Capítulo 6

Resultados

6.1 Resultados de Zombi

Los resultados principales usando el prototipo Zombi sobre datos de circulación de material bibliográfico de la biblioteca de la UDLAP fueron:

1) Uno de los primeros resultados muestra que ha decrecido el requerimiento de libros a lo largo de los últimos cinco años:

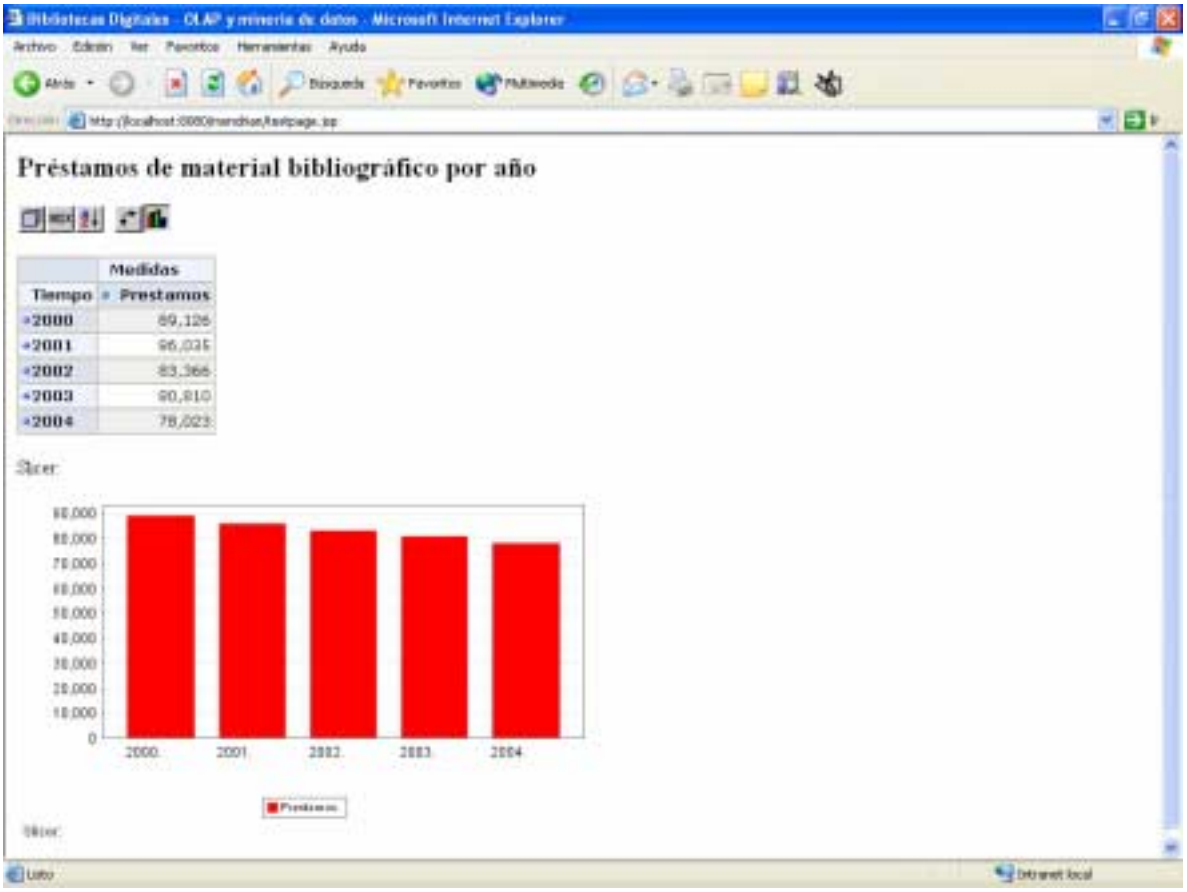


Figura 6.1 Préstamos de material bibliográfico por año

2) Como se ve en la figura 6.2, julio es el mes donde menos se solicitan libros, es un resultado lógico puesto que es un mes de vacaciones.

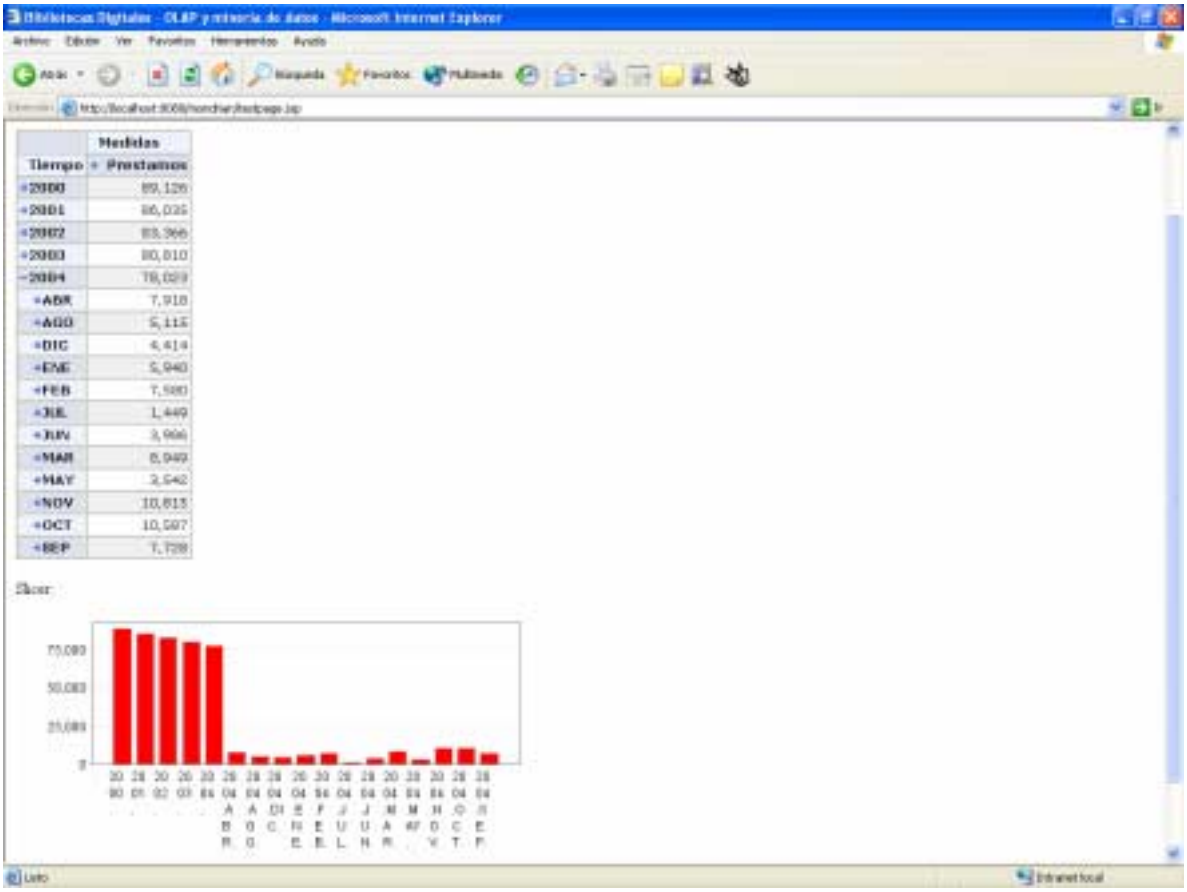


Figura 6.2 Préstamos por mes por año

3) La escuela que más libros ha solicitado en los últimos cinco años, y por lo tanto la que más hace uso de la biblioteca, es la escuela de ciencias sociales, probablemente porque tiene un gran número de estudiantes:

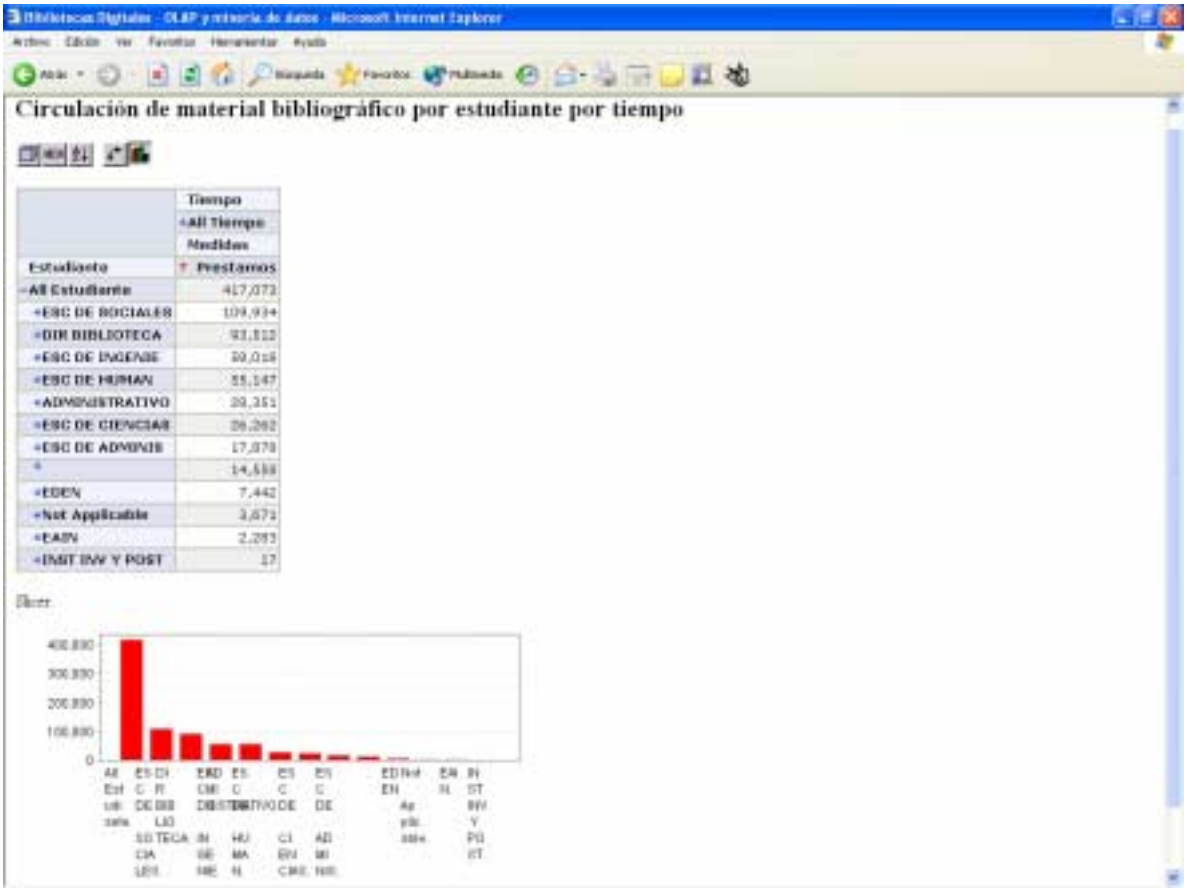


Figura 6.3 Préstamos por estudiante por tiempo

4) Podemos ver navegando interactivamente como se han distribuido los préstamos a lo largo de los últimos cinco años:

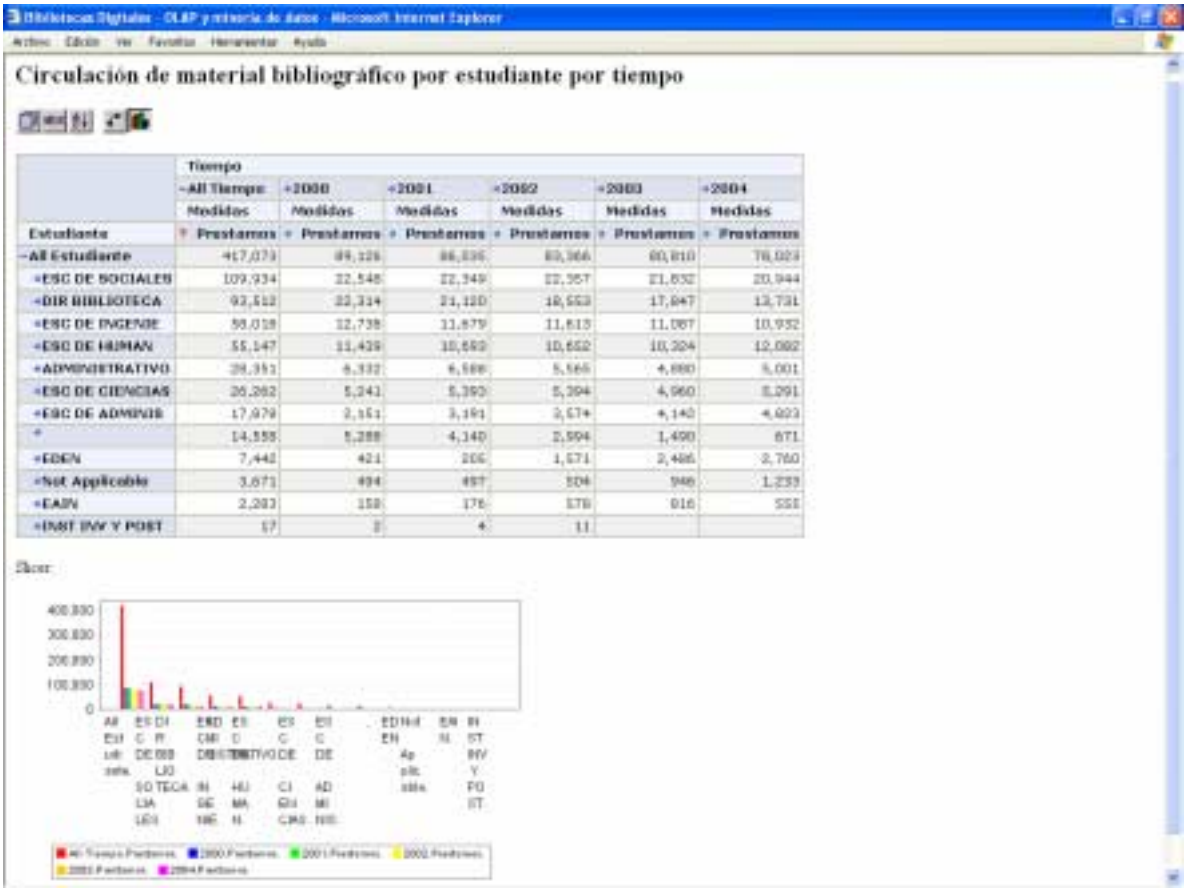
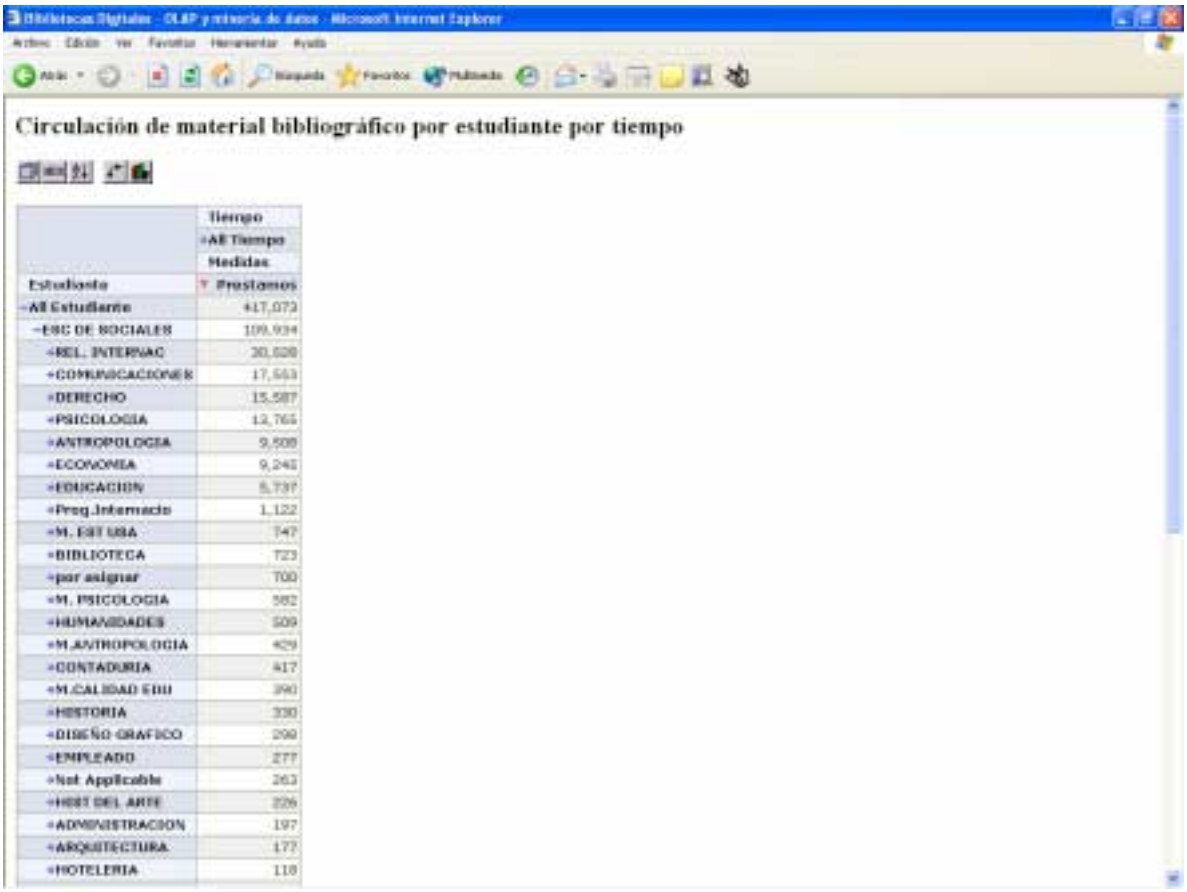


Figura 6.4 Préstamos a los largo de los últimos cinco años por departamento

5) En particular los lectores más asiduos dentro de ciencias sociales pertenecen a la carrera de Relaciones Internacionales.



Circulación de material bibliográfico por estudiante por tiempo

	Tiempo
	All Tiempo
	Medidas
Estudiante	Prestamos
All Estudiante	417,073
ESC DE SOCIALES	109,934
REL. INTERNAC	30,028
COMUNICACIONES	17,553
DERECHO	15,587
PSICOLOGIA	13,765
ANTROPOLOGIA	9,508
ECONOMIA	9,241
EDUCACION	5,737
Prog.Internacio	1,122
M. ESTUDIA	747
BIBLIOTECA	723
por asignar	700
M. PSICOLOGIA	582
HUMANIDADES	509
M.ANTROPOLOGIA	429
CONTADURIA	417
M.CALIDAD EDU	390
HISTORIA	330
DISEÑO GRAFICO	299
EMPLEADO	277
Not Applicable	263
HIST DEL ARTE	226
ADMINISTRACION	197
ARQUITECTURA	177
HOTELERIA	118

Figura 6.5 Escuelas que solicitan el mayor número de libros a la biblioteca

6) Como otro resultado de interés, vemos que en la escuela de ingeniería de la UDLAP los alumnos que hacen mayor uso de la biblioteca (al menos en los últimos cinco años) son los de ingeniería industrial, los alumnos de ingeniería en sistemas ocupan el quinto lugar:

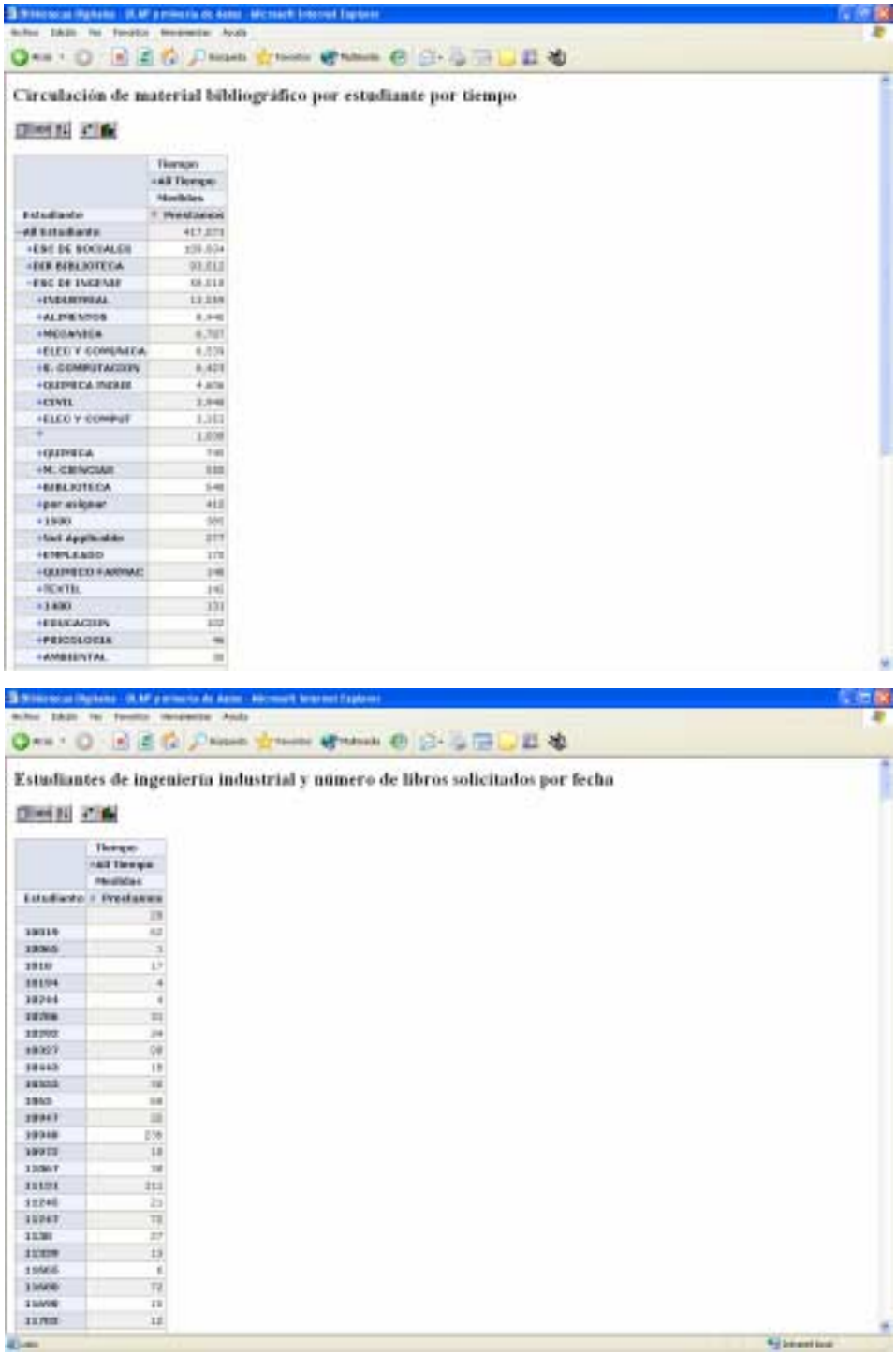


Figura 6.6 Préstamos en el departamento de ingeniería

7) Prueba con grandes volúmenes de información.

Se cargaron los datos de circulación de material bibliográfico de los últimos 5 años obteniendo el total de registros siguiente:

TABLA	**NUMERO DE REGISTROS**
Tabla de hechos	417,073
Dimensión estudiante	15,050
Dimensión clasificación	84,820
Dimensión tiempo	1,852

De un extracto original obtenido de Sydney de tamaño 950 Mb, al almacenarlo en la forma de cubo, la información de interés para análisis solo ocupa 45.3 Mb.

Por otro lado, se encontró que se tienen problemas de memoria insuficiente en la ejecución de consultas que solicitan datos de detalle sobre grandes volúmenes de información, esto implica una plataforma de hardware mayor, es decir más memoria física y probablemente un procesador adicional. En el caso de cubos esparcidos se requiere el desarrollo de un módulo de software que los almacene y administre.

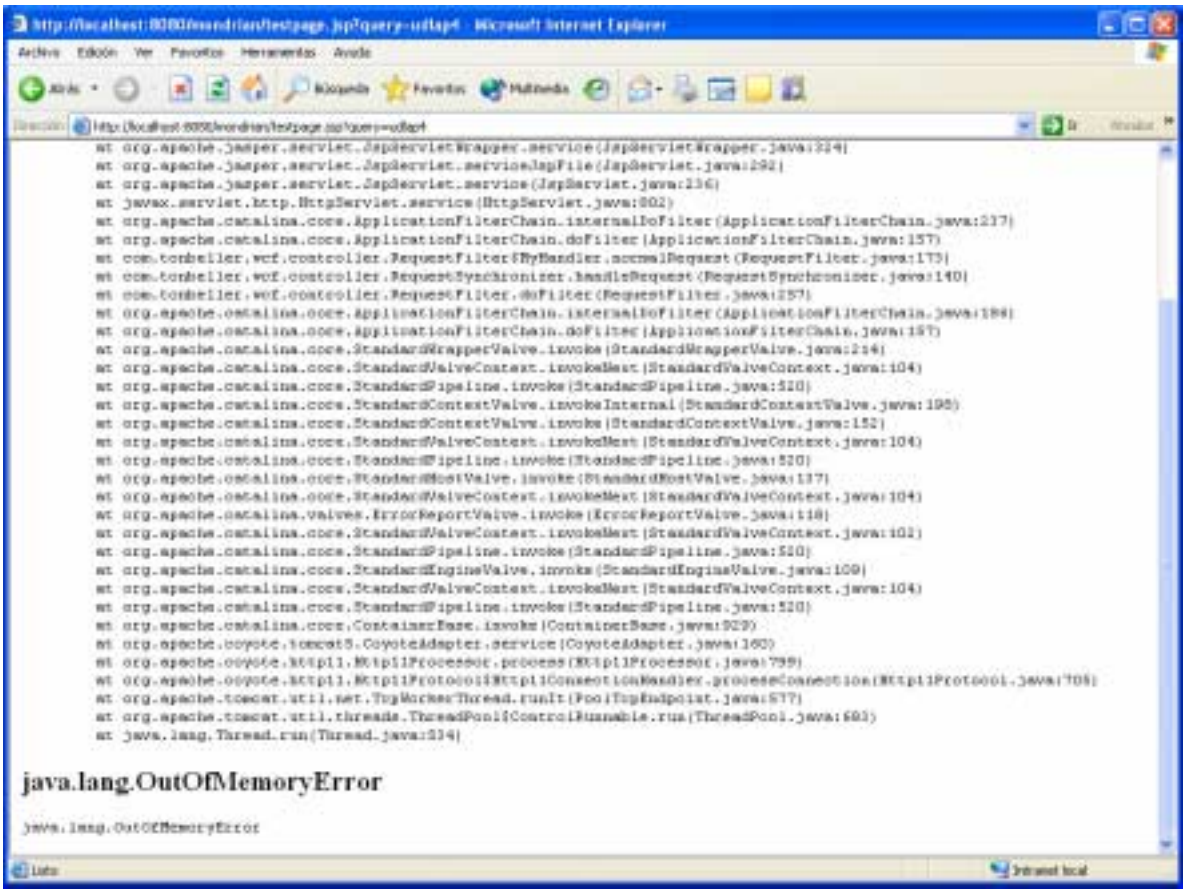


Figura 6.7 Problemas de memoria por gran volumen de datos

8) Problema con limpieza y calidad de datos en Sydney.

El problema principal que se tiene es que la relación división-departamento no se encuentra bien definida en la base de datos, un departamento debe pertenecer a una sola división. Sin embargo se encontraron varios registros en una situación como esta:

DIVISION	**DEPARTAMENTO**
Esc. de Sociales	Rel. Internac
Dir. Biblioteca	Rel. Internac

Vemos que este resultado de OLAP y minería de datos sobre biblioteca nos está mostrando un área de oportunidad interesante. Por ejemplo, si se quiere migrar a una versión nueva de Sydney u otro producto de software para administración de la biblioteca, podría corregirse este problema ya detectado durante el proceso de migración de los datos al nuevo sistema.

11) Respecto a los datos, se tuvo que desarrollar un programa para transformar y cargar los datos en MySql a partir de archivos texto extraídos de Sydney, debido al volumen de datos (que incluyen información histórica de los últimos cinco años), aproximadamente 950 Mb, la ejecución de este proceso tomó aproximadamente 48 horas. Sin embargo este proceso arrojó una base de datos de 45.3 Mb que incluye únicamente la información valiosa para toma de decisiones y eliminó un alto porcentaje de redundancia. Se vio que este nuevo conjunto de datos ofrece posibilidades interesantes:

a) Debido a que es una fuente de información limpia podemos utilizarla para cualquier reporte o gráfica que se requiera.
b) Como se encuentra en una nueva base de datos (MySQL) entonces cualquier actividad de consulta no impacta la operación diaria de Sydney.
c) Vemos que el almacén de datos construido, siendo un conjunto de datos que no tiene redundancia y que únicamente incluye información de valor, se puede considerar una base de datos depurada de Sydney y de esta forma se podría tomar la decisión de eliminar la información histórica de Sydney, recuperando espacio de almacenamiento y convirtiendo al almacén de datos en la fuente oficial de consulta de datos históricos.

6.2 Opiniones de la comunidad de usuarios

Se realizaron pruebas de usabilidad del prototipo con el personal administrativo y de la dirección de la biblioteca de la UDLAP, fueron en total seis usuarios con quienes se realizaron las pruebas: tres profesores bibliotecarios, el director de la biblioteca, el director del área de bibliotecas digitales y.

El procedimiento consistió en:

1) Explicar a cada usuario como utilizar el prototipo y como interpretar los resultados.
2) Pedir al usuario que encontrará información usando el prototipo
3) Después cada usuario pudo usar a discreción el prototipo para pudiera encontrar funcionalidad que en particular fuera de utilidad para él.
4) Finalmente se le entregó un formato de encuesta a cada usuario, el cual se encuentra en el apéndice D, para que lo contestara con base en su experiencia con el prototipo.

Como resultado principal, en general encontraron la herramienta muy útil para la administración y toma de decisiones en la biblioteca, la queja más importante fue el desempeño del prototipo, sin embargo están conscientes del volumen de datos que se esta manipulando y de que de otra forma, generar los reportes de forma manual tomaría varios días.

De la encuesta que se realizó al finalizar las pruebas con cada usuario se obtuvieron las siguientes respuestas.

1) La mayoría está de acuerdo en que Zombi presenta los elementos necesarios para descubrir conocimiento usando los datos históricos de la biblioteca.

2) Están todos de acuerdo en que Zombi les permite tomar decisiones respecto al material bibliográfico.
3) Todos coinciden en que una herramienta como Zombi les permite revisar de manera más fácil la información que el propio Sydney y consideran que el nuevo sistema que sustituirá a Sydney debería incluir funcionalidad como la de Zombi.
4) Muchos empezarán a usar Zombi para reportes en lugar de Sydney, sólo algunos seguirán utilizando algunos reportes de Sydney.
5) Coinciden en que la forma de navegar los datos es fácil una vez que se les explica como hacerlo.
6) El significado de los reportes es fácil de entender.
7) La mayoría quisiera poder desarrollar sus propios reportes en Zombi, aunque hubo quien prefiere que se los construyan porque considera que su actividad es de toma de decisiones.
8) Coinciden en que la interfaz de Zombi es agradable.
9) Todos los usuarios piensan que es fácil encontrar cualquier dato en Zombi.
10) El desempeño de Zombi les parece adecuado aunque quisieran mayor velocidad de respuesta.
11) En general Zombi les pareció una herramienta de utilidad.

Se obtuvieron los siguientes comentarios y sugerencias adicionales:

1) Los usuarios sugieren que Zombi también se utilice para:
    a. Adquisiciones.
    b. Multas.

c. Horarios de servicio.

d. Presupuesto.

e. Cruce de estadísticas.

f. Gráficos diversos.

g. Exportar los resultados a formatos como MS Excel.

h. Reportes por libros con respecto a usuarios y de menor movimiento.

2) Todos los usuarios encuentran potencial a una herramienta como Zombi para apoyarlos en la toma de decisiones y sugieren continuar haciendo desarrollos sobre ella. Decisiones que van desde la limpieza de datos del software operacional hasta decisiones sobre los movimientos del material bibliográfico dentro de la biblioteca.

3) Sobre características negativas que vieron en Zombi se tiene lo siguiente:

   a. Las gráficas son muy sencillas, pequeñas y difíciles de manipular. No se ven bien los valores de los ejes.

   b. Un botón para regresar al menú de reportes.

   c. El significado de los íconos no es muy obvio.

   d. La presentación en general se considera que puede mejorarse.

4) Las características positivas que encontraron fueron:

   a. La posibilidad de ordenar datos al nivel de las columnas.

   b. No requiere de mucha capacitación para entenderlo y usarlo.

   c. En general amigable.

   d. La posibilidad de hacer generalización y especialización sobre los datos.

5) Finalmente las sugerencias adicionales que se mencionaron fueron:

   a. Que los usuario puedan crear vistas y almacenarlas para su uso personal.

b. Acceder directamente a resultados de datos que ya fueron previamente calculados puesto que se cada operación es recalculada.
c. Incluir funciones estadísticas.
d. Explorar más técnicas de visualización y mejorar el tiempo de respuesta.

6.3 Resumen de actividades realizadas para lograr los objetivos

Para poder lograr los objetivos de la tesis se tuvo que investigar a profundidad sobre minería de datos y procesamiento analítico en línea, principalmente sobre los trabajos relacionados a su operación integrada, una vez teniendo la visión general del problema y como ha sido resuelto en otras situaciones, así como sus beneficios y limitaciones, se procedió a conseguir dos prototipos uno para minería de datos y otro para procesamiento analítico en línea. Desafortunadamente el campo de investigación sobre procesamiento analítico en línea es muy limitado y solo se encontró el prototipo Mondrian el cual se adapto para usarlo en este proyecto de tesis, un problema fuerte es que no existe mucha documentación sobre Mondrian a pesar de que esta disponible el código fuente.

En el caso de minería de datos se encontraron más opciones, pero se vio que el minero Weka cumplía con las expectativas que se requerían de análisis además de estar muy bien documentado en un libro.