmente se realizó el diseño del sistema por medio de UML del cual se muestra los diagramas de casos de uso, de clases, de secuencia y de despliegue.

5

Pruebas y resultados de la aplicación VisJ

5 Pruebas y resultados de la aplicación VisJ

Como se mencionó al principio del capítulo 3 el sistema desarrollado tiene como tareas principales:

- Visualizar en un "mapa" las rutas de los puntos de interés o anomalías en la jerarquía de la dimensión de interés.
- Visualizar desde diferentes perspectivas el conjunto de anomalías encontradas.
- Navegación interactiva en el cubo de datos por medio de tableros de control.
- Aprovechar el sistema de percepción humano permitiendo así la exploración visual del cubo de datos.

En las siguientes secciones se presentan las características de los conjuntos de pruebas que servirán para la obtención de resultados y se demuestra que los objetivos del sistema visualizador de anomalías propuesto son alcanzados.

5.1 Conjunto de datos

La herramienta es ejecutada y probada sobre 2 dominios de datos. El primer dominio consiste de una base de datos de un supermercado descrita en 4.1.1 a la cual de ahora en adelante se le denomina "domino comercial" y el segundo dominio es una base de datos de tesis del Centro de Investigación en Computación del IPN (CIC-IPN) que describe la clasificación ACM (Association for Computing Machinery) de estas, analizada en 4.1.2 a la cual se le llama "dominio científico".

El cubo de datos del dominio comercial consiste de dimensiones como son: producto, almacén, tiempo, ubicación, cada una de estas con una jerarquía interna. Como medidas o hechos se tienen: unidades vendidas, ventas por almacén, costos de almacén, por mencionar solo algunas. La tabla de hechos contiene 251,395 registros mientras que el dominio científico está formado por la dimensión clasificación y tiempo. El conjunto de datos contiene 253 registros o tesis.

5.2 Mapa de situaciones de interés en un ambiente comercial y científico

Como se mencionó anteriormente se visualizará un "Mapa de situaciones de interés" como resultado de la consulta de negocio establecida por el usuario, en la cual se explora la lattice de las dimensiones que involucra la consulta de negocio.

De esta manera se plantea la consulta de negocio para los dominios comercial y científico. Supóngase que la consulta de negocio en un dominio comercial es la siguiente: "*saber en qué nivel de la clasificación (jerarquía) de productos se tienen bajos niveles de ventas, digamos abajo del 20% con respecto al año anterior*", como resultado visual se obtiene el mapa de la figura 5.1. En el cual se muestra la ruta de cada producto que presenta una anomalía, cada nivel de la jerarquía está representado por un color, siendo los nodos en

color rojo los puntos de interés. Ejemplo: el producto: “Washington Apple Drink” pertenece a la marca: “Washington”, la cual pertenece a la subcategoría: “Flavored Drinks” que pertenece a la categoría: “Drinks”, esta categoría pertenece al departamento “Beverages” y finalmente esta pertenece a la familia “Drink”. En resumen el producto “Washington Apple Drink” cumple con las características de los productos buscados.

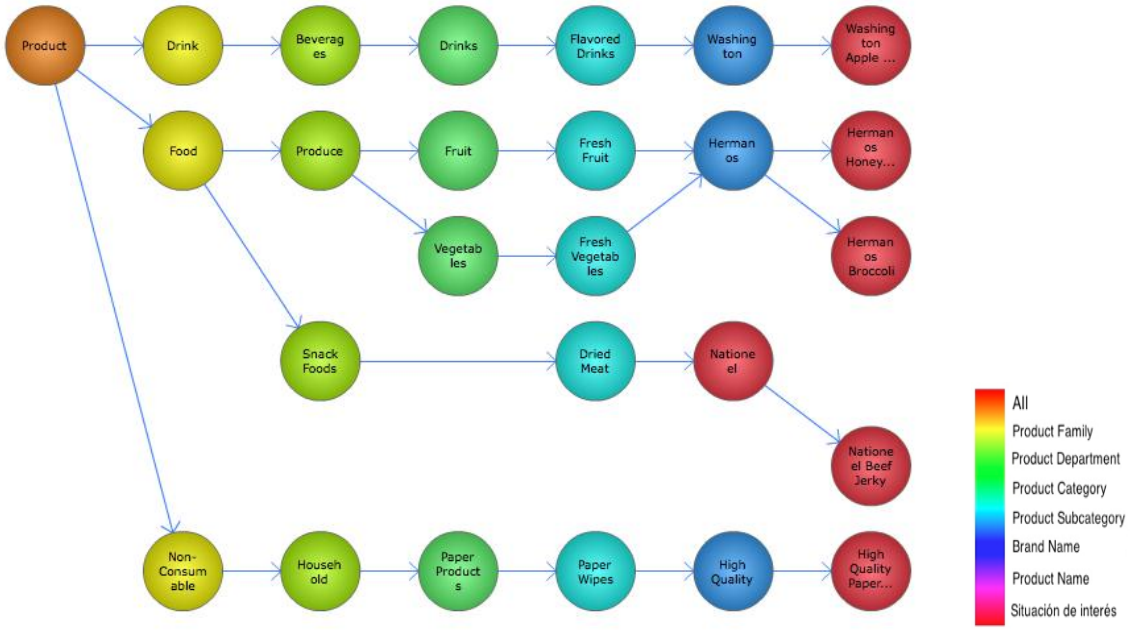


Figura 5. 1 - Mapa de nodos de situaciones de interés en un dominio comercial

Tal y como se mencionó, uno de los objetivos de la aplicación es poder trabajar en distintos dominios de datos, por lo cual se hace uso también del dominio científico y se plantea como ejemplo la siguiente consulta “saber en qué nivel de la clasificación ACM de tesis del CIC-IPN existe un incremento en número en 2 años determinados (2008, 2009)”, visualizando los 7 con mayor crecimiento, como respuesta visual a esta consulta se presenta el mapa de situaciones de interés de la figura 5.2, en el cual de la misma manera que en el dominio comercial se presenta la ruta de las anomalías encontradas que cumplen con los parámetros de la consulta.

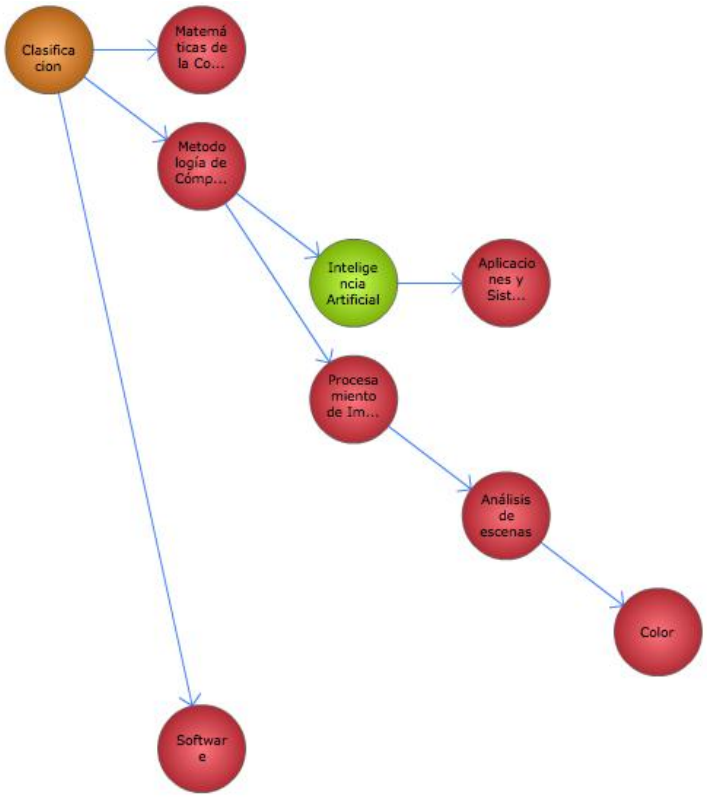


Figura 5. 2 - Mapa de nodos de situaciones de interés en un dominio científico

De la misma manera que en el dominio comercial, se obtiene como resultado un mapa con 7 anomalías con sus respectivas rutas. Por ejemplo la clasificación de nivel 4: "Color" presenta un crecimiento del 500% del año 2008 a 2009. Sin embargo no solo se encontró un crecimiento en el nivel 4, sino también en el nivel 3: "Análisis de escenas", en el nivel 2: "Procesamiento de análisis" y en el nivel 1: "Metodología de computo", todas estas anomalías en la misma ruta. Este resultado indica que la situación de interés comienza a presentarse desde el nivel 1 hasta el nivel 4.

La aplicación desarrollada cumple con el "cambio de perspectivas visuales" que define a una aplicación visual Analítica [Hanrahan, 2009], por lo que se ofrece también otros tipos de visualizaciones como son los estudiados en 3.4.2 para el mismo conjunto de anomalías. Los mapas alternos al mapa de nodos son: "Mapa de calor" figura 5.3 y el "Mapa pastel multi-nivel" figura 5.4.

Figura 5. 3 - Mapa de calor

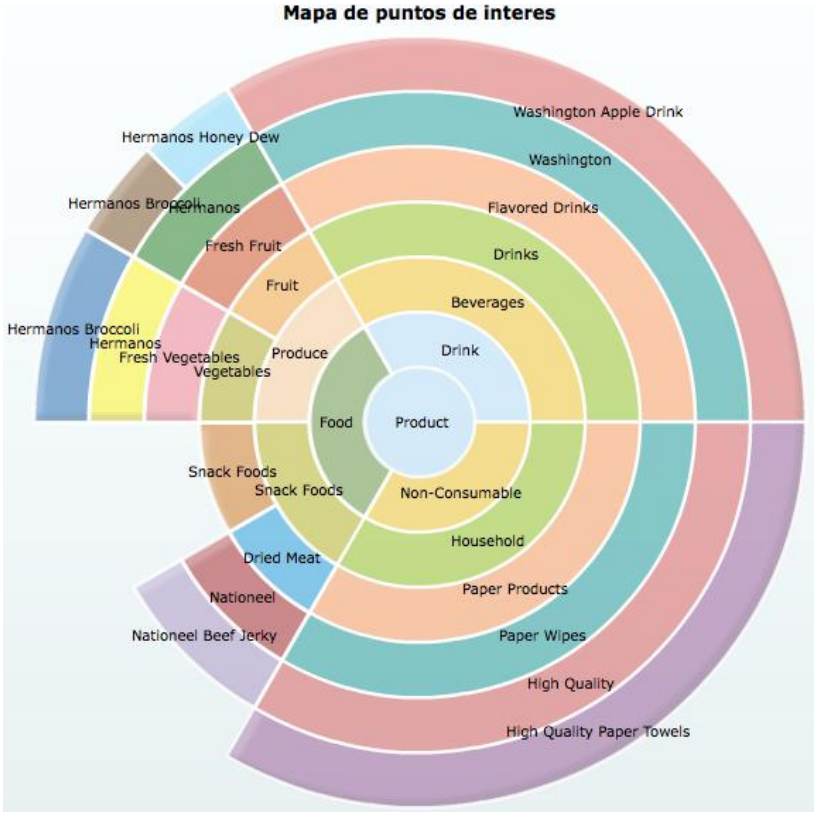


Figura 5. 4 - Mapa pastel multi-nivel

El mapa de calor representa cada ruta de la anomalía como un renglón, mientras que cada columna representa el nivel de la jerarquía de la dimensión de interés. Por ejemplo la ruta para alcanzar el producto "Hermanos Broccoli" es: Product → Food → Produce → Vegetables → Fresh Vegetables → Hermanos → Hermanos Broccoli.

El mapa Pastel-Multi Nivel representa el árbol de la jerarquía de la dimensión analizada en forma de pastel. Cada nivel del pastel representa un nivel, siendo el nivel interior el nivel de menor jerarquía y el nivel exterior el de mayor jerarquía.

5.3 Navegación usando tableros de control

Además de los mapas de situaciones de interés o anomalías, la aplicación ha sido diseñada para permitir al usuario una navegación por las dimensiones de forma visual e interactiva a través de tableros de control (Dashboard) los cuales muestran visualizaciones dinámicas en las dimensiones definidas en el cubo de datos que involucran operaciones OLAP (On-Line Analytical Processing) como son drill-down y roll up. Este tablero de control es ejecutado dentro de los mapas de situaciones de interés al seleccionar un nodo, una celda o un sub-pastel en los mapas de nodos, calor y pastel multi-nivel respectivamente. Por ejemplo al seleccionar el nodo "Food" en el mapa de nodos de la figura 5.1 se presenta la gráfica de ventas por año de cada departamento de la familia "Food" (Figura 5.5), equivalente a una operación drill-down al siguiente nivel de la jerarquía. Al seleccionar un departamento (Ej. Produce - 1998) se presenta de forma dinámica 2 visualizaciones en las dimensiones tiempo (time) y almacén (store), referentes a las ventas en cada trimestre de ese elemento (Figura 5.6) y las ventas por país de cada categoría del departamento (Figura 5.7). Al seleccionar una categoría se despliega un mapa geográfico ubicando las ventas por estado de la categoría (Figura 5.8) y al seleccionar un estado en particular se muestra las ventas por ciudad (Figura 5.9).

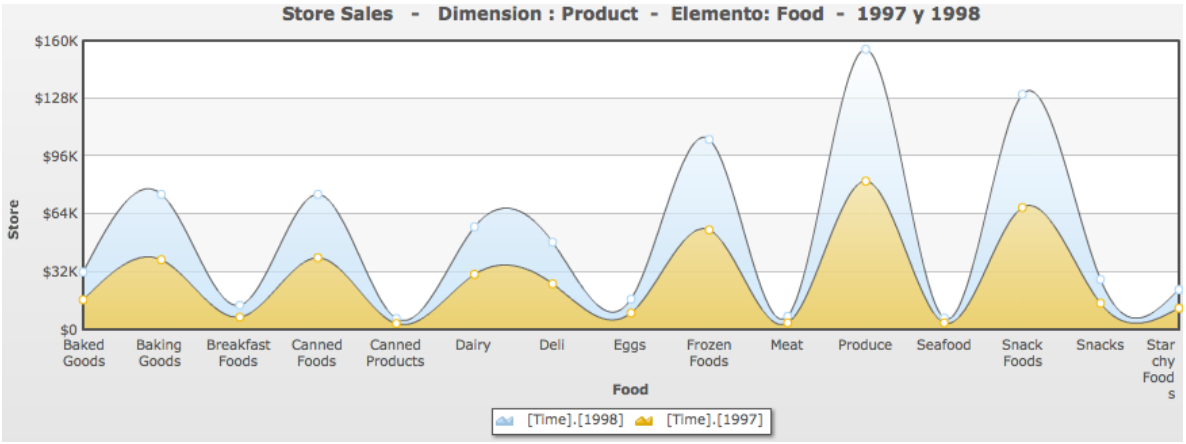


Figura 5. 5 - Tablero de control: Drill down sobre dimensión de interés

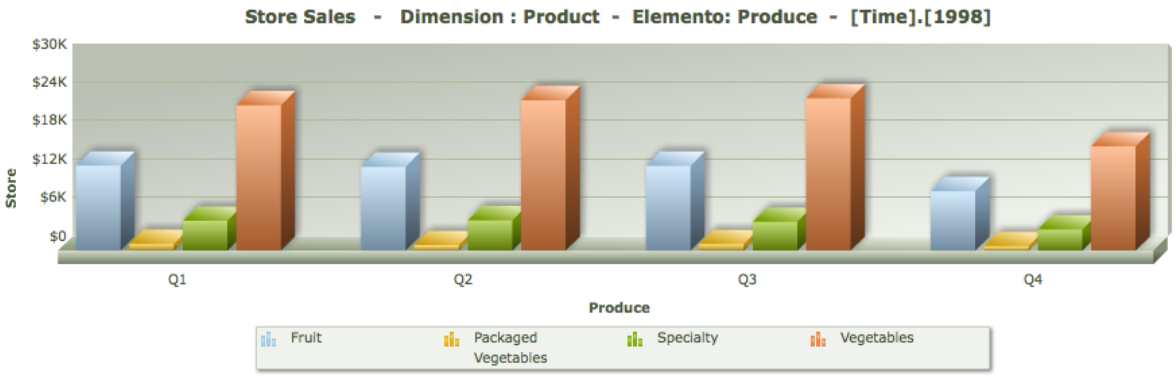


Figura 5. 6 - Tablero de control: Ventas por trimestre

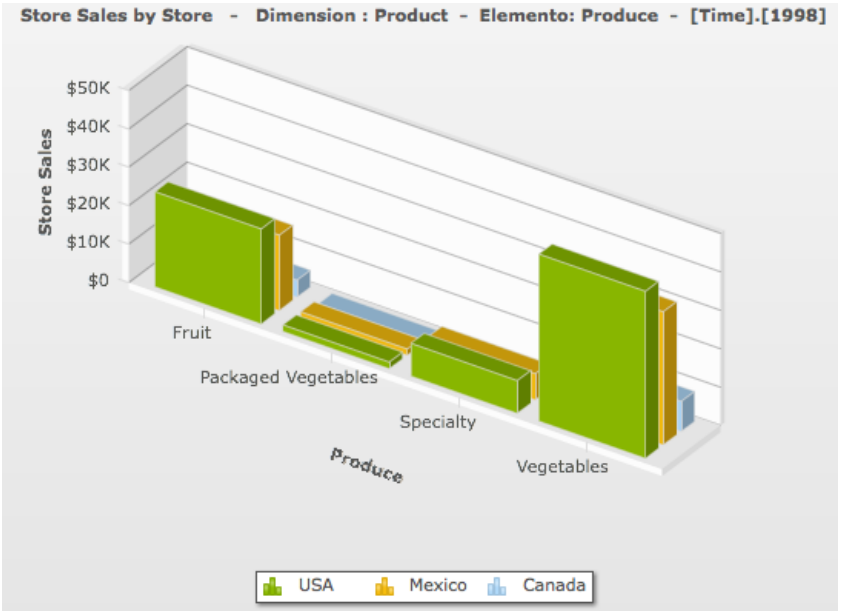


Figura 5. 7 - Tablero de control: Ventas por país

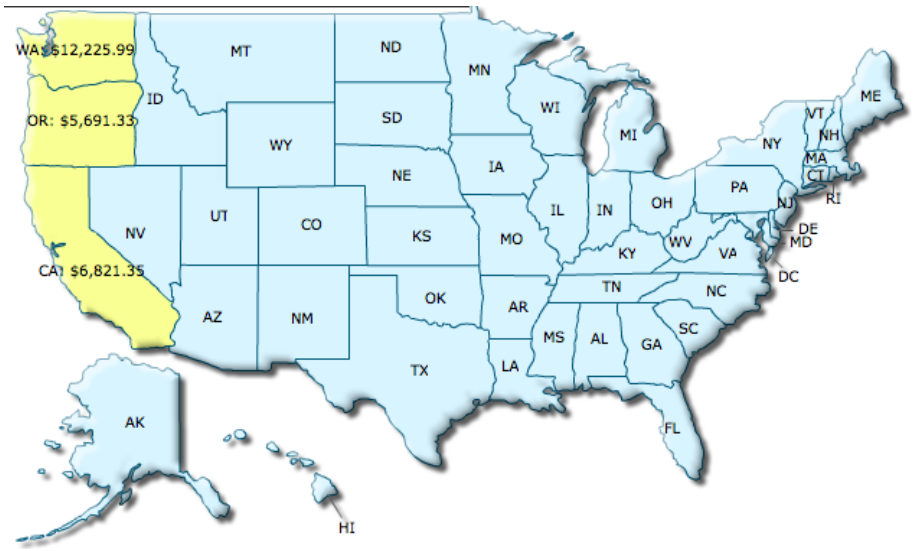


Figura 5. 8 - Tablero de control: Ventas por estado

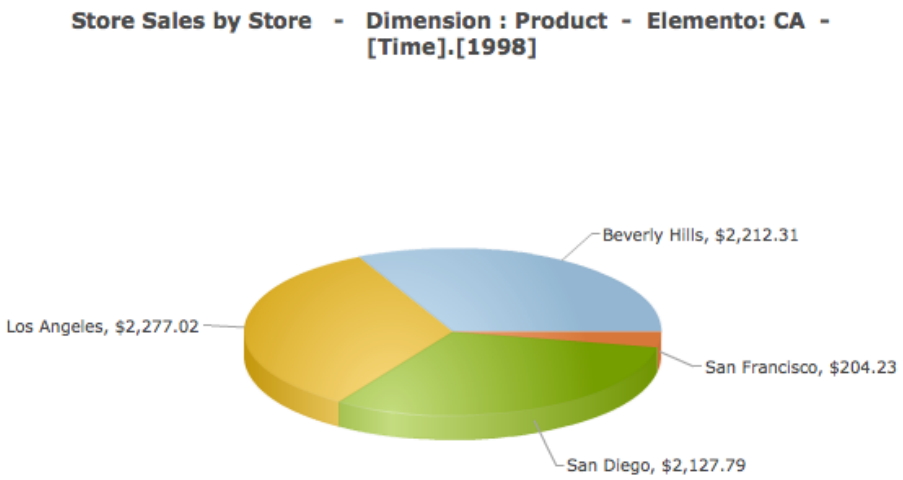


Figura 5. 9 - Tablero de control: Ventas por ciudad

5.4 Ambiente de pruebas

La herramienta ha sido desarrollada en lenguaje Java en su versión Java EE (Java Enterprise Edition), permitiendo su acceso desde cualquier navegador Web. La conexión a los cubos de datos se realiza por medio del API OLAP4J (OLAP for Java), mientras que las consultas a los cubos son escritas en lenguaje MDX (MultiDimensional eXpressions). Las tablas del esquema multidimensional son almacenadas en una base de datos MySQL 5, definiendo los cubos de datos en el motor ROLAP (Relational On-Line Analytical Processing) Mondrian. La base de datos y el motor ROLAP son ejecutados desde un equipo con sistema operativo MAC OS X con CPU a 2.4 Ghz Core 2 Duo, 4 GB de memoria y 360 GB de disco duro.
La interfaz de usuario en la cual se especifica los parámetros de la consulta ha sido implementada por medio del Framework ZK. El sistema tiene una arquitectura MVC (Modelo Vista Controlador), haciendo uso del Framework Spring.

5.5 Tiempo de respuesta

La prueba de tiempo de respuesta consiste en medir el tiempo en la exploración del cubo de datos, buscar los elementos que satisfacen los parámetros de una consulta de negocio, transformar la lista de elementos recuperados a una estructura árbol y posteriormente visualizar los resultados en una representación visual como son los mapas analizados en 3.4.2.

Se trabaja con la base de datos del dominio comercial, de manera que se crea un conjunto de datos superior al conjunto original de la tabla 5.1 y que fue estudiada en 4.1.1, replicando 2,3 y 4 veces el conjunto original para obtener un conjunto final de 1,005,580 registros y así medir el tiempo de respuesta del proceso entero.

La réplica de registros se lleva a cabo por medio de un proceso ETL (Extraction, Transformation and Loading) que se encarga de poblar la tabla de hechos (Figura 5.10), como resultado el número de registros incrementa (Figura 5.11).

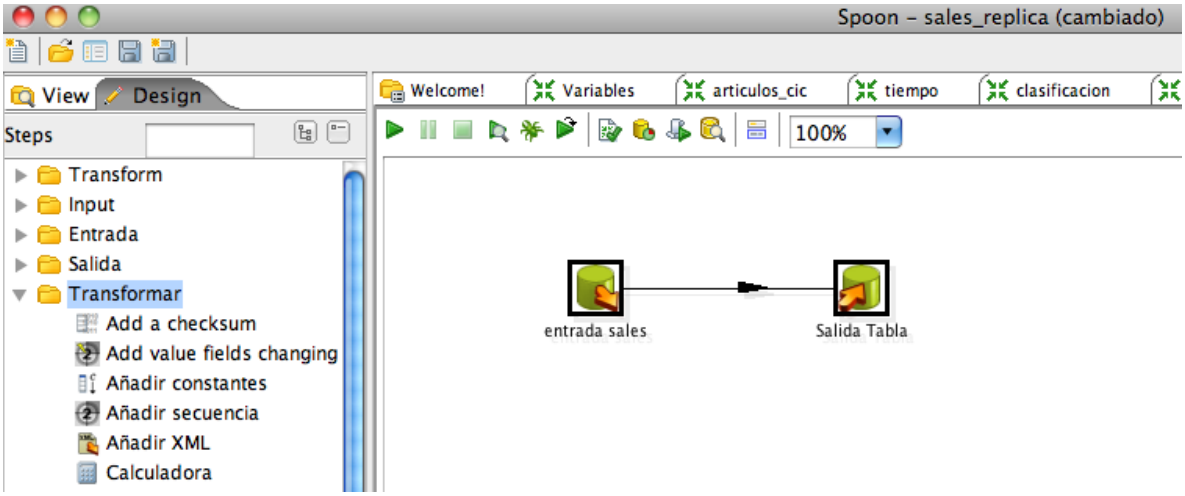


Figura 5. 10 - Proceso ETL

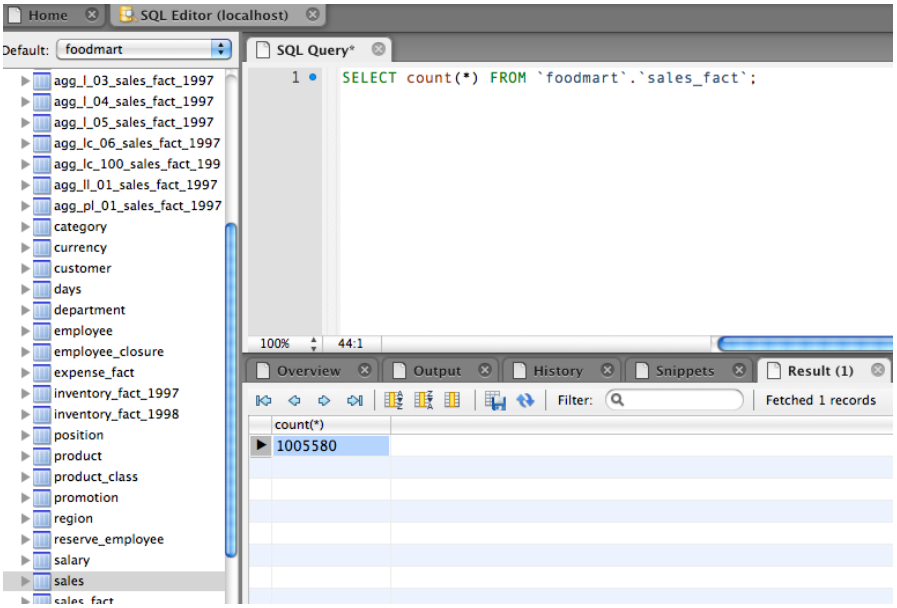


Figura 5. 11 - Número de registros en la tabla de hechos: sales_fact

Tabla	Registros
product	1,560
time_by_day	730
store	25
promotion	1,864
sales_fact	251,395

Tabla 5. 1 - Una tabla de hechos y 4 dimensiones

Para la realización de esta prueba se presentan 2 escenarios los cuales son:

a) Registros en disco

Consiste en la creación o cálculo del cubo de datos en tiempo de ejecución. Los registros que forman parte del cubo de datos son leídos del disco. Este escenario se presenta generalmente al iniciar la aplicación.

b) Registros en memoria cache

Una vez que el cubo de datos ha sido calculado, los registros que forman el cubo son almacenados en una memoria cache, este proceso es realizado por el motor ROLAP, de manera que es posible aprovechar este cubo para agregar nuevas dimensiones. (Ver 4.2.3)

Se ejecutan consultas MDX que involucran 2, 3 y 4 dimensiones del cubo de datos "Sales", estas son: "time", "product", "store", "promotion".
Las consultas a los cubos de datos se realizan por medio del lenguaje MDX cuya estructura es la siguiente:

```
select {[Measures].[parametros.getHecho()]} ON COLUMNS,
        {[parametros.getDimension()].[parametros.getNivel()].members} ON ROWS
                from [parametros.getCubo()]
where {parametros.getYear()};
```

Consultando como columnas a los hechos y como filas a las dimensiones. La palabra "from" indica el cubo de datos a consultar y la palabra "where" la rebanada del cubo o también conocida como slicer.

5.5.1 Resumen de pruebas

La tabla 5.2 presenta los resultados promedio al consultar el cubo de datos "Sales". Los tiempos de cada prueba son medidos en segundos (Figura 5.12 y 5.13).

Escenario	**251,395 registros**	**502,790 registros**	**754,185 registros**	**1,005,580 registros**
Registros en disco	11 seg.	24 seg.	38 seg.	50 seg.
Registros en memoria cache	.42 seg.	.95 seg.	1.38 seg.	2.3 seg.

Tabla 5. 2 - Resultados del tiempo de ejecución

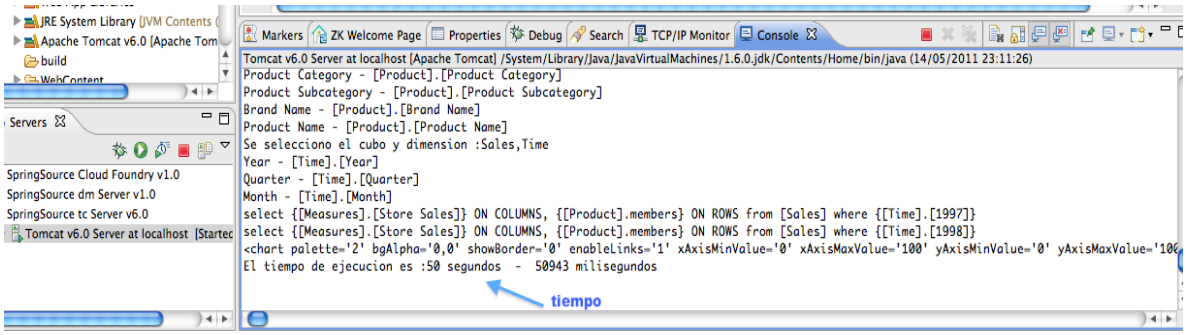


Figura 5. 12 - Tiempo de respuesta para 1, 005,580 registros en escenario “a)”

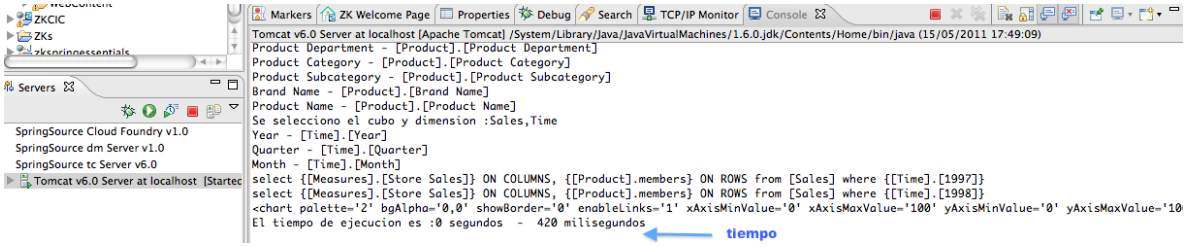
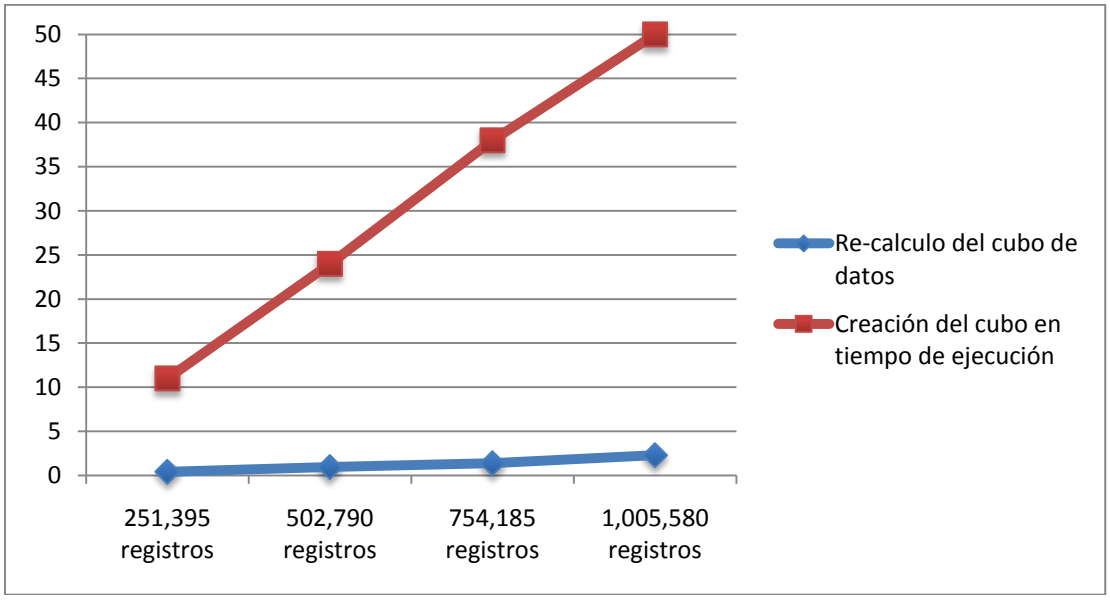


Figura 5. 13 – Tiempo de respuesta para 251,395 registros en escenario “b)”

A partir de los resultados de la tabla 5.2 se concluye que la aplicación tiene mayores tiempos de respuesta cuando se realiza la creación del cubo en tiempo de ejecución a comparación de los tiempos de respuesta en el recálculo del cubo de datos. Esto es debido en parte a la memoria cache del motor ROLAP la cual tiene un tamaño del 90% de la memoria de la máquina virtual Java JVM (Java Virtual Machine). Para este caso el tamaño asignado a la maquina virtual es de 128 Mb por lo tanto el tamaño de la memoria cache del motor es de aproximadamente 115.2 Mb.



Gráfica 5. 1 - Resultados del tiempo de ejecución

5.6 Resumen del capitulo

En este capitulo se obtuvieron los resultados del análisis de la pregunta de negocio planteada en 3.2, se obtienen visualizaciones llamadas mapas de situaciones de interés que muestran el conjunto de anomalías encontradas de acuerdo a los parámetros de la consulta de negocio. Cada anomalía consta de una ruta, que permite ver los elementos de los niveles de la dimensión involucrados. Se presentaron 3 tipos de mapas que proporcionan distintas perspectivas del conjunto de anomalías hallado, además de un tablero de control que permite una navegación interactiva sobre las dimensiones del cubo de datos. Finalmente, se realizaron una serie de pruebas de tiempo de respuesta al sistema VisJ, para conocer su comportamiento con distintas cantidades de registros.

6

Conclusiones y trabajos futuros

6 Conclusiones y trabajos futuros

6.1 Conclusiones

Los beneficios de la visualización de la información brindan un apoyo al proceso del descubrimiento del conocimiento, el cual es extraído de un gran volumen de datos. El uso apropiado de los elementos de la visualización como son colores, formas, tamaños y texturas en las graficas y mapas son de vital importancia para lograr alcanzar un conocimiento.

Los recientes avances en las tecnologías de la información hacen posible el desarrollo de herramientas de análisis de los datos aprovechando las características de la visualización. En este trabajo se han utilizado herramientas y lenguajes de uso libre como son Java, MDX, Mondrian, FusionChart entre otros. Logrando así construir una herramienta (de uso libre) de visualización de situaciones de interés sobre cubos de datos.

6.2 Metas alcanzadas

Al comienzo del proyecto fueron planteados los objetivos principales que debía cumplir este trabajo, los cuales han sido alcanzados de la siguiente manera.

- Se ha construido un sistema visualizador de situaciones de interés usando mapas de nodos, calor y pastel Multi-nivel, los cuales son ideales para la representación de datos con jerarquías. Logrando así visualizar anomalías en un conjunto de datos.

- Se ha definido tableros de control que permiten la exploración y análisis de forma interactiva al usuario sobre el conjunto de anomalías encontradas.

- El sistema visualizador de situaciones de interés desarrollado tiene la posibilidad de trabajar en distintos dominios de datos, siempre y cuando las dimensiones presenten una estructura interna. En resumen el sistema tiene la capacidad de generalización de dominios.

- Se demuestra que con el uso adecuado de herramientas y algoritmos de los campos de minería de datos, OLAP (On-Line Analytical Processing) y de la visualización de la información es posible crear soluciones computacionales que resuelvan el tipo de problemas planteados en este trabajo.

6.3 Aportaciones

La aportación de este trabajo es el desarrollo de un sistema visualizador que explota fuertemente las jerarquías en los datos.

Además del diseño de un paquete de clases que permite extender la funcionalidad del API (Application Programming Interface) de visualización utilizada, por medio de un paquete de clases, que facilitan la comunicación entre el sistema y el motor de visualizaciones comercial.

6.4 Trabajos futuros

6.4.1 Trabajar con dimensiones sin una jerarquía previamente establecida

Las bases de datos con las cuales se ha probado la herramienta desarrollada, tienen una característica en común y es que sus dimensiones tienen ya una jerarquía establecida, la cual ha sido previamente construida por los usuarios de negocio o administradores de la base de datos. Sin embargo si se cuenta con una base de datos sin una estructura interna en sus dimensiones, el sistema visualizador no trabajará de manera adecuada. Por esta razón es necesario entregar los datos al sistema con una jerarquía en las dimensiones.

De manera que el trabajo futuro a desarrollar consiste en un módulo de agrupamiento automático de valores numéricos y categóricos de un atributo, lo que dará como resultado una jerarquía virtual. Esto se logra a través de técnicas de discretización basadas en un análisis estadístico de la distribución de los datos. Por lo tanto la discretización reduciría el número de valores de un atributo, remplazando una gran cantidad de valores de un atributo por un número pequeño de etiquetas conforme de realiza el agrupamiento. Este proceso de discretización seria aplicado como un pre-procesamiento antes de la etapa de minado de nuestro sistema.

6.4.2 Cumplir con la visualización colaborativa

Otro trabajo futuro del sistema desarrollado es cumplir con el séptimo elemento que define a una aplicación visual analítica, recordando que el sistema visualizador ha cumplido con los seis restantes. La visualización colaborativa consiste en permitir que las visualizaciones puedan también ser creadas, modificadas y mejoradas interactivamente por los usuarios del sistema, a través de la red.

6.4.3 Ampliar el tipo de preguntas de negocio

Este trabajo consiste en que el sistema sea capaz de resolver preguntas como las planteadas en [Han, 2006], [Martínez, 2007], logrando así una variedad de consultas de negocio disponibles para el usuario.

6.4.4 Ampliar el dominio de visualizaciones

Finalmente, el sistema podría mejorar su interacción con el usuario expandiendo el tipo de mapas disponibles a otras visualizaciones por ejemplo en 3D.

6.5 Divulgación del trabajo de investigación

Al momento de terminar este trabajo se encuentra en elaboración un artículo, titulado "Mapa de anomalías por medio de jerarquías (Visualización reducida de anomalías de puntos de interés por medio de jerarquías)" el cual presenta un análisis de los elementos involucrados en la búsqueda de situaciones de interés sobre cubos de datos, aprovechando las características de la visualización de la información.

Bibliografía

[Agrawal, 1998] Agrawal, Rakesh., Megiddo, Nimrod.&Sarawagi, Sunita. (1998). Discovery-driven Exploration of OLAP Data Cubes.*EDBT '98 Proceedings of the 6th International Conference on Extending Database Technology: Advances in Database Technology. Lecture Notes In Computer Science,* 168-182.

[Bao, 2003] Bao, Xiaofeng., Li, Qing., North, Chris., Song, Chen. & Zhang, Jinfei. (2003). Dynamic Query Sliders vs. Brushing Histograms.CHI '03: CHI '03 extended abstracts on Human factors in computing systems. 147-153.

[Bayer, 1999] Bayer, Rudolf., Markl, Volker. & Ramsak, Frank. (1999). Improving OLAP Performance by Multidimensional Hierarchical Clustering. *IDEAS `99 International Symposium Proceeding. Database Engineering and Applications 1999.* 165-177.

[Bostock, 2010] Bostock, Michael., Heer, Jeffrey.&Ogievetsky, Vadim.(2010, June). ATour through the Visualization Zoo.*Communications of the ACM*, 53. Retrieved from http://queue.acm.org/detail.cfm?id=1805128

[Carbonell, 2007] Carbonell, Sergio. (2007). Producción de autores cubanos en las revistas sobre ciencias de la computación registradas en el Journal Citation Report en el período 1990-2005. *ACIMED Editorial de Ciencias Médicas*, 39. Retrieved from http://bvs.sld.cu/revistas/aci/vol15_05_07/aci03507.htm

[Chen, 2002] Chen, Chaomei., Lin, Xia., MacCain, Katherine. & White, Howard. (2002). Mapping Scientometrics (1981-2001). *Wiley Subscription Services*, 39. Retrieved from http://www.pages.drexel.edu/~cc345/papers/asis2002.pdf

[Chen, 2006] Chen, Chaomei. (2006).*Information Visualization Beyond the Horizon* (2nd Edition). Philadelphia: Springer.

[Chen, 2009] Chen, Zhibo. &Ordonez, Carlos. (*2009*).Exploration and Visualization of OLAP Cubes with Statistical Tests. *Knowledge Discovery and Data Minig. Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, 46-55.

[Chen, 2010] Chen, Chaomei. (2010, July/August). Information Visualization. *Interdisciplinary Reviews Computational Statistics*, 2. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/wics.89/full

[Chignell, 2005] Chignell, Mark H., McGuffin, Michael J.& Zhao, Shengdong. (2005).Elastic Hierarchies: Combining Treemaps and Node-Link Diagrams.*INFOVIS 2005 IEEE Symposium on. Informtion Visualization 2005*. 57-64.

[Deng, 2007] Deng, Shengchun., He, Zengyou.& Xu, Xaiofei. (2007).Attribute Value Weighting in K-Modes Clustering. *Information Technology (ITSim), 2010 International Symposium in*, 1531-1536.

[Feng, 2002] Feng, Jianlin. &Wang, Wei.(2002). Condensed Cube: An Effective Approach to Reducing Data Cube Size.*Data Engineering, 2002. Proceedings. 18th International Conference on*, 155-165.

[Fowler, 1999] Fowler, Martin.& Scott Kendall (1999). *UML Gota a Gota* (1a Edición). MAssachusetts: Pearson.

[Galati, 2006] Galati, David G.& Simaan, Marwan A. (2006). Automatic descomposition of time series into step, ramp, and impulse primitives.*Journal Pattern Recognition*, 39 (11), 2166-2174.

[Guzmán, 2008] Guzmán, Adolfo. &Martínez,Gilberto L. (2008). Antecumem, Prototipo de Herramienta para el Análisis de Datos con Cubos en Memoria Principal. *Conferencia Mundial sobre Tecnologias de la Informacion y Comunicaciones 2008*, Pachuca Hidalgo, México.

[Han, 2006] Han, Jiawei. & Kamber, Micheline. (2006). Data Warehouse and OLAP Technology: An Overview. In J. Gray & Microsoft Research (Eds.), *Data Mining Concepts and Techniques* (105-155). University of Illinois at Urbana-Champaign: Morgan Kaufmann.

[Hanrahan, 2009] Hanrahan, Pat., Mackinlay, Jock.&Stolte, Chris. (2009).*Selecting a Visual Analytics Application. Tableu.*

[He, 2006] He, Zengyou. (2006). Approximation Algorithms for K-Modes Clustering.*ICIC'06 Proceedings of the 2006 international conference on Intelligent computing: Part II. Lecture Notes In Computer Science*, 296-302.

[Korth, 2010] Korth, Henry F., Silberschatz,Abraham. &Sudarshan, S. (2010). *Database System Concepts* (6a Edition). USA: McGraw−Hill.

[Kriegel, 1996] Kriegel,Hans-Peter., Ester, Martin., Sander, Jorg.& Xu Xiaowei (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.*2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).*

[Li,2003] Li, Qing. &North, Chris. (2003). Empirical Comparison of Dynamic Query Sliders and Brushing Histograms. *INFOVIS'03 Proceedings of the Ninth annual IEEE conference on Information visualization*. 147-153.

[MacEachren, 2004] MacEachren, Alan M. (2004). Taking a Scientific Approach to Improving Map Representation and Design. In Guilford Press (Eds.), *How Maps Work: Representation, Visualization, and Design* (1-20). USA.

[Malinowski, 2006] Malinowski, E.& Zimányi, E. (2006, November). Hierarchies in a multidimensional model: From conceptual modeling to logical representation. *Journal Data & Knowledge Engineering - Special issue: WIDM 2004*, 59. Retrieved from http://portal.acm.org/citation.cfm?id=1225824

[MansmannFlorian, 2006] Mansmann, Florian. &Vinnik, Svetlana. (2006). From Analysis to Interactive Exploration: Building Visual Hierarchies from OLAP. Cubes.*Extending Database Technology - EDBT*. 496-514.

[MansmannSvetlana, 2006] Mansmann, Svetlana. & Scholl, Marc H. (2006). Extending Visual OLAP for Handling Irregular Dimensional Hierarchies.*In DaWaK 2006: Proceedings of 8th International Conference on Data Warehousing and Knowledge Discovery*.

[Martínez, 2007] Martínez, Gilberto L. (2007). Latices y otras estructuras para acelerar búsquedas en minería de datos. Tesis de Doctorado, Centro de Investigación en Computación CIC-IPN. DF México.

[Norman, 1990] Norman, DA. (1990). The Psychology of Everyday Actions. In Doubleday Business (Eds.*), The Design of Everyday Things* (34-53).London.

[Rozeva, 2007] Rozeva, Anna. (2007).Dimensional Hierarchies - Implementation in Data Warehouse Logical Scheme Design. *International Conference on Computer Systems and Technologies - CompSysTech'07*.

[Rozeva, 2008] Rozeva, Anna. (2008). Dimension Updates and Hierarchy Maintenance in OLAP Database.*International Scientific Conference Computer Science*, 914-919.

[Schulz, 2006] Schulz, Hans-Jorg. &Schumann, Heidrun. (2006). Visualizing Graphs - A Generalized View. *IV '06 Proceedings of the conference on Information Visualization. IEEE Computer Society Washington*, 163-173.

[Tegarden, 1999] Tegarden, David P.(January, 1999). Business Information Visualization. *Communications of the Association for Information Systems*, 1. Retrieved from http://www.acis.pamplin.vt.edu/faculty/tegarden/wrk-pap/BusInfoVizTut.pdf

[Vassiliadis, 1998] Vassiliadis, Panos. (1998). Modeling Multidimensional Databases, Cubes and Cube Operations. *SSDBM '98 Proceedings of the 10th International Conference on Scientific and Statistical Database Management*, 53-62.

[Wei, 2008] Wei, Zhi. (2008). A Cluster Algorithm Identifying the Clustering Structure. *2008 International Conference on Computer Science and Software Engineering CSSE*, 288-291.

Referencias Electrónicas

[RE01]http://www.pentaho.com/

[RE02] http://mondrian.pentaho.com/

[RE03] http://kettle.pentaho.com/

[RE04] http://www.zkoss.org/

[RE05] http://www.springsource.org/

[RE06] http://www.olap4j.org/

[RE07] http://www.fusioncharts.com/

[RE08] http://jquery.com/

[RE09] http://tomcat.apache.org/

[RE10] http://www.mysql.com/

[RE11] http://www.oracle.com/index.html

[RE12] http://www-958.ibm.com/software/data/cognos/manyeyes/

[RE13] http://www.tableausoftware.com/

[RE14] http://www.ugr.es/~rruizb/cognosfera/redes_2005/index.htm

[RE15] http://www.cs.uiuc.edu/~hanj/bk2/

[RE16] http://db-book.com/

Anexo A - Glosario

ACM (Association for Computing Machinery).- Es la primera sociedad científica y educativa acerca de la computación, publica revistas, journals y libros.

***API (Application Programming Interface*).-** Es un conjunto de clases que ofrece funcionalidades a otros sistemas.

Anomalía.- Desde el enfoque de minería de datos una anomalía es una situación de interés para el usuario que se distingue por ser una característica o un suceso no común

DAO (Data Access Object).- Es el objeto que tiene acceso a la base de datos, en un modelo MVC los objetos DAO se encuentran definidos e implementados en la capa "Modelo".

DML (Data Manipulation Language).- Es un lenguaje que permite a los usuarios llevar a cabo operaciones como: select, insert, delete y update.

DMQL (Data Mining Query Language). - Es un lenguaje de consulta en minería de datos.

Drill-down. - Es una operación OLAP que permite al usuario ver los datos en una dimensión con mayor grado de detalle.

ETL (Extraction, Transformation and Loading). - Es un proceso sobre las bases de datos que involucra la extracción, transformación y carga de datos.

Framework. - Es una estructura tecnológica de soporte con módulos de software desarrollados que establecen una arquitectura, que puede ser implementada por otros proyectos de software.

JDBC (Java Database Connectivity).- Es un conjunto de clases que permite la ejecución de operaciones SQL desde el lenguaje de programación Java.

Jedox PALO.- Es un servidor de datos multidimensional (MOLAP) orientado a celdas, específicamente desarrollado para almacenamiento y análisis de datos en hojas de cálculo.

JVM (Java Virtual Machine).- Es una maquina virtual capaz de interpretar y ejecutar instrucciones expresadas en código binario (Java bytecode).

Java EE (Java 2 Enterprise Edition).- Es parte de la plataforma Java, permite desarrollar y ejecutar software con una arquitectura de N capas distribuidas, las cuales son ejecutadas desde un servidor de aplicaciones.

MDX (MultiDimensional eXpressions).- Es un lenguaje de consultas para bases de datos OLAP.

MOLAP (Multidimensional On-Line Analytical Processing). **–** Esta implementación almacena los datos en estructuras optimizadas para el acceso multidimensional. Los datos son almacenados en arreglos.
Mondrian.- Es un motor ROLAP que forma parte de la suite de Pentaho.

MVC (Modelo-Vista-Controlador).- Se trata de una arquitectura para el desarrollo de sistemas de software Web que separa la interfaz de usuario, la lógica de negocio y la capa de datos.

OLAP (On-Line Analytical Processing**). -** El procesamiento analítico en línea tiene como objetivo agilizar la consulta en grandes cantidades de datos almacenados en estructuras multidimensional o cubos de datos.

OLAP4J (OLAP for Java).- Es un conjunto de clases que permiten la ejecución de operaciones OLAP desde el lenguaje Java.

OLTP (On-Line Transaction Processing). - Es un tipo de sistema que facilita la administración de aplicaciones transaccionales.

RDBMS (Relational Database Management System). - Es un sistema administrador de bases de datos relacionales, ejemplo: MySQL, Oracle, SQL Server.

ROLAP (Relational On-Line Analytical Processing). - Se trata de la implementación OLAP sobre motores de bases de datos relacionales.

Roll-up. - Es una operación OLAP que permite al usuario ver los datos en una dimensión con menor grado de detalle.

SGBD (Sistema de gestión de bases de datos).- Es un aplicación que tiene como función crear, administrar y servir como interfaz entre la base de datos y el usuario.

SQL (Structured Query Language).- Es un lenguaje declarativo de acceso a bases de datos relacionales.

Spring.- Es un framework de código abierto para el desarrollo de aplicaciones Java.

UML (Unified Model Language).- Es un lenguaje de modelado de sistemas de software usado para el análisis y diseño de sistemas.

XML (eXtensible Markup Language).- Es un metalenguaje de etiquetas desarrollado por el W3 Consortium que permite la descripción de la información contenida en Internet a través de estándares y formatos comunes.

ZK.- Es un framework de aplicaciones Web desarrollado en Java que permite el desarrollo de interfaces de usuario.

Anexo B - Manual de usuario de VisJ

1.- Introducción

El sistema visualizador de anomalías por jerarquías VisJ es una aplicación que permite analizar visualmente los datos almacenados en cubos de datos por medio de mapas de situaciones de interés. Este análisis es llevado a cabo al plantear una pregunta de negocio en un dominio determinado de datos (comercial, científico, institución). La pregunta de negocio se plantea alrededor de un objeto de interés y su comportamiento en el tiempo. La base de datos sobre las cuales se realiza la búsqueda de objetos de interés contiene miles o millones de registros. El tipo de pregunta que se plantea en esta aplicación tiene como nombre "Tendencia con niveles jerárquicos".

Tendencia con niveles jerárquicos

Cuando las dimensiones de interés presentan una estructura interna llamada también jerarquía que describe la granularidad de los datos, es necesario plantear una consulta que involucre ambos conceptos: Tendencia-Jerarquía. De modo que sea posible encontrar el comportamiento de los elementos de interés en cualquier nivel de la jerarquía en un periodo de tiempo.

Definición

La tendencia con niveles jerárquicos se refiere a localizar un conjunto de elementos (hechos) que presenten un comportamiento creciente, decreciente o constante a través del tiempo. La localización se realiza en cualquier nivel de la jerarquía de una dimensión, en un periodo de tiempo determinado.
El crecimiento o decremento es determinado comparando los hechos de interés entre 2 cuboides de datos, correspondiente al mismo nivel de la misma dimensión de interés y al mismo periodo de tiempo.

Ejemplo

Un ejemplo de esta pregunta es:

"En una empresa de venta de productos se desea saber en qué nivel de la clasificación (jerarquía) de productos se tienen bajos niveles de ventas, digamos abajo del 20% con respecto al año anterior"

La figura 1.A da un aspecto visual de los 2 cuboides que se comparan, correspondientes a distintos años de ventas.

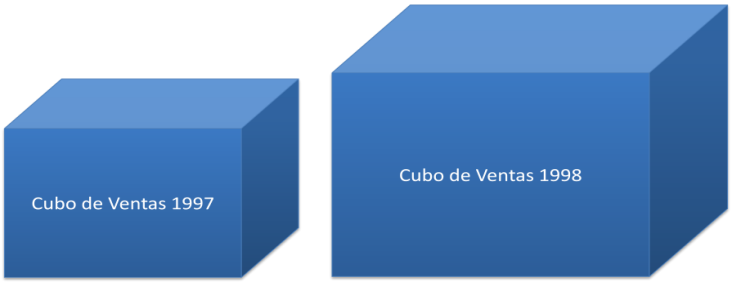


Figura 1.A- Comparación de 2 cubos de datos

El porcentaje de crecimiento o decremento se puede definir matemáticamente como:

$$Crecimiento/Decremento = 100 \times \frac{(Cubo\ 2 - Cubo\ 1)}{Cubo\ 1} \qquad (ecuación\ 1)$$

Donde:

Cubo 1.- El cubo de referencia en la comparación
Cubo 2.- El cubo en el cual se desea saber la situación de interés.

Escenarios de tendencia con niveles jerárquicos

La pregunta de tendencia con niveles jerárquicos puede ser planteada de diferentes maneras, se puede consultar por crecimientos o decrementos (tipo de tendencia) en los hechos de los elementos de la dimensión de interés, indicando la cantidad de elementos, un rango o un porcentaje, además de indicar la unidad de tiempo (mes, trimestre, año).
Una revisión o modelado de las posibilidades de la pregunta de tendencia con niveles jerárquicos se observa en la figura 1.B.

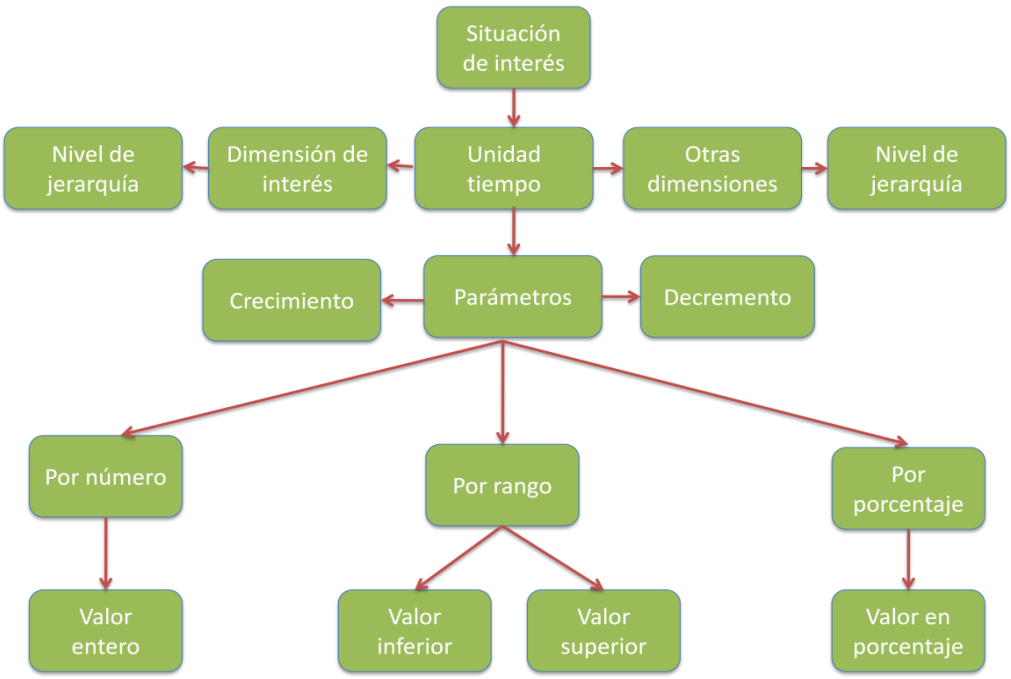


Figura 1. B - Escenarios de la consulta

Para indicar los parámetros de la pregunta se inicia con la selección de nodos en la parte superior, hasta llegar a los nodos en el último nivel, de esta forma se especializa la pregunta, eligiendo una ruta que la describa.

Esto es, si se desea plantear la consulta:

"En una empresa de venta de productos se desea saber en qué nivel de la clasificación (jerarquía) de productos se tienen bajos niveles de ventas, digamos abajo del 20% con respecto al año anterior".

La forma de describirla es:

1.- Seleccionar el nodo "Decremento".
2.- Seleccionar el nodo "Por porcentaje".
3.- Seleccionar el nodo "Valor en porcentaje"

Es posible especificar el espacio de búsqueda de los elementos de interés, indicando un valor y nivel especifico de la jerarquía en la dimensión de interés y dimensiones alternas, además de seleccionar la unidad de tiempo (día, mes, año). De esta manera una variante a la pregunta de eficiencia seria:

¿En qué departamentos de la Familia de productos "Comida" se obtuvo un incremento en ventas (50%), en las sucursales de California, en el primer trimestre del año 1998 con respecto al primer trimestre del año 1997?

2.- Objetivo del manual

Este manual tiene como finalidad explicar el funcionamiento del sistema visualizador de anomalías con jerarquías VisJ a nivel de administración como a nivel usuario.

Nivel administrativo

- Definir los cubos de datos
- Inicialización del software

Nivel de usuario

- Definir los parámetros de conexión a base de datos
- Definir y ejecutar una consulta de negocio
- Revisar los mapas de anomalías que el sistema da como resultado
- Revisar el tablero de control que permiten navegar en los puntos de interés.

3.- Requerimientos y acceso a la aplicación

3.1 Software necesario

1) Maquina virtual de Java (Java Runtime Edition 6).
2) TOMCAT 5.5 o superior.
3) Navegador de Intenet.
4) Instalar el archivo VisJ.war en la carpeta webapps dentro del directorio de Tomcat.

3.2 Acceso a la aplicación

1) Inicializar el servidor Tomcat

Windows XP, Vista, 7
C: \TOMCAT_HOME\bin\startup.bat

Mac OS X
home/TOMCAT_HOME/bin/startup.sh

2) Introducir **http://localhost:8080/VisJ/zk** en su navegador. (El número de puerto dependerá de la configuración previa del servidor).

4.- Descripción de la interfaz de usuario

Al iniciar la aplicación la pagina de inicio es la presentada en la figura 4.A

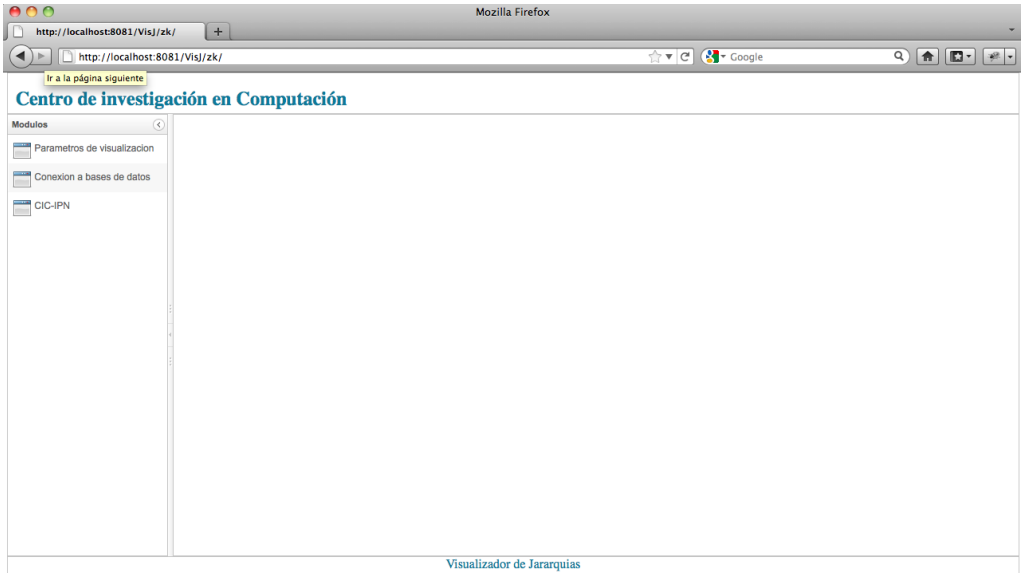


Figura 4.A – Vista inicial de la aplicación

Como primer paso es necesario establecer el conjunto de datos sobre el cual se realizara el análisis. Por lo tanto se accede al módulo “Conexión a bases de datos” (Figura 4.B) en el

cual se definen los parámetros de conexión a la base de datos que almacena los datos, y se registra el esquema XML que define el modelo multidimensional.

Parámetros de conexión	**Descripción**
Tipo de conexión	Se define el tipo de base de datos al cual se desea conectar: MySQL, Oracle.
Nombre de Host	Es el nombre del servidor que almacena la base de datos.
Nombre de la BD	Nombre de la base de datos.
Número de puerto	Número del puerto por el cual se realiza la conexión a la base de datos.
Usuario	Es el usuario que tiene permisos de conexión y consulta sobre la base de datos.
Contraseña	Es la contraseña del usuario de la base de datos.
Esquema	Es el nombre del esquema que almacena el modelo multidimensional.

Tabla 4.A – Parámetros de conexión

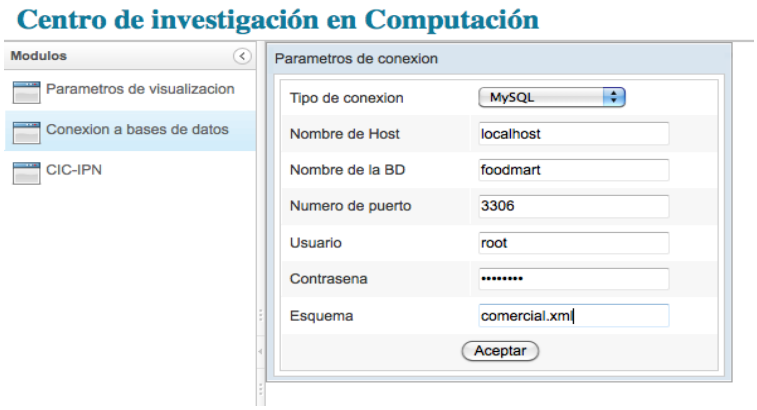


Figura 4.B – Vista del módulo de conexión

Al ser enviados los parámetros, el sistema se conecta a la base de datos que almacena la información. Se selecciona el módulo “Parámetros de visualización” y se observa que del lado izquierdo aparece un navegador de jerarquías (tipo árbol) que define las dimensiones y niveles de los cubos de datos definidos en el esquema multidimensional.

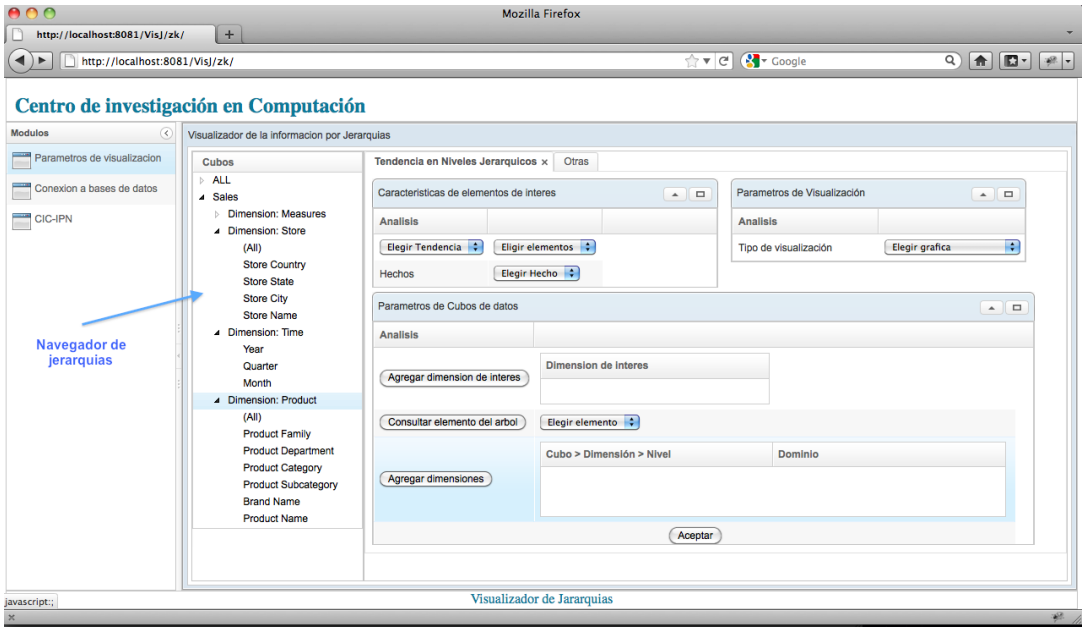


Figura 4.C – Vista de la aplicación después de conectarse a la BD

Una vez que los cubos de datos han sido cargados y visualizados, la aplicación permite la definición de la pregunta de negocio planteada, visualizando los resultados de la exploración en los cubos de datos por medio de un "Mapa de situaciones de interés" y de "tableros de control". El sistema consiste de 3 módulos necesarios para la definición de la consulta de negocio. A continuación se muestra los módulos del sistema necesarios para plantear la consulta.

Módulo de características de elementos de interés

Es el módulo en el cual se define el escenario de la consulta, escenarios descritos en la introducción de este documento.

Figura 4.D - Módulo de características de elementos de interés

Módulo de parámetros de cubo de datos

En el módulo de parámetros de cubos de datos se especifica el espacio de búsqueda en el cubo de datos. Seleccionando la dimensión de interés y el dominio de las demás dimensiones.

Figura 4.E - Módulo de parámetros de cubo de datos

Módulo de parámetros de visualización

A través de este módulo de define el tipo de mapa deseado para visualizar los puntos de interés o anomalías.

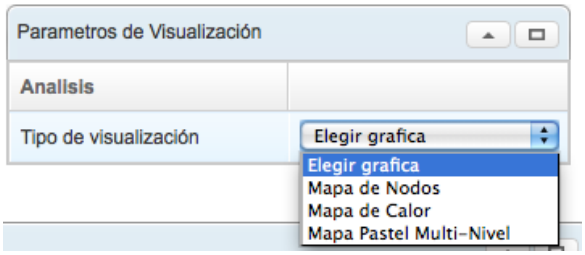


Figura 4.F - Módulo de parámetros de visualización

5.- Descripción del conjunto de datos de prueba

El sistema tiene como característica la generalidad lo cual significa que es posible trabajar en cualquier dominio de datos siempre y cuando se hayan definido previamente y correctamente el esquema multidimensional en un documento XML (eXtensible Markup Language).
La herramienta es ejecutada y probada sobre 2 dominios de datos. El primer dominio consiste de una base de datos de un supermercado a la cual se le llamará "domino comercial" y el segundo dominio es una base de datos de tesis del Centro de Investigación en Computación del IPN (CIC-IPN) que describe la clasificación ACM (Association for Computing Machinery) de estas a la cual se le llamará "dominio científico".

El cubo de datos del dominio comercial consiste de dimensiones como son: producto, almacén, tiempo, ubicación, cada una de estas con una jerarquía interna. Como medidas o hechos se tiene: unidades vendidas, ventas por almacén, costos de almacén por mencionar solo algunas. La tabla de hechos contiene 251,395 registros mientras que el dominio científico está formado por la dimensión clasificación y tiempo. El conjunto de datos contiene 253 registros o tesis.

5.1.- Conjunto de datos comerciales

Se hace uso de una base de datos diseñada de forma multidimensional llamada "FoodMart" cuyo dominio es información de ventas de un supermercado. Esta base de datos está

disponible en la página oficial de Mondrian, como sentencias DML (Lenguaje de definición de datos). La base de datos contiene 37 tablas cuya información es productos, sucursales, clientes, promociones, ventas de almacén, costos de almacén, unidades vendidas, entre otros.
Esta base de datos ha sido diseñada cuidadosamente para su uso con ROLAP, esto es se tienen tablas que contienen agregados (tabla de hechos) los cuales resumen combinaciones de distintas dimensiones (productos, sucursal, cliente) y estas tablas están relacionadas con tablas que almacenan información específica de cada dimensión (tabla de dimensión).

Las tablas de hechos son: sales_fact, sales_fact_dec_.
Las tablas de dimensiones son: product, employee, store, promotion, category, region, days.

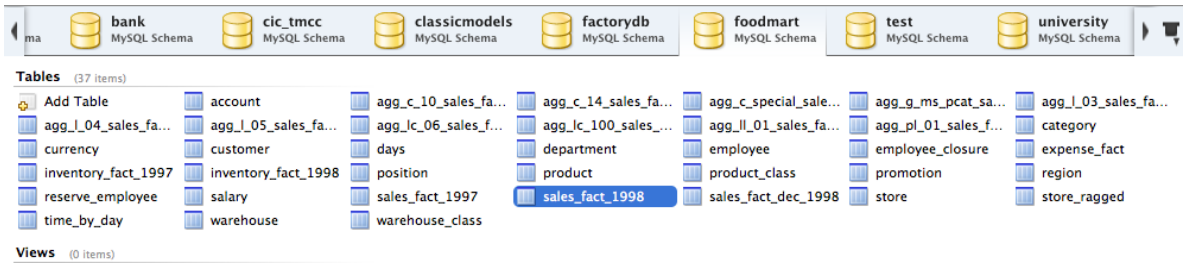


Figura 5.A - Vista de las tablas en la base de datos FoodMart

La figura 5.A muestra el número de tablas que componen a la base de datos foodmart. La figura 5.B presenta una tabla de hechos llamada: sales_fact_ (central), 5 tablas de dimensión relacionadas a la tabla de hechos por medio de una llave foránea y una tabla de detalle de la dimensión producto llamada: product_class. Esta estructura tiene como nombre esquema de copo de nieve. A partir de este diseño se modela un cubo OLAP usando el motor Mondrian.

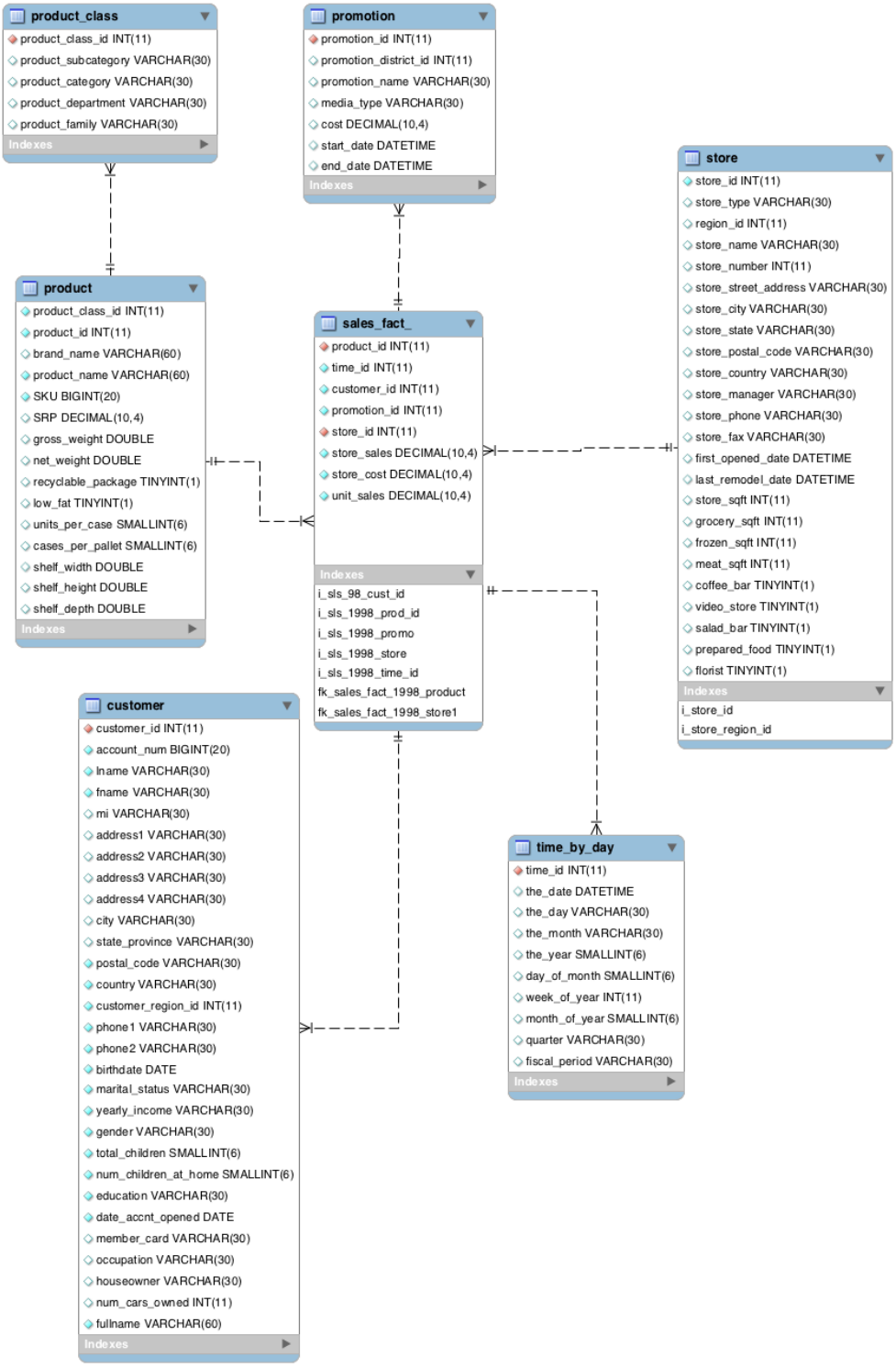


Figura 5.B - Esquema copo de nieve del dominio comercial

Esta base de datos es almacenada en MySQL 5 aunque pudo haberse usado cualquier otro administrador como es Oracle, MSAccess o SQL Server.

5.2.- Conjunto de datos científicos

La base de datos multidimensional del dominio científico almacena información relacionada a un conjunto de tesis y su clasificación ACM. Estas tesis pertenecen al Centro de Investigación en Computación del IPN (CIC-IPN). La base de datos está conformada por una tabla de hechos llamada: tesis_cic y una tabla de dimensión: tiempo. La tabla de hechos contiene la clasificación ACM de cada tesis, de manera que en realidad existen 2 dimensiones dentro del dominio científico: clasificación y tiempo. (Ver figura 5.C)

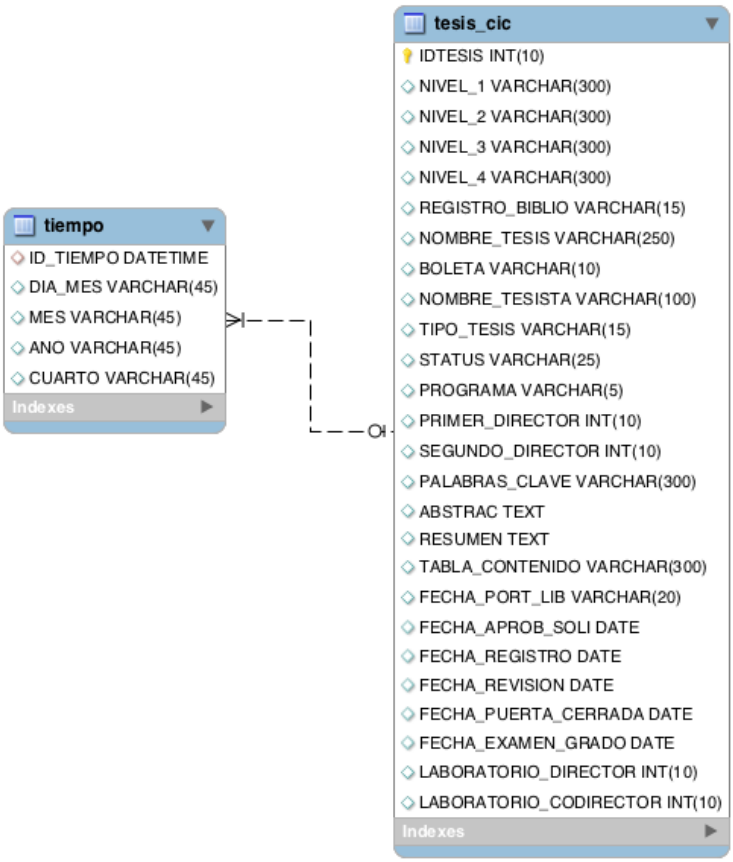


Figura 5.C - Esquema Estrella del dominio científico

De la misma manera que el dominio comercial, esta base de datos fue almacenada en MySQL 5.

6.- Ejecución de pregunta de negocio

Como ejemplo se plantea la consulta de negocio para los dominios comercial y científico. Supongamos que la consulta de negocio en un dominio comercial es la siguiente: "*saber en qué nivel de la clasificación (jerarquía) de productos se tienen bajos niveles de ventas, digamos abajo del 20% con respecto al año anterior*" (Ver figura 6.A).

a) En el "navegador de jerarquías" (lado izquierdo) se selecciona la dimensión de interés:

1) El cubo: "Sales"
2) La dimensión: "Product"
3) El nivel: "ALL"
4) Click en el botón "Agregar dimensión de interés"

b) En el "navegador de jerarquías" (lado izquierdo) se selecciona:

1) La dimensión: "Time"
2) El Nivel: "Year"
3) Click en el botón "Consulta elemento del árbol"
4) Se selecciona del combo del lado derecho del botón anterior, el elemento de tiempo a analizar (Ej. 1997/1998)
5) Click en el botón "Agregar dimensiones"
6) Se repite el paso 4) y 5) para así agregar los elementos de la dimensión tiempo que conforma el espacio de búsqueda.

c) En el módulo de "Características de elementos de interés"

1) Se selecciona del combo de tendencia: "Decremento"
2) Se selecciona del combo de elementos: "Por porciento"
3) Se ingresa el valor del porcentaje: 20%
4) Se selecciona del combo tipo de medida: "Store Sales"

d) En el módulo de "Parámetros de visualización"

1) Se selecciona del combo tipo de visualización: "Mapa de nodos"

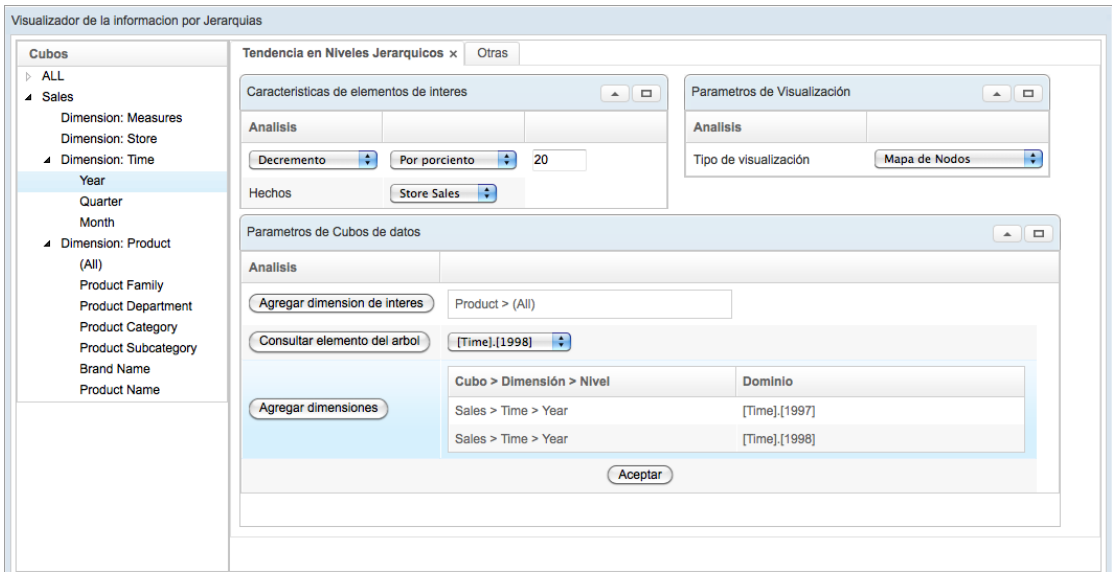


Figura 6.A – Vista de la aplicación con los parámetros del dominio comercial definidos

Como resultado visual se obtiene el mapa de la figura 6.B. En el cual se muestra la ruta de cada producto que presenta una situación de interés, cada nivel de la jerarquía está representado por un color, siendo los nodos en color rojo los puntos de interés. Ejemplo: el producto: "Washington Apple Drink" pertenece a la marca: "Washington", la cual pertenece a la subcategoría: "Flavored Drinks" que pertenece a la categoría: "Drinks", esta categoría pertenece al departamento "Beverages" y finalmente esta pertenece a la familia "Drink". En resumen el producto "Washington Apple Drink" cumple con las características de los productos buscados.

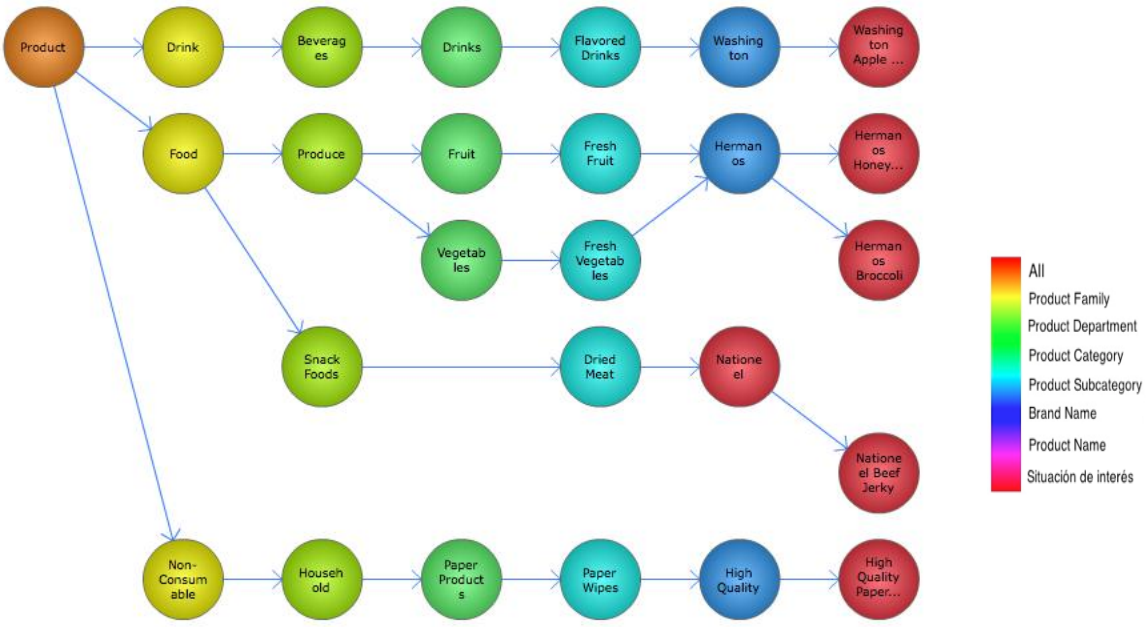
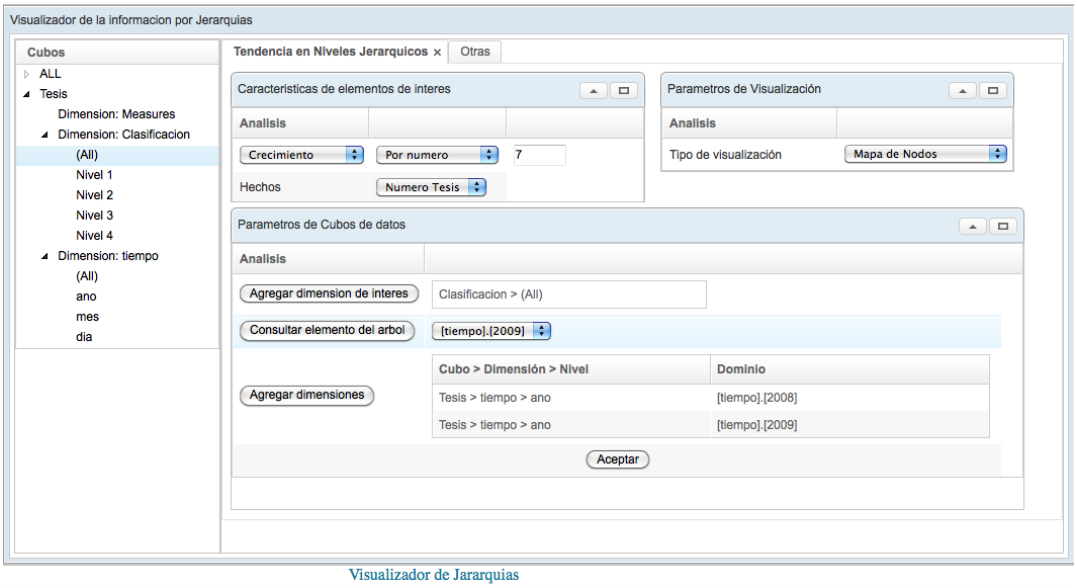


Figura 6.B - Mapa de nodos de situaciones de interés en un dominio comercial

Se hace uso también del dominio científico y se plantea como ejemplo la siguiente consulta "saber en qué nivel de la clasificación ACM de tesis del CIC-IPN existe un incremento en número en 2 años determinados (2008, 2009)", visualizando los 7 con mayor crecimiento (Ver figura 6.C) como respuesta visual a esta consulta se presenta el mapa de situaciones de interés de la figura 6.D, en el cual de la misma manera que en el dominio comercial se presenta la ruta de las anomalías encontradas que cumplen con los parámetros de la consulta.



Visualizador de Jararquias

Figura 6.C - Vista de la aplicación con los parámetros del dominio cientifico definidos

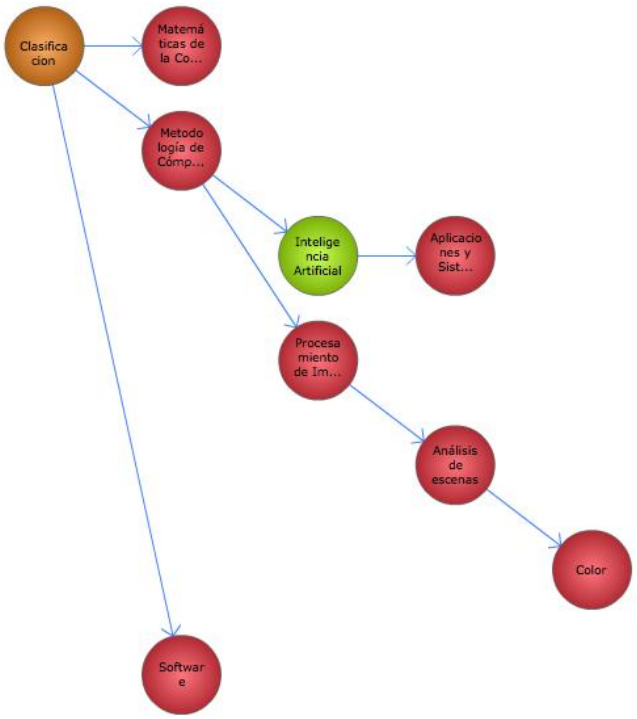


Figura 6.D – Mapa de nodos de situaciones de interés en un dominio científico

Se obtiene como resultado un mapa con 7 anomalías con sus respectivas rutas. Por ejemplo la clasificación de nivel 4: "Color" presenta un crecimiento del 500% del año 2008 a 2009. Sin embargo no solo se encontró un crecimiento en el nivel 4, sino también en el nivel 3: "Análisis de escenas", en el nivel 2: "Procesamiento de análisis" y en el nivel 1: "Metodología de computo", todas estas anomalías en la misma ruta. Este resultado nos indica que la situación de interés comienza a presentare desde el nivel 1 hasta el nivel 4.

La aplicación desarrollada cumple con el "cambio de perspectivas visuales" que define a una aplicación visual analítica, por lo que se ofrece también otros tipos de visualizaciones como son "Mapa de calor" (Figura 6.E) y el "Mapa pastel multi-nivel" (Figura 6.F) para el mismo conjunto de anomalías.

Figura 6.E - Mapa de calor

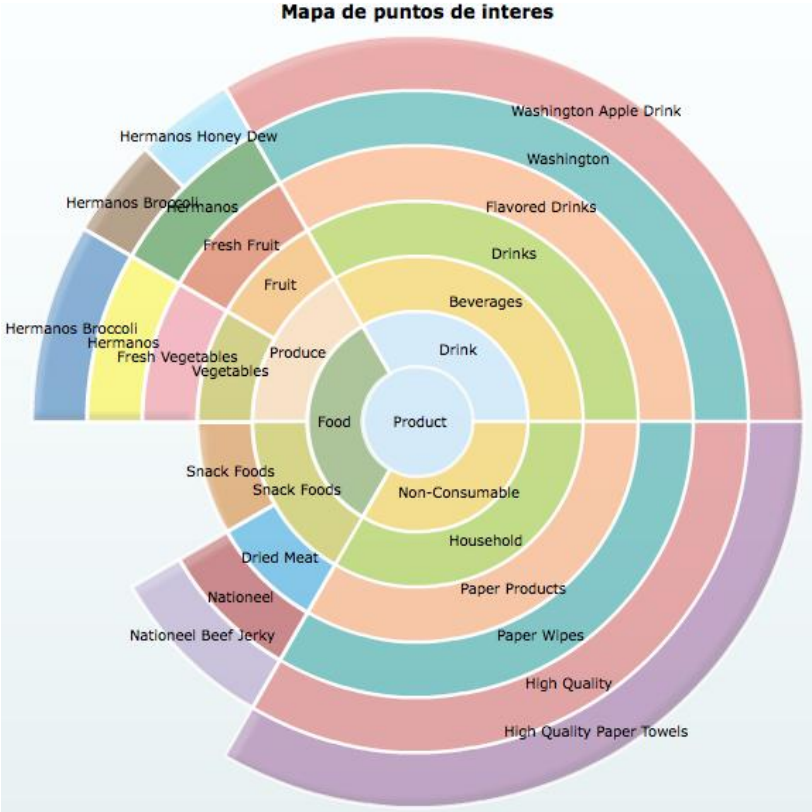


Figura 6.F - Mapa pastel multi-nivel

El mapa de calor representa cada ruta de la anomalía como un renglón, mientras que cada columna representa el nivel de la jerarquía de la dimensión de interés. Por ejemplo la ruta para alcanzar el producto "Hermanos Broccoli" es: Product → Food → Produce → Vegetables → Fresh Vegetables → Hermanos → Hermanos Broccoli.

El mapa Pastel-Multi Nivel representa el árbol de la jerarquía de la dimensión analizada en forma de pastel. Cada nivel del pastel representa un nivel, siendo el nivel interior el nivel de menor jerarquía y el nivel exterior el de mayor jerarquía.

6.1.- Navegación usando tableros de control

Además de los mapas de situaciones de interés o anomalías, la aplicación ha sido diseñada para permitir al usuario una navegación por las dimensiones de forma visual e interactiva a través de tableros de control (Dashboard) los cuales muestran visualizaciones dinámicas en las dimensiones definidas en el cubo de datos, tales visualizaciones involucran operaciones OLAP (On-Line Analytical Processing) como son drill-down y roll up. Este tablero de control es ejecutado dentro de los mapas de situaciones de interés al seleccionar un nodo, una celda o un sub-pastel en los mapas de nodos, calor y pastel multi-nivel respectivamente. Por ejemplo al dar click en el nodo "Food" en el mapa de nodos de la figura 6.B se presenta la gráfica de ventas por año de cada departamento de la familia "Food" (Figura 6.G), equivalente a una operación drill-down al siguiente nivel de la jerarquía. Al hacer click sobre un departamento (Ej. Produce - 1998) se presenta de forma

dinámica 2 visualizaciones en las dimensiones tiempo y almacén, referentes a las ventas en cada trimestre de ese elemento (Figura 6.H) y las ventas por país de cada categoría del departamento (Figura 6.I). Al dar click a una categoría se despliega un mapa geográfico ubicando las ventas por estado de la categoría (Figura 5.J) y al seleccionar un estado en particular se muestra las ventas por ciudad (Figura 5.K).

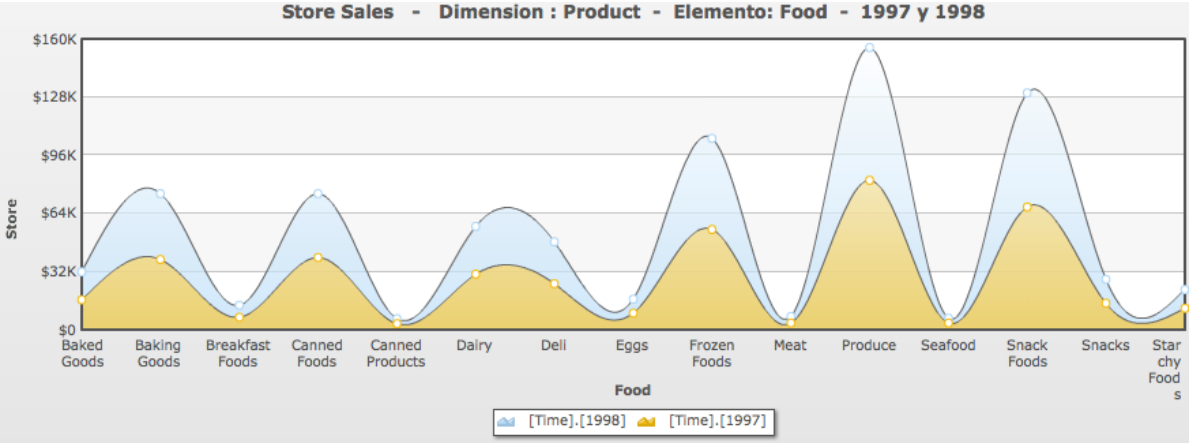


Figura 5.G- Tablero de control: Drill down sobre dimensión de interés

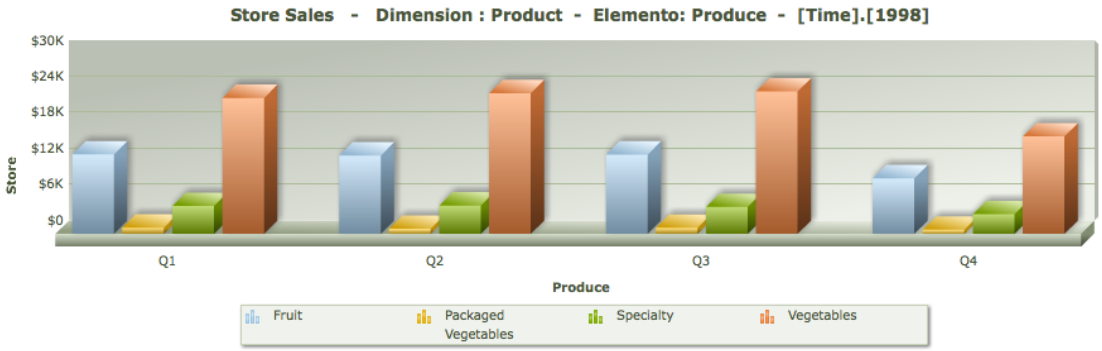


Figura 5.H - Tablero de control: Ventas por trimestre

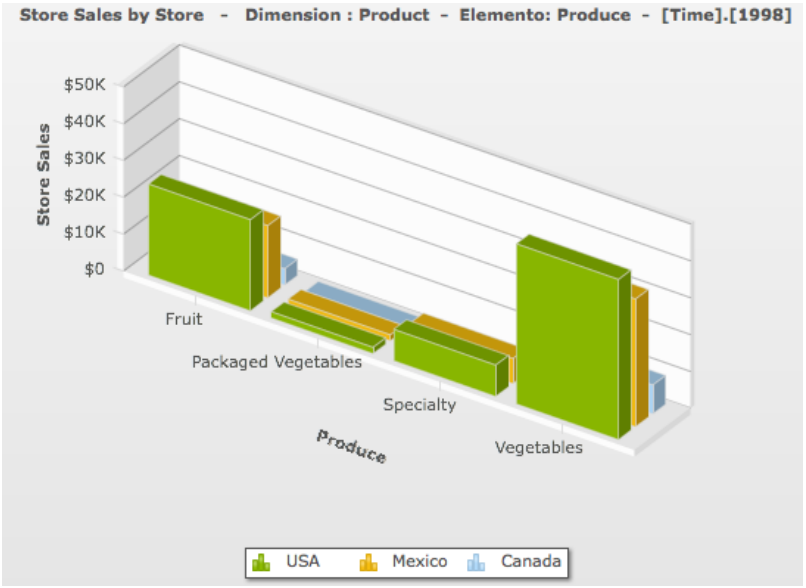


Figura 5.I- Tablero de control: Ventas por país

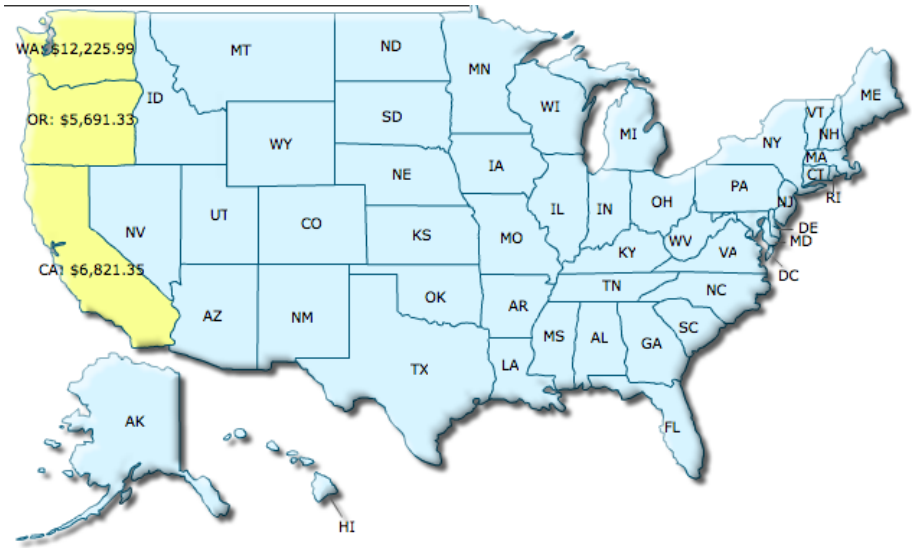


Figura 5.J- Tablero de control: Ventas por estado

Store Sales by Store - Dimension : Product - Elemento: CA - [Time].[1998]

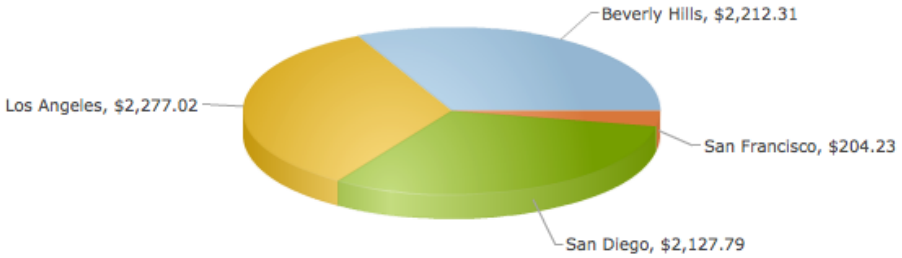


Figura 5.K - Tablero de control: Ventas por ciudad

Anexo C – Escenarios de la pregunta de tendencia con niveles jerárquicos

Este anexo forma parte de las pruebas y resultados del sistema visualizador de anomalías desarrollado (VisJ), en el cual se presenta una serie de ejemplos donde se muestran los diferentes escenarios de la pregunta de negocio "tendencia en niveles jerárquicos". Existen 6 escenarios generales de los cuales es posible generar diversas combinaciones (Figura C.1). En el anexo C se mostró como ingresar los parámetros de cada pregunta en la interfaz de usuario, por lo tanto en esta sección se presenta la ruta de elementos que representa cada pregunta y los resultados obtenidos en un mapa de nodos.

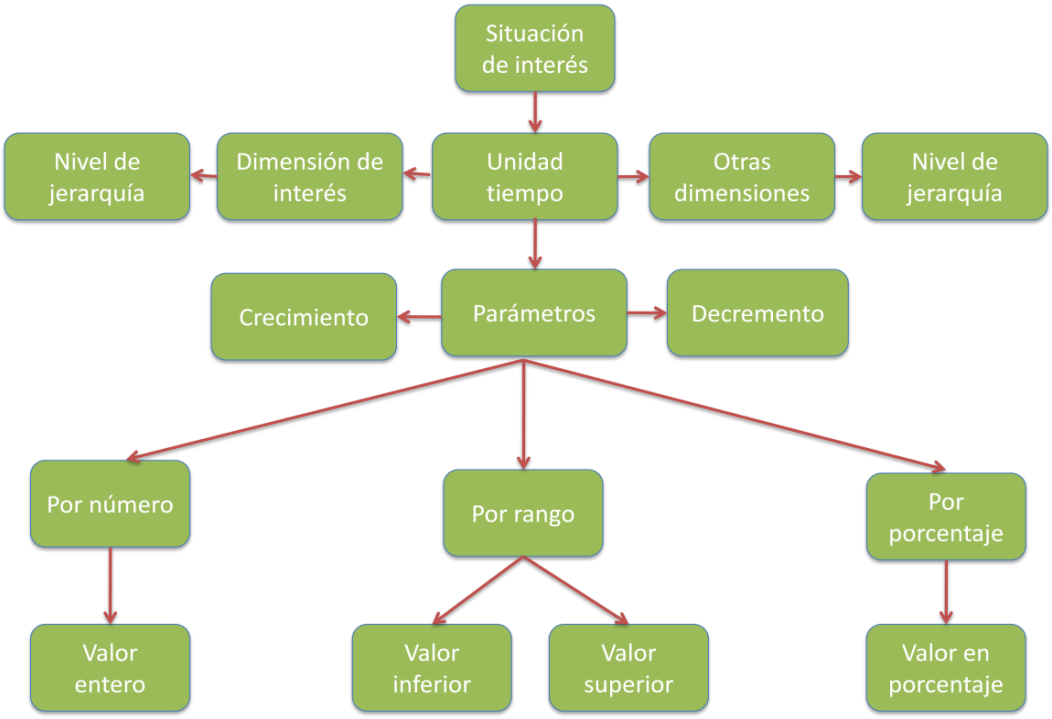
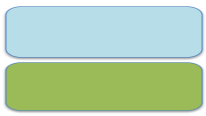


Figura C.1 - Escenarios de tendencia con niveles jerárquicos

Los elementos de la figura D.1 representan los elementos necesarios para poder plantear una pregunta de negocio. Cada elemento puede tener 2 tipos de colores.



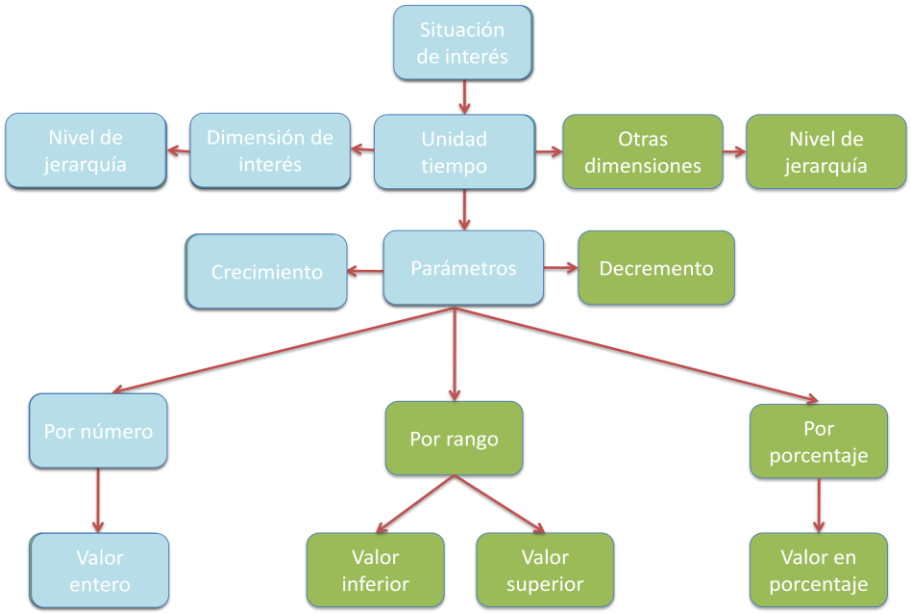
Elemento seleccionado

Elemento no seleccionado

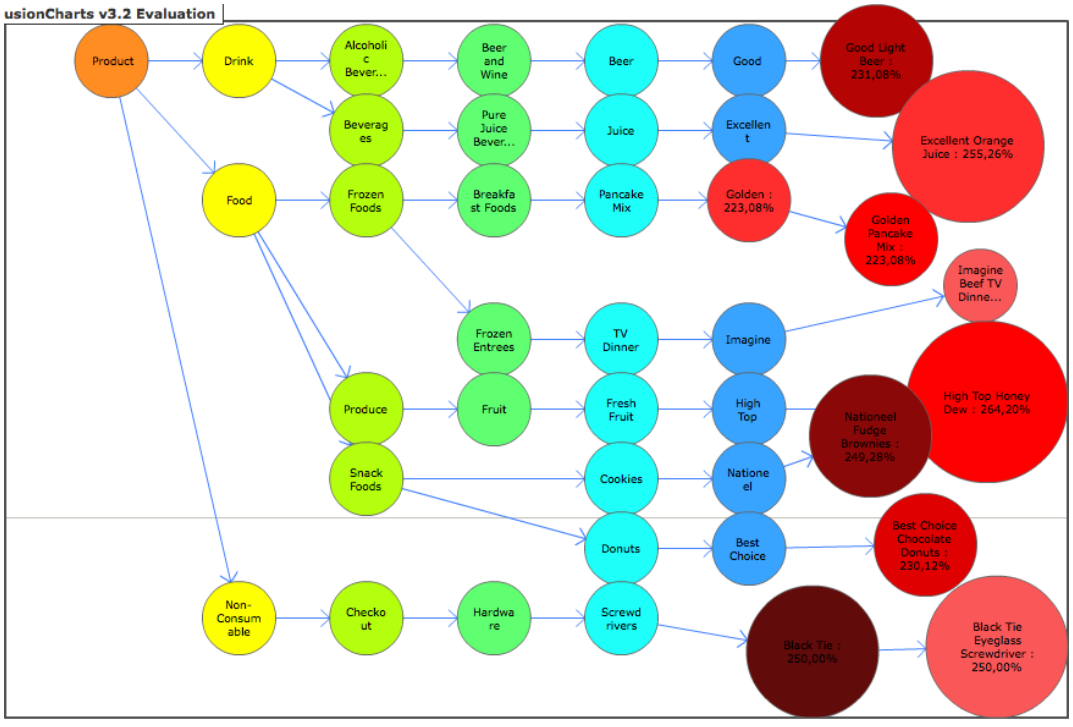
Las pruebas son realizadas en la base de datos comercial estudiada en el capitulo 3.
La visualización de los resultados es presentada en un mapa de nodos 2D el cual tiene la característica de representar cada nodo hoja con diferente tamaño y diferente nivel de color rojo, de acuerdo al valor numérico del nodo.

Pregunta 1.- Los N mejores

"Se desea saber en qué nivel de la jerarquía de productos se tienen los mejores 10 niveles de ventas, con respecto al año anterior"

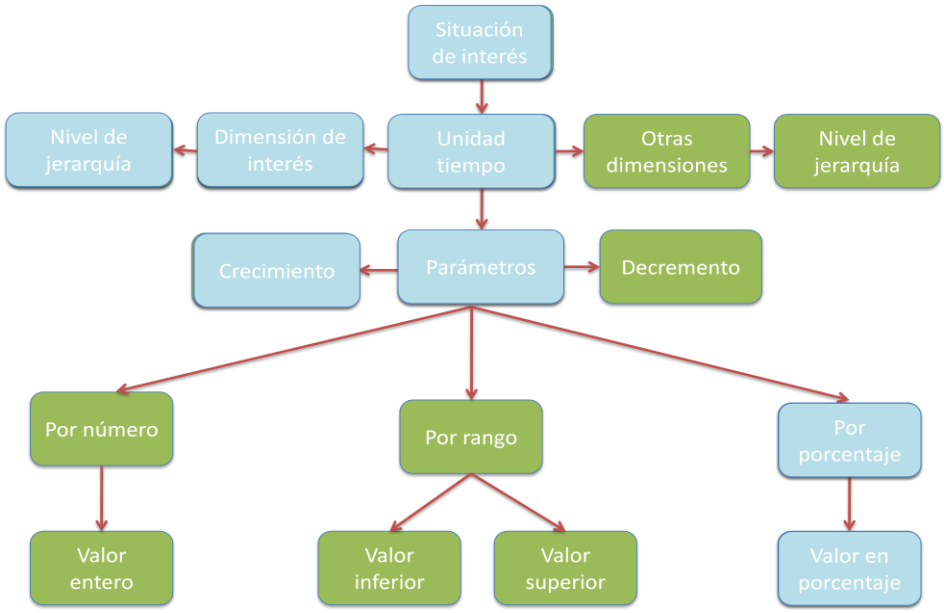


Dimensión de interés	Producto	**Nivel de jerarquía**	Todos (ALL)
Unidad de tiempo	Año (1997 y 1998)		
Tipo de consulta	Crecimiento		
Parámetro	Por número	**Valor entero**	10

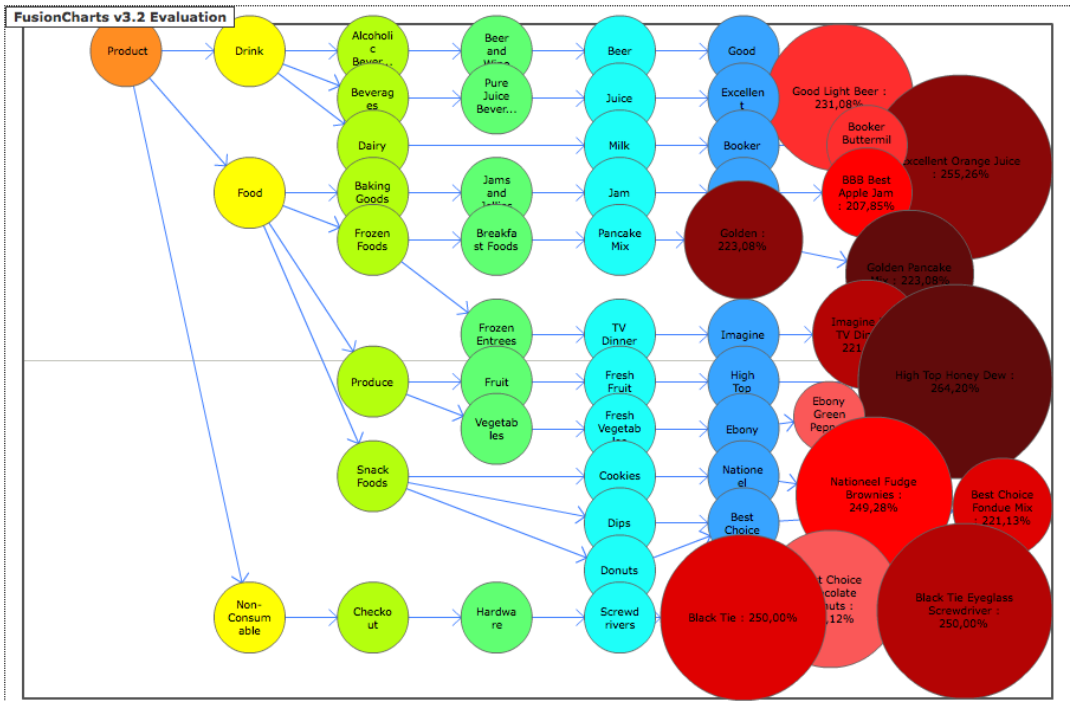


Pregunta 2.- Crecimiento mayor a N%

"Se desea saber en qué nivel de la jerarquía de productos se tienen altos niveles de ventas, digamos mayores al 200% con respecto al año anterior"

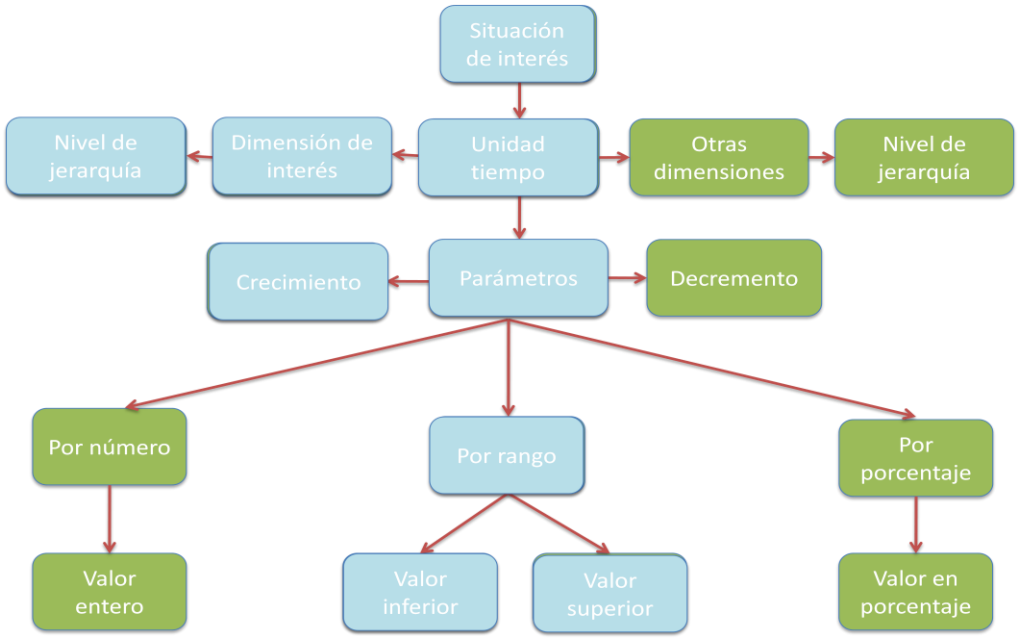


Dimensión de interés	Producto	**Nivel de jerarquía**	Todos (ALL)
Unidad de tiempo	Año (1997 y 1998)		
Tipo de consulta	Crecimiento		
Parámetro	Por porcentaje	**Valor en porcentaje**	200 %

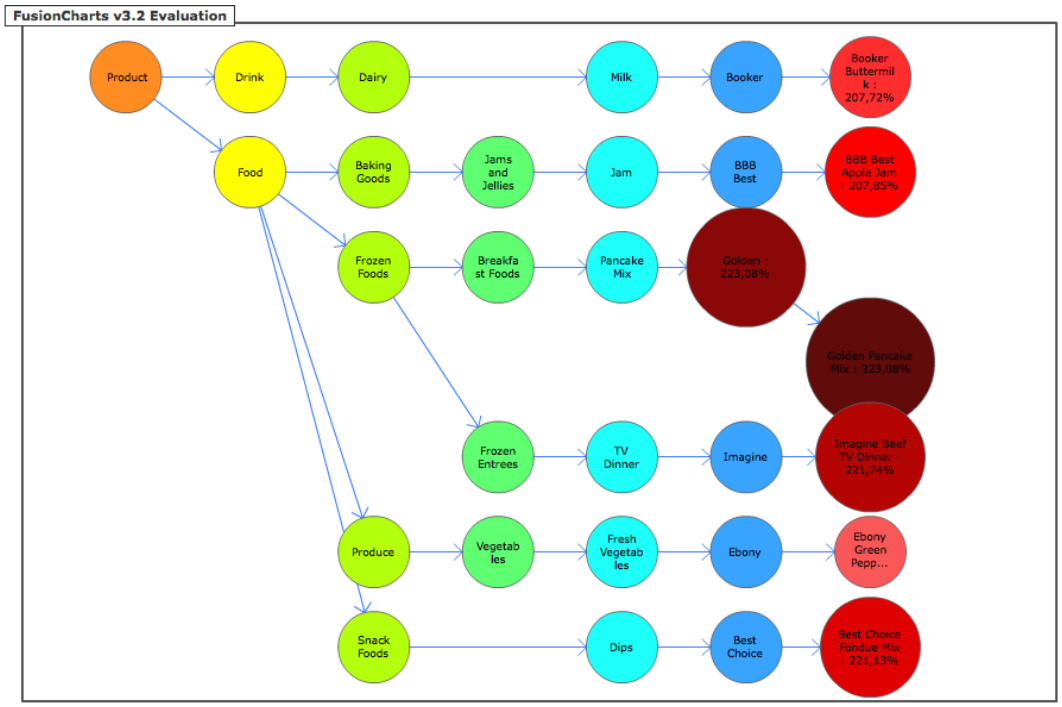


Pregunta 3.- Crecimiento entre N1% y N2%

"Se desea saber en qué nivel de la jerarquía de productos se tienen altos niveles de ventas, digamos entre el 200% y 230% con respecto al año anterior"

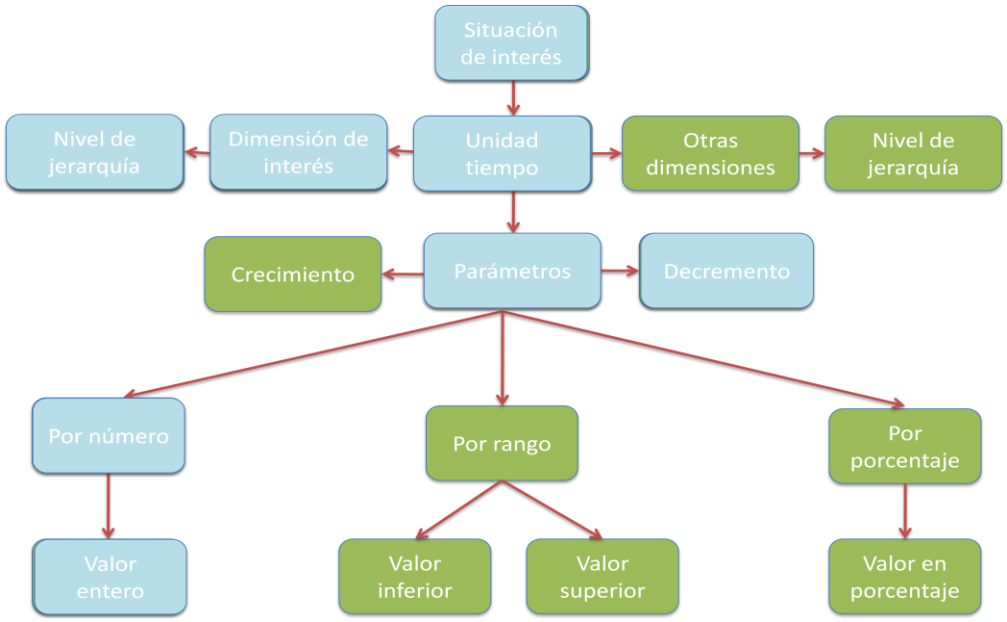


Dimensión de interés	Producto	**Nivel de jerarquía**	Todos (ALL)
Unidad de tiempo	Año (1997 y 1998)		
Tipo de consulta	Crecimiento		
Parámetro	Por rango	**Valor inferior y superior**	200 % y 230%

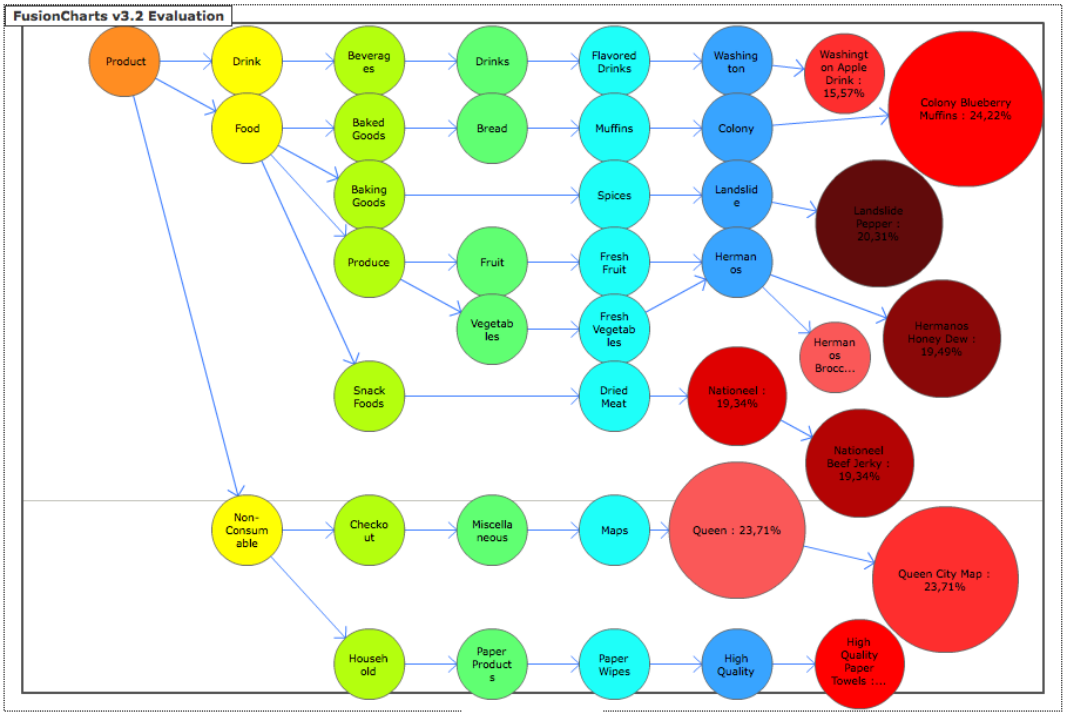


Pregunta 4.- Los N peores

"Se desea saber en qué nivel de la jerarquía de productos se tienen los peores 10 niveles de ventas, con respecto al año anterior"

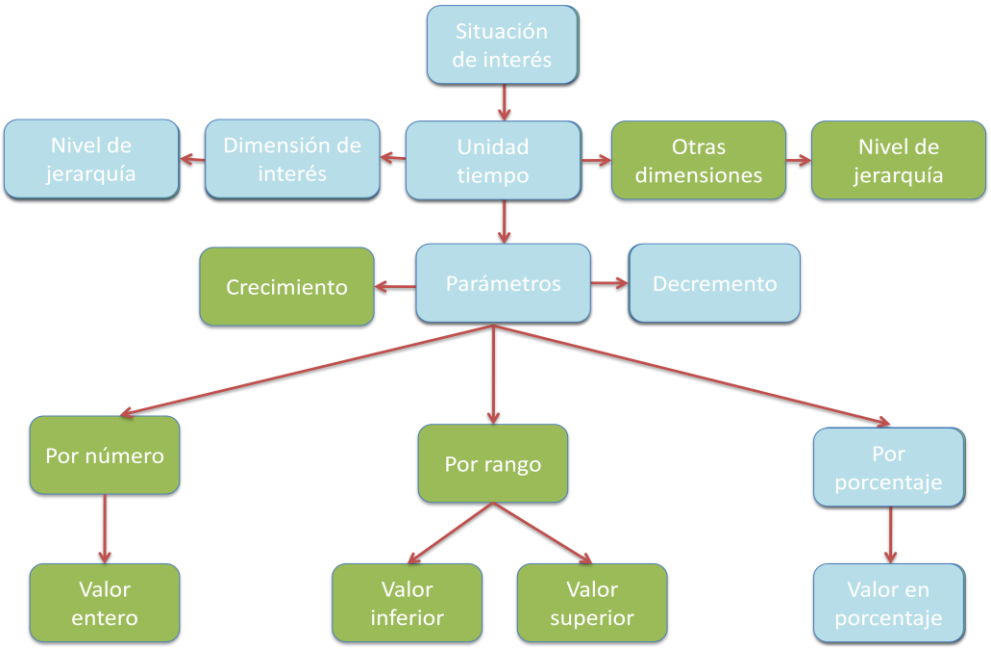


Dimensión de interés	Producto	**Nivel de jerarquía**	Todos (ALL)
Unidad de tiempo	Año (1997 y 1998)		
Tipo de consulta	Decremento		
Parámetro	Por número	**Valor entero**	10

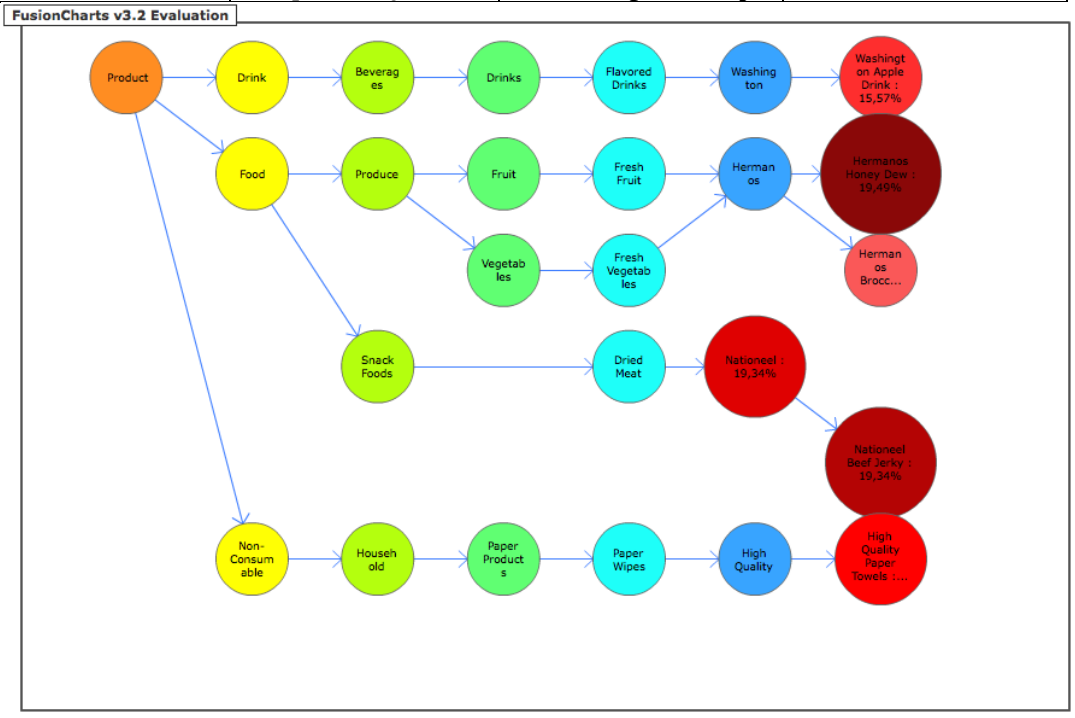


Pregunta 5.- Decremento menor a N%

"Se desea saber en qué nivel de la jerarquía de productos se tienen bajos niveles de ventas, digamos menores al 20% con respecto al año anterior"

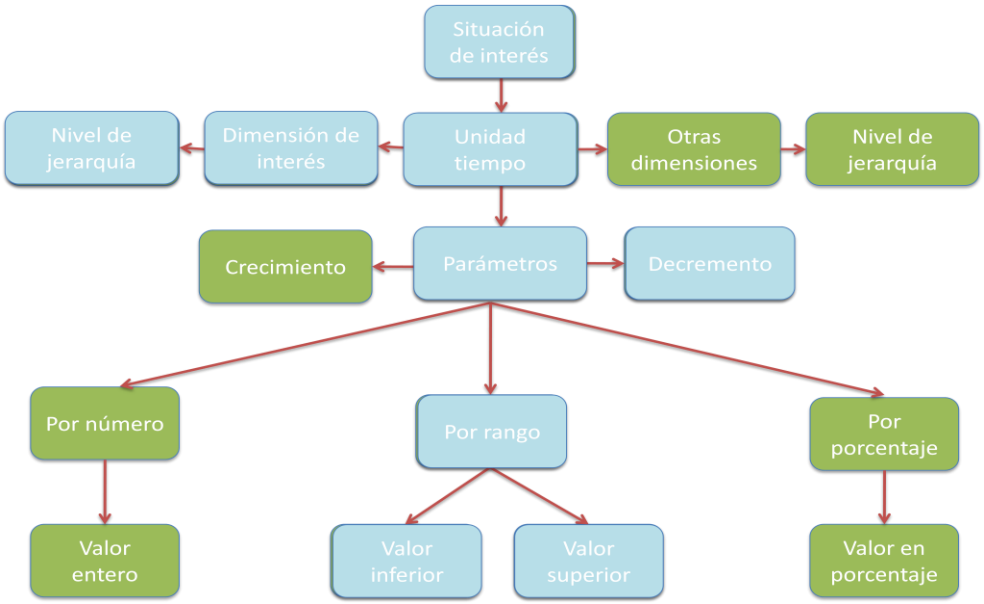


Dimensión de interés	Producto	**Nivel de jerarquía**	Todos (ALL)
Unidad de tiempo	Año (1997 y 1998)		
Tipo de consulta	Decremento		
Parámetro	Por porcentaje	**Valor en porcentaje**	20 %

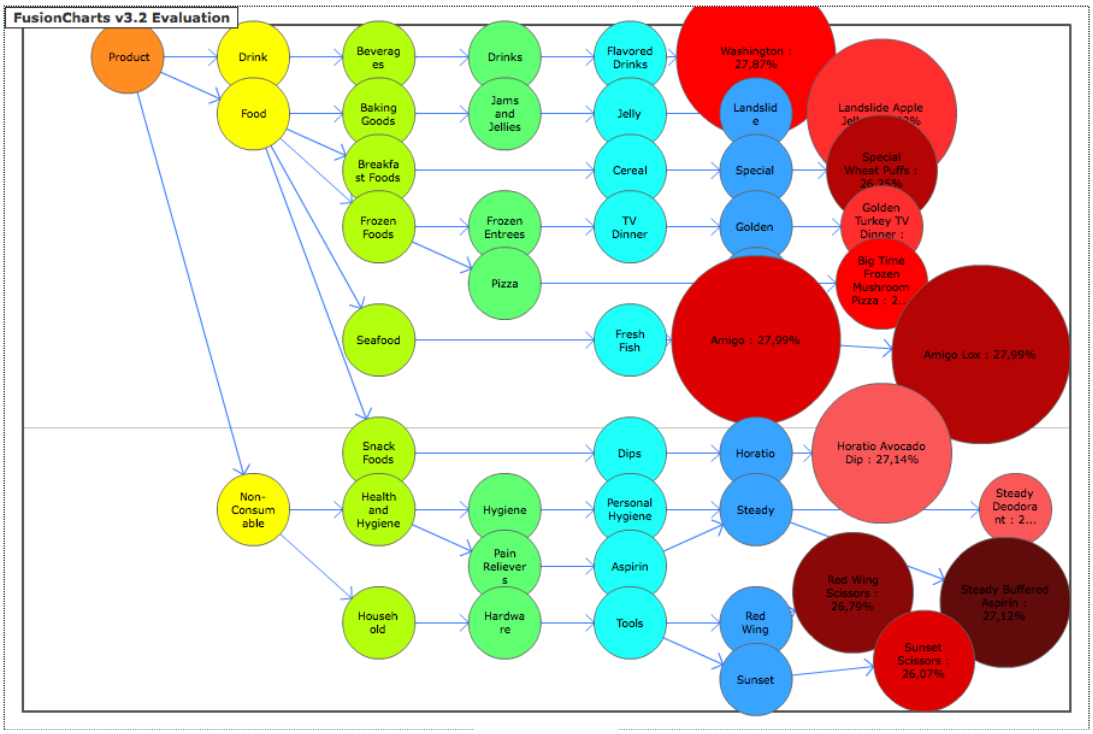


Pregunta 6.- Decremento entre N1% y N2%

"Se desea saber en qué nivel de la jerarquía de productos se tienen bajos niveles de ventas, digamos entre el 25% y 28% con respecto al año anterior"



Dimensión de interés	Producto	**Nivel de jerarquía**	Todos (ALL)
Unidad de tiempo	Año (1997 y 1998)		
Tipo de consulta	Decremento		
Parámetro	Por rango	**Valor inferior y superior**	25 % y 28%



Anexo D - Comparativa con trabajos académicos

En esta sección se realiza un estudio comparativo entre herramientas académicas similares a la propuesta en el presente trabajo. Las cuales tienen como objetivo resolver otros tipos de preguntas de negocio como los planteados en [Chen, 2009] y [Bao, 2003]. Estas herramientas han sido ya estudiadas en el tema 2.9 (Estado del arte), analizando sus principales características.

Las características a evaluar en cada uno de los trabajos son las siguientes:

Jerarquía

Se refiere la estructura interna ó clasificación en los datos que permita consultarlos en distintos niveles de granularidad.

Volumen de datos

Es la cantidad de registros almacenados en la base de datos.

Velocidad

La velocidad se refiere al tiempo de respuesta de cada herramienta en realizar los cálculos necesarios en los cubos de datos y presentar los resultados al usuario. Se debe tener en cuenta que este parámetro depende de varios factores como por ejemplo: los algoritmos utilizados o las características físicas de cada equipo.

Exploración visual

Un usuario podría comenzar con una pregunta básica y basado en las señales visuales o puntos de vista profundizar la investigación.

Aumento de la percepción humana

Se fomenta el pensamiento visual aprovechando los poderes de la percepción humana. Se hace un preciso uso del tamaño, color, forma y texto para hacer las diferencias y cuando son usados de forma adecuada ayudan a la interpretación.

Expresividad visual

Permite visualizar múltiples dimensiones de un problema sin esfuerzos, en formatos que sean fáciles de comprender.

Visualización automática

Incluye la sugerencia automática a visualizaciones efectivas para un problema específico

Cambio de perspectivas visuales

Sugiere una serie de alternativas de visualizaciones, logrando ver el mismo resumen de los resultados desde diferentes puntos de vista.

Enlace de perspectivas visuales

Se refiere a la correlación de la información, lo cual significa que una visualización lleva a otra.

Una visualización podría mostrar un conjunto de anomalías y el usuario puede seleccionar una anomalía e instantáneamente ver otra visualización que despliega el detalle de los datos.

Visualización colaborativa

Es la habilidad de crear interactivamente útiles visualizaciones de la información en equipo. Las personas publican resultados de forma segura interactivamente y disponibles en la red.

A continuación se presenta una tabla que resume las principales características de los trabajos similares y del trabajo propuesto en esta tesis.

Operaciones OLAP

Las operaciones OLAP son las operaciones que puede aplicarse en los cubos de datos tales como: drill-down, roll-up, slice y dice.

Estructura de datos

Las estructuras de datos se refieren a la lattice de dimensiones de un cubo de datos que se guarda en memoria.

Característica	**Discovery-driven Exploration of OLAP Data Cubes**	**Exploration and Visualization of OLAP Cubes with Statistical Test**	**Empirical Comparison of Dynamic Query Sliders and Brushing Histograms**	**Sistema visualizador de mapas de anomalías VisJ**
Jerarquía	Si	No	No	Si
Volumen de datos	Si	Si	No	No
Velocidad	Si	No	No	No
Exploración visual	Si	Si	Si	Si
Aumento de la percepción humana	No	Si	Si	Si
Expresividad visual	Si	Si	No	Si
Visualización automática	No	No	No	Si
Cambio de perspectivas visuales	No	Si	Si	Si
Enlace de perspectivas visuales	No	Si	Si	Si
Visualización colaborativa	No	No	No	No
Operaciones OLAP	Si	Si	No	Si
Estructura de datos	Si	Si	No	Si

Discovery-driven Exploration of OLAP Data Cubes (Exploración de cubos OLAP usando un descubrimiento dirigido)

- Se puede trabajar con datos que presenten jerarquías.
- Es posible consultar millones de registros.
- La interfaz de usuario similar a una hoja de cálculo permite cumplir con la "exploración visual".
- El tiempo de respuesta reportado es inferior a los demás trabajos, por lo cual se le puede considerar el más rápido.
- Se hace uso de una lattice de dimensiones y se aplican operaciones OLAP para navegar en los agregados.

Exploration and Visualization of OLAP Cubes with Statistical Test (Exploración y visualización de cubos OLAP con pruebas estadísticas)

- Se cumple la "exploración visual" la cual permite navegar sobre las celdas del tablero de resultados.
- Se hace uso de un espectro de colores que indica el nivel de diferencia entre cuboides o celdas dentro del tablero, logrando así cumplir con el "aumento de la percepción humana".
- Los resultados de las consultas que involucran varias dimensiones son presentados en un mapa que facilita la comprensión, por lo tanto se cumple la "expresividad visual".
- El "cambio y enlace de perspectivas" son logrados gracias a la relación entre graficas que resumen las diferencias entre los cuboides.
- Se hace uso de una lattice de dimensiones y se aplican operaciones OLAP para navegar en los agregados.

Empirical Comparison of Dynamic Query Sliders and Brushing Histograms (Comparación empírica de deslizadores e histogramas)

- Las preguntas dinámicas que se plantean en este trabajo consisten en consultar y visualizar al mismo tiempo, lo cual forma parte de la "exploración visual".
- El "aumento de la percepción humana" se logra gracias a los cambios de colores presentes en los diferentes estados o ciudades del mapa.
- La posibilidad de usar deslizadores o histogramas satisfacen el criterio de "cambio de perspectivas visuales".
- Cualquier modificación en el valor de un deslizador o histograma se ve reflejado en el mapa inmediatamente, esto es una visualización lleva a otra, por lo tanto existe un "Enlace de perspectivas visuales".

Sistema visualizador de mapas de anomalías VisJ

- Se puede trabajar con datos que presentes jerarquías.
- La posibilidad de navegar en el tablero de control que pertenece a una anomalía del mapa, permite cumplir con la "exploración visual".

- Se hace uso de colores y tamaños en los mapas representando niveles de jerarquías o niveles de diferencia entre anomalías. Esta característica es el "aumento de la percepción humana"
- Es posible definir una pregunta en la interfaz de usuario que involucre mas de 2 dimensiones (producto, tiempo, ubicación), presentando los resultados en mapas de anomalías, sin causar alguna confusión al usuario por haber involucrado varias dimensiones. Se cumple con la "expresividad visual".
- De acuerdo a las ventajas y desventajas de las distintas representaciones visuales estudiadas en 3.3.3, cuando se consulta más de 50 elementos en el mapa de nodos, se presenta automáticamente una visualización por medio de mapas de calor o mapas pastel-multi-nivel ideales para esa cantidad de elementos. Se logra la "visualización automática"
- El "cambio de perspectivas visuales" permite visualizar los resultados de la consulta por medio de 3 tipos de mapas: de nodos, calor, pastel multi-nivel.
- El "enlace de perspectivas visuales" se logra por medio del tablero de control presente en cada nodo u elemento del mapa de anomalías.
- Se hace uso de una lattice de dimensiones y se aplican operaciones OLAP para navegar en los agregados.