Centro de Investigación Científica y de Educación Superior de Ensenada

INDEXADO Y RECUPERACION DE INFORMACION MULTIMEDIA EN UNA BIBLIOTECA DIGITAL DE TESIS DE POSGRADO

TESIS
MAESTRIA EN CIENCIAS

RICARDO ACOSTA DIAZ

Ensenada, Baja Cfa., Mexico. Mayo de 2000

TESIS DEFENDIDA POR

Ricardo Acosta Díaz

Y APROBADA POR EL SIGUIENTE COMITÉ

Dr. Jesús Favela Vara

Director del Comité

Dr. Federico Graef Ziehl

Miembro del Comité

Dr. Andrei Tchenikh

Miembro del Comité

Mc. Jorge E. Preciado Velasco

Miembro del Comité

Dr. Jesús Favela Vara

Jefe del Departamento de Ciencias de la Computación

Dr. Federico Graef Ziehl

Director de Estudios de Posgrado

3 de mayo del 2000

CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN SUPERIOR DE ENSENADA

DIVISIÓN FÍSICA APLICADA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION

Indexado y Recuperación de Información Multimedia en una Biblioteca Digital de Tesis de Posgrado

TESIS

que para cubrir parcialmente los requisitos para obtener el grado de MAESTRO EN CIENCIAS presenta:

RICARDO ACOSTA DÍAZ

Ensenada, Baja California, México, Mayo del 2000.

RESUMEN de la Tesis de RICARDO ACOSTA DÍAZ, presentada como requisito parcial para la obtención del Grado de MAESTRO EN CIENCIAS en CIENCIAS DE LA COMPUTACIÓN. Marzo del 2000, Ensenada, Baja California, México.

INDEXADO Y RECUPERACIÓN DE INFORMACIÓN MULTIMEDIA EN UNA BIBLIOTECA DIGITAL DE TESIS DE POSGRADO.

Resumen aprobado por:

Dr. Jesús Favela Vara
Director del Comité de Tesis

En el proyecto MIND (Mixed-media Networked Digital Library), se ha desarrollado un ambiente para explorar mecanismos eficientes para el indexado y recuperación de información multimedia en una biblioteca digital de tesis de posgrado. Se almacena e indexa texto e imágenes de los acetatos usados por los tesistas durante su defensa de tesis, el audio y video de la presentación así como el documento completo de la tesis. El audio, video y acetatos son capturados automáticamente por el sistema, utilizando un visualizador de Web estándar.

Por cada tesis se realizó el indexado del documento de tesis y del texto contenido en los acetatos de la presentación basánse en el algoritmo TFIDF, así mismo los acetatos y el video se indexaron utilizando eventos y con ello se creo una biblioteca digital la cual permite realizar consultas en diferentes medios y utiliza mecanismos para navegar entre ellos. Por ejemplo, viendo acetatos de la defensa es posible obtener información mas detallada sobre ellos, visualizando el documento de tesis ó el video correspondiente.

Una interfaz gráfica presenta los acetatos en formato HTML, permitiendo al usuario navegar a través de ellos. Cuando el usuario selecciona un acetato su contenido se despliega en un panel y es posible reproducir el video correspondiente. Otro panel despliega automáticamente las secciones de los documentos que contienen información relacionada con el acetato que está siendo visualizado utilizando el formato PDF, así mismo ofrece una lista de documentos que se relacionan con el tema.

Las pruebas realizadas demuestran que el sistema tiene una precisión del 78% para recuperar documentos de la tesis que explican más a detalle el acetato que se está visualizando y posee una precisión del 84% para detectar el documento más relevante.

Keywords: indexado y recuperación de información multimedia, bibliotecas digitales

ABSTRACT of the Thesis presented by RICARDO ACOSTA DIAZ, as partial requirement to obtain the degree of MASTER IN SCIENCE with specialization in COMPUTER SCIENCE. March 2000, Ensenada, Baja California, México.

MULTIMEDIA INFORMATION INDEXING AND RETRIEVAL IN A DIGITAL LIBRARY OF GRADUATE THESIS.

Approved by:

Ph.D. Jesús Favela Vara
Thesis Advisor

In the MIND (Mixed-Media Networked Digital Library) project, we have developed an environment to explore efficient mechanisms for indexing and retrieving multimedia information from a digital library of graduate thesis. We have stored the text from the slides used for the thesis defense, the audio and video from the presentation of the thesis and the thesis document itself. Audio, video and slides are captured automatically by the system using a standard web-browser.

Full-text indexing of the whole document and the text in the slides were done using the TFIDF algorithm, and event-based indexing of the presentation and the video were used to create a rich multimedia library with support for queries that incorporate different media and navigation mechanisms between these media. For instance, looking at the slides of the thesis defense and asking for more detail information the user is directed to the appropriate place in the thesis document or in the video.

A graphical user interface presents the slides in HTML format, allowing users to navigate through them. When the user selects a slide of any presentation its contents is displayed in a panel and he is able to play its corresponding video segment which is stored in Real Media format. Another frame displays a section of the thesis document that is related to the slide being displayed using PDF format, as well as a list of documents related to that theme.

Tests performed to the system show that it has 78% of precision retrieving thesis documents which explain in more detail the slide that is being shown, and it has 84% of precision detecting the most relevant document.

Keywords: multimedia information indexing and retrieval, digital libraries

DEDICATORIA

*Al **P**adre **G**alindo!!!*

AGRADECIMIENTOS

Al Dr. Jesús Favela, mi director de tesis por su apoyo, amistad y orientación a lo largo de estos casi tres años.

A los miembros de mi comité de tesis, Mc. Jorge Preciado, Dr. Federico Graef y Andrei Tchernik por su tiempo y observaciones para mejorar el trabajo realizado.

A Momis y a Danny por su paciencia y comprensión.

A los H. Enfadosos compañeros del Coolab: Haydeé, Manuelillo, Mireles, Octavio, Cesar, Juan, Adrián, Rafa y Memo, por su amistad y por hacer más ameno mi año de tesista.

A Haydeé y Edelmira, mis dos grandes amigas por brindarme su apoyo siempre en las buenas y en las malas ☺.

A los compañeros de la maestría que me ayudaron a realizar las pruebas al sistema.

A la Mtra. Lulú y al Lic. Victórico por el apoyo desinteresado que me han brindado a lo largo de estos años.

A la revista InstanTips y a mis clientes que a lo largo de 9 meses sin beca hicieron posible, mi permanencia en Ensenada y terminar mi tesis.

Al Centro de Investigación Científica y Educación Superior de Ensenada por el apoyo brindado durante mi permanencia en él.

A la Universidad de Colima, por el apoyo económico brindado para realizar estudios de Posgrado.

Al Consejo Nacional de Ciencia y Tecnología por hacer viable económicamente mis estudios de maestría.

A todas las personas, o instituciones que me ayudaron física, mental, intelectual espiritual o económicamente durante mis estudios de maestría.

A Dios por estar siempre a mi lado, aunque yo no lo mereciera.

TABLA DE CONTENIDO

Lista de Figuras

Lista de Tablas

Capítulo I. Introducción.

I.1 Antecedentes

Durante los últimos cinco siglos, la palabra impresa ha sido la forma dominante de comunicación para difundir información, ya que está representa generalmente un alto nivel de preparación, análisis y autoridad. El sistema social para registrar, compilar, organizar, conservar, utilizar y difundir información, tiene sus orígenes en la década de 1940, y adquiere presencia social a partir del desarrollo y aplicación de la idea de crear sistemas dedicados al almacenamiento, análisis y recuperación de información [Lafuente, 1999].

Los grandes avances en las telecomunicaciones y la computación han abierto la posibilidad de transmitir información en formato digital, a bajo costo y estructurada conforme a nuevos esquemas de escritura, edición, almacenamiento, publicación y recuperación, aumentando su valor al ofrecerla en forma inmediata.

Un ejemplo de estos grandes avances es el "Web", que desde sus inicios a principios de los noventa, se ha convertido en el medio mas accesible para publicar cualquier tipo de material sin tener que recurrir a una casa editorial. Sin embargo, en muchos casos, la información presentada tiende a ser superficial, es por ello que muchos usuarios siguen considerando la biblioteca tradicional como su recurso primario.

Para enriquecer el mundo electrónico, habría que trasladar los libros y revistas impresas a formato digital, pero dada la magnitud de la tarea y los altos costos involucrados, la solución práctica tiene dos aspectos: primero, ser selectivos, ofreciendo en formato digital aquellos materiales que tengan la mayor demanda actual; segundo, comercializar los productos de modo que se cubran los costos de conversión.

El creciente aumento en la publicación y circulación de documentos digitales con texto, sonido e imágenes ha dado lugar a la creación de una categoría denominada *Biblioteca Digital* para aludir a la idea de la creación y manejo de acervos en formatos digitales [Lesk97].

A continuación se presentan el objetivo y la motivación para realizar una investigación en el área bibliotecas digitales.

I.2 Planteamiento del problema

Las bibliotecas juegan un papel central en la educación y el aprendizaje, siendo el mismo caso para las bibliotecas digitales. Entre los principales roles de la biblioteca en el aprendizaje se encuentran [Marchionini, 1995]:

1. Compartir recursos.
2. Preservar y organizar artefactos e ideas.
3. Tienen un rol social e intelectual al acercar ideas y gente.

Uno de los principales recursos de información con que cuentan las universidades son las tesis, principalmente de posgrado, que en ellas se producen. En éstas se detalla buena parte del trabajo de investigación que se produce en la institución. Además sirven como apoyo a cursos y como introducción a temas de actualidad, ya que generalmente presentan una descripción del estádo del arte del tema de estudio.

Los grandes avances de las telecomunicaciones y la computación han causado una revolución en el ámbito educativo y de investigación, por lo que en la actualidad es de gran utilidad estár enterado de los proyectos que se han realizado en los diferentes centros educativos y de investigación.

Dados estos acontecimientos, sería de gran utilidad la creación de una biblioteca digital de tesis de posgrado, la cual proporcionaría información útil para los centros de investigación y al mismo tiempo sería de gran interés para estudiantes a nivel licenciatura y posgrado, ya que ofrecería las siguientes ventajas:

- Permitiría incluir información en distintos medios (los libros no pueden incluir animaciones, audio y video).
- Se podría indexar el texto completo de los documentos contenidos en la biblioteca digital y utilizar máquinas de búsqueda para su acceso.
- La información almacenada en un sólo lugar podría ser accesada al mismo tiempo por varias personas, a diferencia de un libro del cual se requiere una copia por cada usuario simultáneo.
- La presentación del material podría ser adaptada a cada usuario.
- Ofrecería información que puede ser utilizada sin las restricciones de propiedad intelectual.
- Mayor circulación del material permitiendo su acceso por medio de Internet.

- La gran mayoría de las tesis hoy en día son escritas en computadora con lo que se disminuye el costo de captura.
- Las instituciones educativas están interesadas en dar difusión a sus productos.
- Algunas áreas de investigación como es la Ciencia de la Computación cuentan con pocos investigadores en México los cuales se encuentran relativamente aislados [INEGI, 1994]. Los estudiantes de posgrado en México, conocen mas lo que se hace en otros países que los trabajos relacionados que se desarrollan en su propio país.

La posibilidad de incluir medios distintos al texto en bibliotecas digitales, ha originado en los últimos años investigación relativa a la captura indexado y recuperación de texto, imágenes, audio y video. Si bien se han obtenido resultados prometedores en la mezcla de medios para la recuperación de información, es claro que aún quedan muchos problemas abiertos, como se describe en el capítulo 3.

I.3 Objetivo

El objetivo principal de está investigación es proponer una arquitectura para y un ambiente que permita realizar en forma automática y eficiente, la captura, indexado y recuperación de información multimedia en una biblioteca digital de tesis de posgrado.

Para ello se construirá una biblioteca digital, se evaluarán algoritmos existentes para el indexado y recuperación de información y se adaptará e implementará uno de ellos en está biblioteca digital.

I.4 Alcances y limitaciones

La presente investigación se limitará a desarrollar una arquitectura y un ambiente que permita capturar, indexar y recuperar automáticamente información multimedia usando eventos y palabras claves. Como primera instancia se usarán palabras clave para alinear los acetatos con el documento de tesis y eventos para alinear el acetato con el audio y el video. Dicho prototipo permitirá al usuario utilizar el medio más conveniente (acetatos, texto, audio y video) de tal manera que pueda encontrar información mas detallada sobre un tema visualizado en un acetato en el documento de tesis.

I.5 Organización

El capítulo II inicia con una introducción a las bibliotecas digitales. El capítulo III hace una revisión del estádo del arte de las bibliotecas digitales multimedia y los métodos tradicionales para el almacenamiento y recuperación de información. El capítulo IV presenta los requerimientos del sistema MIND (Mixed-media Networked Digital Library) un ambiente para la captura, indexado y recuperación de información multimedia en una biblioteca digital de tesis de posgrado, desarrollado como parte de está investigación. El capítulo V describe los procedimientos para el indexado, recuperación y visualización de medios en el sistema desarrollado. El capítulo VI presenta los resultados obtenidos de la evaluación del sistema. Finalmente, el capítulo VII presenta las conclusiones, logros obtenidos, aportaciones y hace recomendaciones para trabajo futuro.

Capítulo II. Bibliotecas Digitales.

II.1 Introducción

El origen del concepto de *biblioteca digital* puede ubicarse en los escritos de visionarios como Vannevar Bush [1945, 1959], en los que se planteaba la posibilidad y la necesidad de utilizar los adelantos tecnológicos acelerados por la Segunda Guerra Mundial para generar nuevas formas de almacenar y mantener conocimientos, así como nuevos ambientes y prácticas para el trabajo intelectual. Muchas de estás ideas han servido de inspiración para conceptos hoy populares como hipertextos [Nelson, 1977], hipermedios y sistemas para trabajo cooperativo. En los años sesenta, J. C. R. Licklider [1965] desarrolla una visión ya sobre el uso de las computadoras como un elemento que revolucionaría el funcionamiento de las bibliotecas tradicionales, introduciendo entre otras funciones el indexado inteligente de documentos. No es sino hasta la presente década cuando se identifica a las bibliotecas digitales como un factor clave de competitividad y varios países inician programas encaminados a apoyar su desarrollo y a buscar avances dramáticos en los medios de recolectar, almacenar, organizar, hacer disponibles y explotar grandes volúmenes de datos en forma digital. En este sentido, sobresalen las siguientes iniciativas, programas y proyectos [Lesk 1997]:

1. En 1993 fue lanzada la primera fase de La Iniciativa de Bibliotecas Digitales (DLI, por sus siglas en inglés) en los Estádos Unidos, auspiciada conjuntamente por la Fundación Nacional para la Ciencia (NSF), la Agencia de Proyectos Avanzados de Investigación

(ARPA), y la Administración Nacional de Aeronáutica y del Espacio (NASA) brindaron apoyo a seis grupos de investigación integrados por universidades e industrias con alrededor de 24 millones de dólares para realizar proyectos de bibliotecas digitales. Los aspectos a investigar incluyeron: procesamiento y consulta de imágenes, indexado y recuperación de datos espaciales, video digital, procesamiento de lenguaje natural, uso en ambientes de aprendizaje e infraestructura de redes y bases de datos. Actualmente [DLI-P2, 1998] en la segunda fase de dicha iniciativa se tiene como propósito proveer liderazgo en los fundamentos para la investigación de la próxima generación de bibliotecas digitales, proponer el uso y usabilidad de los recursos de información que se encuentran distribuidos globalmente en la red y estimular a las comunidades nuevas y existentes a enfocarse en áreas de aplicaciones innovadoras. Hasta la fecha se ha triplicado la cantidad de propuestás recibidas con relación a la primera fase (230 de las cuales requieren mas de 400 mil dólares cada una) [DLib, 1999].

2. La iniciativa de la Biblioteca del Congreso de Estados Unidos para recaudar 60 millones de dólares destinados a la digitalización de alrededor de 5 millones de documentos de su colección.
3. El programa británico E-LIB, que asignó 20 millones de libras esterlinas para financiar 35 proyectos relacionados con digitalización de documentos, publicación bajo demanda, desarrollo de infraestructura y estudios socio-tecnológicos en el área.
4. Los esfuerzos en Francia para construir una Biblioteca Nacional digital, que contiene 10,000 libros en texto ASCII y 100,000 libros en formato de imágenes digitales. Francia también sobresale por su participación en el proyecto *Bibliotheca Universalis,*

el cual involucra la digitalización de recursos culturales claves de muchos países, enfocándose a la preservación de la cultura francesa y a la difusión de sus logros.

5. La gran variedad de proyectos (desde nuevos sistemas de catalogación hasta el desarrollo de diccionarios multimedios) apoyados por la Unión Europea
6. El proyecto de digitalización de la Biblioteca Nacional del Congreso (Dieta) del Japón, para el cual se han asignado 50 millones de dólares.
7. Los esfuerzos de digitalización en las instituciones culturales de Australia para lo que se asignaron 10 millones de dólares australianos en 1995.
8. El proyecto "Isla Inteligente" de Singapur, el cual incluye la interconexión de bibliotecas mediante redes de alta velocidad.

La competencia por el financiamiento asignado para bibliotecas digitales en 1993 por DLI generó una serie de actividades y alianzas entre integrantes de la academia e industria en los Estádos Unidos, de tal suerte que la convocatoria para un Primer Congreso Internacional sobre Bibliotecas Digitales (DL'94) obtuvo una respuestá muy entusiasta [Schnase *et al.* 1994]. En este evento, realizado en College Station, Texas, se discutieron los diferentes aspectos de investigación, enfoques y proyectos en desarrollo y se formalizaron los esfuerzos para estáblecer una comunidad internacional de investigación y desarrollo en bibliotecas digitales. Se presentaron ahí también los resultados del Taller sobre Acceso Inteligente a Bibliotecas en Línea (IEEE CAIA'94) [Gladney *et al.* 1994], financiado por NSF.

DL'94 dio lugar a DL'95 [Shipman *et al.* 1995] y a la serie de conferencias (ACM DL'96-'99) que en 1996 fue adoptada por la ACM como el foro internacional por excelencia en el área. Paralelamente, la Sociedad de Computación de IEEE ha auspiciado la serie de congresos denominada "Avances en Bibliotecas Digitales" (ADL'94-'99) [IEEE 1998]. Otros foros para el trabajo actual en bibliotecas digitales incluyen el Simposio Internacional sobre Bibliotecas Digitales, celebrado anualmente en Japón [ISDL 1997], y el Congreso Anual sobre Bibliotecas Digitales del Reino Unido [ELVIRA 1997].

Entre las revistas arbitradas importantes que han surgido al establecerse bibliotecas digitales como área de investigación, sobresalen: "International Journal on Digital Libraries" [IJDL,1998], "Journal of Digital Information" [JoDI 1998] y D-Lib Magazine [D-Lib,1998].

II.2 Componentes de una Biblioteca Digital

El término biblioteca digital es tomado con un significado diferente dependiendo del tipo de usuario que se trate. Para algunos simplemente sugiere la computarización de las bibliotecas tradicionales, para otros, quienes han estudiado la ciencia de las bibliotecas, se llama a la ejecución de las funciones de la biblioteca tradicional en una nueva forma, la cual contiene nuevos tipos de recursos de información nuevas formas de adquisición, nuevos métodos de almacenamiento y preservación, mas confianza en sistemas electrónicos y redes, y un cambio dramático en las prácticas económicas, organizacionales e intelectuales.

Para muchos profesionales en computación, una biblioteca digital es simplemente un sistema de información distribuido basado en texto, una colección de servicios de información distribuidos, un espacio distribuido de información interligado, o un sistema de información multimedia en red [Slonim94]. Para los usuarios del Web, una biblioteca digital sugiere más de lo mismo, con mejoramiento en el desempeño, organización y funcionalidad y usabilidad.

Una biblioteca digital está formada por tres clases de elementos: datos, metadatos y procesos [Nurnberg, 1995]. Los datos constituyen el material de la biblioteca (Hypernovela, visualización científica y programas de computadora). Los metadatos son información acerca de las bibliotecas y sus datos (Índices dinámicos, estructuras personales, anotaciones). Y los procesos son funciones activas que se ejecutan sobre los elementos de la biblioteca (Búsquedas en todo el texto, búsquedas en imágenes y videos, recuperación por agentes).

II.3 Ventajas de una Biblioteca Digital

Aunque las bibliotecas digitales heredan las funciones de sus análogos convencionales y comparten algunas de sus características, su importancia radica en lo que las hace diferentes. El medio digital hace posible, entre otras cosas, liberar a los usuarios de las limitaciones de los objetos físicos y sus contenedores (edificios, pastas, libreros), y hace posible nuevas prácticas y oportunidades, como el trabajo cooperativo remoto, la

visualización de información desde múltiples perspectivas y la personalización de recursos y servicios de la biblioteca.

Los usuarios de documentos tradicionales, por ejemplo libros, reciben información de documentos estáticos mientras que los documentos digitales pueden actualmente interactuar con el usuario y responder a sus necesidades. Está interacción convierte a los documentos digitales en herramientas poderosas, Por ejemplo, los lectores que tienen una visión pobre pueden requerir que el documento incremente el tamaño de la letra y de las imágenes, o quizá el usuario quiera que el documento le sea presentado en un estilo diferente donde le sea mas fácil entenderlo o navegarlo. Algunos documentos interactivos también permiten a los usuarios agregarles directamente nueva información u otro formato (gráficos o ilustraciones).

Los documentos digitales son baratos para almacenarlos y distribuirlos, en consecuencia disminuyen los costos de publicación, creación y mantenimiento de la biblioteca. No se requieren inventarios ni se requieren enormes edificios para almacenar las colecciones. Los documentos digitales pueden cambiar dinámicamente, por lo tanto la versión actualizada del documento puede reemplazar instantáneamente la anterior cada vez que se desee.

Los documentos digitales son almacenados y manejados digitalmente pero pueden ser producidos como "hard copy" en medios tal como libros, CDs o como una representación

(música o movimiento). Los documentos digitales no eliminan la conveniencia de la salida "hard copy", ellos la enriquecen.

La biblioteca digital reduce las restricciones de tiempo ya que los usuarios pueden navegar, buscar y seleccionar documentos publicados las 24 horas del día los 365 días del año, mientras que los usuarios de las bibliotecas tradicionales están limitados a las horas de oficina. La biblioteca digital permite también acceso virtual inmediato a los documentos que se necesitan sin tener que esperar en una cola.

La biblioteca digital, no importa donde se localice físicamente siempre es local para el usuario. El autor está siempre solo a unas cuantas teclas o "clicks" del usuario. Un individuo o institución puede publicar en un servidor localizado en cualquier lugar del mundo y la biblioteca puede ser accesada por usuarios que se encuentren en cualquier lugar del planeta. La biblioteca digital reduce la necesidad de estár físicamente cerca de los documentos.

Las bibliotecas digitales permiten a los usuarios personalizar la manera en que los documentos están organizados, donde y cuando se almacenarán, que documentos aparecerán y bajo que condiciones, el formato y apariencia de cada documento. Esto permite actualmente a la biblioteca digital aparecer como personal y única.

En una biblioteca tradicional, el contenido de información y de conocimiento disponibles a cualquier usuario está limitado en varios aspectos:

1. Por el espacio físico.
2. Por el número de publicaciones que alberga.
3. Por la poca información disponible en cuanto a su contenido (tema, título, autor e índice, a lo sumo).
4. Por la cantidad de personal que la atiende y por el conocimiento que éste tiene sobre las características y contenido del acervo.
5. Por el número de usuarios que comparten el material de su interés.

En contraste, el contenido de información y de conocimiento disponible en una biblioteca digital es mucho más extenso, ya que:

1. No está condicionado por el espacio físico sino por las capacidades de la computadora en que se almacena.
2. Permite compartir recursos de terceros.
3. Se puede extender a toda la información y el conocimiento que está disponible en Internet.
4. Se dispone de información sobre la información (metadatos), que permite una localización más eficiente de información más relevante, en un contexto mucho más amplio de posibilidades.
5. La cantidad de personal que atiende a la biblioteca no es un factor relevante ni se depende de su conocimiento sobre el acervo para hacer búsquedas más eficientes.
6. Un gran número de usuarios pueden tener acceso simultáneo a la misma información.

La biblioteca digital conserva las funciones básicas de toda biblioteca: la recopilación, organización y catalogación de información, así como los servicios de búsqueda, localización y reproducción de información se conservan en las bibliotecas digitales, aunque la forma de realizarlas es diferente a la tradicional.

Tabla I. Funciones básicas en la Biblioteca Tradicional y en la Biblioteca Digital

OBTENCIÓN DE INFORMACIÓN RELEVANTE	BIBLIOTECA TRADICIONAL	BIBLIOTECA DIGITAL
Patrón de búsqueda	Editorial, tema, título, autor, fecha	Cualquier palabra clave o combinación de ellas
Localización	Apoyo del personal bibliotecario	Automatizada utilizando metadata
Búsqueda	Manual o automatizada en el catálogo de la biblioteca	Automatizada en acervos distribuidos
Recuperación	Referencia de localización	Referencias al texto completo con descripción de contenidos
Análisis	Revisión física del material identificado	Revisión de información relativa a contenido y posibilidad de explorar el texto completo
Selección	Páginas o secciones de las publicaciones	Referencias al texto completo con descripción de contenidos
Reproducción	Fotocopiado	Impresión o almacenamiento del texto completo

II.4 Clasificación de Bibliotecas

Las bibliotecas tradicionales basadas predominantemente en materiales impresos, convencionalmente se clasifican en cinco grupos [Berkeley, 1998]:

Comprensiva: Biblioteca con acervos que incluyen todo material significativo (publicaciones, manuscritos, etc.) y el conocimiento en todas las lenguas en un campo determinado.

Investigación: Biblioteca con acervos integrados por las principales publicaciones requeridas para disertaciones e investigación independiente, incluyendo material en el que se reportan resultados de investigaciones, nuevos descubrimientos, resultados científicos experimentales y toda la demás información de utilidad para los investigadores, así como material de referencia y monografías especializadas, revistas y publicaciones periódicas, *journals* e índices en un campo determinado de conocimiento.

Estudio: Biblioteca cuyo acervo está constituido para atender las necesidades de estudiantes de nivel profesional y de posgrado, de acuerdo a los programas de estudio para una área determinada de conocimiento.

Básica: Biblioteca constituida por una selección de publicaciones que permiten la introducción a un tema determinado

Mínima: Biblioteca que contiene pocas selecciones sobre un tema específico.

Está distinción es importante no solo por las diferencias existentes en cuanto a sus alcances, propósitos y poblaciones atendidas, sino porque cada una de ellas requiere de una organización e infraestructura diferente, pero que es similar en cada caso a todas las bibliotecas que integran las diferentes categorías.

Además de la categorización anterior, en el caso de las biblioteca digitales se ha propuesto distinguir otras cuatro categorías:

Archivo: La biblioteca almacena las colecciones y se compromete a mantener disponibles sus acervos en forma permanente.

Servicio: La biblioteca colecciona material pero sin el compromiso de mantenerlo disponible en forma permanente.

Espejo: La biblioteca almacena una copia de material coleccionado en otra biblioteca, sin compromiso respecto a su organización y a mantenerlo disponible en forma permanente.

Referencial: La biblioteca solo colecciona referencias a los sitios en donde reside la información de otras colecciones, sin ningún tipo de compromiso respecto a la permanencia del material.

También en este caso cada tipo de biblioteca demanda recursos y organizaciones diferentes para servir a sus usuarios.

En base a estos dos criterios se construye la matriz de posibilidades para clasificar bibliotecas digitales (Tabla II).

Tabla II. Matriz de posibilidades para clasificar una biblioteca digital [AMC, 1998]

	ARCHIVO	SERVICIO	ESPEJO	REFERENCIAL
COMPRENSIVA				
INVESTIGACIÓN				
ESTUDIO				
BÁSICA				
MÍNIMA				

II.5 Bibliotecas Digitales en México

Hasta ahora, los proyectos en el área se han dado en forma un tanto aislada. Por parte de la comunidad de bibliotecas, la mayoría de los esfuerzos se han orientado al uso de la tecnología para la automatización de catálogos y su consulta a través de la red global.

Sin embargo existe un interés creciente por participar en el desarrollo de bibliotecas digitales, por lo que recientemente se realizó el primer "Workshop on Digital Libraries" [WDL,1999] cuyo principal objetivo fue reunir investigadores de México y Estádos Unidos para fomentar la participación bilateral y el intercambio de experiencias.

Por parte de la comunidad mexicana de ciencias de la computación, los proyectos que se están realizando son variados y algunos apenas empiezan a insertarse al contexto de bibliotecas digitales. Sin embargo, se ha propiciado un aceleramiento en el intercambio y la integración de los investigadores del área.

II.5.1 Proyectos en desarrollo en México

A continuación se describen algunos de los proyectos mas importantes que se encuentran en desarrollo en México.

- *Proyecto Phronesis [Phron,1999]*

Con el apoyo del CONACYT (proyectos REDII) y del ITESM Campus Monterrey, en 1988 inició el proyecto Phronesis, el cual plantea desarrollar herramientas de dominio público que, permitan la fácil creación de acervos digitales y su acceso a través de Internet. El trabajo realizado se ha enfocado en los niveles de manejo de datos, abstracción y

servicios de Biblioteca Digital. En particular las áreas en las que se trabaja involucran aspectos de almacenamiento y recuperación de información.

Como resultado parcial se ha desarrollado el Servidor Phronesis, el cual permite que cualquier persona (bibliotecario, autor o editor) con acceso a Internet pueda almacenar documentos enviándolos vía WWW. y especificando los metadatos necesarios para facilitar su recuperación. Los metadatos se basan en el estándar internacional *Dublin Core* definido por *OCLC* (On Line Computer Library Center). El Servidor Phronesis permite que los usuarios realicen búsquedas booleanas y categorizadas (*ranked search*) . Las búsquedas se realizan en el contenido completo del documento (*full text search*) y en sus metadatos. Una vez que el usuario localiza el documento, es posible tener acceso directo al mismo. Es posible contar con varios Servidores Phronesis donde cada instancia podría almacenar documentos con una cierta temática. Los usuarios pueden realizar búsquedas en diferentes Servidores Phronesis en paralelo de una manera rápida y transparente.

- *Biblioteca Digital Latinoamericana y Caribeña [BDIC, 1999]*

Con un presupuesto de 237,200 dólares y con el objetivo de conformar una colección básica de aproximadamente 100 obras, libres de derecho de autor, de cada uno de los países de Ibero América y el Caribe representativas de sus respectivas culturas, la cual será editada en discos digitales y distribuidas entre las instituciones culturales y educativas de la región, La Universidad de Colima y la UNESCO, con la participación de expertos del Centro Nacional de Información de Ciencias Médicas de Cuba y del Instituto Brasileño de Investigación Científica y Tecnológica (IBICT), han trabajado en un proyecto piloto de Biblioteca Digital Latinoamericana y del Caribe. Como producto inicial se ha presentado

en un CD-ROM, un ejemplo de biblioteca digital, cuyo objetivo principal es entregar a las bibliotecas de la región Iberoamericana y del Caribe interesadas en sumarse a está gran biblioteca regional y pública, una metodología general para su creación. El disco, por lo tanto, sirve para que cualquier biblioteca interesada disponga de las instrucciones para digitalizar, catalogar y clasificar la información, así como para su ubicación en Internet. Con este propósito explica detalladamente y demuestra la utilización de las herramientas mas contemporáneas, tales como: los lenguajes HTML, SGML, XLS; los clasificadores Global Information Locator Service (GILS), Dublin-Core Metadata; y el protocolo Z39.50, ISO23950. El CD-ROM, contiene una muestra de 300 documentos en audio, video, texto o imagen aportados por los archivos nacionales y bibliotecas de 27 países latinoamericanos y del Caribe. En principio, la biblioteca digital puede ser consultada sólo a través de discos compactos; posteriormente estará disponible en Internet.

- *Biblioteca Digital Florística [BDF,1999]*

En 1996 se estábleció un grupo que ha venido trabajando en la definición de una arquitectura, modelo de datos y ambientes e interfaces de usuario para una biblioteca digital que apoye las actividades de la comunidad interesada en biodiversidad. La Biblioteca Digital Florística ha recibido apoyo del Centro de Informática Botánica del Jardín Botánico de Missouri por más de 200 mil dólares incluyendo equipo de cómputo, becas para estudiantes, sueldos para asistentes de investigación y viáticos. Los resultados incluyen el planteamiento de una arquitectura distribuida orientada a servicios de usuario, interfaces basadas en agentes para agilizar la introducción masiva de datos, ambientes de

colaboración para grupos virtuales, interfaces para visualización de datos taxonómicos y para personalización de grandes espacios de información.

Capítulo III. Bibliotecas Digitales y Recuperación de Información Multimedia

III.1. Servicios Básicos de una Biblioteca Digital

Con el advenimiento de Internet, ha llegado una nueva ola de tecnologías de búsqueda y recuperación de información y muchos usuarios tradicionales que antes solo hacían uso de la información ahora se han convertido en generadores, recopiladores y organizadores de la misma. Todo esto coincide también con el abaratamiento de los medios necesarios para digitalizarla y almacenarla.

Los servicios que brindan la mayoría de las bibliotecas digitales son aquellos que permiten al usuario; buscar, localizar, seleccionar, recuperar y reproducir información. Es por ello que en Estádos Unidos la Iniciativa de Bibliotecas Digitales en su segunda fase, está apoyando preferentemente tres categorías de investigación [NSFII, 1998]:

1. *Creación de acervos y contenido digital.* Eficiencia en la captura, representación, preservación y almacenamiento de información.
2. *Sistemas.* En cómputo y en sistemas se están apoyando proyectos e iniciativas relacionadas con arquitecturas abiertas, agentes inteligentes, interoperabilidad, captura, representación y digitalización de información multimedia avanzada.

3. *Búsqueda y recuperación de información e interfaces con el usuario.* Dentro de está área se financian proyectos relacionados con el mejoramiento y desarrollo de nuevas herramientas y tecnologías para búsqueda y recuperación de información.

A continuación se presenta una descripción de los principales mecanismos utilizados para recuperar información multimedia.

III.2 Definición de Multimedia

El término multimedia es definido de diferentes formas dependiendo del área de aplicación. Para algunos [Guojun, 1996] la multimedia es el resultado de combinar medios estáticos (texto, gráficas, imágenes) y dinámicos (video, animaciones) para atraer y estimular más los sentidos del usuario ya que no están limitados únicamente a imágenes impresas y estáticas. Ellos pueden incluir imágenes con movimiento, animación, sonido, reconocimiento y síntesis de voz, datos numéricos, bases de datos, representaciones de procesos y cualquier cosa que pueda ser representada en formato digital. Otros definen el término como la unión entre la computadora y la televisión. La definición mas apropiada de multimedia dentro del área de bibliotecas digitales y manejo de información digital, es la siguiente [Fluckiger,1995]:

"El campo relacionado con la integración controlada por computadora, de texto, gráficas, video estático y dinámico, sonido, y cualquier otro medio donde cada tipo de información pueda ser representado, almacenado, transmitido y procesado digitalmente".

III.3 Indexado y Recuperación de Información

Con el gran crecimiento que ha tenido la información digital, se ha hecho necesaria la creación de biblioteca digitales para organizarla y esto ha generado en los últimos años investigación en el área de recuperación de información.

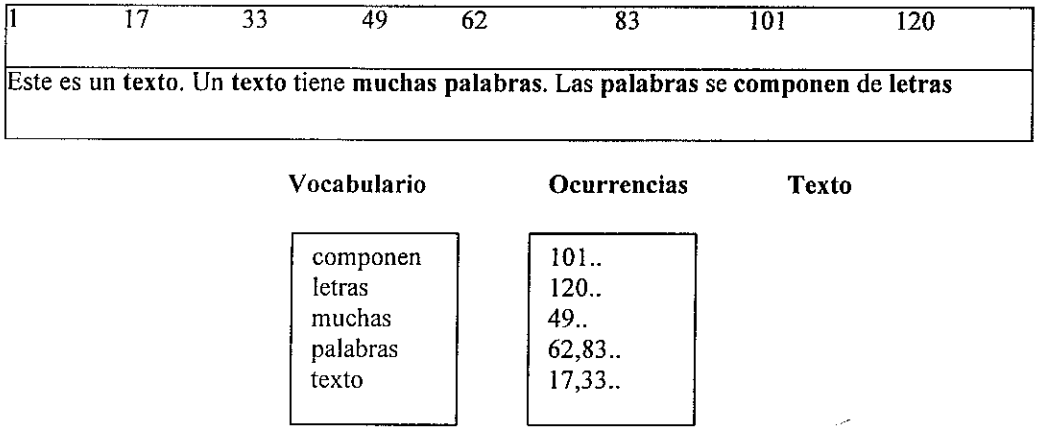
III.3.1 Indexado y Recuperación de Texto

Para hacer el indexado automático de texto, cada uno de los documentos almacenados se procesa en base a las palabras contenidas en ellos, algunas veces complementadas por sus metadatos los cuales describen el contenido del documento [Salton, 1986] y se realiza el siguiente proceso:

1. Se obtienen todas las palabras que constituyen los documentos.
2. Se eliminan las palabras con una frecuencia muy alta (artículos, pronombres personales, etc). Estás palabras son comúnmente llamadas "stop words"
3. Se identifican los mejores términos de indexado y se asignan al documento correspondiente, durante este paso se hace la eliminación de sufijos y opcionalmente pueden utilizarse sinónimos lo cual reduciría aun más el índice que representará al documento.
4. Se eligen las palabras que serán utilizadas para el indexado.

Un índice invertido es un mecanismo orientado a palabra para indexar documentos de texto, su estructura se compone básicamente de dos elementos: el vocabulario y las ocurrencias. El vocabulario es el conjunto de todas las posibles palabras contenidas en

el texto. Por cada palabra se almacena la posición donde aparece en el texto. El conjunto de todas esas listas es llamado ocurrencias, a continuación se muestra un ejemplo:



El algoritmo de búsqueda en un archivo invertido sigue tres pasos:

- *Búsqueda en el vocabulario*, las palabras de la consulta son buscadas en el vocabulario.
- *Recuperación de ocurrencias* se recupera la lista de todas las ocurrencias encontradas
- *Manipulación de ocurrencias*, las ocurrencias son procesadas.

La frecuencia de las palabras es clave dentro de un documento, es un indicador de importancia de dichos términos [Luhn, 1958]. Un estándar de alto desempeño y costo modesto es el utilizar la función inversa de la frecuencia de las palabras en el documento para obtener el peso o factor de importancia que indica la relevancia del término en el documento. Los términos con pesos más altos, pueden ser asignados a los documentos de la colección con o sin los pesos de los términos [Salton, 1983]. Las consultas son realizadas utilizando palabras clave para recuperar todos aquellos documentos que las contengan.

III.3.2 Indexado y Recuperación de Imágenes y Video

Los primeros trabajos de recuperación de imágenes basada en su contenido tomaron dos direcciones [Grosky, 1989; Jain, 1993; Narasimhalu, 1995]. En el primero los contenidos de las imágenes son modelados como un conjunto de atributos extraídos manualmente y manipulados con sistemas manejadores de bases de datos convencionales. Las consultas son realizadas usando estos atributos. En la segunda alternativa se depende de un sistema encargado de la extracción automática de atributos para luego insertar la imagen con sus atributos a la base de datos.

En los últimos años el área de recuperación de imágnes por contenido [Niblack,1993] ha estudiado mecanismos para buscar imágenes digitales dando como entrada atributos de la imagen, (colores, forma, textura, etc.), o bien otras imágenes semejantes a aquéllas que se desean encontrar. Para esto se han utilizado algoritmos de procesamiento de imágenes como la comparación de histogramas [Falatous,1994] y la transformada wavelet [Jacobs, 1995], entre otras.

Las técnicas de procesamiento de imágenes también han sido usadas para el indexado de video [Brown, 1995, Pentland, 1997], ya que el video está formado por un número de imágenes tomadas a cierta velocidad, las cuales forman escenas, por lo que sería inadecuado procesar individualmente cada imagen. Afortunadamente el video puede ser dividido en segmentos. Estos son un conjunto de imágenes que tienen características similares, como la descripción de una misma escena, la presencia de un objeto o persona,

documento de tesis al que hace referencia a dicho acetato, así como el audio o el video.

- El usuario puede navegar y manipular independientemente los medios (textos, acetatos, video).

En el siguiente capítulo se presenta un análisis de los requerimientos planteados para el desarrollo de este sistema.

Capítulo IV. Requerimientos del Sistema MIND.

IV.1 Introducción

Tomando en cuenta los elementos que componen una biblioteca digital y sus principales características funcionales, para diseñar un modelo de biblioteca digital adecuado deben responderse al menos las siguientes interrogantes [Stanford, 1998]:

1. ¿Cuál es la arquitectura adecuada de la biblioteca digital de tesis, en términos del contenido, equipamiento y servicios?
2. ¿Cuáles son los formatos más adecuados para almacenar documentos e información?
3. ¿Qué protocolos se requieren para la integración de repositorios y servicios?
4. ¿Cómo se puede localizar la información más relevante entre los repositorios distribuidos?
5. ¿Cómo debe ser la interface con el usuario?

Para poder responder estás interrogantes será indispensable desarrollar una plataforma de prueba con características adecuadas para evaluar las opciones técnicas y operativas que se vayan identificando en el proceso de búsqueda de respuestá a estás interrogantes.

IV.2 Alcance del sistema

MIND (Mixed Media Networked Digital Libray) debe ser un prototipo de un software fácil de usar, que trabaje sobre el protocolo TCP/IP, el cual permita a individuos y

organizaciones con recursos computacionales limitados y pocos conocimientos sobre bibliotecas digitales iniciar operando sus propias bibliotecas y publicaciones digitales. El software debe permitir realizar búsquedas, visualización, selección, impresión y distribución de documentos digitales a través de un visualizador de "Web".

IV.3 Requerimientos del sistema

El sistema debe ofrecer un ambiente mediante el cual el usuario pueda crear una biblioteca digital de tesis, realizando en forma automatizada la captura y el indexado de audio, video de la defensa de tesis, el texto del material usado para la defensa, así como el texto completo del documento de tesis, las imágenes y tablas contenidas en este. Además debe ofrecer mecanismos para la recuperación y manejo de ésta información de manera eficiente y flexible.

Se consideran básicos los siguientes requerimientos en el diseño de la biblioteca digital:

1. *Abierta al público-* para que todos los estudiantes, profesores e investigadores puedan tener acceso a sus servicios y recursos.
2. *Distribuida-* porque será imposible almacenar toda la información en un solo servidor conforme crezca el número de tesis almacenadas.
3. *Integradora-* de modo que sea posible acceder a sus servicios y recursos distribuidos, desde los equipos convencionales que tengan enlace a Internet.

4. *Dinámica-* para que se puedan seguir incorporando nuevas tesis.
5. *Extensible-* de modo que puedan añadirse nuevos elementos (tipos de datos, servicios, etc.);
6. *Modular-* en el sentido de que se pueden añadir o remover algunos componentes sin afectar su funcionamiento
7. *Escalable-* porque la infraestructura de cómputo y de telecomunicaciones es muy heterogénea y reducida aún en la mayoría de las instituciones.
8. *Contar con audio y video de calidad aceptable-* Dada la experiencia que se tiene al recuperar información del "Web" en ocasiones se prefiere tener un buen audio en lugar del video, por lo que se debe incluir la opción para que el usuario pueda seleccionar si desea audio y video o solo audio
9. *Sincronía entre los acetatos y el audio o video-* Las partes más importantes al momento de realizar consultas a las presentaciones de tesis es poder ver y saber que fue lo que se dijo sobre el acetato que se está visualizando. Por esto es muy importante poder sincronizar la reproducción del audio o video con el despliegue de los acetatos.
10. *Una captura sencilla-* La captura de las presentaciones se debe realizar de manera automática cuidando que no se vea afectada la forma natural de realizar las presentaciones.
11. *Consultas por palabra clave-* Utilizando palabras clave, el usuario deberá poder recuperar información almacenada en los documentos de la presentación o del documento de tesis.

IV.4 Contexto de uso del Sistema

La implementación del sistema para la captura, indexado y recuperación de información multimedia servirá inicialmente para almacenar las defensas de tesis que se realicen en el posgrado de Ciencias de la Computación en CICESE, y con ello se creará una Biblioteca Digital de Tesis de Posgrado del área de Computación. Posteriormente podrán incluirse tesis de otras áreas.

IV.5 Restricciones de acceso y seguridad

En el primer prototipo no se contemplará ninguna restricción de acceso y seguridad adicional a la que proporcionen los servidores de Web que servirán de repositorios de la información, por lo que los usuarios tendrán acceso a la información en cualquier momento.

En el siguiente capítulo se presenta la descripción de cada uno de los módulos del sistema.

Capítulo V. Indexado, Recuperación y Visualización de Medios en el Sistema MIND.

En este capítulo se describe la arquitectura del sistema MIND (Mixed-media Networked Digital Library), así como cada una de las partes que lo integran y los motivos por los cuales se tomó la decisión de incluirlas.

Como se mencionó anteriormente, el sistema MIND se apoya, en SICREP (Sistema para la Captura y Reproducción de Cursos Electrónicos) [Garcilazo, 1998] para realizar la captura de la información. A continuación se da una breve introducción a la arquitectura de este sistema así como la descripción completa del sistema MIND y la interacción entre ambos sistemas.

V.1 Arquitectura del sistema SICREP

En la figura 5, se presenta un esquema general de la arquitectura del sistema SICREP, la cual consta de cuatro elementos principales: el presentador, la audiencia, el servidor SICREP y por último el servidor Web.

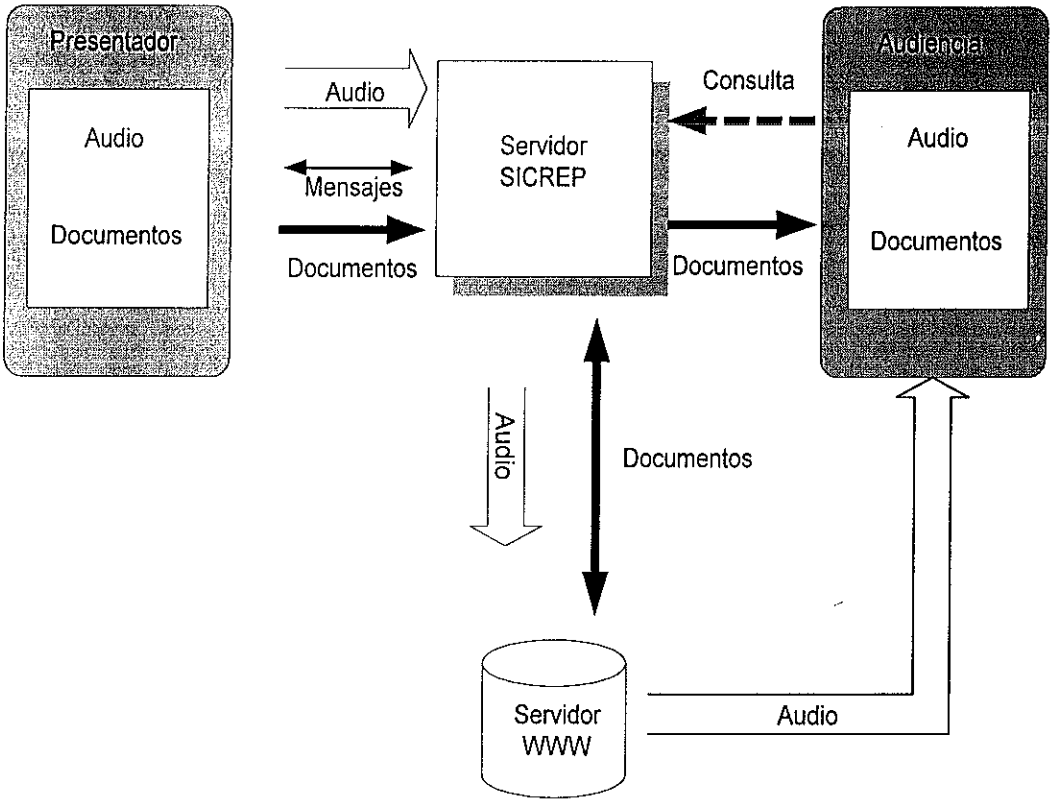


Figura 5. Vista general de la arquitectura del sistema SICREP

El servidor SICREP recibe el audio del presentador y de los acetatos que en este caso son páginas HTML. Además de está interacción, se tiene entre estos dos elementos un protocolo de mensajes el cual es una parte importante en la captura de la información. Por otro lado, la audiencia tiene conexión tanto con el servidor SICREP como con el servidor de Web. Del primero, la audiencia recibe como respuestá a la consulta que envió, los documentos del servidor SICREP. Con el segundo, se tiene una conexión para accesar los archivos de audio que son parte esencial de la consulta realizada por la audiencia. Entre los dos tipos de servidores la conexión que existe permite intercambiar la información, es decir, los archivos de audio y páginas HTML.

V.2 Arquitectura del sistema MIND

La arquitectura de MIND está diseñada para una rápida implementación, facilidad de uso, bajo costo y para evolución futura en componentes, funciones y estándares.

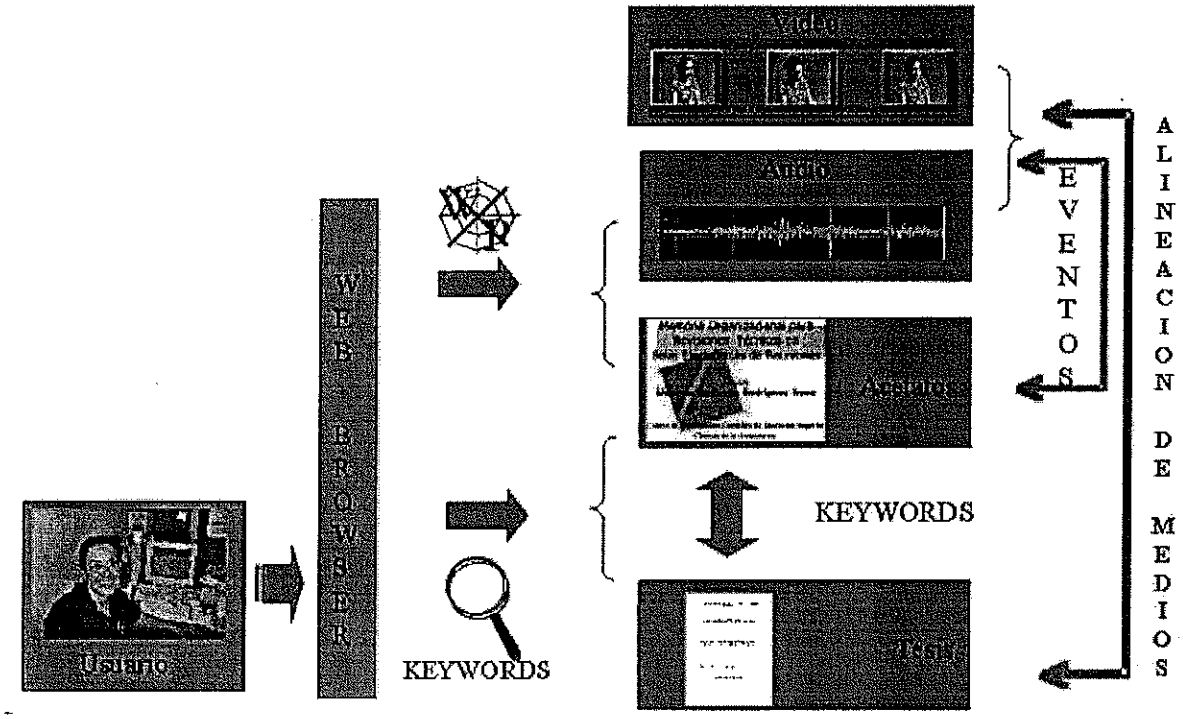


Figura 6. Arquitectura de MIND

MIND permite almacenar y recuperar documentos (imágenes, documentos PDF y HTML) y sincronizarlos con audio y video sobre Internet, poniéndolos a disposición de usuarios que cuenten con una computadora que contenga tarjeta de audio y bocinas, utilizando un visualizador de Web.

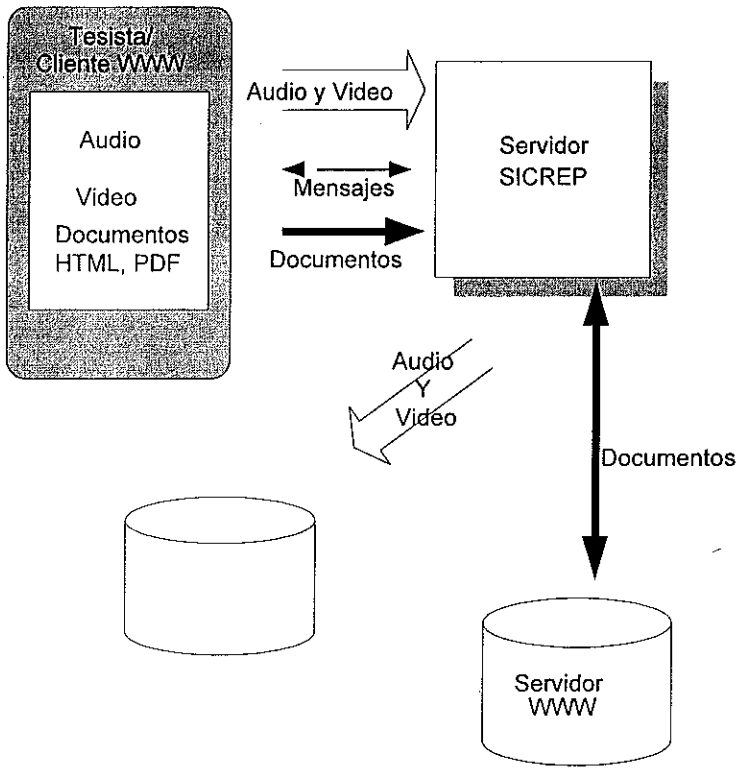


Figura 6. Arquitectura del sistema MIND para la captura de información

La arquitectura de MIND para la captura y la recuperación se muestra en las figuras 6 y 7 respetivamente. Su funcionamiento se describe a detalle en los siguientes párrafos.

V.3 Componentes del sistema MIND

MIND cuenta con cuatro procesos principales: captura, indexado, recuperación y visualización de medios. Cada uno de los componentes así como la relación entre ellos, se describe a continuación.

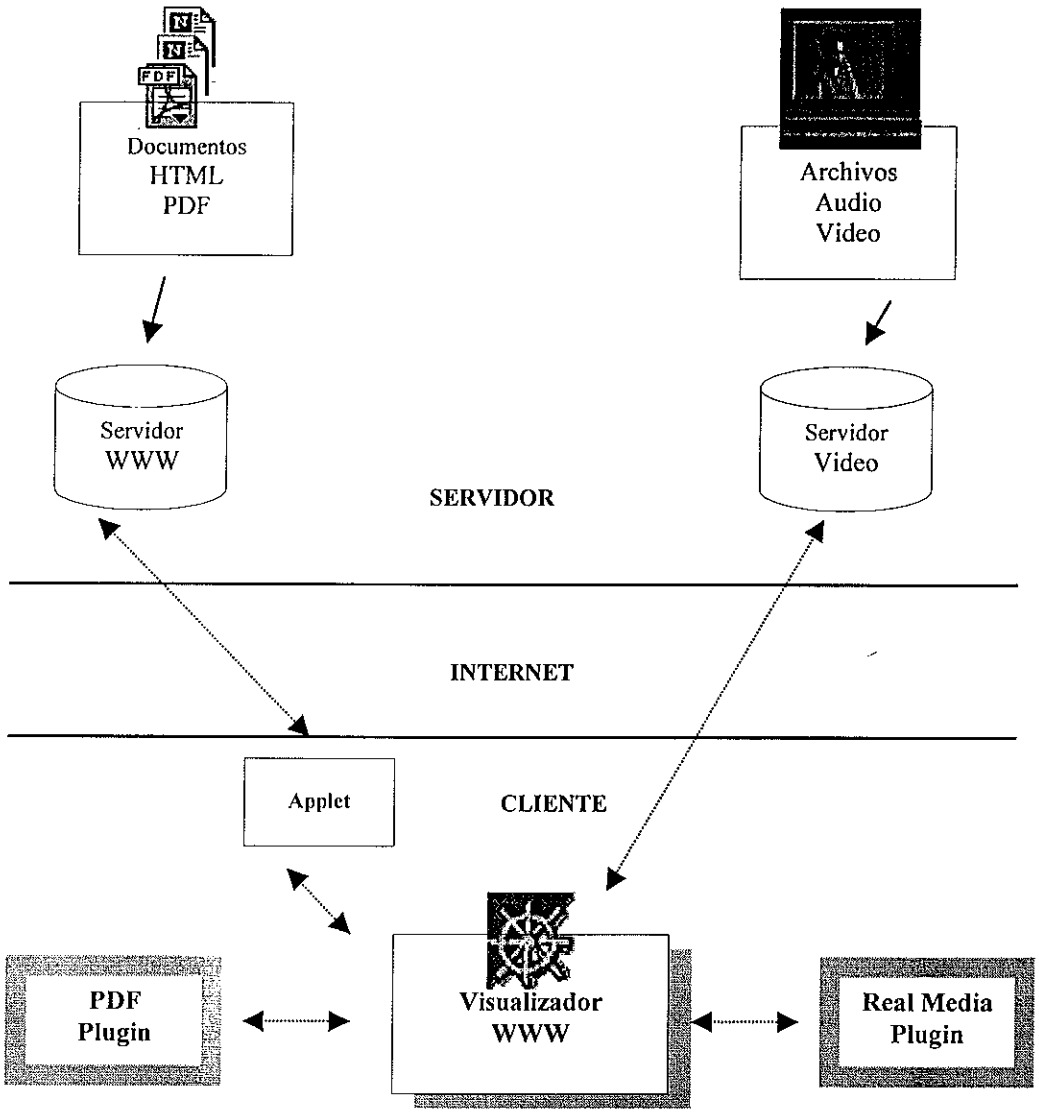


Figura 7. Arquitectura del sistema MIND para la Recuperación de información

V.3.1 Captura

El procedimiento de captura de datos multimedios es por lo general laborioso y puede abrumar al usuario [Stern., *el. al., 1997,* Abowd, *et, al.*, 1998], por lo tanto se

requiere en la mayoría de los casos de personal capacitado para el manejo de audio y video, ya que por lo general se requiere realizar trabajos de post-producción para llevar el material a la calidad adecuada, realizar cambios de formatos etc. Una de las principales ventajas del sistema MIND sobre los sistemas que capturan multimedia es que automatiza el procedimiento de captura de audio, texto e imágenes, sin necesidad de personal calificado así como una labor de post-producción del audio. El video se captura automáticamente y el trabajo que se debe realizar para segmentarlo es mínimo y la conversión al formato deseado es semiautomática.

En MIND se capturan dos partes que describen el material de la presentación: la información general sobre ésta (título de tesis, nombre del autor, etc) y la información sobre su contenido (audio, video, acetatos, el documento de tesis y su ficha bibliográfica).

V.3.1.1 Descripción de la presentación

Para cada una de las presentaciones que se capturan se debe proporcionar información general que ayude a identificarlas fácilmente, esto se hace desde el navegador una vez que se ha tenido acceso a la interfaz del presentador (Figura 8).

La captura de los datos de la información sobre la presentación consta de tres pasos:

1. El presentador selecciona la opción configuración y se despliega una interfaz para capturar información sobre la presentación, está información incluye el nombre de la

tesis, nombre del autor, etc. Está información será de utilidad para el usuario al momento de realizar las consultas y recuperación de las presentaciones.

2. Iniciar la grabación presionando la opción "Record", al ejecutar está acción el sistema empieza a capturar el audio del presentador y lo relaciona automáticamente con el acetato que se está visualizando, creando un archivo de audio por cada acetato visualizado.
3. Detener la grabación presionando seleccionando la opción "Stop" una vez que ha terminado la presentación.

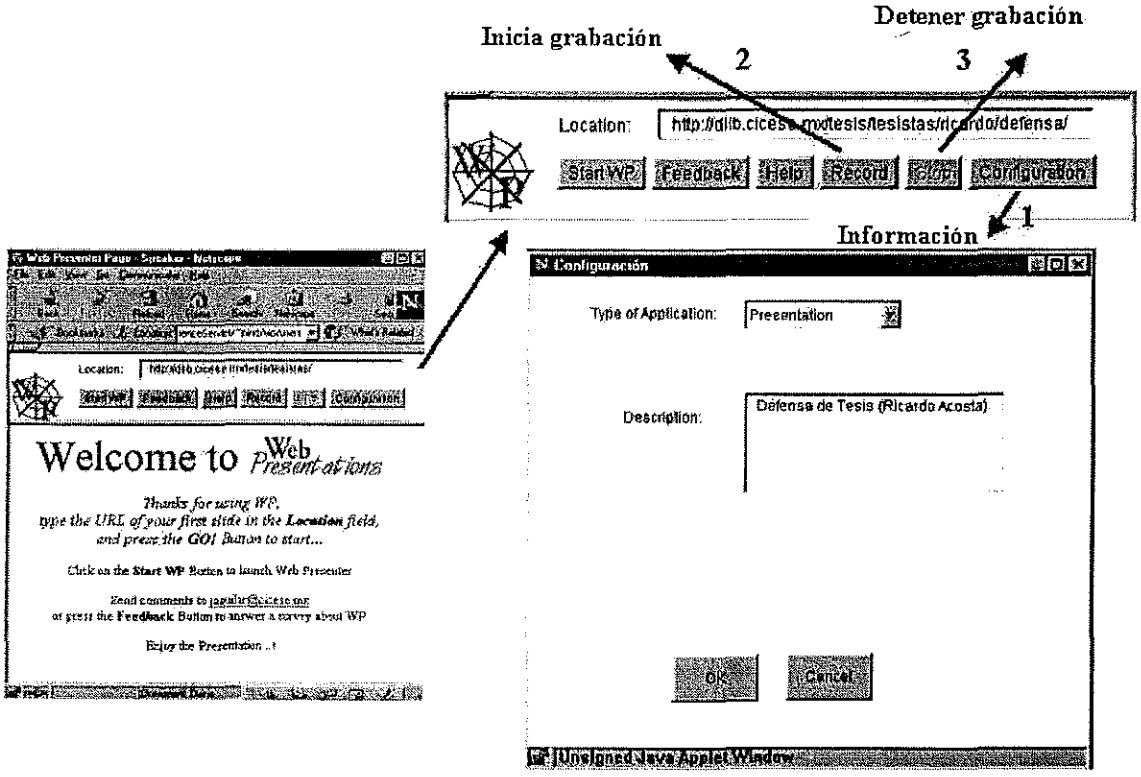


Figura 8. Interfaz para capturar presentaciones

Los datos que describen a la presentación son enviados al servidor de captura. Este se encargará de comunicarse con el servidor de Web, para asignarle un espacio en el mismo, donde se almacenará toda la información referente a la presentación (figura 9).

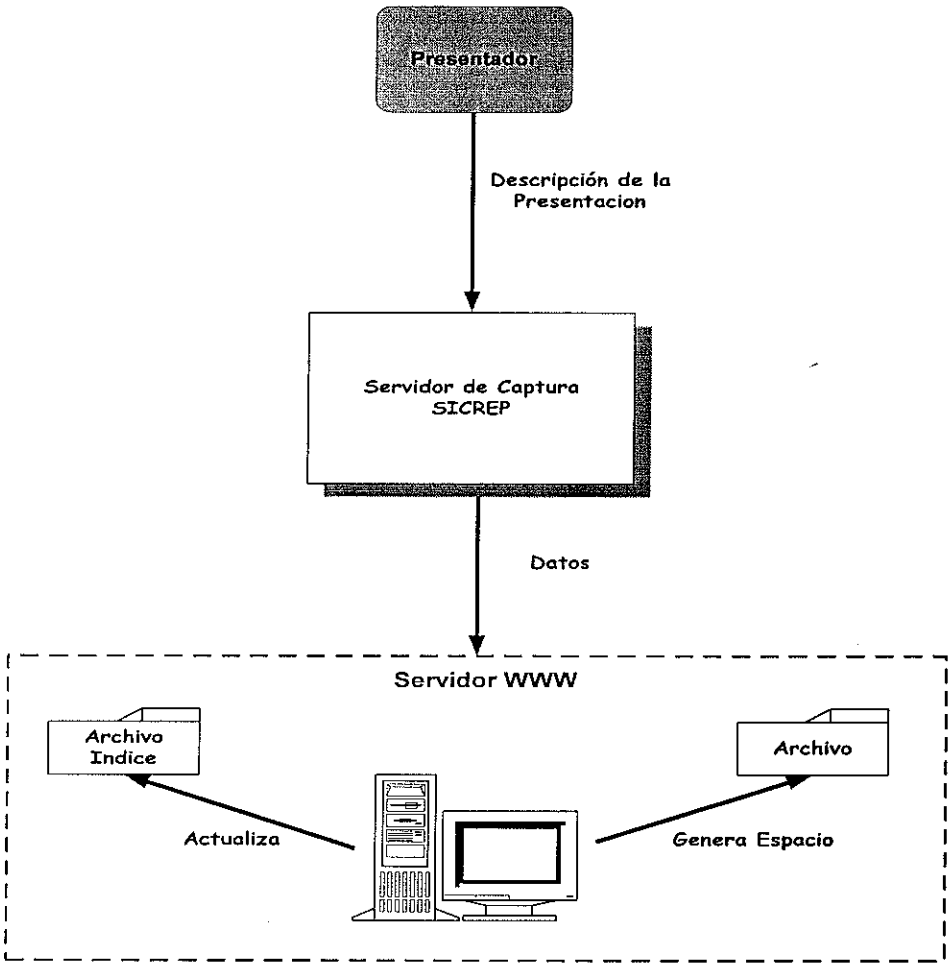


Figura 9. Procedimiento de captura de la información sobre la presentación

En el servidor de Web se crea un archivo, el cual contiene la información proporcionada por el usuario, con lo cual a partir de este momento el sistema estará listo para recibir los mensajes de cambio de acetatos enviados por el presentador, agregando de está manera información al archivo creado.

V.3.2 Captura de audio, video y acetatos

En el Sistema MIND a cada acetato le corresponde un archivo de audio y video de acuerdo a la explicación del ponente cuando se estába proyectando el acetato al momento de la grabación, por lo tanto se necesita comunicar al servidor cuando el presentador ha cambiado de acetato. Al recibir el servidor este mensaje inmediatamente se comunica con la aplicación del cliente, que se encarga de realizar las grabaciones del audio y almacenarlas en un directorio temporal. Como no es posible grabar archivos de audio y de video por separado en una sola máquina al mismo tiempo, el video es capturado en otra máquina y en forma continua, por lo que al terminar la grabación será necesario segmentar en forma manual el video, de acuerdo a la duración de los segmentos de audio. Al final de la sesión se envían al servidor que actúa como repositorio de la biblioteca digital todos los archivos de audio y video correspondientes a la presentación.

Además del audio y el video, MIND obtiene las direcciones de los acetatos utilizados durante la presentación cada vez que el presentador realiza un cambio de acetato. Al final de la presentación el servidor de presentaciones hace una revisión de los acetatos utilizados y los indexa de acuerdo a las palabras relevantes que contienen (figura 10).

Por último, la dirección de los documentos HTML de la presentación, los archivos de audio y video del presentador y la información general sobre la presentación, son almacenados en un directorio en el servidor de Web o repositorio de la biblioteca digital.

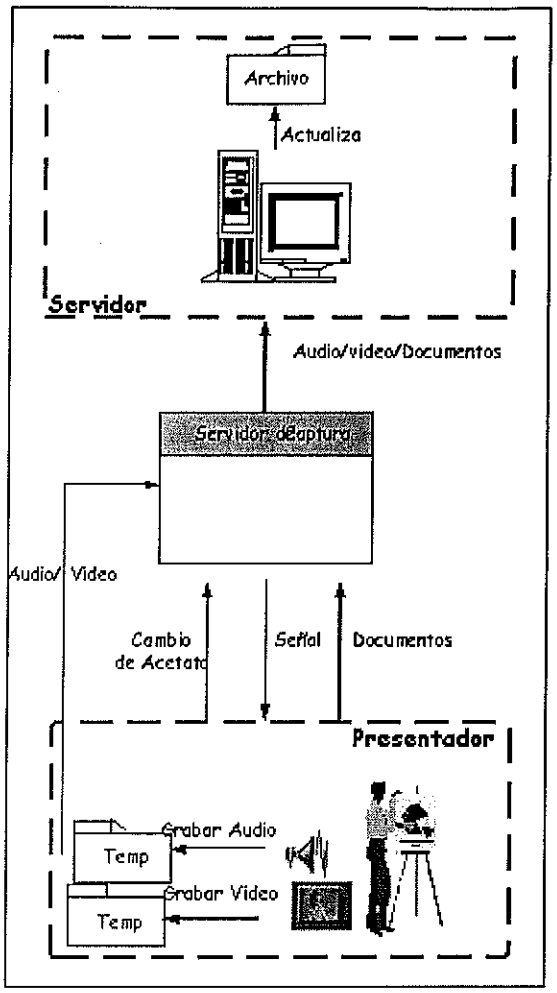


Figura 10. Esquema de la captura del audio, video y texto.

V.3.3 Ficha bibliográfica de la tesis

El sistema MIND proporciona también una interfaz (Figura 11), mediante la cual se captura la ficha bibliográfica de cada una de las tesis que se almacenarán en la biblioteca digital, para ello el administrador del sistema debe dar de alta la tesis, proporcionando los datos correspondientes a la misma tales como: nombre de la tesis, fecha, autor, director de tesis, institución, las palabras claves que identifican el contenido de esa tesis, las

direcciones del repositorio donde se encuentran almacenados; el documento de tesis, los acetatos y el video. Estos datos serán de utilidad para llevar un control de las tesis existentes y para permitir acceso directo a una tesis en particular o alguno de los medios que la integran.

Figura 11. Interfaz de captura de ficha bibliográfica de la tesis

V.4 Indexado

Una componente determinante de la calidad de los servicios de una biblioteca digital, está constituida por los criterios utilizados para organizar la información almacenada en formato digital, por lo tanto una vez que se tiene toda la información, está se organiza de manera que el usuario pueda recuperarla fácilmente. A continuación se describe el proceso seguido para indexar cada uno de los medios y poder recuperarlos posteriormente en forma eficiente o para poder alinearlos entre si.

V.4.1 Indexado del audio y video

El indexado del audio se realiza al momento de estár grabando la presentación, este es un indexado por eventos, ya que como se mencionó anteriormente por cada acetato existirá un archivo de audio que le corresponde o en su caso mas de un archivo de audio si es que el presentador visitó ese acetato en mas de una ocasión. La información de correspondencia entre los archivos de audio y acetatos se almacena en un archivo en el servidor que está siendo utilizado como repositorio de la biblioteca digital. Este archivo contiene información general sobre la presentación, el número total de acetatos, la dirección donde se encuentran almacenados cada uno de ellos y los archivos de audio que les corresponden.

Para relacionar el video con el acetato no es necesario realizar un indexado adicional, es decir se puede aprovechar el indexado del audio para relacionar los segmentos de video con el acetato que les corresponde ya que lo único que varía es el

nombre del video o mas específicamente la extensión del archivo que contiene el segmento de video almacenado.

V. 4.2 Indexado de los acetatos

Después de la captura, cada acetato es procesado para obtener de él las palabras relevantes que contiene. El procedimiento para realizar lo anterior se basa en el indexado automático de texto descrito en el capítulo III.

Los pasos realizados por el sistema MIND son los siguientes:

- Análisis del acetato (archivo HTML).- Durante este proceso, se extrae el texto contenido en el archivo, para luego poder aplicar el indexado automático.
- Aplicación del proceso de indexado automático de texto.- Al realizar este proceso se obtienen las palabras relevantes de cada uno de los acetatos y con ellas se crea un índice.

El proceso de indexado automático se aplica para cada uno de los acetatos (documento HTML), y las palabras extraídas se colocan en un índice invertido por palabras, de tal forma que una palabra puede tener uno o mas acetatos asociados a ella.

V.4.3 Indexado del documento de tesis

Para realizar el indexado de los archivos del documento de tesis, los cuales se encuentran en formato HTML se tomó como base el Script ICE que está basado también en el proceso de indexado automático. Este script fue desarrollado en 1993 en El Fraunhofer Institute en Dinamarca por Cristiane Neuss [Weil, 1997]. Se tomó ICE como base para realizar el indexado porque ofrece ventajas como: regresar ligas a las páginas que satisfacen los criterios de búsquedas, realiza búsquedas en múltiples directorios, permite realizar búsquedas booleanas; como está codificado en el lenguaje Perl puede operar en diferentes plataformas, bajo los sistemas operativos UNIX, Windows o Mac. Es un código libre, que puede ser utilizado y modificado.

V.5 Recuperación de información

Una vez que se han generado los índices del audio, el video, los acetatos y el documento de tesis se puede recuperar la información fácilmente. MIND ofrece dos formas principales para tener acceso a la información:

La primera es realizando búsquedas en las fichas bibliográficas de las tesis, de está forma el usuario recupera las fichas bibliográficas de las tesis que le interesen y desde la ficha bibliográfica, tiene un acceso directo al medio (Audio, Video, acetatos, documento) que el desee (Figura 12).

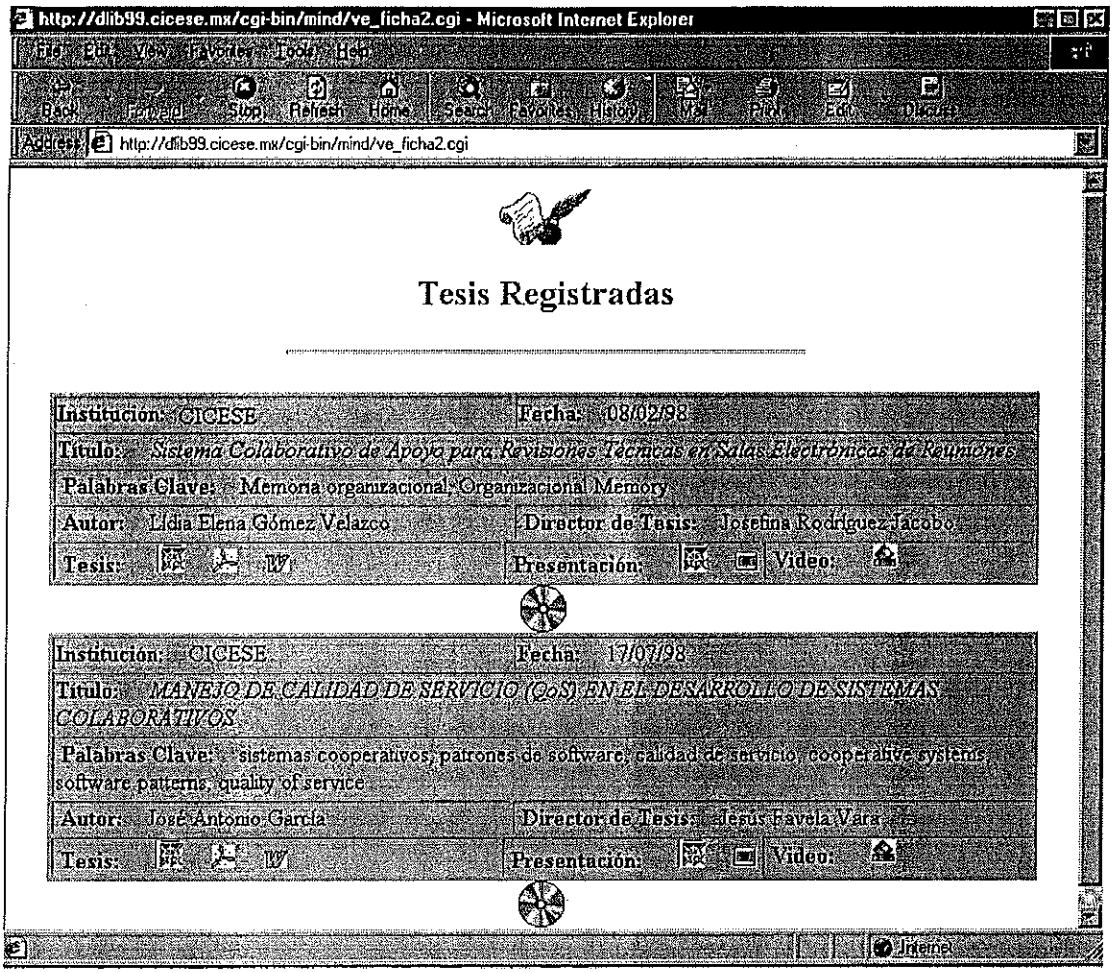


Figura 12. Interfaz de recuperación de información utilizando fichas bibliográficas

La segunda manera de tener acceso a la biblioteca digital es utilizando una interfaz más sofisticada (Figura 13) la cual permite hacer uso de la alineación de medios para recuperar la información.

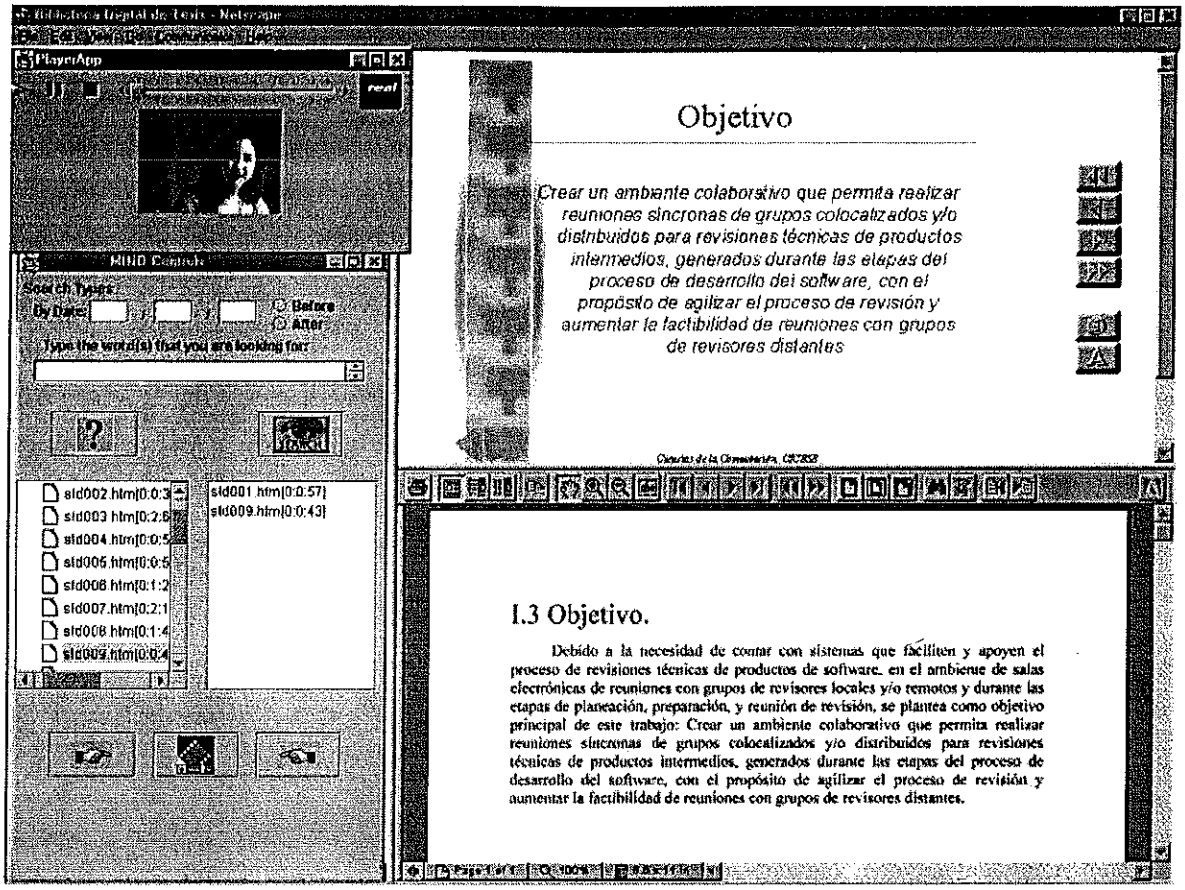


Figura 13. Interfaz del sistema MIND que permite la consulta y alineación automática de medios

V.6 Procedimiento a seguir en la recuperación de la información multimedia

MIND realiza varias tareas que son transparentes al usuario para realizar la recuperación de la información. A continuación se describen dichas tareas:

- El usuario puede consultar la biblioteca digital utilizando un navegador de Web, la figura 14 muestra la pantalla inicial del sistema de recuperación de información.

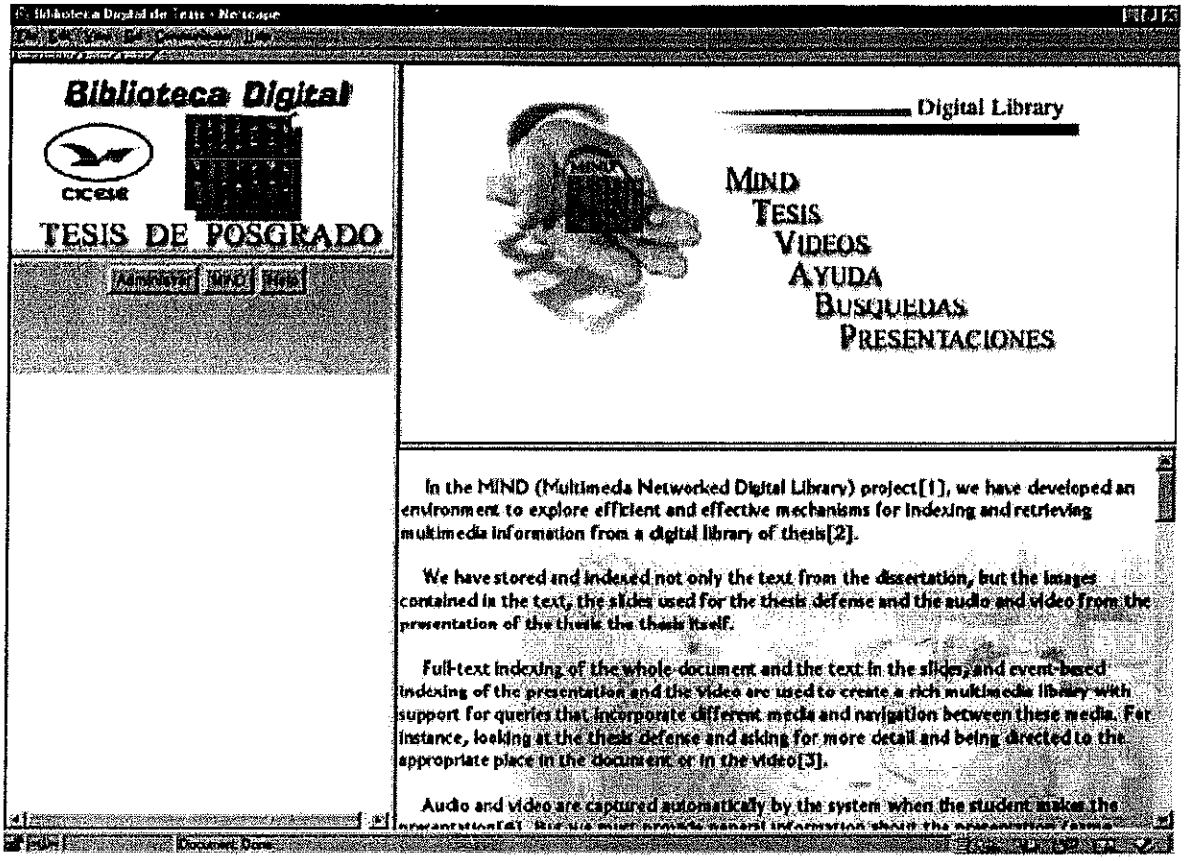


Figura 14. Interfaz inicial del sistema de recuperación de información

- El usuario formula su búsqueda mediante un applet[1] cuya interfaz es mostrada en la figura 15.
- El applet envía la consulta al servidor de consultas.
- El servidor de consultas procesa dicha consulta.
- El servidor de consultas se conecta al servidor de Web que sirve como repositorio de la biblioteca digital para obtener el resultado de la búsqueda.

[1] *Applet es un programa en Java que es incluido en una página HTML y su código es interpretado en un visualizador de Web.*

- El servidor de consultas regresa un listado de las presentaciones que pueden ser de interés para el usuario.
- El usuario revisa la información que se le ha mostrado, selecciona y reproduce los acetatos que son de su interés.

Al ser ejecutada está acción por el usuario se inicia el applet reproductor el cual recupera y despliega el acetato correspondiente, así mismo carga y reproduce el audio o video asociado con él y en el panel inferior derecho (Figura 13) haciendo uso de la alineación de medios, despliega la parte del documento de tesis que da información mas detallada sobre el acetato que se está visualizando.

A continuación se describen los tipos de consultas que se pueden realizar utilizando está interfaz.

V.7 Tipo de consultas que ofrece la interfaz

La interfaz para realizar las consultas (Figura 15) es sencilla y fácil de manejar, está interfaz permite realizar consultas en las presentaciones por fechas, por palabras claves o con una combinación de ambas.

Las consultas basadas en fecha permiten a los usuarios recuperar presentaciones que se realizaron antes o después de una fecha señalada.

Cuando se realiza una búsqueda basada en palabras clave, el usuario debe introducir palabras que indiquen el tema que desea consultar, para que el sistema realice la búsqueda entre las presentaciones almacenadas.

Así mismo, es posible hacer una combinación de los dos tipos de búsquedas, es decir proporcionando una fecha y palabras claves, esto aumenta la efectividad y acota la información en la cual se debe realizar la búsqueda, aunque el tiempo de resolución de la consulta resulta mayor

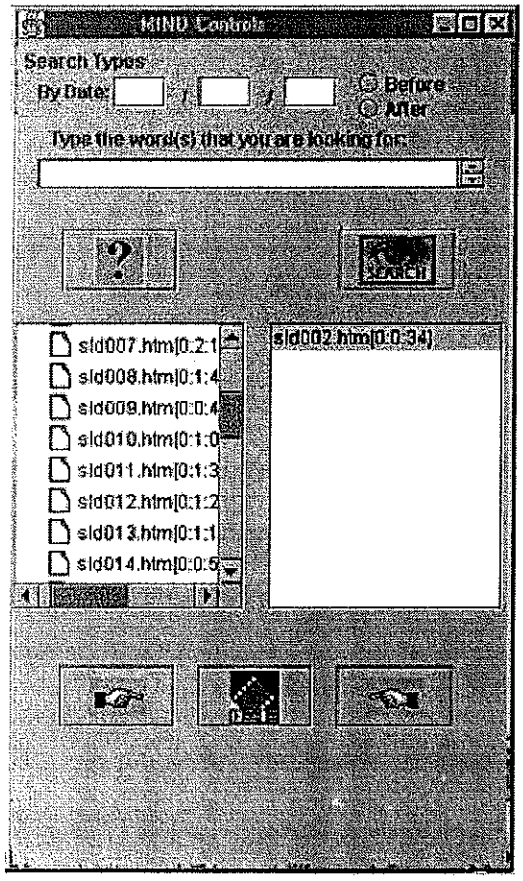


Figura 15 Interfaz utilizada para realizar los diferentes tipos de búsquedas

Capítulo VI. Evaluación de la recuperación de información en MIND.

VI.1 Introducción

Aplicar pruebas a un sistema es un proceso minucioso que se realiza con el propósito de encontrar errores y verificar que el sistema haga lo que se supone debe hacer [Myers, 1979]. Existe un número de condiciones, las cuales deben reunirse en la ejecución de un software para que sea considerada como prueba [Jacobo, 1996].

1. *Un medio ambiente controlado/observado*. Esto es esencial en la prueba para poder reproducirla en cualquier otro momento.
2. *Entradas simples*. En las pruebas sólo se utiliza un ejemplo de todas las posibles entradas.
3. *Resultados parecidos.*. Esto ofrece la ventaja de comparar los resultados esperados y los obtenidos.
4. *Resultados analizados*. Los resultados deben ser estudiados y procesados para interpretarlos y reportarlos.

En este capítulo se describen las métricas utilizadas para evaluar la recuperación de información en MIND, mediante la cual se alinean los acetatos de la defensa con los documentos de la tesis, así como las pruebas realizadas y los resultados obtenidos.

VI.2 Métricas utilizadas

En los Sistemas de Recuperación de Información se evalúa por lo general la precisión del conjunto de documentos que se obtienen como respuestá a una consulta.

Considérese un ejemplo de una consulta C y un conjunto R de documentos relevantes. Sea |R| el número de documentos relevantes en este conjunto. Supóngase que el algoritmo que se está probando procesa la consulta C y genera como respuestá un conjunto A. Sea |A| el número de documentos en ese conjunto. Llámese entonces |Ra| al número de documentos en la intersección de los conjuntos R y A. La Figura 16. ilustra estos conjuntos.

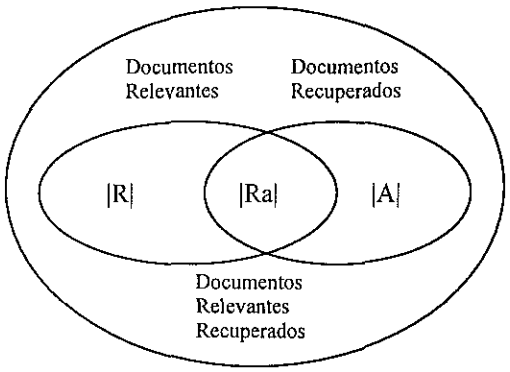


Figura 16. Precisión y Recuerdo

Las métricas Recuerdo y Precisión se definen de la siguiente manera [Baeza, 1999]:

- *Recuerdo* es la fracción de documentos relevantes (conjunto R) que han sido recuperados:

$$\text{Recuerdo} = \frac{|\,Ra\,|}{|\,R\,|} \qquad (1)$$

- Precisión es la fracción de documentos recuperados (conjunto A) los cuales son relevantes:

$$\text{Precisión} = \frac{|Ra|}{|A|} \qquad (2)$$

Los documentos del conjunto A se ordenan y son presentados al usuario iniciando por el de mayor relevancia.

Considérese ahora un ejemplo de recuperación donde se formula una consulta q. Supóngase que el conjunto Rq contiene 10 documentos relevantes para la consulta q.

Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

Considérese ahora que el algoritmo que se está probando regresa para la consulta q, un conjunto de documentos ordenados como se muestra a continuación:

1. d23●	6. d9●	11. d38
2. d84	7. d511	12. d48
3. d56●	8. d129	13. d250
4. d6	9. d187	14. d113
5. d8	10. d25●	15. d3●

Los documentos que son relevantes para la consulta q están marcados con una viñeta después del número de documento. Si se examina está lista, empezando por el documento mas relevante, se observan varios puntos:

1. El documento d23 que ocupa el primer lugar de los documentos ordenados, es relevante. Por lo tanto, a este documento le corresponde el 10% de los documentos

relevantes en el conjunto Rq. Por lo tanto se puede decir que se tiene una precisión del 100% con 10% de recuerdo.

2. El documento d56 que ocupa el lugar número tres en la lista es el siguiente documento relevante. Entonces, hasta este punto se tiene una precisión de aproximadamente 66% (dos de tres documentos son relevantes) con un 20% de recuerdo (dos de los diez documentos relevantes han sido vistos).
3. Si se continúa examinando los documentos ordenados es posible graficar la curva de precisión/recuerdo, como se muestra en la Figura 17. Donde se puede observar que la precisión a niveles de recuerdo mayores a 50% tiende a 0 porque no se han recuperado todos los documentos relevantes.

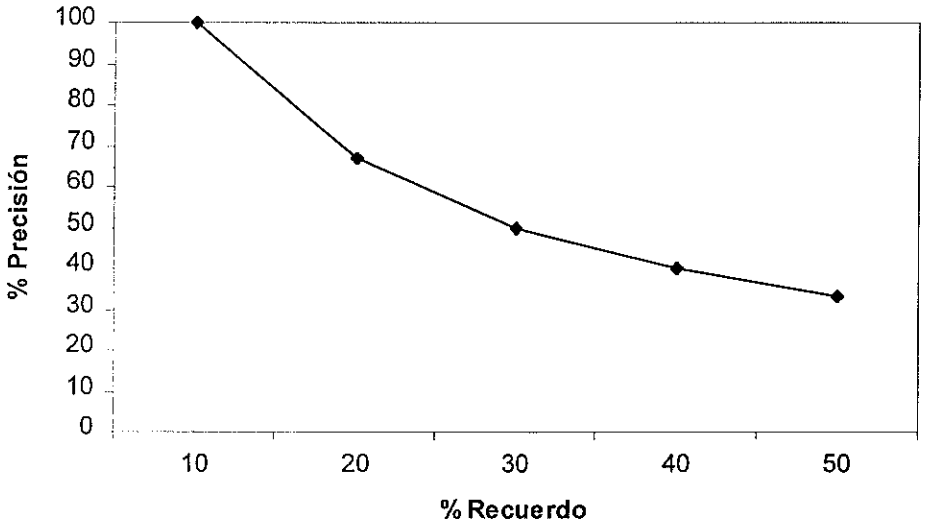


Figura 17. Precisión vs. Recuerdo

VI.3 Pruebas realizadas

Para el caso del sistema MIND se aplicó solamente *pruebas de precisión*, ya que se desconoce el número exacto de documentos relevantes para cada acetato y no se cuenta con los expertos (autores de la tesis) para definir dicho conjunto. En el caso bajo estudio se consideró de mayor importancia las pruebas de precisión porque el objetivo principal es encontrar el documento más relevante, y no obtener el total de documentos relevantes. Las pruebas realizadas tuvieron como objetivo determinar la precisión del algoritmo para recuperar documentos de una tesis recibiendo como entrada el texto de los acetatos de la presentación.

Para realizar estás pruebas se apoyó en un grupo de 5 estudiantes del área de computación de CICESE, a los cuales se les indicó una lista de tareas a realizar, y con los resultados obtenidos se les pidió contestár un cuestionario (Anexo A). Posteriormente esos datos se procesaron, se sacó una media aritmética de los resultados y con ello se obtuvo la evaluación del algoritmo.

Las tareas realizadas fueron las siguientes:

1. Visitar 10 acetatos de una presentación (los cuales fueron preseleccionados al azar).
2. Leer el acetato y entender el contenido del mismo.
3. Realizar una consulta utilizando el algoritmo de recuperación.
4. Revisar cada uno de los documentos que regresa el algoritmo.

5. Con los resultados obtenidos y de acuerdo a su criterio contestár el cuestionario del apéndice A el cual contiene 4 preguntas para cada acetato:

    a) Número de documentos que regresó el algoritmo.
    b) Lugar que ocupó el documento más relevante dentro del conjunto ordenado.
    c) Lugar que ocupó el segundo documento más relevante.
    d) Número de documentos relevantes recuperados.

Dado que la recuperación de información es dinámica, es decir para cada acetato que es visitado debe ejecutarse el algoritmo de recuperación para encontrar el documento que lo explique a detalle, se realizaron dos tipos de pruebas para evaluar la precisión como se muestra en la figura 18.

- *Prueba 1.* Se dio como entrada en la interfaz de consulta el título del acetato y se utilizó el operador booleano AND, está prueba fue aplicada en base a la suposición de que el título del acetato viene a ser como un resumen del contenido del mismo, además de que el título contiene pocas palabras, lo cual aumenta la velocidad de respuestá del algoritmo.
- *Prueba 2.* Se utilizó el texto completo del acetato y el operador OR. La decisión de aplicar está prueba fue en base a los resultados obtenidos al aplicar la prueba anterior, ya que había ocasiones que en el acetato se hacia referencia a temas que no estában relacionados con el título y no eran considerados.

	AND	OR
Título del acetato	✓	X
Texto completo del acetato	X	✓

Figura 18. Parámetros de las pruebas realizadas

VI.4. Discusión de Resultados Obtenidos

VI.4.1. Prueba No. 1

En está sección se presentan los resultados obtenidos, dando como parámetro de búsqueda el título del acetato y utilizando el operador AND.

Como resultado se tiene que el 70% de las veces el documento más relevante aparece en el primer lugar dentro del conjunto ordenado de documentos que regresa el algoritmos de recuperación, el 26% de las veces en segunda posición y el 4% en tercera posición como puede verse en la Figura 19. También se puede observar que con un máximo de tres documentos que regrese el algoritmo se puede obtener el documento más relevante, el cual es el objetivo principal de este algoritmo.

La figura 20 muestra la posición que ocupa el segundo documento más relevante dentro del conjunto recuperado. El cual si existe, aparece la mayor parte de las veces en la segunda posición o en la tercera, también puede observarse en la figura que el 68% de las veces regresa un segundo documento relevante (suma de las cinco primeras posiciones).

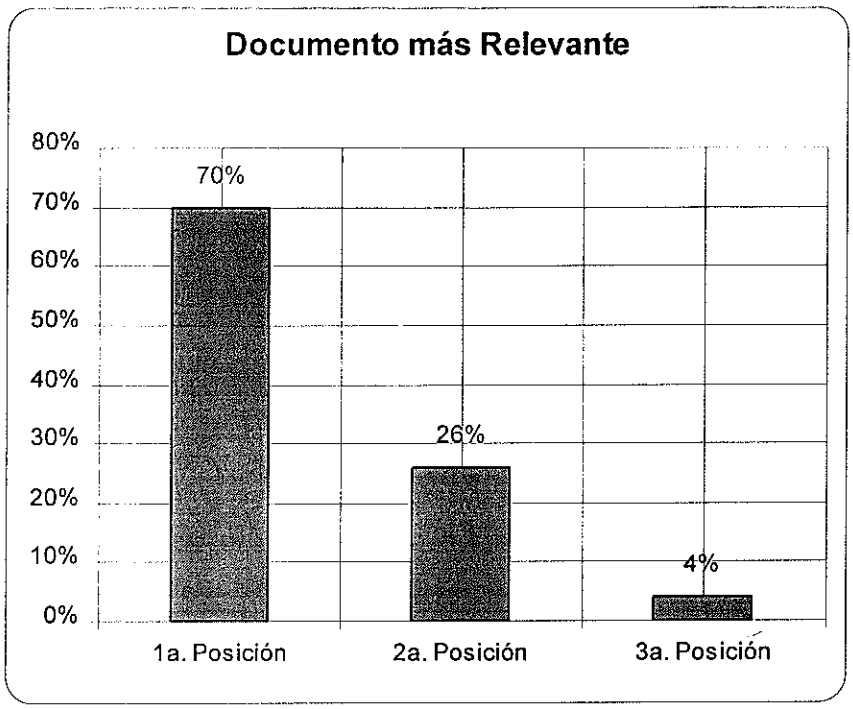


Figura 19. Lugar que ocupó el documento más relevante

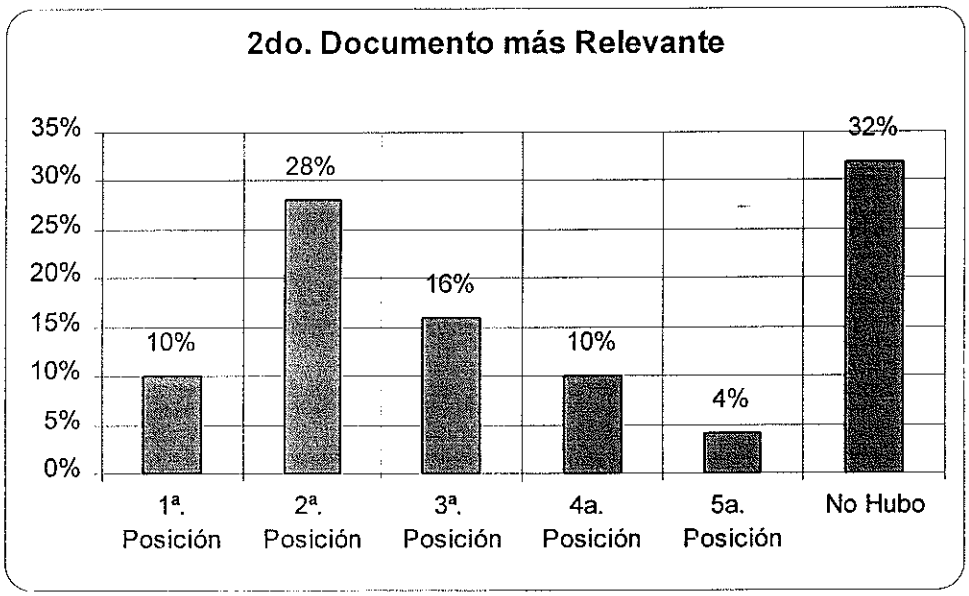


Figura 20. Lugar que ocupó la Segunda liga más relevante.

Como resultado al ejecutar las 10 consultas se recuperaron 27 documentos de los cuales un promedio de 19.4 fueron relevantes por lo tanto utilizando estos parámetros para el algoritmo se logró una precisión del 72% como lo ilustra la figura 21.

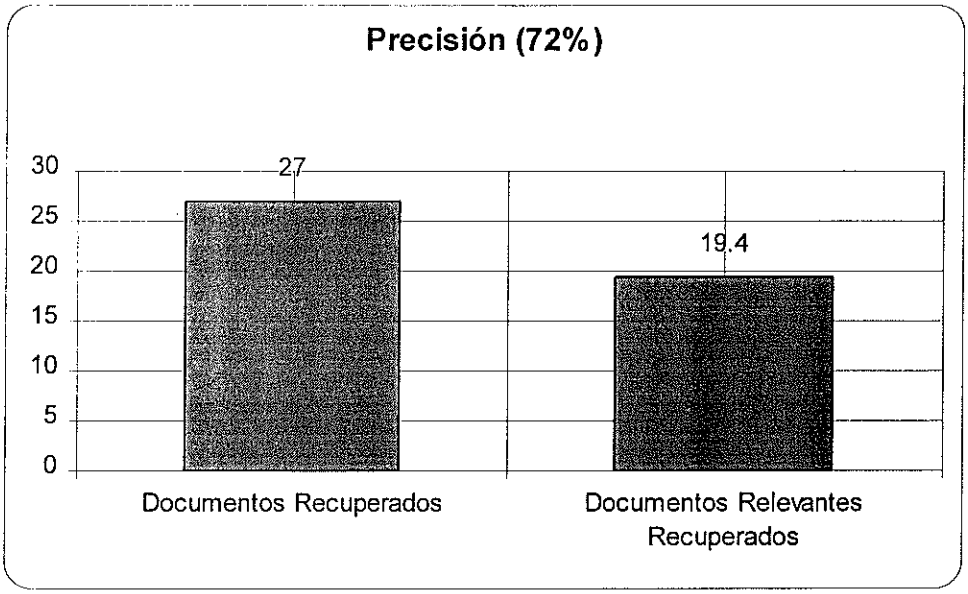


Figura 21. Precisión obtenida con el operador AND

VI.4.2. Prueba No. 2

A continuación se presentan los resultados obtenidos, al dar como parámetros de entrada el contenido total del acetato al algoritmo de recuperación utilizando el operador OR.

La figura 22 muestra la posición que ocupó el documento más relevante dentro del conjunto de documentos recuperados por el algoritmo, como puede observarse el 84% de las veces el documento ocupó la primera posición y en porcentajes menos importantes la segunda, tercera o cuarta posición.

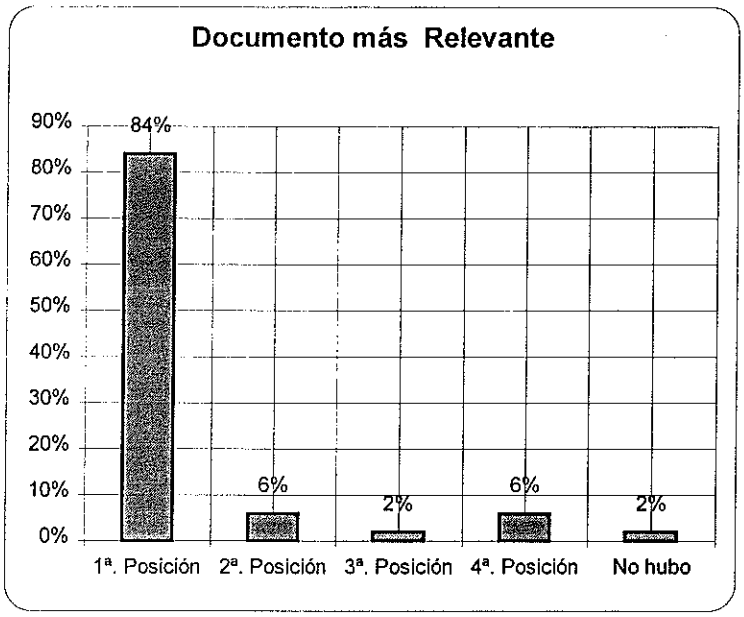


Figura 22. Liga más relevante usando OR

En La figura 23 se observa que en los casos que exista un segundo documento que es relevante este aparece el 56% de las veces en la segunda posición dentro del conjunto de documentos obtenidos y un 14% en la tercera posición.

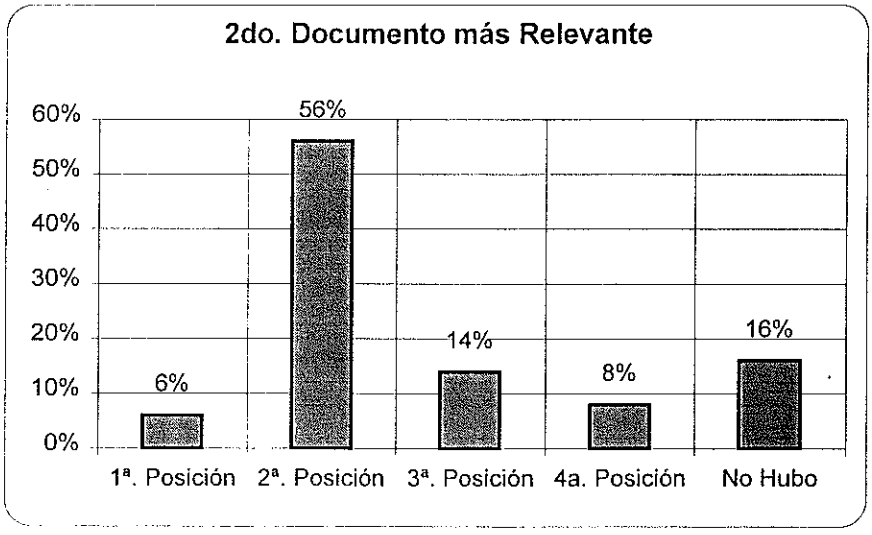


Figura 23. Segunda Liga más relevante usando OR

La figura 24 indica que ejecutando el algoritmo con estos parámetros regresa un total de 40 documentos de estos, 31 de ellos fueron relevantes, por lo tanto la precisión es del 78%.

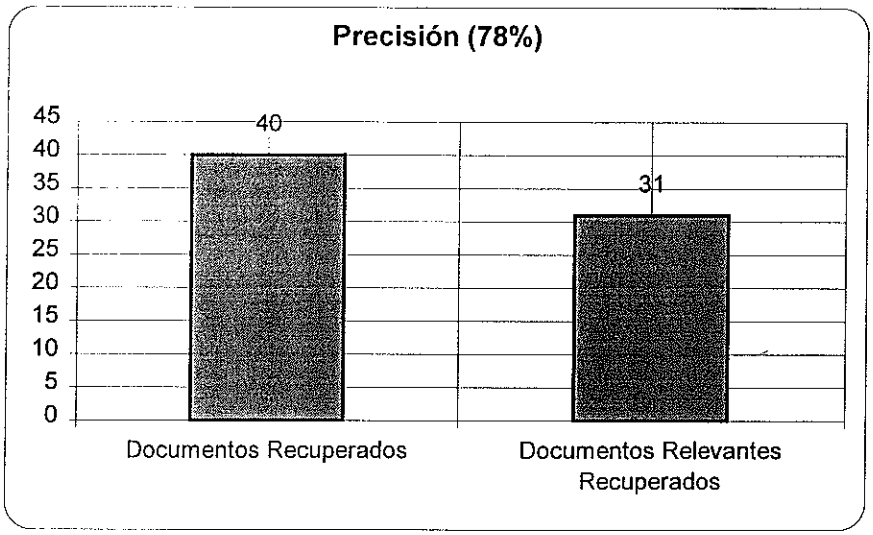


Figura 24. Precisión usando el operador OR

VI.4.3. Recomendaciones

Como el objetivo principal del algoritmo es recuperar el documento más relevante o que contenga información que describa mas a detalle un acetato, se recomienda utilizar la segunda opción, es decir dar como parámetros de entrada para el algoritmo el texto completo del acetato utilizando el operador OR, ya que ofrece un 6% más de precisión que el primero, además de que es 14% mas preciso para presentar el documento más relevante en la primera posición y 28% para presentar el segundo más relevante en la segunda posición. Así mismo se recomienda visualizar automáticamente el documento que aparece en la primera posición como descripción del acetato que está siendo visualizado y

opcionalmente desplegar un documento HTML que contenga ligas a los documentos que ocupen las cuatro primeras posiciones, por si el usuario desea obtener mas información relacionada con el acetato.

Capítulo VII. Conclusiones

Se presentó una descripción del trabajo realizado el cual consistió en diseñar una arquitectura y un ambiente para crear y consultar una biblioteca digital multimedia de tesis de posgrado, la cual puede ser consultada a través de Internet, permitiendo a los usuarios manipular la información de acuerdo a sus necesidades.

Para finalizar la descripción de este trabajo de investigación, a continuación se comentan en forma breve los logros obtenidos, las aportaciones que se hicieron y las recomendaciones para trabajo futuro.

VI.1 Logros obtenidos

A continuación se describen en forma resumida los logros obtenidos al realizar este trabajo:

- Se hizo una revisión bibliográfica del estádo del arte de la recuperación de información multimedia y las bibliotecas digitales y se aplicaron estos conceptos en el desarrollo de una biblioteca digital multimedia de tesis de posgrado.

- Se diseñó una arquitectura y un ambiente que permite la captura e indexado de información multimedia para alimentar automáticamente las bases de datos de dicha biblioteca digital.

- Se mejoró e implementó un algoritmo de recuperación de información para realizar la alineación automática entre los acetatos y los documentos de tesis en la biblioteca digital

VI.2 Aportaciones

Al desarrollar este trabajo se realizaron dos aportaciones principales:

- Se desarrolló un ambiente que permite la alineación automática de información multimedia.

- Se creó una Biblioteca Digital Multimedia de Tesis de Posgrado que ofrece las siguientes ventajas:

    - Realiza la alineación automática entre los acetatos, el audio, video y el documento de tesis.

- Puede ser accesada a través de un navegador WWW

- La información puede ser manipulada de acuerdo a las necesidades de cada usuario.

- Permite realizar búsquedas por palabras claves, fechas o utilizando la ficha bibliográfica de la tesis.

VI.3 Trabajo Futuro

Algunos puntos que pueden considerarse para mejorar los servicios que ofrece este prototipo de biblioteca digital son descritos a continuación.

- Agregar un módulo para que realice la captura y segmentación automática de video, así como automatizar la conversión de texto al formato correspondiente.

- Ofrecer una interfaz de usuario que permita realizar consultas usando lenguaje natural escrito y hablado.

- Realizar el indexado y recuperación de información que no sea texto basado en contenido.

- Agregar perfiles de usuario y medios de comunicación que permitan a los visitantes de la biblioteca digital comunicarse entre sí [Llamas, 2000].

Literatura Citada

Abowd G., Brotherton J., Bhalodia J., 1998. Classroom 2000: A system for capturing and accessing multimedia classroom experiences, ACM CHI'98 Demonstration paper.

AMC. 1998 Fideicomiso SEP-UNAM Academia Mexicana de Ciencias, "Componentes Básicos y Funcionalidad de una Biblioteca Digital"

Baeza Ricardo, Neto, Ribeiro. Modern Information Retrieval. Ed. Addison Wesley, 1999.

BDF, 1999. http://ict.udlap.mx/projects/fdl

BDIC, 1999. http://infolac.ucol.mx/proyectos/bdigital.html

Berkeley Digital Library SunSITE. "Digital Library SunSITE Collection and Preservation Policy".(http://sunsite.berkeley.edu/Admin/collection.html)

Bob Weil and Chris Baron, "Drag and Drop CGI", Addison-Wesley Pub Co, 1997

Brown, M. Foote, J., Sparck-Jones K. And Young, S., "Automatic Content-Based Retrieval of Broadcast News". AAAI, Working Notes, Stanford University, Sanford, California,1997.

Proc ACM Multimedia 95, 35-43, ACM Press. Nov. 1995.

Bush, V. 1945. "As we may think". Atlantic Monthly 176, 1 (Julio), 101-108.

David Garza, Martha Sordia, Yolanda Martínez, Phronesis: Una herramienta práctica y eficiente para la creación de bibliotecas digitales distribuidas en Internet. IX Coloquio de Automatización de Bibliotecas, Universidad de Colima, 1999.

[Dienst, 1999] http://www.cs.cornell.edu/NCSTRL/protocol.html,

D-Lib. 1998. D-Lib Magazine. http://www.dlib.org/.

D-Lib. 1999. Stephen M. Griffin, Digital Libraries Initiative - Phase 2. http://www.dlib.org/dlib/july99/07griffin.html

DLI-P2, 1998. http://www.nsf.gov/cgi-bin/getpub?nsf9863

ELVIRA. 1997. Proceedings of Electronic Library & Visual Information Research 4 (ELVIRA4, Milton Keynes, UK, May) http://www.iielr.dmu.ac.uk/ELVIRA/ELVIRA4/proceedings.html).

Falatous, C *et all.*, "Efficient and Effective Querying and Image Content". Journal of Intelligent Information Systems, Vol. 1, No. 3, 231-262 1994.

Fluckiger, F. 1995. "Understanding networked multimedia". Prentice Hall.

Garcilazo, J., "Sistema para la captura y recuperación de cursos electrónicos"., Tesis de maestría. CICESE. Octubre 1998.

Glendford J. Myers. "The Art of Sofware Testing". Ed. John Wiley & Sons, 1979.

Grosky W. y R. Mehratra 1989. Guest editor, Special issue on image database management. Computer 22(12).

Guojun L., 1996. Communication and computing for distributed multimedia systems. Artech House, Inc. Primera Edicion.

Hauptman, A., Witbtrock, M. Rudnicky, A y Reed, S, "Speech for multimedia information retrieval". Proc. of User Interface ,Sofware Technology (UIST-95), Pittsburgh, PA, ACM.

IEEE. 1998. 5th Advances in Digital Libraries Conference (ADL '98, Santa Barbara, Calif., Abril). IEEE Computer Society, Los Alamitos, Calif.

IJDL. 1998. International Journal of Digital Libraries. Springer-Verlag, Heidelberg. http://link.springer.de/link/service/journals/00799/index.htm).

INEGI, Plan Nacional de Informática 1995-2000 México.

Jacobo Josefina, Jesús Favela, Rubén Martínez, Ingeniería de la programación: Verificación y Validación del Proyecto EFICAZ, Comunicaciones Académicas, CICESE, 1996.

Jacobs, C., Finkelstein, D. And Salestin, D., "Fast multiresolution image querying". Proc. ACM SIGGRAPH´95 pp. 277-286, Agosto 1995.

Jain R., 1993. NSF Workshop on visual information management systems. SIGmod Record. Marzo 1995. 57-75 p.

JoDI. 1998. Journal of Digital Information. The British Computer Society & Oxford University Press. (http://jodi.ecs.soton.ac.uk/).

Laboratory, Texas A&M University, College Station, Tex., 1994 (http://www.csdl.tamu.edu/DL94).

Lafuente Ramiro "Biblioteca Digital y Orden Documental" CUIB/UNAM, 1999.

Llamas Contreras Rafael, 2000, "Interacción Casual en Bibliotecas Digitales", Tesis de Maestría, CICESE.

Lesk, M. 1997. Practical Digital Libraries: Books, bytes and bucks. Morgan Kaufmann Publishers, San Francisco, Calif.

Licklider, J. C. R. 1965. Libraries of the Future. MIT Press, Cambridge, Mass.

Machine, edited by James M. Nyce and Paul Kahn (1991). Academic Press. 165-184.

Managing Gigabytes: "Compressing and Indexing Documents and Images", 1999 http://www.cs.mu.OZ.AU/mg/

Marchioni, G y Maurer, H, 1995 "The roles of Digital Libraries in Teaching and Learning". CACM, Vol. 38, No. 4, pp 67-75.

Narasimhalu A. 1995. Guest editor, Special issue on content-based retrieval. ACM Multimedia Systems.

[NCSTRL] http://www.ncstrl.org/Dienst/UI/2.0/ListPublishers, 2000.

Nelson, T. 1977. Computer Lib/Dream Machines. Tempus Books (reprinted by Microsoft Press., Redmond, Wash.).

Niblack, W. *Et all.*, "The QBIC project: Query image by content using color, texture and shape". Storage and Retrieval for Image and Video Datebases, pp. 173-187, San Jose 1993. SPIE.

NSFII, National Science Foundation. "Digital Libraries Initiative- Phase II". 1999. (http://www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm).

Pentland, A., "Machine Understanding of Human Behavior in Video, In Intelligent Multimedia Information Retrieval" Maybury, M. (ed.), MIT Press, 175, 190, 1997.

[Phronesis], 1999. http://dgicii.mty.itesm.mx/~phron

Proc. of the Workshop on Intelligent Integration and Use of Text, Image, Video and Audio Corpora, AAA-97 Spring Symposium Series, Marzo 1997, Stanford, CA.

Proc. of the Workshop on Mixed Media Databases. Conf. on Automated Learning and Discovery, Pittsburgh, PA, Junio 1997.

Research Laboratory, Texas A&M University, College Station, Tex. 1995 (http://www.csdl.tamu.edu/DL95).

Schnase, J., Leggett, J., Furuta, R., and Metcalfe, T. (Eds.). 1994. Proceedings of Digital Libraries '94 (College Station, Tex., June).

Shipman, F., Furuta, R., and Levy, D. (Eds.). 1995. Proceedings of Digital Libraries '95 (Austin, Tex., June). Hypermedia

Stanford University - The Digital Library. "Digital Libraries?". 1999 (http://walrus.stanford.edu/diglib/pub/abstract.html)

Stanford University. "Stanford Digital Library Project: Information Bus Infrastructure and Models - II.B.3 Information Finding Services - II.B.3.2: A network finding service",1996 http://walrus.stanford.edu/diglib/pub/proposal/partII/node16.html).

Stanford University-The Digital Library. "Digital Libraries?" 1999 (http://walrus.stanford.edu/diglib/pub/abstract.html)

Stern M., J. Steinberg, H.I. Lee, J. Padhye y J. Kurose 1997. MANIC: Multimedia asynchronous networked individualized courseware. Proceeding of Educational Multimedia and Hypermedia.

Trace. 1998. Modelo de biblioteca con acervos digitales y bases para desarrollar una plataforma nacional de redes de alta velocidad. Informe de Trabajo. Parte II: Investigación sobre bibliotecas digitales y proyectos en marcha en México y en el extranjero. Trace, S. C. Consultores. México, (http://ict.pue.udlap.mx/dl/docs/INVESTIGACION.ps).

WATERS, 1999 http://www.cs.cornell.edu/NCSTRL/waters.html

WDL99 "First workshop on Digital Libraries". http://www.istec.org/liblink/jerome/dlform.html

Apendice A

Formato para aplicar pruebas al algoritmo de recuperación de información

Nombre:_____

Fecha:_____

No. de Prueba: _____

Interfaz de consultas: http://dlib.cicese.mx/tesis/consultas/

#Acetato	#L	LMR	2LMR	#LR
1				
4				
5				
7				
14				
21				
26				
27				
30				
32				

- #L No. de documentos que recuperó L >= 0 <= 5
- #LMR Lugar que ocupa la liga más relevante.
- #2LMR Lugar que ocupa la segunda liga más relevante
- #LR No. de ligas relevantes recuperadas

Apéndice B. Instalación y Ejecución del Sistema MIND.

B.1 Introducción

En este apéndice se presenta la lista de requerimientos de software y hardware necesarios para instalar y ejecutar el sistema MIND, así mismo se detallan los pasos a seguir para realizar la instalación y configuración del mismo.

B.2 Requerimientos de Hardware y Software

Los requerimientos mínimos de hardware y software para que el sistema funcione se muestran en la tabla III, aunque estos requerimientos aumentarán dependiendo del número de tesis que se desee almacenar así como del número de usuarios que visiten simultáneamente la biblioteca digital.

	Servidor	Cliente
HARDWARE	• PC Pentium o WorkStation • PC pentium con tarjeta de captura de video. • Cámara de Video • Memoria 64 MB • Disco Duro 4GB • Micrófono	• Macintosh, PC o Estación de trabajo • Memoria 32 MB • Tarjeta de audio • Disco Duro 1 GB
SOFTWARE	• JDK 1.1.X • JavaWebServer 1.X • JavaMediaFramework 1.X • JavaFundationClass 1.X • Swing 1.02 • Real Producer 6.0 • Real Server Basic • Ulead Media Capture Server • Adobe Premier y Acrobat	• Visualizador de Web (Netscape, Explorer) • JavaMediaFramework 1.X • JavaPlugIn 1.1.X • Adobe Acrobat Reader 3.0 • RealPlayer 6.0

Tabla III. Requerimientos de Hardware y Software para el sistema MIND

B.3 Estructura de directorios

El software necesario para instalar, configurar, compilar y ejecutar el sistema puede obtenerse en "http://dlib.cicese.mx/~mind/mind/instalar/".

Una vez que se instale el servidor de Web se deben copiar dentro de los directorios public_html, cgi-bin y servlets (generados automáticamente al instalar el servidor de Web) el contenido de los directorios con el mismo nombre obtenidos del URL mencionado anteriormente.

En la PC que utilizará el presentador (servidor de captura) deberá copiarse el software del servidor de captura obtenido de la distribución.

B.3.1 Archivos a modificar

Una vez que el software ha sido copiado en sus respectivos directorios se debe proceder a configurar la dirección de los archivos de índices, el archivo maestro y la dirección IP del servidor que se estará utilizando.

B.3.1.1 Modificar IP

La siguiente es una lista de archivos en los cuales debe especificar (modificar) la dirección IP o el URL del servidor que será utilizado como repositorio de la biblioteca digital.

1. \wp\SpeakerApplet.java
2. \wp\Recuperar.java
3. \wp\HrefParser.java
4. \wp\AudienceApplet.java
5. \wp\SimplePlayerApplet.java
6. \servlets\PreferenceServlet.java
7. \servlets\UserServlet.java
8. \wp\util\Junta.java

9. \wp\prominence\chat\Registry.java
10. \servlets\HrefParserSevlet
11. \servlets\UserServlet.java
12. \wp\speaker_frames.html
13. \wp\register.html
14. \wp\feedback.html
15. \wp\audiende_frames.html
16. \wp\audience.html
17. Cliente2.java (En el servidor de captura)

B.3.1.2 Modificar ruta de archivos

En los archivos que se listan a continuación debe especificarse la ruta de los directorios donde se almacenarán los archivos *maestro.prf* e *Invertido.prf.* El primero lleva el registro de las presentaciones existentes y el segundo es el archivo de índices de los acetatos de las presentaciones:

1. \wp\Indexar.java
2. \wp\BuscaDocumento.java
3. \wp\Handler2.java
4. \wp\HandlerBusqueda.java
5. \wp\Procesar.java
6. \wp\ProcesarPresentacion.java
7. \wp\servlet\PreferenceServlet.java

B.4 Ejecutar el sistema

Una vez que se ha realizado la configuración descrita en los puntos B.2 y B.3 y los archivos se han compilado, se procede a ejecutar el sistema, el cual esta compuesto de cuatro servidores, los cuales se ejecutan de la siguiente manera:

- Servidor de Búsqueda. java Busqueda
- Servidor de Captura. java CaptureServer
- Servidor de acetatos. java SlideServer 9840
- Servidor WP. java prominence.chat.WPSystemServer 9830

Apéndice C. Procedimiento para agregar tesis a la Biblioteca Digital

C.1 Introducción

El propósito de este apéndice es describir los procedimientos que deben seguirse para grabar una presentación, segmentar el video y el documento de tesis, registrar la ficha bibliográfica e indexar el documento de tesis.

C.1 Grabar presentación

Para grabar una presentación se realizan los siguientes pasos:

1. Convertir el archivo de presentación que se encuentra en formato de Power Point a Formato HTML, utilizando el convertidor automático que trae integrado el Power Point.
2. Almacenar la presentación en el servidor de la biblioteca digital siguiendo la estructura: http://dlib.cicese.mx/~mind/tesis/tesistas/nombre_del_tesistas/defensa/
3. Ejecutar el programa Cliente2.class en el servidor de captura.
4. Accesar mediante un visualizador de Web el sistema Wp entrando como "Presenter" en la dirección: http://pc-acosta.cicese.mx/wp/register.html.

5. Proporcionar la descripción de la presentación, así como la dirección donde se encuentra almacenada la presentación y realizar la grabación como se indica en el subcapítulo V.3.1.1 y figura 8.
6. Paralelo al paso número 5 con otra computadora y una cámara de video se debe iniciar la grabación de la presentación.
7. Una vez terminada la presentación los archivos de audio generados en c:\windows\temp en el servidor de captura se transfieren al servidor de la biblioteca digital al directorio generado por el servidor de captura, y el archivo de video una vez que es convertido al formato "Real Media" utilizando el sistema "Real Producer" se debe almacenar en la dirección http://dlib.cicese.mx/~mind/tesis/tesistas/nombre_del_tesistas/video/

C.3 Segmentar Video

El archivo video es segmentado en fragmentos con la misma duración que los archivos de audio generados en la presentación y se guardan con el mismo nombre pero (con la extensión .rm) , esto se realiza utilizando el software Adobe Premier. Una vez que se ha terminado de segmentar el video, los fragmentos se almacenan en el mismo directorio que los archivos de audio.

C.4 Segmentar documento de tesis

El documento completo de la tesis que originalmente se encuentra en formato de Word es convertido a formato PDF utilizando el Acrobat Distiler y el archivo resultante se almacena en la biblioteca digital siguiendo la estructura:

http://dlib.cicese.mx/~mind/tesis/tesistas/nombre_del_tesistas/tesis/.

El documento de tesis que se encuentra en formato de Word es segmentado por subcapítulos y cada segmento es guardado como un archivo por separado en formato HTM y PDF, los cuales son almacenados bajo la estructura:

http://dlib.cicese.mx/~mind/tesis/html//nombre_del_tesistas/.

C.5 Indexar Documentos de tesis

Cada vez que se agregue una tesis debe generarse el indexado de documentos para lo cual debe ejecutarse el script ./ice3-idx.pl que se encuentra en el directorio *cgi-bin* en el servidor de la biblioteca digital.

C.5 Registrar tesis

Para registrar una nueva tesis, se debe proporcionar los datos de la ficha bibliográfica de la misma, para ello se sigue el procedimiento descrito en el subcapítulo V.3.3 llenando la forma que se muestra en la figura 11, en el URL:

http://dlib.cicese.mx/~mind/tesis/ficha/