VNIVERSIDAD NACIONAL AVTNOMA DE MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

ESTUDIO BIBLIOMÉTRICO SOBRE INFORMACIÓN DE LA BASE DE DATOS DE TESIS DIGITALES PERTENECIENTE A LA DIRECCIÓN GENERAL DE BIBLIOTECAS

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRA EN INGENIERÍA (COMPUTACIÓN)

P R E S E N T A:

SILVIA SOCORRO BALLESTEROS ESTRADA

DIRECTORA DE TESIS: DRA. AMPARO LÓPEZ GAONA

México, D.F. **2008.**

Dedico especialmente

A mi niña,

Por aceptar involuntariamente mi ausencia, aun estando presente, y a pesar de ello esperarme cada día con lo brazos abiertos para darme un abrazo "fuerte". Por tomar del tiempo que te pertenecía y que nunca me exigiste. Por ser "feliz". Por tus sueños e ilusiones. Por tu fortaleza. Por ser una lección de vida.

Para ti chiquita, "Un sueño imposible "…

Con fe lo imposible soñar
al mal combatir sin temor
triunfar sobre el miedo invencible
en pie soportar el dolor

Amar la pureza sin par
buscar la verdad del error
vivir con los brazos abiertos
creer en un mundo mejor

Es mi ideal
la estrella alcanzar
no importa cuan lejos
se pueda encontrar
luchar por el bien
sin dudar ni temer
y dispuesto al infierno llegar si lo dicta el deber

Y yo sé
que si logro ser fiel
a mi sueño ideal
estará mi alma en paz al llegar
de mi vida el final

Será este mundo mejor
si hubo quien despreciando el dolor
combatió hasta el último aliento

Con fe lo imposible soñar
y la estrella alcanzar

Silvia Socorro Ballesteros Estrada

Agradezco

A quien fue parte fundamental para cumplir un objetivo más en mi vida, la maestría, y que sin su apoyo me hubiese sido muy difícil cumplirlo…

A mis padres,
Por el infinito apoyo que me han brindado. Por que en esta etapa de mi vida han compartido y asumido mis compromisos y obligaciones más que siempre.

A mi madre,
Por ayudarme con la otra profesión, ser madre. Por el tiempo, la dedicación y la paciencia que has entregado a mi hija en mi ausencia (y también en mí presencia). Por obtener las fuerzas y agallas que a mí me han faltado.

A Eve y a Juan,
Por estar junto a mí, por ser piezas fundamentales del rompecabezas de mi vida.

A Natalia,
Por cada "mami" que ha salido de tu boca y que quizá yo no he merecido. Por tu infinidad de ocurrencias que me arrancan la sonrisa. Por que no me quieres, por tus palabras –No te quiero… solamente I love you.

A M.I. Marcial,
Por las múltiples facetas en que me ha permitido conocerle. Por el profesor que me trasmitió su conocimiento. Por el jefe que me enseño a ser profesionista. Por el amigo que me ha escuchado y aconsejado.

Al Dr. Boris,
Por la oportunidad de iniciar y culminar el posgrado.

A la Dra. Amparo,
Por su valioso apoyo. Por confiar en mí. Por la dedicación a este trabajo.

A la Dra. Hanna, al Dr. Voutssas y a la M.C. Guadalupe Ibarguengoita,
Por el tiempo dedicado a la revisión y observaciones a este trabajo.

A Alex,
Por compartir esta etapa conmigo. Por el apoyo y la amistad siempre incondicional.

Silvia Socorro Ballesteros Estrada

Agradezco

A quien me acompaño en este camino…

A Luís, a David y a Oswaldo por su amistad.

A Ramiro y a Lulú por las palabras de aliento cuando sentí vencerme, palabras que me ayudaron a retomar la confianza para continuar en el camino.

A quien ha estado conmigo incondicionalmente…

A Paco R. R. por los conocimientos transmitidos, por los comentarios siempre oportunos, y por esas muy gratas y reconfortantes visitas de doctor.

A mis abuelos Moy, Lupita, Pedro y Eustolia.

A mis tías Mary, Lucy y Lupita.

A mis primos Martín, Luis, Irma, Verónica, Norma, Claudia, Gabriel, Oscar, Dalila, Eva, Vale y Evelia.

Al resto de mi familia.

A Emmanuel.

A todos y a cada uno de mis amigos.

Silvia Socorro Ballesteros Estrada

Índice

página

Introducción

Antecedentes

La función de las bibliotecas universitarias es apoyar directamente la vida académica a través de selección, adquisición, procesamiento, difusión, circulación, control y preservación del material bibliográfico. En la UNAM, la Dirección General de Bibliotecas (DGB) es la dirección encargada de coordinar el sistema bibliotecario, en la actualidad está conformada por 142 bibliotecas, 16 para bachillerato, 50 de licenciatura y posgrado, 20 de humanidades, 32 de investigación científica, y 24 de extensión y administración universitaria, a través de las cuales da servicios bibliotecarios a la comunidad, entre los que se encuentran la consulta a los catálogos de libros, tesis, revistas y mapas.

Con el avance de la tecnología en las últimas décadas y el empleo de ésta para automatizar los procesos de diferentes actividades del quehacer humano, las bibliotecas de las universidades no han sido la excepción al adoptar o hacer uso de software para automatizar los servicios que ofrecen a los usuarios. Las bibliotecas ahora tienen la necesidad de mantener una constante en el uso de las tecnologías de la información.

A principios de 1994 el Comité Asesor de Cómputo de la Dirección General de Bibliotecas en coordinación con la Dirección General de Servicios de Computo Académico de la UNAM y conforme a los lineamientos para el desarrollo bibliotecarios e informático, deciden modernizar las bases centrales de la DGB como lo es LIBRUNAM que contiene registros bibliográficos de libros, SERIUNAM que contiene registros bibliográficos de publicaciones periódicas y TESIUNAM que contiene registros bibliográficos de tesis. Por lo anterior se dan a la tarea de avaluar sistemas integrales para la administración de bibliotecas y como resultado se decidió adquirir e implantar el sistema ALEPH.

En el año 1996, la Dirección General de Bibliotecas adquiere el sistema ALEPH en su versión 300 el cual es implementado en el año 1997 en el sistema bibliotecario de la UNAM, las características de la versión eran:

- Información almacenada bajo un sistema de archivos de texto plano.
- No contaba con una interfaz gráfica

- Conexión remota a través de TELNET[1] o SSH[2]
- Uso de comandos para realizar las actividades: "pt" para préstamo, "dv" para devolución, "ren" para renovación, "l" o "le" para lista de lectores, etc.

En el año 2001, la DGB, adquiere la actualización de ALEPH a su versión 500 y en 2002 comienza la migración, sus características son:

- La arquitectura está basada en multicapas, siguiendo el modelo cliente/servidor.
- Diseño de base de datos flexible.
- Está compuesto por siete módulos interrelacionados: autoridades, bibliográfico, acervos, administrativo, préstamo interbibliotecario y un módulo general de administración
- Los módulos funcionales del sistema ALEPH 500 se trabajan mediante el cliente GUI (Graphic User Inteface).

La base de datos TESIUNAM es una de las bases de datos que originalmente estaba bajo ALEPH 300 y que posteriormente se migra a ALEPH 500.

Breve reseña histórica de TESIUNAM

- **1986** Inicia TESIUNAM.
- **1992** Se edita en disco compacto.
- **1996** Aparece la segunda edición y se proyecta la digitalización de las tesis.
- **1997** A partir de este año, se administra a través de ALEPH 300 y el catálogo está disponible a través del sitio Web de la DGB.
- **2006** Se libera en el sistema ALEPH 500

TESIUNAM contiene los registros bibliográficos correspondientes a las tesis que son generadas por los alumnos egresados de la UNAM, así como de algunas universidades incorporadas, desde 1900 hasta la fecha.

TESIUNAM es una base de datos, que cuenta con más de 363,000 registros bibliográficos de tesis en formato impreso y más de 40,000 registros bibliográficos de tesis en formato electrónico, ya que a partir del año 2002 se cuenta con texto completo. Siendo así, uno de los acervos más grandes que se tienen en México y Latinoamérica, y con ello se cuenta también con una gran fuente de información que puede ser aprovechada para estudios cuantitativos.

[1] Servicio otorgado por un servidor utilizando protocolo de comunicación TCP, por el cual se establecen sesiones remotas por parte de los clientes y la información que viaja por la red puede ser codificada fácilmente.
[2] Servicio otorgado por un servidor utilizando protocolo de comunicación TCP, cuando se establece una conexión cliente – servidor, los datos utilizados en este protocolo utilizan tecnologías de cifrado.

Definición del problema

El tratamiento y manejo de la literatura científica por medios cuantitativos sirve no sólo para analizar el volumen de publicaciones, la productividad de autores, revistas o materias, sino también en un sentido más amplio, para el conocimiento de los procesos y la naturaleza de las ciencias.

Es de gran importancia conocer el material bibliográfico con el que cuenta la DGB a través de sus catálogos disponibles, sin embargo, como dependencia encargada de coordinar el sistema bibliotecario de la UNAM, surge la necesidad de obtener indicadores bibliométricos, es decir, medidas basadas habitualmente en recuentos de publicaciones que persiguen cuantificar resultados.

Uno de los problemas que se tienen en la actualidad en la Dirección General de Bibliotecas es que no cuenta con un sistema automatizado para la obtención de medidas cuantitativas de la información contenida en sus bases de datos.

Considerando el volumen actual de la base de datos TESIUNAM y su perspectiva de crecimiento, la Biblioteca Central, la cual se encarga de la recepción de tesis, se ve en la necesidad de obtener información cuantitativa de la producción de tesis llevando un "c*ontrol*" en bitácoras en papel, lo cual es un procedimiento obsoleto, poco fiable ya que no se asegura de que todas y cada una de las tesis sean registradas, además que es un procedimiento manual, tedioso y redundante debido a que la catalogación bibliográfica se realiza en el sistema ALEPH.

La obtención de indicadores bibliométricos no es tarea fácil, ya que los registros bibliográficos almacenados en la base de tesis no cuentan con el formato para acceder de manera directa a la información y la forma de catalogar los registros bibliográficos no está unificada.

El formato en el que los registros bibliográficos se encuentran almacenados en la base de datos es bajo una norma internacional de catalogación, dicho formato presenta una gran ventaja para el almacenamiento ya que los registros bibliográficos son formados por una secuencia de caracteres y son almacenados en campos de longitud variable debido a que sus campos no cuentan con una estructura y longitud fija, por ejemplo pueden tener uno o mas autores, uno o mas asesores, etc. Sin embargo en el acceso a los datos se presenta una desventaja ya que al ser almacenados como una estructura secuencial de caracteres no es posible acceder de forma directa.

Otro de los obstáculos en la obtención de indicadores bibliométricos es que, debido a la edad de la base de datos, gran cantidad de personal la ha alimentado bajo distintas reglas de catalogación, debido a ello es posible encontrar información que no se encuentra estandarizada.

Por lo tanto en el presente trabajo se muestra la implementación de un procedimiento automatizado para la clasificación y recuperación de la información para la obtención e interpretación de indicadores bibliométricos.

Objetivo primario del proyecto

- El objetivo del presente trabajo consiste en normalizar, indizar y almacenar la información de la base de datos de tesis para optimizar la recuperación de información, a la cual se le puedan aplicar modelos matemáticos para la obtención de indicadores bibliométricos que muestren el comportamiento en la producción de tesis, y que estos a su vez se presenten de forma amigable a quien está interesado en estudios bibliométricos.

Objetivos secundarios del proyecto

- Tener los datos normalizados de acuerdo a la optimización de procedimientos de análisis, de almacenamiento, de recuperación de información.
- Clasificar los registros bibliográficos de tesis dentro de las siguientes áreas del conocimiento: físico matemáticas, ciencias biológicas y de la salud, ciencias sociales y humanidades y arte.
- Obtener indicadores bibliométricos necesarios para la interpretación de la tendencia de investigación en la UNAM.

Metodología

- Estudiar el formato en que los registros bibliográficos son almacenados en el sistema ALEPH en su versión 500.
- Estudiar y aplicar técnicas de indización de texto en lenguaje natural.
- Estudiar y aplicar técnicas de recuperación de información.
- Aplicar la teoría de redes neuronales artificiales para clasificar de forma automática los registros bibliográficos.

Alcances

Desarrollo y aplicación de un algoritmo para la clasificación automática de registros conforme a las áreas establecidas.

Desarrollo y aplicación de un algoritmo para extracción, limpieza y normalización de datos almacenados en la base de tesis, TESIUNAM.

Desarrollo de un sistema para la obtención de indicadores bibliométricos que permita medir e interpretar la tendencia en la investigación de los trabajos recepcionales que se realizan en la UNAM, para con ello poder evaluar y descubrir las tendencias de la producción.

El presente trabajo muestra el desarrollo y aplicación de algoritmos de extracción, limpieza, normalización, almacenamiento y recuperación de información bibliográfica de tesis de alumnos egresado de la UNAM y escuelas incorporadas, con la finalidad de obtener indicadores bibliométricos que muestren el comportamiento de la investigación.

Se divide en 5 capítulos:

Capítulo 1. Marco teórico. Se describen los fundamentos teóricos para el desarrollo del presente trabajo.

Capítulo 2. Recuperación de información. Se presentan los factores que intervienen en la recuperación de información, la estructura de los documentos bibliográficos y algunos modelos matemáticos que sirven de base para la obtención de resultados y su interpretación.

Capítulo 3. Análisis de texto e indexación. Se presenta la forma en que el leguaje natural puede ser procesado y sus inconvenientes, así como las etapas del proceso de indización aplicadas a los registros bibliográficos.

Capítulo 4. Redes neuronales artificiales. Se presentan los fundamentos de redes neuronales necesarios para clasificar por áreas de forma automática los registros bibliográficos.

Capítulo 5. Herramienta para la obtención de indicadores. Se presenta el análisis, diseño y la implementación de la herramienta.

Apéndice A. Se presenta un ejemplo de la estructura de un registro bibliográfico con gran cantidad de información, tal cual es almacenado en la base de datos.

Capitulo 1
Marco teórico

1.1. Introducción

El capítulo 1 contiene una serie de fundamentos teóricos auxiliares para la comprensión y desarrollo del presente trabajo.

Involucra temas relacionados con las bases de datos y sistemas de bases de datos. Hoy en día es difícil pensar en una aplicación, empresa y/o área que no hagan uso de las bases de datos para la optimización de sus procesos.

En el presente trabajo el área involucrada es la bibliometría, por ello también dentro de los fundamentos teóricos se encuentran aquellos que hablan del área y su importancia.

1.2. Las bases de datos

Un sistema de bases de datos es básicamente un sistema computarizado, cuya finalidad es almacenar, recuperar y actualizar información a través de peticiones.

Uno de los principales propósitos de un sistema de bases de datos es proporcionar a los usuarios una visión abstracta de los datos. Es decir, el sistema esconde detalles de cómo se almacenan y mantienen los datos, a través de niveles de abstracción, tal como lo muestra la figura 1.1, para simplificar la interacción de los usuarios con la información:

- Nivel físico: es el nivel más bajo de abstracción y describe *cómo* se almacenan realmente los datos, se describen en detalle las estructuras de datos complejas de bajo nivel.
- Nivel lógico: es el siguiente nivel de abstracción, describe *qué* datos se almacenan en la base de datos y *qué* relaciones existen entre esos datos.
- Nivel de vistas: es el nivel más alto de abstracción, describe sólo parte de la base de datos completa. Para que el usuario simplifique su interacción con el sistema se define la abstracción a nivel de vistas.

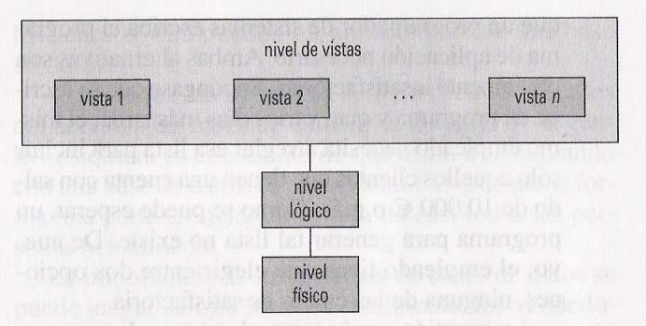


Figura 1.1 Niveles de abstracción de datos

En el nivel físico, un registro se puede describir como un bloque de posiciones almacenadas consecutivamente (por ejemplo, palabras o bytes). El sistema de bases de datos esconde detalles de almacenamiento de nivel inferior a los programadores de bases de datos. Los administradores de la base de datos pueden ser consientes de ciertos detalles de la organización física de los datos.

En el nivel lógico cada registro se describe mediante una definición de tipo y la relación entre esos tipos de registros. Los programadores, cuando usan un lenguaje de programación trabajan en este nivel de abstracción, los administradores de la base de datos lo trabajan habitualmente.

Finalmente, en el nivel de vistas, los usuarios ven un conjunto de programas de aplicación que esconden los detalles de los tipos de datos, además, también proporcionan un mecanismo de seguridad para evitar que los usuarios accedan a ciertas partes de la base de datos.

El diseño completo de la base de datos se llama el *esquema* de la base de datos los sistemas de bases de datos tienen varios esquemas divididos de acuerdo a los niveles de abstracción.

1.2.1 Modelos de datos

Bajo la estructura de la base de datos se encuentra el modelo de datos: una colección de herramientas conceptuales para describir los datos, las relaciones, la semántica y las restricciones de consistencia. Existen modelos relacionales y no relacionales, en el modelo relacional se ven tablas, en contraste, en un modelo no relacional se ven otras estructuras. Por ejemplo, en un sistema jerárquico, los datos son representados ante el usuario como un conjunto de estructuras de árbol (jerárquicas), y los operadores que se proporcionan para manipular dichas estructuras incluyen operadores para apuntadores de recorrido. [2]

Los primeros productos relacionales comenzaron a aparecer a finales de los años setenta y principios de los ochenta. Hasta el momento, la gran mayoría de los sistemas de bases de datos son relacionales y operan prácticamente en todo tipo de plataforma de hardware y software.

1.2.1.1 El modelo relacional

El modelo relacional es sin lugar a dudas el fundamento de la tecnología moderna de bases de datos. En el modelo relacional se utiliza un grupo de tablas para representar los datos y las relaciones entre ellos.

El modelo relacional se ocupa de tres aspectos principales de la información: la *estructura* de datos, la *manipulación* de datos y la *integridad* de los datos. Cada aspecto tiene su propia terminología. Los términos **estructurales** más importantes son: el propio término relacional, tupla, cardinalidad, atributo, grado, dominio y clave primaria. Figura 1.2

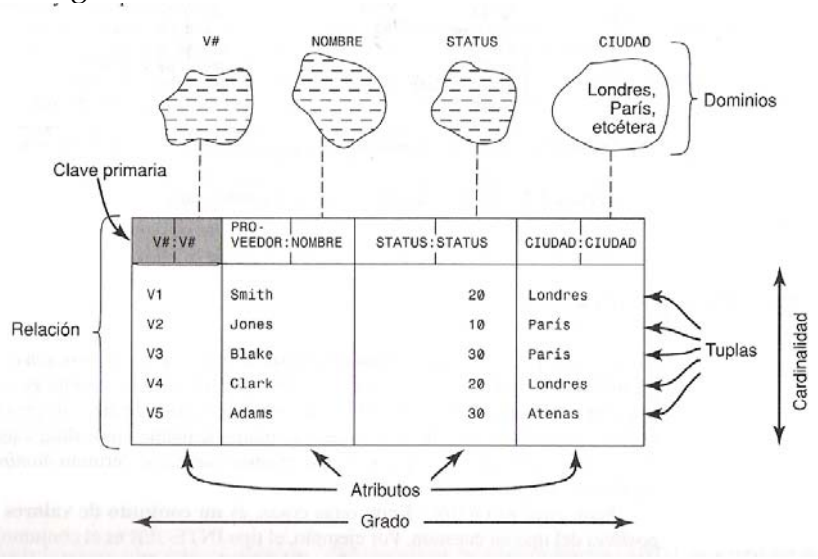


Figura 1.2. Terminología estructural

La relación es una tabla, la tupla corresponde a una fila de dicha tabla y un atributo a una columna; al número de tuplas se le llama cardinalidad y al número de atributos se le denomina grado; y un dominio es un conjunto de valores, de donde se toman los valores de atributos específicos de relaciones específicas.

La parte de **manipulación** del modelo relacional ha evolucionado considerablemente, sin embargo, todavía se da el caso de que el componente principal de esa parte de manipulación es lo que se denomina **álgebra relacional**, que básicamente es sólo el conjunto de operadores que toman relaciones como sus operadores y regresan una relación como resultado. Tres de estos operadores son: *restringir, proyectar y juntar.*

El álgebra relacional constaba de ocho operadores, figura 1.3., en dos grupos de cuatro cada uno:

1. Operadores tradicionales sobre conjuntos *unión, intersección, diferencia y producto cartesiano*.
2. Los operadores relacionales especiales *restringir (seleccionar), proyectar, juntar y dividir.*

Restringir. Regresa una relación que contiene todas las tuplas de una relación especificada que satisfacen una condición.
Proyectar. Regresa una relación que contiene todas las tuplas o subtuplas que quedan en una relación especificada después de quitar los atributos especificados.
Producto. Regresa una relación que contiene todas las tuplas posibles que son una combinación de dos tuplas, una de cada una de dos relaciones especificadas.
Unión. Regresa una relación que contiene todas las tuplas que aparecen en una o en las dos relaciones especificadas.
Intersección. Regresa una relación que contiene todas las tuplas que aparecen en las dos relaciones especificadas.
Diferencia. Regresa una relación que contiene todas las tuplas que aparecen en la primera pero no en la segunda de las dos relaciones especificadas.
Juntar (Join). Regresa una relación que contiene todas las tuplas posibles que son una combinación de dos tuplas de cada una de dos relaciones especificadas, tales que las dos tuplas que contribuyen a cualquier combinación dada tengan un valor común para los atributos comunes de las dos relaciones.
Dividir. Toma dos relaciones unarias y una relación binaria y regresa una relación que contiene todas las tuplas de una relación unaria que aparecen en la relación binaria y que a la vez coinciden con todas las tuplas de la otra relación unaria.

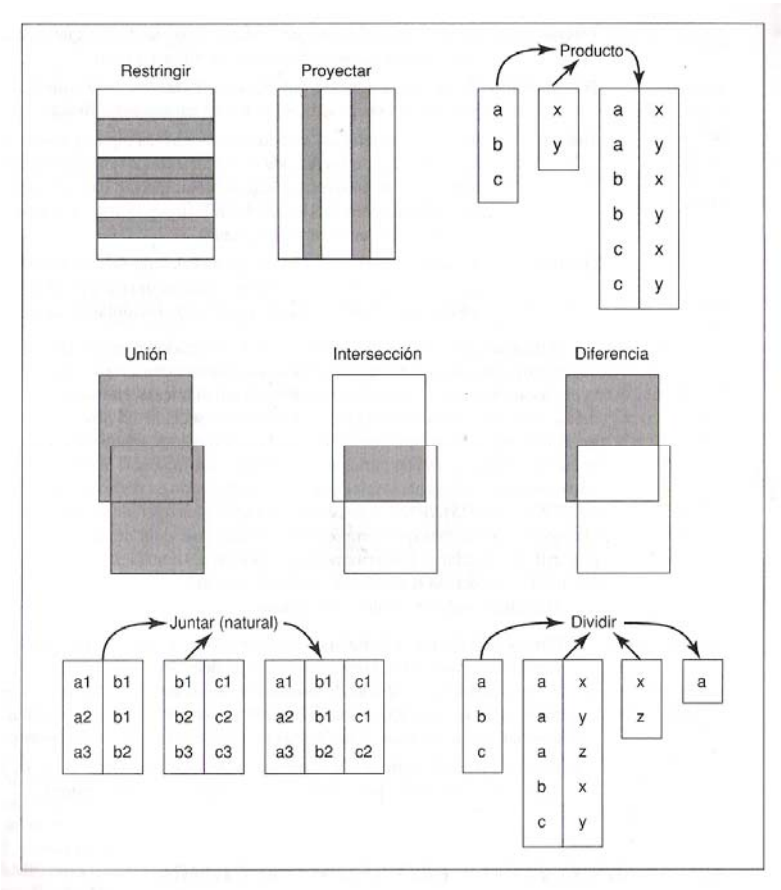


Figura 1.3. Panorama de los ocho operadores originales.

El término **integridad** se refiere a la *exactitud* o *corrección* de los datos en la base de datos. Las restricciones se pueden clasificar en general en cuatro grandes grupos:

- Una restricción de *tipo* especifica los valores válidos para un tipo dado.
- Una restricción de *atributo* especifica el valor válido de un atributo dado.
- Una restricción de *tabla* especifica los valores válidos de una tabla determinada.
- Una restricción de *base de datos* especifica el valor válido de una base de datos dada.

1.2.2. Lenguajes de Bases de Datos

Un lenguaje de bases de datos es un lenguaje para acceso a las bases de datos, realizando funciones de definición, control y administración de bases de datos. La mayoría de los sistemas de bases de datos los soportan, proporcionando:

- Un lenguaje de definición de datos (DDL) para especificar el esquema de la base de datos, además, actualiza un conjunto especial de tablas denominado diccionario de datos.
- Un lenguaje de manipulación de datos (DML) para expresar las consultas a la base de datos y las modificaciones.

En la práctica, los lenguajes de definición y manipulación de datos no son dos lenguajes separados, en su lugar simplemente forman partes de un único lenguaje de bases de datos, tal como el lenguaje SQL. [1]

Los programas de aplicación son programas que se usan para interaccionar con la base de datos. Los programas de aplicación se describen usualmente en un lenguaje anfitrión, tal como C, C++, Pro*C o Java con consultas de SQL embebido que acceden a la base de datos, para ello sentencias DML (Data Manipulation Language) necesitan ser ejecutadas desde el lenguaje anfitrión. Hay dos maneras de hacerlo:

- Proporcionando una interfaz de programas de aplicación (conjunto de procedimientos) que se pueden usar para enviar instrucciones DDL o DML a la base de datos. El estándar de conectividad abierta de bases de datos (ODBC, Open Data Base Connectivity) definido por Microsoft para el uso con el lenguaje C es un estándar de interfaz de programas de aplicación usado comúnmente. El estándar de conectividad de Java con bases de datos (JDBC, Java Data Base Connectivity) proporciona características correspondientes para el lenguaje Java con la peculiaridad de ser independiente de cualquier plataforma.

- Extendiendo la sintaxis del leguaje anfitrión para incorporar llamadas DML dentro del programa del lenguaje anfitrión. Usualmente, un carácter especial precede a las llamadas DML, y un preprocesador, llamado precompilador DML convierte las sentencias DML en llamadas normales a procedimientos en el lenguaje anfitrión.

Las bases de datos forman una parte esencial de casi todas las empresas, no siendo la excepción las educativas, y particularmente, las bibliotecas.

Las bases de contenido bibliográfico, constituyen una de las principales fuentes de información sobre las publicaciones periódicas, libros, tesis, mapas, etc. A continuación se mencionan algunas ventajas que tienen las bases de datos para la elaboración de estudios bibliométricos.

- Su gran capacidad de almacenamiento permite actuar sobre grandes unidades de datos en cantidad suficiente para una evaluación correcta.

- La estructura y organización de los datos posibilita la presentación homogénea de las referencias bibliográficas.
- El gran número de campos posibles (autores, título, editorial, año de publicación, clasificación, descriptores o resumen) permite una gran variedad de elementos de recuperación e índices sobre los que aplicar los parámetros o indicadores con suficientes garantías de fiabilidad.

El estudio de la información es frecuentemente usado en bibliotecas científicas y ciencias de la información.

En un sentido extenso, cualquier sistema de información diseñado para auxiliar el estado del conocimiento utiliza conceptos y procedimientos de almacenamiento y recuperación de la información.

La recuperación de información, IR por sus siglas en inglés, está relacionada con la representación, almacenamiento, organización y acceso a datos.

Antes de utilizar una base de datos para realizar un estudio bibliométrico hay que analizar su cobertura temática, geográfica y documental, sus criterios de indización, etc.

Cada registro es clasificado de acuerdo con los procedimientos establecidos e incorporado en una colección. Los procedimientos son establecidos para formular preguntas diseñadas para satisfacer las necesidades de información, o *consultas*, con la descripción de los registros almacenados.

Los procedimientos de selección permiten hacer la discriminación de los campos innecesarios para el estudio, estos procedimientos son realizados en bibliotecas donde los registros catalográficos son la principal herramienta.

En la práctica, la relevancia de registros de información específicos a una respuesta articulada no está determinada directamente. Los registros son convertidos a una forma especial usando procesamiento, clasificación y/o indización.

1.2.3. El Sistema de Administración de Bases de Datos

El sistema manejador de bases de datos, DBMS por sus siglas en inglés (Data Base Management System), es el software que maneja todo acceso a la base de datos. De manera conceptual, lo que sucede es lo siguiente:

1. Un usuario emite una petición de acceso, utilizando algún sublenguaje de datos específico (por lo regular SQL)

2. El DBMS intercepta esa petición y la analiza.
3. El DBMS inspecciona, en su momento, (las versiones objeto de) el esquema externo para el usuario, la transformación externa/conceptual correspondiente, el esquema conceptual, la transformación conceptual/interna y la definición de la estructura de almacenamiento.
4. El DBMS ejecuta las operaciones necesarias sobre la base de datos almacenada.

Funciones de DBMS

- *Definición de datos.* El DBMS debe ser capaz de aceptar definiciones de datos en la forma fuente y convertirlas a la forma objeto correspondiente. En otras palabras, el DBMS debe incluir entre sus componentes un procesador DDL, o compilador DDL, para cada uno de los diversos DDL's.
- *Manipulación de datos.* El DBMS deber ser capaz de manejar peticiones para recuperar, actualizar o eliminar datos existentes en la base de datos o agregar nuevos datos a ésta. En otras palabras, el DBMS debe incluir un componente procesador DML o compilador DML para tratar con el DML.
- *En general, las peticiones DML pueden ser "planeadas" o "no planeadas":*
  a. Una petición planeada es aquella cuya necesidad fue prevista antes del momento del ejecutar la petición.
  b. En contraste, una petición no planeada es una consulta *ad hoc*; es decir, una petición para la que no se previó por adelantado su necesidad.
- *Optimización y ejecución*. Las peticiones DML, planeadas o no planeadas, deben ser procesadas por el componente optimizador, cuya finalidad es determinar una forma eficiente de implementar las peticiones.
- *Seguridad e integridad de los datos.* El DBMS debe vigilar las peticiones del usuario y rechazar todo intento de violar las restricciones de seguridad y de integridad definidas por el administrador de la base de datos.
- *Recuperación de datos y concurrencia.* El DBMS o probablemente, algún otro componente de software relacionado, denominado comúnmente administrador de transacciones o monitor de procesamiento de transacciones (monitor PT) debe imponer ciertos controles de recuperación y concurrencia.
- *Diccionario de datos.* El DBMS debe proporcionar una función de diccionario de datos. Este diccionario puede ser visto como una base de datos por derecho propio, contiene "datos acerca de los datos"; es decir, definiciones de otros objetos del sistema. En particular, los diversos esquemas y transformaciones y las diversas restricciones de seguridad y de integridad, serán almacenadas en el diccionario.
- *Rendimiento.* El DBMS debe realizar todas las tareas antes identificadas de la manera más eficiente posible.

1.2.4. Independencia física de los datos

Las aplicaciones implementadas en sistemas más antiguos tienden a ser *dependientes de los datos,* esto significa que la forma en que físicamente son representados los datos en el almacenamiento secundario y la técnica empleada para su acceso, son dictadas por los requerimientos de la aplicación en consideración, y más aún, significa que el conocimiento de esa representación física y esa técnica de acceso están integrados dentro del código de la aplicación.

En un sistema de bases de datos, sería inconveniente permitir que las aplicaciones fuesen dependientes de los datos en el sentido descrito; por lo menos por las dos razones siguientes:

- Las distintas aplicaciones requerirán visiones diferentes de los mismos datos.
- El administrador de base de datos debe tener la libertad de cambiar las representaciones físicas o la técnica de acceso en respuesta a los requerimientos cambiantes, sin tener que modificar las aplicaciones existentes.

De aquí que dar independencia a los datos sea un objetivo principal de los sistemas de bases de datos. Se puede definir la independencia de los datos *como la inmunidad de las aplicaciones a cambios en la representación física y la técnica de acceso*; lo que implica desde luego que las aplicaciones involucradas no dependan de ninguna representación física o técnica de acceso en particular. [2]

1.2.5. Vistas

Una vista es básicamente sólo una expresión del álgebra relacional que no es evaluada sino que es "recordada" por el sistema, el cual la guarda con un nombre especificado. Existen muchas razones por las cuales es necesario el soporte de vistas:

- *Las vistas proporcionan seguridad automática para datos ocultos.* "Datos ocultos" se refiere a los datos que no son visibles a través de alguna vista determinada. Existe la seguridad de que estos datos no serán accedidos a través de esa vista en particular. Por lo tanto, obligar a los usuarios a acceder a las base de datos a través de vistas constituye un mecanismo de seguridad simple pero efectivo.
- *Las vistas ofrecen una posibilidad de forma abreviada o "macro".* Sólo se recupera información necesaria.
- *Las vistas permiten que los datos sean vistos de distinta forma por diferentes usuarios al mismo tiempo.* Las vistas permiten a los usuarios concentrarse

solamente en la porción de las base de datos que les interesa e ignorar el resto.

- *Las vistas pueden ofrecen independencia lógica de los datos.*

1.2.6. Sistemas de bases de datos frente a sistemas de archivos

Una manera de mantener la información en una computadora es almacenarla en archivos del sistema operativo. Para permitir a los usuarios manipular la información, el sistema tiene un número de programas de aplicación que manipula los archivos, los cuales se escriben por programadores de sistemas en respuesta a las necesidades de la organización.

Mantener información de la organización en un sistema de procesamiento de archivos tiene una serie de inconvenientes importantes:

- *Redundancia e inconsistencia de datos.* Debido a que los archivos y programas de aplicación son creados por diferentes programadores en un largo período de tiempo, los diversos archivos tienen probablemente diferentes formatos y los programas pueden estar escritos en diferentes lenguajes. Más aún, la misma información puede estar duplicada en diferentes archivos. Esta redundancia conduce a un almacenamiento y coste de acceso más alto. Además, puede conducir a inconsistencia de datos; es decir, las diversas copias de los mismos datos pueden no coincidir.
- *Dificultad en el acceso a los datos.* El entorno de procesamiento de archivos convencional no permite que los datos necesarios sean obtenidos de una forma práctica y eficiente. Se deben desarrollar sistemas de recuperación de datos más interesantes para un uso general.
- *Asilamiento de datos.* Debido a que los datos están dispersos en varios archivos, y los archivos pueden están en diferentes formatos, es difícil escribir nuevos programas de aplicación para recuperar los datos apropiados.
- *Problemas de integridad.* Los valores de los datos almacenados en la base de datos deben satisfacer cierto tipo de restricciones de consistencia.
- *Problemas de atomicidad.* Un sistema está sujeto a fallo. En muchas aplicaciones es crucial asegurar que, una vez que un fallo ha ocurrido y se ha detectado, los datos se restauran al estado de consistencia que existían antes del fallo. Es decir, la operación debe ser *atómica:* debe ocurrir por completo o no ocurrir en absoluto. Es difícil asegurar esta propiedad en un sistema de procesamiento de archivos convencional.
- *Anomalías en el acceso concurrente.* Para protegerse de estado incorrectos, el sistema debe mantener alguna forma de supervisión.
- *Problemas de seguridad.* Como los programas de aplicación se añaden al sistema de una forma *ad hoc*, es difícil garantizar restricciones de seguridad.

Estas dificultades, entre otras, han motivado el desarrollo de los sistemas de bases de datos. [1]

1.2.7. Beneficios del enfoque de bases de datos

- *Los datos pueden compartirse*. Compartir no sólo significa que las aplicaciones existentes puedan compartir la información para operar sobre los mismos datos. En otras palabras, es posible satisfacer los requerimientos de datos de aplicaciones nuevas sin tener que agregar información a la base de datos.
- *Es posible reducir la redundancia*. En sistemas que no son de bases de datos, cada aplicación tiene sus propios archivos exclusivos. A menudo este hecho puede conducir a una redundancia considerable de los datos almacenados, con el consecuente desperdicio de almacenamiento.
- *Es posible, hasta cierto grado, evitar la inconsistencia*. Como alternativa, si no se elimina la redundancia pero se *controla*, entonces en DBMS puede garantizar que la base de datos nunca será inconsistente, asegurando que todo cambio realizado a cualquiera de las entidades será aplicado también a la otra en forma automática. A este proceso se le conoce como propagación de actualizaciones.
- *Es posible brindar un manejo de transacciones*. Una transacción es una unidad de trabajo lógica, que por lo regular comprende varias operaciones de la base de datos (en particular, varias operaciones de actualización). El ejemplo común es el de transferir una cantidad de efectivo de una cuenta A a otra cuenta B, es claro que aquí se necesitan dos actualizaciones, una para retirar el efectivo de la cuenta A y la otra para depositarlo en la cuenta B. si el usuario declara que las dos operaciones son parte de la misma transacción, entonces el sistema puede en efecto garantizar que se hagan ambas o ninguna de ellas, aún cuando el sistema fallara a la mitad del proceso.
- *Es posible mantener la integridad*. El problema de la integridad es el de asegurar que los datos de la base de datos estén correctos. La inconsistencia entre dos entradas que pretenden representar el mismo "hecho" es un ejemplo de la falta de integridad; desde luego, este problema en particular puede surgir sólo si existe redundancia en los datos almacenados. No obstante, aún cuando no exista redundancia, la base de datos podría seguir conteniendo información incorrecta. El control centralizado de la base de datos puede ayudar a evitar estos problemas permitiendo que el administrador de datos defina y el administrador de la base de datos implemente las restricciones de integridad (también conocidas como reglas de negocio) que serán verificadas siempre que se realice una operación de actualización.
- *Es posible hacer cumplir la seguridad*. Al tener la completa jurisdicción sobre la base de datos, el administrador de base de datos puede asegurar que el

único medio de acceso a la base de datos sea a través de los canales adecuados y por lo tanto puede definir las reglas o restricciones de seguridad que serán verificadas siempre que se intente acceder a datos sensibles. Es posible establecer diferentes restricciones para cada tipo de acceso (recuperación, inserción, eliminación, etc.) para cada parte de la información de las bases de datos.

- *Es posible equilibrar los requerimientos en conflicto*. Al conocer los requerimientos generales de la empresa (a diferencia de los requerimientos de los usuarios individuales), el administrador de base de datos puede estructurar los sistemas de manera que ofrezcan un servicio general que sea "el mejor para la empresa". Por ejemplo, es posible elegir una representación física de los datos almacenados que proporcione un acceso rápido para las aplicaciones más importantes.
- *Es posible hacer cumplir los estándares*. Con el control central de la base de datos, el administrador de la base de datos puede asegurar que todos los estándares aplicables en la representación de los datos sean observados. Estos estándares podrían incluir alguno o todos los siguientes: departamentales, de instalación, corporativos, de la industria, nacionales e internacionales. Es conveniente estandarizar la representación de los datos en particular como un auxiliar para el *intercambio de datos* o para el movimiento de datos entre sistemas. En forma similar, los estándares en la asignación de nombre y en la documentación de los datos también son muy convenientes como ayuda para compartir y entender los datos. [2]

1.3. Estándar para intercambio de información

El almacenamiento de los registros bibliográficos se rige bajo normas internacionales de catalogación, las cuales presentan considerables ventajas para ser almacenados sin embargo se tienen diversas desventajas para la extracción de la información.

Z39.50 es uno de los estándares usados para la implementación de la automatización para almacenar, transferir y recuperar la información bibliográfica. Dicho estándar a lo largo de los años ha sufrido diversas modificaciones lo cual conlleva al estándar internacional ISO 2709, formato de documentación para el intercambio de información bibliográfica en cinta magnética. Fue aprobado por el subcomité de ANSI por la *Nacional Information Standards Organization.*

Este estándar especifica los requisitos para un formato generalizado del intercambio de información que acomoda muchos tipos de datos, especialmente de descripción bibliográfica de todas las formas de materiales y de datos relacionados tales como autoridad, *holdings*, circulación, etc.

1.4. Bibliometría

La bibliometría es la aplicación de métodos estadísticos y matemáticos dispuestos para definir comportamientos y tendencias de la comunicación escrita así como la naturaleza y el desarrollo de las disciplinas científicas mediante técnicas de recuento y análisis de dicha comunicación.

A través de la bibliometría es posible conocer la actividad, estructura y evolución de una ciencia, cuantificar sus resultados y aplicarlos en campos como la biblioteconomía, la historia de las disciplinas, la sociología de las ciencias o la política científica.

La bibliometría en la práctica se orienta a estudios bien definidos entre los que destacan:

- Los aspectos estadísticos del lenguaje y la frecuencia de uso de palabras, tanto en textos redactados en lenguaje natural como en otros medios.
- Las características de la productividad autoral, medida por el número de documentos publicados o por el grado de colaboración.
- Las características de las fuentes publicadas, incluyendo la distribución de los documentos por disciplinas.
- Los análisis de citas, teniendo en cuenta la distribución por autores, por tipo de documento, por instituciones o países.
- El uso de la información registrada, a partir de su demanda y circulación.
- La obsolescencia de la literatura, en virtud de la medición de su uso y de la frecuencia con que se cita.
- El incremento de la literatura por temas. [13]

El inicio de un estudio bibliométrico requiere tomar una serie de decisiones metodológicas que van a repercutir de forma importante en los resultados del análisis. Dicho estudio tiene tres pasos importantes:

- Recolección de datos
- Síntesis o reducción de datos
- Interpretación de datos

1.4.1. Indicadores bibliométricos

Los indicadores bibliométricos de producción científica son medidas, basadas habitualmente en *recuentos* de publicaciones, que persiguen cuantificar los resultados científicos atribuibles bien a unos *agentes* determinados o bien a agregados significativos de esos agentes. Los agentes elementales son los

investigadores, pero es más frecuente calcular indicadores de producción referidos a agregados como instituciones, regiones, países o disciplinas.

El número de publicaciones es el indicador de producción más sencillo y seguramente el primer indicador bibliométrico empleado concientemente como tal.

El objetivo primario de los indicadores de producción, como el de otros indicadores bibliométricos, es permitir la comparación entre un conjunto de agentes o de agregados científicos con la finalidad de detectar diferencias relevantes que sirvan para caracterizar el comportamiento de cada unos de ellos o del sistema del que pueden formar parte. [9]

El procedimiento por el cual se define una colección de agentes sobre el cual aplicar un indicador de producción debe incluir un criterio claro y explícito que especifique cuáles son los rasgos compartidos, cuando se dispone de un conjunto homogéneo de agentes se configura el marco de referencia que posibilita la interpretación de los valores del indicador.

Los indicadores, afirman algunos autores, sólo son aplicables a aquellas fuentes que sean un buen reflejo de la actividad del área. Hay revistas de ciencias sociales de tipo divulgativo y trabajos coyunturales, de escaso nivel científico, que podrían trasgiversar, de no ser contrastados, la realidad de una disciplina.

En la UNAM es de interés saber cual es la producción temática en las distintas disciplinas, descubrir la corriente de los investigadores, analizar la evolución cronológica de cierta disciplina, en particular de las tesis.

1.4.2. Estadística

Muchas disciplinas son involucradas en los sistemas de información, incluyendo, teoría de la información, teoría de probabilidad y estadística, semántica computacional, teoría de programación y álgebra. Técnicas de dichas disciplinas son usadas para obtener sistemas de información.

Cualquier estudio en el que se utilice la estadística se refiere a un conjunto de entidades, conocido con el nombre de población. Una población es un conjunto de individuos (personas, animales, cosas o entidades abstractas) con propiedades comunes, sobre los que se realiza una investigación de tipo estadístico. Por ejemplo, si se quiere investigar el número de páginas de los libros de una biblioteca, entonces la población está constituida por todos los libros de dicha biblioteca; pero si se quiere estudiar el número de líneas de todas y cada una de las páginas de un libro, entonces la población está constituida por las páginas de dicho libro.

La población en el presente trabajo son las tesis generadas en la UNAM, registradas y resguardadas por la Dirección General de Bibliotecas.

Cuando se estudia una población puede resultar difícil o costoso investigar a cada uno de los individuos de dicha población. De ahí, que en muchas ocasiones, sólo se estudie una parte de la población. A esa parte de la población se le llama muestra y al número de individuos que la conforman se llama tamaño muestral. Si se está interesado en estudiar el número de páginas de los libros de una biblioteca que consta de 10,000 libros, entonces la población esta constituida por esos 10,000 libros. Si parece difícil observar el número de páginas de todos esos libros, entonces se observa el número de páginas de unos cuantos de estos libros; por ejemplo, de 30 libros. Estos 30 libros constituyen una muestra.

La muestra ofrece una serie de datos que se pueden ordenar, simplificar y describir. Pero el objetivo fundamental es el de poder describir la población de partida mediante lo que se pueda encontrar en la muestra. Y lo más importante para poder describir la población es que las muestras sean representativas. La muestra en este trabajo son los registros bibliográficos de las tesis digitales.

Cuando se estudian los individuos de una población, se interesa por alguna de sus propiedades, precisamente, se llama variable a cualquier propiedad o cualidad que puede manifestarse bajo dos o mas formas distintas de un individuo de una población. La propiedad en este trabajo es la información bibliográfica.

1.5. Sistema ALEPH

Automated Library Expandable Program Hebrew, ALEPH, es un sistema adquirido por la UNAM para administrar las tareas de una biblioteca, tales como son la catalogación y la circulación de materiales.

La primera versión adquirida por la UNAM fue ALEPH 300, la cual:

- No cuenta con una interfaz gráfica
- Hace uso de comandos para realizar las actividades de préstamo, devolución renovación, lista de lectores, etc. (pt, dv, ren, l, le, br, xp, brd, etc.).
- La información almacenada en el sistema se guarda en archivos de texto plano.
- La interfase de ALEPH es a través de programas de conexión remota entre el servidor y la PC tales como telnet, Secure Shell, Win telnet, etc.

- ALEPH trabaja bajo plataformas VAX/VMS y UNIX, haciéndolo disponible en una amplia variedad de equipos, satisface aplicaciones que requieren pocas terminales y a instituciones grandes con centenares de terminales.

Para 1996 Ex-Libris Ltd. lanzó su última tecnología en cuanto a automatización de bibliotecas se refiere, ALEPH 500. Fue un rediseño del sistema ALEPH en plataforma UNIX. Está basada en una arquitectura cliente/servidor con un fuerte soporte del Sistema Manejador de Bases de Datos relacional ORACLE.

ALEPH maneja el concepto de *biblioteca* para referirse a un esquema de trabajo, hay 6 tipos de bibliotecas y cada biblioteca hace referencia a una base de datos según el tipo de información que maneja y se identifica por un código, formado por 3 caracteres seguidos de 2 dígitos, los dígitos identifican el tipo de biblioteca.

- Una biblioteca *bibliográfica*, que contiene registros bibliográficos, se identifica por un número entre 01 – 09 (por ejemplo. USM01)
- Una biblioteca *administrativa*, que contiene datos acerca de adquisiciones, circulación y usuarios, se identifica por un número entre el 50 y 59 (por ejemplo, USM50).
- *Holding*, que contiene datos agrupados de publicaciones periódicas, por ejemplo; se identifica por un número ente el 60 y 69 (por ejemplo, USM60).
- *Autoridad*, que contiene datos de autores y de materia, se identifica por un número ente el 10 y 19 (por ejemplo, USM11).
- *Préstamo inter-bibliotecario*, que contiene datos de proveedores que suministran material de préstamo, se identifica por un número ente el 20 y 29 (por ejemplo, USM20).
- *Course Reading*, que contiene datos de material reservado o utilizado en sala, se identifica por un número ente el 30 y 39 (por ejemplo, USM30)

ALEPH usa como manejador de bases de datos ORACLE ya que permite almacenar y manipular gran cantidad de información, siendo ésta una de las principales necesidades de la UNAM y en particular la base de datos TESIUNAM, a cargo de la Dirección General de Bibliotecas.

1.6. ORACLE

Los sistemas de base de datos como ORACLE están entre las mayores compañías de software del mundo.

ORACLE es un sistema de bases de datos relacional extremadamente potente y flexible. Esta potencia y flexibilidad, sin embargo, implican también cierta complejidad. Para poder diseñar aplicaciones basadas en ORACLE es

necesario entender cómo manipula ORACLE los datos almacenados en el sistema. PL/SQL es una herramienta de gran importancia diseñada para la manipulación de datos, tanto internamente dentro de ORACLE como externamente en aplicaciones.

PL/SQL es un sofisticado lenguaje de programación que se utiliza para acceder a bases de datos ORACLE desde distintos entornos, PL/SQL está integrado con el servidor de bases de datos, de modo que el código PL/SQL puede ser procesado de forma rápida y eficiente.

Los objetos de gran tamaño (Large Object, LOB) son una característica útil de cualquier base de datos relacional y se utilizan para almacenar grandes cantidades de información, ya sea en forma binaria o de texto. Un objeto de gran tamaño (LOB) es meramente un campo de la base de datos que contiene una gran cantidad de información, tal como un archivo gráfico o un archivo de texto de gran longitud. En ORACLE, una columna VARCHAR2 puede contener hasta un máximo de 2000 bytes. Además de la limitación de tamaño, una columna VARCHAR2 puede contener únicamente caracteres y no datos binarios. Una columna LONG puede contener hasta 2 GB y puede almacenar datos de tipo carácter. Junto con el tipo de datos LONG RAW, que puede almacenar datos binarios, LONG y LONG RAW son los tipos de datos más adecuados para el almacenamiento de información LOB en ORACLE. Sin embargo, las columnas LONG y LONG RAW tienen bastantes restricciones, incluyendo el hecho de que únicamente puede existir una columna LONG o LONG RAW por tabla.

ORACLE ha agregado soporte SQL para un amplio rango de funciones analíticas, incluyendo cubos, abstracciones, conjuntos de agrupación, clasificaciones, agregación de traslado, funciones *led* y *lag*, cajones de histograma, regresión lineal y desviación estándar, conjunto con la capacidad de optimizar la ejecución de dichas funciones en el motor de la base de datos.

ORACLE ha extendido las vistas materializadas para permitir funciones analíticas, en particular los conjuntos de agrupaciones. La capacidad de materializar partes o todo el cubo es primordial para el rendimiento de un sistema de administración de bases de datos multidimencionales y las vistas materializadas proporcionan al sistema de administración de bases de datos relacionales la capacidad de realizar lo mismo.

ORACLE tiene soporte extensivo para constructores relacionales orientados a objetos, incluyendo:

- Tipos de objetos. Se soporta un único modelo de herencia para las jerarquías de tipos.

- Tipos de colecciones. Soporta varrays, que son arrays de longitud variable, y a tablas anidadas.
- Tablas de objetos. Se utilizan para almacenar objetos mientras se proporciona una vista relacional de los atributos del los objetos.
- Funciones de tablas. Son funciones que producen conjuntos de filas como salida y se pueden utilizar en la cláusula *from* de una consulta.
- Vistas de objetos. Proporciona una vista de tablas de objetos virtuales de datos almacenados en una tabla relacional normal.
- Métodos. Se pueden escribir en PL/SQL, Java o C.
- Funciones de agregación definidas por el usuario. Se pueden utilizar en instrucciones SQL de la misma forma que las funciones incorporadas tales como *sum* y *count*.
- Tipos de datos XML. Se pueden utilizar para almacenar e indexar documentos XML.

1.6.1. SQL

Structured Query Language, SQL, es el lenguaje estándar para trabajar con bases de datos relacionales y es soportado prácticamente por todos los productos del mercado.

La mayoría de los productos SQL permiten la ejecución de instrucciones SQL de manera directa (es decir, en forma interactiva desde una terminal en línea) y también como parte de una programa de aplicación (es decir, las instrucciones SQL pueden estar incrustadas, lo que significa que pueden estar entremezcladas con las instrucciones del lenguaje de programación de dicho programa).

El principio fundamental subyacente al SQL incrustado, es que *toda instrucción SQL que puede ser usada en forma interactiva, también puede ser usada en un programa de aplicación*. Por supuesto, hay varias diferencias de detalle entre una determinada instrucción SQL interactiva y su contraparte incrustada en especial, las instrucciones de recuperación requieren de un tratamiento más amplio en un entorno de programa anfitrión.

En SQL incrustado:

- Las instrucciones SQL están precedidas pro EXEC SQL
- Una instrucción ejecutable puede aparecer en cualquier parte en donde aparezca una instrucción ejecutable del lenguaje anfitrión.
- Las instrucciones de SQL pueden incluir referencias a variables anfitrión; estas referencias deben incluir un prefijo de dos puntos para distinguirlas de los nombres de columnas de SQL. También pueden aparecer en una

cláusula INTO del SELECT o de FETCH para designar destinos de operaciones de recuperación.

- Todas las variables anfitrión a las que se hace referencia en instrucciones SQL deben estar declaradas dentro de una sección de declaración de SQL incrustado, la cual es delimitada por las instrucciones BEGIN y END DECLARE SECTION.
- Todo programa que contenga instrucciones de SQL incrustado debe incluir una variable anfitrión denominada SQLSTATE. Después de ejecutar cualquier instrucción de SQL, un código de estado devuelto al programa en dicha variable significa que la instrucción se ejecutó con éxito o no.
- Las variables anfitrión deben tener un tipo de datos apropiado de acuerdo con los usos para los que son propuestas.[2]

1.6.2. PL-SQL

ORACLE tiene dos lenguajes procedimentales principales, PL/SQL y Java. PL/SQL fue el lenguaje original de ORACLE para los procedimientos almacenados y tiene una sintaxis similar al utilizado en el lenguaje Ada. Java se soporta mediante una máquina virtual Java dentro del motor de la base de datos. ORACLE proporciona un paquete para encapsular procedimientos, funciones y variables relacionadas en unidades únicas. ORACLE soporta SQLJ (SQL incorporado en Java) y JDBC además proporciona una herramienta para generar las definiciones de clases Java correspondientes a tipos de la bases de datos definidos por el usuario.

ORACLE proporciona varios tipos de disparadores y varias opciones para el momento y forma en que se invocan. Los disparadores pueden escribirse en PL/SQL, Java o como llamadas a C. Para los disparadores que se ejecutan sobre instrucciones DML tales como *insert*, *update* y *delete*, Oracle soporta disparadores de fila (row) y disparadores de instrucciones (statement). Los disparadores de fila se pueden ejecutar una vez por cada fila que se vea afectada por la operación DML. Un disparador de instrucción se ejecuta solamente una vez por instrucción. En cada caso, el disparador se puede definir tanto como un disparador *before* o *alter* dependiendo de si se va a invocar antes o después de que se lleva a cabo la operación DML.

1.6.2.1. Disparadores

Los disparadores (o *triggers*) son parecidos a los procedimientos y funciones, en el sentido de que son bloques nominados de PL/SQL con secciones declarativa, ejecutable y de tratamiento de excepciones. Como los paquetes, los disparadores tienen que almacenarse en la base de datos como objetos independientes y no pueden ser locales a un bloque o paquete. Los disparadores se ejecutan de forma

implícita cuando ocurre el suceso que lo activa y no acepta argumentos. El acto de ejecutar un disparador se conoce como disparo. El suceso que lo activa puede ser una operación DML (INSERT, UPDATE o DELETE) sobre una tabla de la base de datos o ciertos tipos de vistas. ORACLE 8i amplía esta funcionalidad permitiendo que se activen con sucesos del sistema, como el inicio o la desconexión de la base de datos, y con ciertos tipos de operaciones del lenguaje de definición de datos DDL.
Los disparadores pueden utilizarse para varias finalidades, incluyendo:

- Mantener restricciones de integridad complejas que no pueden hacerse a través de restricciones declarativas especificadas durante la creación de la tabla.
- Auditar la información contenida en la tabla, registrando las modificaciones y el autor de las mismas.
- Indicar automáticamente a otros programas que es necesario realizar alguna acción cuando se modifica una tabla.

Los disparadores DML se activan a través de instrucciones DML y el tipo de instrucción determina el tipo de disparador. Los disparadores DML pueden definirse para las operaciones INSERT, UPDATE o DELETE y pueden dispararse antes o después de la operación. Dichos disparadores también pueden dispararse en operaciones de fila o de instrucción. La combinación de estos factores determina el tipo del disparador. Existen un total de 12 tipos posibles: 3 instrucciones por 2 opciones temporales por 2 niveles.

Un disparador de nivel de fila se activa una vez por cada fila que procese la instrucción de disparo. Dentro del disparador es posible acceder a los datos de la fila que se está procesando, lo que se consigue a través de dos identificadores de correlación **:old** y **:new.** Un *identificador de correlación* es un tipo especial de variable de acoplamiento de PL/SQL. El carácter dos puntos que los precede indica que son variables de acoplamiento, igual que las variables host utilizadas en PL/SQL incrustado, y que no son variables PL/SQL normales.

ORACLE 8i proporciona un tipo de disparador adicional. Los *disparadores de sustitución* (instead-of triggers) pueden definirse únicamente sobre vistas. A diferencia de los disparadores DML, un disparador de sustitución se ejecuta en lugar de la instrucción DML que lo activa. Los disparadores de sustitución tienen que ser de nivel de fila.

ORACLE 8i proporciona un tercer tipo de disparador. Un *disparador del sistema* se activa cuando tiene lugar un suceso del sistema, como la conexión o desconexión de la base de datos. Un disparador de sistema también puede dispararse con operaciones DDL, como la creación de tablas.

*1.6.3. Precompilador Pro*C/C++*

Otra forma de manipular los datos de ORACLE es a través de su precompilador Pro*C/C++. Un precompilador de ORACLE es una herramienta de programación que permite insertar sentencias SQL en un programa fuente de alto nivel. En la figura 1.4 se muestra como el precompilador acepta el programa fuente como entrada, traduce las sentencias SQL en llamadas a librerías de ORACLE, y genera un programa fuente modificado que se pueda compilar, ligar y ejecutar.

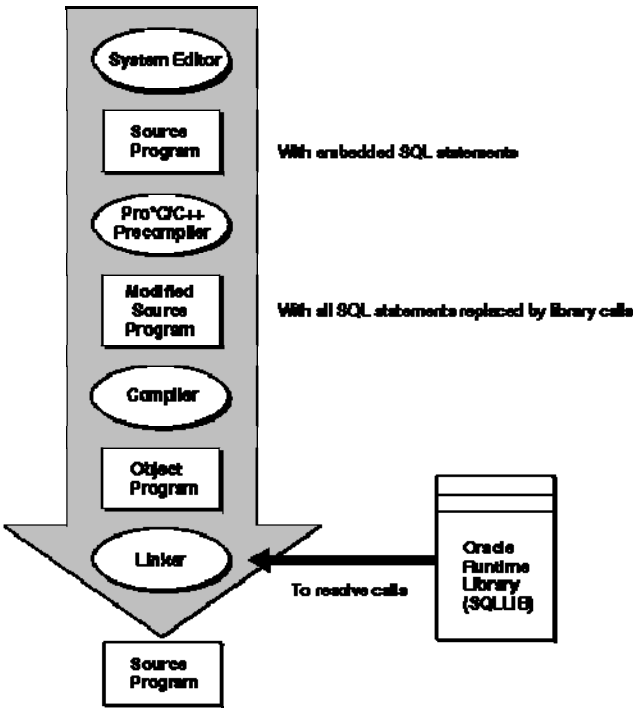


Figura 1.4. Desarrollo de un programa de SQL embebido

El precompilador de ORACLE Pro*C permite hacer uso del poder y flexibilidad de SQL en los programas de aplicación. Permite a la aplicación acceder directamente a ORACLE. Tal como se muestra en la figura 1.5, Pro*C ofrece muchas características beneficios que ayudan el desarrollo de aplicaciones eficaces y confiables.

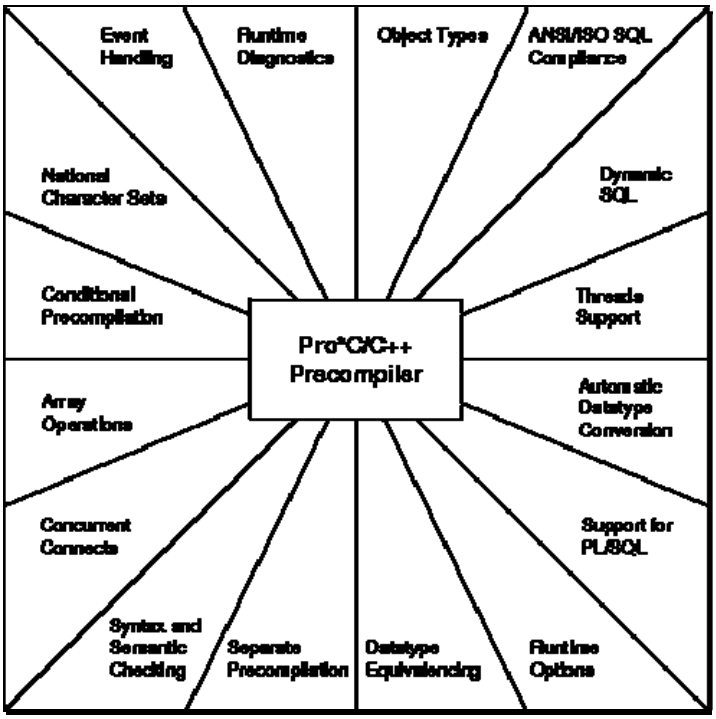


Figura 1.5. Características y beneficios de Pro*C

Pro*C/C++ permite:

- Escribir aplicaciones en C o C++
- Sigue estándares ANSI/ISO para sentencias de SQL embebido en un lenguaje de alto nivel.
- Toma ventajas de SQL dinámico, una técnica de programación avanzada que permite aceptar o construir cualquier sentencia válida en tipo de ejecución.
- Permite crear aplicaciones con necesidades específicas.
- Automáticamente convierte tipos de datos entre ORACLE y un leguaje de alto nivel.
- Mejora el funcionamiento de bloque de procesamiento de transacciones PL/SQL embebido.
- Especifica opciones útiles de precompilación *inline* y en la línea de comandos y cambia sus valores en la precompilación.
- Permite el acceso concurrente a las bases de datos de ORACLE.
- Condicionalmente precompila secciones de código las cuales pueden correr en diferentes entornos.
- Maneja errores y advertencias con el área de comunicaciones de SQL (SQLCA) y sentencias *WHENEVER* o *DO*.
- Trabaja con tipos de objetos definidos por el usuario en la base de datos.
- Usa LOBs (Large Objects) en la base de datos.

Capítulo 2
Recuperación de información

2.1. Introducción

En este capítulo se da un panorama de lo que son los sistemas de recuperación de información, así como parámetros involucrados, considerando el mundo social y tecnológico, ya que dichos sistemas son usados para almacenar elementos de información que necesitan ser procesados, buscados, actualizados y diseminados a varias poblaciones de usuarios.

La recuperación de información es un estudio interdisciplinario, donde disciplinas como la lingüística, documentación, estadística e informática trabajan en conjunto en busca de información.

2.2. Recuperación de información

En 1986, Van Rijsbergen sugirió un modelo de un sistema de recuperación de información basado en lógica, porque el uso de una lógica adecuada proporciona todos los conceptos necesarios para representar las diversas funciones de un sistema IR (*Information Retrieval*). En un modelo IR lógico, el contenido de información de un documento es representado por una fórmula *d* y la información necesaria, según lo expresado en la pregunta, es representada por una fórmula *q*. La "*verdad*" de $d \rightarrow q$ significaría que la fórmula de la pregunta se puede deducir la fórmula del documento, es decir, la información capturada por *d* es suficiente para deducir la información representada por *q*.

La lógica es una formalización de la manera en que se utiliza la información en la vida diaria para pensar, deducir, concluir, adquirir conocimiento, tomar decisiones y así sucesivamente. En este sentido, la lógica pretende modelar *información* y el *flujo*, por ejemplo, en sistemas IR textuales, en el flujo permite leer (reconocimiento de letras, palabras, y oraciones); el flujo permite entender que se está leyendo (semántica); y el flujo permite que, con respecto al conocimiento del tema, se derive información adicional de lo que se ha leído (pragmática).

La *teoría de situación* representa la información sin indicar específicamente cuál es la información. Considera la información como entidad fundamental. La teoría de la situación se refiere a la información de la forma:

Una propiedad P pertenece / no pertenece a un conjunto de objetos $a_1,\ldots a_n$

2.2.1. Señal

La información es conocimiento sobre una fuente; es comunicado por una señal a un receptor. En IR, la fuente es el documento y el receptor es cualquiera que observa el documento (leyendo un texto, escuchando una cinta magnética para audio u observando una imagen). La señal es cualquier medio por cuál se entrega la información sobre una fuente al receptor. Por ejemplo, si el documento es un texto, la señal es una mezcla de la capacidad de la visión del lector o la comprensión de la información leída, y de su conocimiento general sobre su tema. Una señal puede también ser el proceso de la indización de direcciones que entrega una representación del contenido de información del documento.

2.2.2. Percepción

La cantidad de información contenida en documentos es generalmente muy grande, se podría comparar el contenido de información (físico) del documento con una experiencia sensorial, que incorpora a menudo una gran variedad de detalles. La *percepción* de una experiencia sensorial es el proceso por el cual la información es entregada a un agente cognoscitivo para su uso selectivo. Se identifica con una señal que lleve la información sobre una fuente, la cual es codificada en forma analógica. Hasta que la información se ha extraído de esta señal, nada que correspondía al reconocimiento, a la clasificación o a la identificación ha ocurrido. Es la conversión acertada de la información en la forma digital apropiada que constituye la esencia de la actividad cognoscitiva.

Un proceso de percepción incorpora regularmente la información sobre una variedad de detalles que, si estuvieron transportados en total al agente cognoscitivo, requerirían capacidades inmensas del almacenaje y de recuperación. Por otra parte, hay más información que puede ser extraída y/o ser explotada por el agente cognoscitivo.

Un proceso de percepción es determinado no por qué información lleva, sino por el camino que la lleva. El ver, el oír o el leer no son procesos distintos por la información que llevan (la información puede ser la misma), sino debido al medio por el cual esta información es entregada. Dos conceptos diversos están implicados aquí: *cómo* se entrega la información y *qué* información representa. La teoría de la situación se refiere al último, porque una situación puede ser un texto, una imagen o un discurso. Por lo tanto, un modelo basado en teoría de la situación podía incorporar eventualmente sistemas multimedia de IR.

2.2.3. Cognición

Se describe la cognición como la conversión de la información que un agente cognoscitivo recibe en forma analógica a forma digital. La información que resulta es usualmente calificada como conocimiento con respecto al agente cognoscitivo. La conversión, referida por Dretske como *digitalización*, implica una pérdida de información porque se transforma de una estructura de mayor contenido de información a una de menor contenido de información.

El proceso de la indexación en IR se puede comparar con un proceso de digitalización. El sistema IR es el agente cognoscitivo, el documento es una situación que contiene la información en forma analógica, la información que es (exitosamente) digitalizada constituye la representación del documento. La meta es reducir al mínimo la pérdida de información implicada en la conversión mientras que al mismo tiempo se obtiene la suficiente representación del documento para su almacenamiento y rápida recuperación. El resultado del proceso de cognición es a menudo una representación parcial de la información obtenida de la fuente, que es un documento en IR.

2.2.4. Información contra significado

La información y el significado son dos diversos conceptos. De hecho, no hay razón de asumir que la información que una señal lleva es idéntica a su significado, pero regularmente, la información contenida en una señal *excede* su significado.

2.2.5. Flujo de la información

Durante el proceso de la indización en IR, la información que es digitalizada constituye la representación inicial del documento, la cual no es una descripción exhaustiva del contenido de información de ese documento. Esta representación muestra regularmente el contenido *explícito* de información del documento, el cual es determinado en la mayoría del significado de las oraciones del documento. La información adicional puede ser regularmente identificada como parte del contenido de información del documento, esta información adicional constituye la información implícita del contenido del documento, esta es cargada en forma analógica y permanece hasta ser extraída (o digitalizada) por el proceso de la indización. Después de esto, es cargada en forma digital, es decir, como parte del contenido de información (conocido) del documento. La existencia de esta información es debido al flujo de información que se presenta en el contenido explícito de la información del documento.

El propósito de un sistema IR es proporcionar la información acerca de una pregunta, la cual es una representación de una necesidad de información que un sistema IR procura satisfacer. Por lo tanto, la determinación de importancia consiste en el procesar la información contenida en un objeto (por ejemplo, un documento) sobre otro (por ejemplo, una pregunta). Este problema es la identificación de un flujo de la información entre el objeto del documento y el objeto de la pregunta.

La información contenida es descrita generalmente por relaciones entre los elementos de la información y el objeto afectado de esta información.

2.2.6. Procesamiento de lenguaje natural

El uso de la teoría de la situación proporciona una ventaja adicional porque ha sido utilizado en el desarrollo de un marco para el proceso de lenguaje natural. Este marco se llama *Situation Semantics*, este modela la elocución de una oración con tres entidades: el tipo que representa el contenido de información de la oración, la situación que la oración describe, y la situación en la cual se pronuncia la oración. Todos los componentes de una oración se definen en términos de estos tres tipos de entidades, que se combinan para formar las tres entidades de la oración. Un modelo basado en teoría de la situación puede utilizar la semántica de la situación como el proceso del lenguaje natural para identificar los tipos que son apoyados por la situación que modela el documento.

2.2.7. Sistema de recuperación de información

Hay cuatro elementos clave para diseñar un sistema con la organización y la recuperación de hechos en los dominios relativamente libres (por ejemplo, eventos actuales, información bibliográfica, resúmenes científicos):

- Primero, el sistema debe poder entender automáticamente el texto en lenguaje natural (entrada a la base de datos y a las preguntas al sistema).
- En segundo lugar, la información en el sistema debe estar formateada y organizada (automáticamente) de una manera tal que el contenido o el significado conceptual de la entrada se pueda utilizar para la recuperación. Las representaciones producidas por el programa de análisis deben ser adecuadas para esta tarea y se deben integrar automáticamente en la organización de la memoria.

- Tercero, el sistema necesita las reglas para tener acceso a la base de datos, para el ingreso de nueva información y recuperación. Esas reglas del acceso necesitarán utilizar conocimiento sobre la información de los dominios.
- El cuarto elemento tratará el grado y la organización de ese conocimiento.

El sistema se debe basar en una teoría de análisis, de la inferencia y de la organización del contenido que se pueda ampliar a otros dominios. Algunos de los elementos que la teoría tendrá que tratar son los siguientes:

- Análisis automático del texto de lenguaje natural (para la pregunta que contesta y agrega nuevos datos a la base de datos).
- Organización del contenido de ese texto según su contenido conceptual.
- Acceso a la base de datos para ingreso y recuperación.
- Organización y extensión del conocimiento sobre el dominio necesario para entender, actualizar y recuperar.
- Actualización automática de la base de datos.
- Creación automática de nuevas categorías.
- Construcción automática de la llave de búsqueda del texto de lenguaje natural.
- Estrategias para buscar en la base de datos.

Estos elementos caen en cuatro categorías importantes:

- Organización de los datos.
- Texto de entrada entendible.
- Actualización automática de la base de datos.
- Respuesta a preguntas.

2.2.8. Organización de los datos

En un sistema de recuperación la organización debe de poder hacer accesible el contenido para la recuperación y la inferencia.

Los datos deben de estar organizados tal que cualquier factor puede ser extraído de la base de datos.

2.2.9. Análisis de texto

Extraer el significado del texto en leguaje natural es un obstáculo para construir un sistema de recuperación y organización.

El analizador de texto en un sistema de recuperación de información debe poder ampliar automáticamente la base de datos. Para hacer esto, debe emplear un sistema que 'no caiga aparte' cuando encuentre palabras nuevas o desconocidas. Es decir, el sistema debe poder procesar cualquier nuevo acontecimiento sin la intervención humana

2.2.10. Actualización automática de la base de datos

Después de ser formateado, un nuevo elemento de datos debe ser agregado a la base de datos. El sistema debe poder agregar automáticamente cada elemento.

En sistemas IR tradicionales las categorías para las nuevas entradas son elegidas extrayendo las palabras contenidas de la entrada y haciendo un análisis estadístico para considerar qué categoría tiene la mayoría de las palabras claves referentes a él, los nuevos datos entonces se ponen en esa categoría. Hay un número de problemas importantes con este esquema. Primero, el análisis estadístico de las palabras claves no dará lugar siempre a una categoría correcta que es elegida. En segundo lugar, algunas entradas pudieran ser relevantes en más de una categoría, y asignarla solamente a una disminuirá la oportunidad de ser recuperable.

2.2.11. Respuestas a preguntas

En un sistema de recuperación de datos se debe poder formular preguntas en lenguaje natural, y el sistema debe poder analizar las preguntas, extraer sus significados, para crear automáticamente llaves de búsqueda de la entrada de lenguaje natural, recuperar y generar respuestas apropiadas. Este proceso que entiende requiere inferencia, un almacén grande del conocimiento sobre los dominios que se entenderán y una comprensión de la estructura de la base de datos (las mismas clases de conocimiento necesarias para el mantenimiento de la clasificación y de la organización).

2.3. La estructura del documento

Los documentos, aquellos escritos mediante los que tiene lugar la comunicación formal de los resultados alcanzados en una investigación, han adquirido una configuración estable, tras un proceso de evolución relativamente rápido.

El análisis bibliométrico de la literatura científica tiene como fundamento que la presencia del documento en el proceso científico no es algo contingente; pero depende de un modo más inmediato de la constancia de los rasgos de la literatura que pueden tomarse como base de información estadísticamente significativa. Varias de las partes que constituyen la estructura del documento han recibido una considerable atención en los análisis bibliométricos. Atienden a una muestra de esas partes, que categorizar del siguiente modo: marcas de identificación (que incluyen título, lista de autores, afiliación institucional, agradecimientos, palabras clave y resumen); cuerpo del texto (se ocupa de los siguientes aspecto: formato, estilo discursivo y uso léxico).

Todas las partes del documento tienen una función informativa más o menos explícita, pero habitualmente clara. A continuación se comentarán las diferentes partes que conforman la estructura típica.

- Titulo. Frase que encabeza y presenta el trabajo mediante una descripción breve de su contenido. Sitúa el trabajo en un foco temático, ayuda a clasificarlo de algún modo.
- Lista de autores. La mención explícita de autoría es una marca muy significativa para determinar interés por el documento. La lista de nombre puede ser, en especial con investigadores afamados, un indicativo del calibre del trabajo.
- Afiliación institucional de los autores. Hacer constar el lugar de trabajo es indicativo de una dedicación profesional.
- Resumen. Facilita una comunicación rápida de los puntos esenciales del contenido del documento. Suele describir el problema y enumerar los resultados más destacados que se reivindican en el texto principal.
- Palabras clave. Pueden ser libres o de vocabulario controlado. Su objetivo es destacar los puntos por los que el trabajo se conecta con la investigación de su disciplina, por lo que resaltan la relevancia. Pueden ser otorgadas por indexadores o especificadas por los autores, pero en general puede considerarse mas una herramienta usada por el órgano emisor del servicio de consulta que por el propio autor.

2.4. El acopio de datos

Las bases de datos constituyen la principal fuente de acopio para la construcción de los datos empíricos de los estudios bibliométricos. Los datos consistentemente disponibles en las bases de datos incluyen información bibliográfica básica: nombre del autor, título, paginación. Esta clase de información es la utilizada con mayor frecuencia a la hora de investigaciones bibliométricas.

Las bases de datos bibliográficas tienen limitaciones, algunas de ellas son las siguientes:

- Variaciones en la cobertura
- Discrepancias en la forma de asentar a los autores
- Errores en la captura de los elementos bibliográficos
- Inconsistencia en la calidad
- Información no siempre disponible en campos específicos de búsqueda.

2.5. Indicadores bibliométricos de producción

Los indicadores bibliométricos de producción son medidas, basadas habitualmente en recuentros de publicaciones, las publicaciones que se tienen en

cuenta son documentos persistentes, lo cual asegura su adecuación a determinadas características formales y de contenido.

2.5.1. Los recuentos de publicaciones

Contar el número de trabajos científicos que ha publicado un determinado laboratorio, región o país es, seguramente, el procedimiento más accesible para obtener una cuantificación con una base objetiva que describa la actividad o el peso de esos agregados. La relativa sencillez de la obtención de recuentos, especialmente desde la consolidación de bases de datos que recogen literatura científica, y la engañosa objetividad de sus cifras han favorecido la proliferación de estudios de variada índole.

La validez de los recuentos pueden cuestionar varios problemas, en primer lugar, debe abordarse una cuestión básica: ¿de que modo cabe entender el resultado de una recuento, cuál es su *significado*? Si los recuentos carecieran de él, es obvio que no podrían ser empleados en el cálculo de ningún tipo de indicador. Algo menos conceptual que lo anterior se plantea en torno al problema de la *atribución* de resultados. La *métrica* de los recuentos es la tercera cuestión, se refiere a la naturaleza de las unidades empleadas en éstos: ¿son o no lineales? ¿y constantes?.

La atribución de los resultados a sus agentes productores encuentra su fundamento en el sistema de publicaciones, y puede ser problemática cuando se presenta una colaboración ente autores. Existen varios procedimientos para asignar resultados que pueden aplicarse tanto a nivel investigador individual como a nivel colaboración. A nivel de individuo: ¿cómo se puede asignar crédito a casa uno de los coautores? la alternativa básica es fraccionar o no el resultado. La variedad se produce en los modos posibles de fraccionamiento, ya que son concebibles diferentes maneras de repartir el resultado entre sus coautores.

2.6. Modelos matemáticos

Los motores de búsqueda del Web como Google o AltaVista proveen a sus usuarios un campo simple del texto para incorporar algunas palabras, responden proporcionando una lista según su importancia de referencias a las páginas Web que son esperanzadamente relevantes al usuario. La manera que se produce esta graduación se puede modelar razonablemente bien con un lenguaje de modelado, incluyendo por ejemplo el *page rank* (o importancia) de Google usando estadística sobre hiperligas [20]

A menudo, la importancia de una palabra para el *rank* final del documento es reflejada por su distribución de frecuencia en la colección.

Para cada documento *D* el lenguaje define la probabilidad $P(T_1,\ldots,T_n|D)$ de una secuencia de los *n* términos de la pregunta $T_1,\ldots,T_n$ y los documentos son alineados por esa probabilidad. La tarea esencial es reevaluar las probabilidades, para asignar una cierta probabilidad diferente a cero a los términos de la pregunta que no ocurren en un documento. Como tal, la valoración de la probabilidad es una alternativa para la valoración de la toda probabilidad.

La probabilidad de la importancia de un término $\lambda\iota = 0$ implica la certeza que el término no es importante. Las palabras de la consulta que son ignoradas durante la búsqueda son usualmente referidas como *stop words* o palabras sin significado, palabras que pueden ser ignoradas porque ocurren muy frecuentemente en la colección, palabras frecuentes no pueden contribuir significativamente en la cuenta final del documento.

Así, $\lambda_\iota = 1$ asegura que todos los documentos recuperados contengan el ι-*esimo* término de la pregunta. Esto hace sospechar que para los valores cerca de 1, el *rank* esté en un nivel de coordinación. La coordinación del *rank* es un *rank* parcial de los documentos tales que los documentos que contienen términos de la pregunta de *n* están alineados siempre sobre los documentos que contienen n-1 términos de la pregunta.

A. Pritchard en 1969 empleó el término "Bibliometrics" (Bibliometría) por primera vez, para denotar una disciplina que la define como "la aplicación de los métodos matemáticos y estadísticos a los libros y otros medios de comunicación"

Para determinar el comportamiento de las tendencias y regularidades de la producción y comunicación científica se utiliza una gama muy variada de métodos y modelos matemáticos y estadísticos que se aplican a la información científica y técnica.

Las denominadas "leyes" clásicas de la bibliometría son distribuciones estadísticas de tipo hiperbólicas. Su aplicación se caracteriza por la utilización de series cronológicas retrospectivas, cuyo comportamiento se representa o formula a partir de modelos matemáticos que, bajo determinadas circunstancias unas veces condicionadas por el nivel de especialización de área temática a estudiar, otras por el tamaño o cobertura cronológica de la muestra, alcanzan cierta verificación y comprueban el planteamiento de sus postulados teóricos y gráficos.

Los modelos matemáticos de Lotka, Bradford y Zipf constituyen, en buen juicio, la columna vertebral de la bibliometría, debido a que identifican el comportamiento de tres de las principales regularidades cuantitativas presentes en el flujo de la información documental, tales como la productividad de autores científicos, la concentración y dispersión de la información por fuentes y el

volumen de los textos a partir de las frecuencias y rango de las palabras en el mismo, respectivamente. [12]

En las tablas 2.1-2.3 se muestran los modelos matemáticos de estudios métricos de información en la especialidad métrica de bibliométria y regularidad de producción científica.

Nombre del modelo	**Formulación matemática**	**Resultado obtenido**
Modelo matemático de Lotka	$Y_{(n)} = C / n^2$ $Y_{(n)}$ = cantidad de autores que producen n documentos c = constante para cada temática n^2 = cuadrado de la frecuencia de los autores	Núcleo de autores más productivos en temas específicos
Modelo de elitismo de Price	$E = (N)^{½}$ E = élite de autores que publican el 50% de los trabajos N = población total de autores	Identifica la élite de autores más productiva

Tabla 2.1. Aspecto o regularidad que identifica productividad científica de autores

Nombre del modelo	**Formulación matemática**	**Resultado obtenido**
Índice de autoria	Ic = Caf/Cd Caf = cantidad de autores firmantes Cd = Cantidad de documentos	Media de autores por documentos
Tasa de documentos con coautorados	Tdc = Cta/Ctd Cta = cantidad total de documentos con autoría múltiple. Ctd = cantidad total de documentos	Proporción de documentos con autoría múltiple
Índice de colaboración de lawani	$IC = \Sigma_{i=1}^{N} j_i n_j / N$ N = total de documentos j_i = número de firmas por documento n_j = cantidad de documentos con autoría múltiple	Peso promedio del número de autores por documento

Tabla 2.2.. Aspecto o regularidad que identifica autoría y colaboración entre autores

Nombre del modelo	**Formulación matemática**	**Resultado obtenido**
Grado de colaboración de Subrayaman	$GC = N_m / N_m + N_s$ N_m = total de documentos con autoría múltiple N_s = total de documentos escritos por un solo autor	Proporción de documentos con autoría múltiple idem a la tasa de documentos coautorados
Coeficiente de colaboración de Ajiferuke, Burrel Y Tague	$CC = 1 = \Sigma_{i=1}^{N} [(1/j_i)n_{ji}]/N$ N_{ji} = cantidad de documentos j para cada frecuencia de firma i	Nivel de colaboración en los documentos, atendiendo a sus relaciones de autoría

Tabla 2.2. (Continuación) Aspecto o regularidad que identifica autoría y colaboración entre autores

Nombre del modelo	**Formulación matemática**	**Resultado obtenido**
Modelo matemático de Zipf	$(r)(f) = C$ C = constante extraída del principio del menor esfuerzo r = posición de la palabra en el texto f = frecuencia de aparición de la palabra en le texto.	Frecuencia de aparición de palabras y relación entre las palabra y su posición en el texto
Razón vocabulario-palabra (Chen y lemkuhler)	$V/p + \ln V/\ln p = 1$ V = cantidad de palbras distintas en el texto P = total de palabras en el texto Ln = logaritmo natural	Relación entre el vocabulario y las palabras usadas en un texto

Tabla 2.3. Aspecto o regularidad que identifica: Antecedentes y nexos lingüísticos y de contenido interdisciplinario de los documentos.

Las tablas anteriores dan una idea de los modelos que se pueden aplicar para la obtención de indicadores bibliométricos.

Muchos de los modelos presentados se aplican bajo la autoría del documento y considerando que el campo de estudio es la base da datos de tesis de la UNAM, entonces, incluso antes de cualquier estudio, se puede decir que a lo más un autor publica 3 tres tesis de grado, licenciatura, maestría y doctorado, lo cual no sería un dato relevante, sin embargo, en vez de aplicar los modelos a la

autoría bien pueden ser aplicados a la dirección de la tesis, es decir en vez de hacer el estudio sobre los autores se hace sobre los directores de tesis, lo cual arrojaría resultados relevantes al estudio bibliométrico.

En el análisis de la productividad científica confluye un conjunto de regularidades que contribuyen a identificar no sólo la proporción cuantitativa de producción de literatura científica de los autores, enfoque del cual partieron Lotka y Price para proponer dos de los modelos matemáticos considerados hoy clásicos de la bibliometría, sino también dentro de la regularidad se hallan otras características de la productividad más asociadas con los problemas sociales y que entrañan la forma o estructura de la población que produce o publica.

El estudio de la autoría múltiple en documentos científicos, suele ser empleado para analizar relaciones de colaboración entre autores, instituciones y colectivos científicos, pero para el estudio de la colaboración también se requieren otras condicionantes que no siempre se dan en las relaciones de coautoría.

Así, de los modelos clásicos que utilizan como parámetro la autoría se pueden derivar modelos que hagan uso de otro parámetro o variable de estudio, tal como, la dirección de las tesis, siendo dato relevante la frecuencia de dirección, incluso codirección.

Capítulo 3
Análisis de texto e Indización

3.1. Introducción

Los documentos y las consultas de usuario en la recuperación de información regularmente se encuentran en lenguaje natural, este capítulo describe el procesamiento automático de texto en lenguaje natural y varios niveles de métodos lingüísticos con énfasis en aplicaciones de recuperación de información.

3.2. Lenguaje Natural

El procesamiento de texto en lenguaje natural permite que un sistema explore textos originales para recuperar información particular o para derivar estructuras de conocimiento.

3.2.1. Niveles de procesamiento de lenguaje

- Nivel *fonológico* se ocupa del tratamiento de los sonidos.
- Nivel *morfológico* se refiere al proceso de las formas individuales de la palabra y porciones reconocibles de la palabra. El recogimiento, la eliminación de sufijos y prefijos de la palabra se basan en este nivel.
- Nivel *léxico* se ocupa de los procedimientos que operan en palabras completas. En la recuperación de información esto cubre operación como cancelación de palabras comunes, aplicación de diccionario en palabras individuales, y el reemplazo de palabras por clases de tesauro. En aplicaciones de análisis sintácticos donde se hace una tentativa de obtener una descripción estructural de una oración, una operación léxica preliminar identifica normalmente un sistema de características lingüísticas (por ejemplo, sustantivos, los adjetivos, preposición, etc.) para cada palabra del texto que se utilizará más adelante en el proceso sintáctico principal del análisis.
- Nivel *sintáctico* está diseñado para agrupar las palabras de una sentencia en unidades estructurales tal como frases preposicionales, y agrupaciones sujeto-verbo-objeto que representan colectivamente la estructura gramatical de la oración. Un análisis sintáctico se basa normalmente en la estructura

circundante en la cual la palabra individual encaja en una oración y en el uso de las características sintácticas que caracterizan la palabra individual.
- Nivel *semántico* agrega conocimiento del contexto al proceso puramente sintáctico para reestructurar el texto en unidades de significado real.
- Nivel *pragmático* usa información adicional acerca del entorno social en el cual un documento dado existe y acerca de las relaciones que normalmente prevalecen entre varias entidades. [3]

Las operaciones del nivel léxico serán examinadas, en el punto siguiente de este capítulo, como parte de las operaciones de indización automática.

3.2.2. Dificultades en el procesamiento de leguaje natural

- A nivel léxico una palabra puede tener varios significados, y la selección del significado apropiado se debe deducir a partir del contexto, en el caso del presente trabajo mucho ayuda la clasificación del área temática.
- A nivel estructural se requiere de la semántica para eliminar la ambigüedad de la dependencia.
- A nivel pragmática, a menudo, no significa lo que realmente se está diciendo.

Muchos sistemas gramaticales incluyen operaciones basadas en notación de autómata finito representado como grafo. Una gráfica de estados finitos consiste de nodos y ramos entre paredes de nodos. Cada nodo simboliza un estado de la máquina, y las ramas representan las transiciones de un estado a otro.

Entre las principales aplicaciones del procesamiento del leguaje natural están la recuperación y extracción de información. La primera busca información en documentos, información de documentos que describan los documentos, tal como se vio en el capítulo previo, la segunda es una especie de recuperación de información cuyo objetivo es extraer automáticamente información estructurada o semiestructurada, una de sus tareas es el reconocimiento de nombre o expresiones numéricas.

3.3. Indización

Durante la indización los registros son preparados para ser usados por los sistemas de recuperación de información.

Esto es, la preparación de la colección de registros en una representación fácilmente accesible. Preparar un registro en índices involucra el uso de:

- Un conjunto de expresiones regulares
- Análisis (parsers)
- Una lista de palabras no significativas (stop words)
- Algunos otros filtros

Uno de los procedimientos usados en el entorno del procesamiento de registros más difíciles de realizar es el *análisis*. En principio, el proceso de análisis es redundante si la colección de registros en pequeña. En el entorno de bibliotecas, las operaciones de análisis son conocidas indistintamente como catalogación, clasificación, indización y abstracción. Sin embargo en el presente trabajo dichos términos son utilizados como etapas de un proceso, la catalogación la realiza el personal bibliotecario bajo la interfaz del sistema ALEPH y cuyos registros son almacenados en tablas de una base de datos bajo los estándares de catalogación.

La tarea de indización consiste primero en asignar a los términos que representan a cada registro, y en segundo lugar asignar a cada término el peso, o el valor, reflejando su importancia presumida para los propósitos de la identificación de contenido.

La indización comienza con la observación de la frecuencia de ocurrencia de un tipo de palabra individual (esto es, de las distintas palabras) en lenguaje natural del texto.

Así, preparar un registro se realiza normalmente en varios pasos: eliminación de formato (*markup & format removal),* extracción de palabras *tokenization,* filtración (*filtration),* determinación de la raíz de la palabra *(stemming)* y determinación de pesos (*weighting)*.

3.3.1. Eliminación de formato (markup)

Durante esta fase, todas las marcas y formatos especiales son eliminadas del registro, tal como las etiquetas HTML.

En la actualidad se puede observar, en el ámbito bibliotecológico mundial, un alto incremento en la generación de bases de datos de contenido bibliográfico. Las de mayor uso y prestigio son aquellas en la que se contemplan criterios de almacenar e intercambiar su información bajo normas de carácter nacional o internacional. Entre ellas, se pueden señalar las bases de datos de la Biblioteca de Congreso de los Estados Unidos de América, de la Biblioteca Nacional Británica y de la Biblioteca Nacional de Canadá, entre otras.

Con base en la documentación analizada se puede señalar que para almacenamiento e intercambio de registros bibliográficos los formatos que se utilizan con frecuencia son: USMARC, UNIMARC, INTERMARC, el de referencias del UNISIST, el de la CEPAL y el CCF. [11]

La identificación de un modelo bibliográfico obedece a los propósitos de contar con una instrumento de investigación en el terreno del control bibliográfico, a fin de utilizarlo tanto para la explicación de fenómenos propios de éste, como para el desarrollo de estructuras y sistemas aplicados a la solución de problemas vinculados con el control bibliográfico.

Se pretende que el modelo facilite el manejo de los distintos niveles de significación implicados en el uso de formatos bibliográficos para la automatización de información bibliográfica, haciendo más comprensible su aplicación y, por lo tanto, puede servir como instrumento de orientación en el diseño de bases de datos y registros bibliográficos de intercambio.

Para el caso del presente trabajo el formato que se maneja es un estándar de catalogación, donde los registros se encuentran almacenados en una tabla del esquema del sistema ALEPH.

3.3.2. Extracción de palabras

El primer obstáculo en la extracción de palabras (*tokenization)* está en la existencia de incertidumbre en la noción de *palabra/token* y en el reconocimiento de palabras y/o tokens en el contexto, por que la misma sentencia en un contexto diferente puede ser tokenizada diferente, ya que esta puede o no tener un significado gramático u ortográfico. Para fines prácticos, en el presente trabajo palabra y token serán manejados indistintamente.

Los leguajes son clasificados en dos grupos: segmentado y no segmentados. En un lenguaje segmentado, tal como el español, los símbolos de puntuación y secuencias de caracteres pueden ser palabras del diccionario [18]. Por ejemplo para español:

1. Tipo de lenguaje: segmentado
2. Delimitadores: espacios en blanco, tabulador, y saltos de línea
3. símbolos de puntuación: [.][,][:][;]['][“]…[0][1][2]…
4. símbolos que no representan información y que podrían descartase, como el caso de los símbolos [¿],[¡],[%],[&]

Normalmente, todos los símbolos de puntuación y los caracteres alfanuméricos extraños son eliminados en esta fase.

Básicamente cada carácter es procesado, como lo muestra la tabla 3.1, para aplicar el proceso de extracción de palabras.

Sentencia	Modelo sistemático de la atención medica familiar en el consultorio No.5 de la Clínica-Hospital Dr. Miguel Trejo Ochoa Colima, Colima.
Caracteres	|M|o|d|e|l|o| |s|i|s|t|e|m|á|t|i|c|o| |d|e| |l|a| |a|t|e|n|c|i|ó|n| |m|e|d|i|c|a| |f|a|m|i|l|i|a|r| |e|n| |e|l| |c|o|n|s|u|l|t|o|r|i|o| |N|o|.|5| |d|e| |l|a| |C|l|í|n|i|c|a|-|H|o|s|p|i|t|a|l| |D|r|.| |M|i|g|u|e|l| |T|r|e|j|o| |O|c|h|o|a| |C|o|l|i|m|a|,| |C|o|l|i|m|a|.|
Palabras gráficas	|Modelo| |sistemático| |de| |la| |atención| |medica| |familiar| |en| |el| |consultorio| |No|.|5| |de| |la| |Clínica-Hospital| |Dr|.| |Miguel| |Trejo| |Ochoa| |Colima|,| |Colima|.|

Tabla 3.1. Tokenización

Cabe señalar que la definición depende de la naturaleza de la información y análisis previo de ella. Por ejemplo, el definir cual o cuales serán los caracteres que fungen como delimitadores. Inicialmente se consideraba el guión "-", adicional al espacio en blanco, salto de línea y tabulador, resultado de un análisis sobre una pequeña muestra de información, sin embargo, y como se ha mencionado, esto no sería del todo factible. Como se muestra en los ejemplos siguientes, del *a* al *d* si se tomara como delimitar el guión las palabras resultantes de cierta forma aportan significado, por el contrario para los ejemplos del *e* al *g*, los cuales nos arrojarían tokens sin valor semántico. Es por ello que para el caso de este trabajo el guión "-" no se toma como delimitador ni como un carácter sin significado, siempre y cuando no esté al principio o al final del token.

Ejemplos

a. Características de la migración interna en la región noroeste de México estudio de caso (Los municipios costeros) 1970-1990
b. Aproximaciones teóricas sobre la problemática urbano-ambiental
c. La demografía como uno de los factores determinantes del desempleo en México y su repercusión en las relaciones obrero-patronales (1988-1996)
d. Correlación clínico-endoscópica en niños con ulcera peptica :
e. Caracterización de propiedades al desgaste de una aleación electrolítica Ni-Co-B
f. Comparación de los efectos entre la reexpresión por Boletín B-10 y la actualización por medio de la Ley del impuesto sobre la renta
g. Estudio del cerámico superconductor La1-xSrx CuO4

3.3.3. Filtrado

El filtrado se refiere al proceso de decidir qué palabras y/o etiquetas (para el caso presentado) serán utilizadas para representar a los registros.

Con frecuencia los términos que no son relevantes para representar el registro (stop words) son removidos. Esto se logra en base a una lista de dichos términos.

Las características fundamentales que se observan de la información documental que se procesa en bases de datos bibliográficas y en registros de intercambio son las siguientes:

Sistemas de clasificación. Determinan el o los temas que se tratan en los documentos que se analizan, se expresan en forma de notaciones representativas de esquemas de clasificación; su función es identificar el tema o los temas que trata la obra, con objeto de construir símbolos unívocos con base en los cuales es factible ubicar físicamente los documentos para su posterior recuperación.

3.3.4. Stemming

Refiere al proceso de reducir palabras a su raíz o tema. Ayuda en la recuperación de mayor número de documentos en sistemas de recuperación. Por ejemplo para las palabras "biblioteca", "bibliotecología" y "bibliotecario" su *stem* es "bibliotec"

3.4. Asignación de pesos

Es la etapa final en el proceso de recuperación de información. Los términos son pesados de acuerdo a un modelo de pesos dado, el cual puede incluir pesos locales o globales o ambos. Si los pesos son locales, los pesos de los términos son normalmente expresados como frecuencia del término, *tf*. Si son usados pesos globales, el peso de un término está dado por valores de frecuencia. El más común (y básico) esquema de pesos es uno en el cual pesos globales y locales son usados, el peso de un término = *ft*IDF.*

3.4.1. Modelos de Pesos

3.4.1.1. Salton Vector Space

Sistemas de recuperación de información asignan pesos a los términos considerando:

1. Información local de registros individuales
2. información global de la colección de registros

En estudios IR, el esquema clásico de pesos es el modelo *Salton Vector Space*, el cual está dado por w_i [30]

$$w_i = tf_i * \log (D / df_i) \qquad \text{Ecuación A}$$

Donde

- tf_i es la frecuencia del término o el número de veces que el término *i* ocurre en el registro.
- dfi es la frecuencia o número de documentos que contienen el término *i*.
- D es el número de registros en la base de datos.
- Log (D/df_i) es conocido como la inversa de la frecuencia (IDF_i), una medida de volumen de información asociada al término *i* con un conjunto de registros.
- El cociente de df_i/D es la probabilidad de recuperar de D un registro que contiene el término *i*.

3.4.1.2. Pesos locales

La ecuación A demuestra que el peso w_i aumenta con tf_i. Esto hace el modelo vulnerable para abusos de repetición de términos.

3.4.1.3. Pesos globales

En la ecuación A el término $log(D/df_i)$ es conocido como la frecuencia inversa del documento, una medida (IDF_i), una medida del volumen de información asociada a un término *i* dentro de un conjunto de documentos. Examinado el cociente *df/D*, es la probabilidad de recuperar de *D* un documento que contiene el término *i*.

Por ejemplo, si en 1000 registros bibliográficos de la base de datos sólo 10 contienen el término *sistema,* el IDF para este término es IDF = log(1000/10) = 2. Sin embargo, si sólo un registro lo contiene IDF = log(1000/1) = 3.

Así, los términos que aparecen en muchos registros (por ejemplo, las *stopwords*) reciben un peso bajo, mientras que términos no comunes los cuales aparecen en pocos registros reciben un peso alto. Los extremos no son recomendados en un trabajo rutinario de recuperación. Los términos con pesos aceptables son los que no son demasiado comunes o demasiado raros. Figura 3.1.

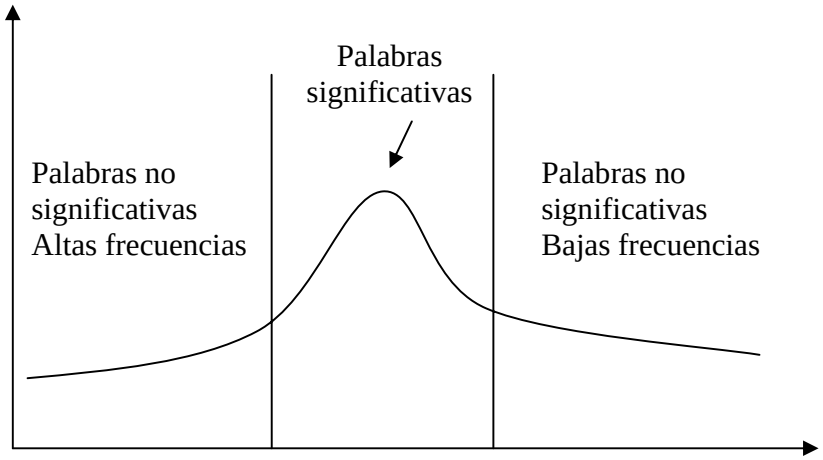


Figura 3.1. Frecuencia de palabras en orden decreciente

Cuando las distintas palabras en el texto son arregladas en orden decreciente según la frecuencia de sus ocurrencias, tabla 3.2, la ocurrencia característica del vocabulario puede ser caracterizada por la ley de *Rank-Frequency:* Frequency * Rank = constant

(Número de palabras N = 1,000,000)

Fila (R)	**Término**	**Frecuencia (F)**	**R*(F/1,000,000)**
1	El	69,971	0.070
2	De	36,411	0.073
3	Y	28,852	0.066
4	A	26,149	0.104
5	Un	23,237	0.116
6	En	21,341	0.128
7	Que	10,595	0.074
8	Es	10,099	0.081
9	Fue	09,816	0.088
10	El	09,543	0.095

Tabla 3.2. Ley de File-Frecuencia (Rank-Frequency)

Usando la ley como punto de partida, es posible derivar los factores de significado basados en las características de frecuencia en las palabras individuales en el texto. Consideraciones:

1. Dada una colección de registros, calcular para cada registro la frecuencia de cada término único en el documento. Esto es la frecuencia del termino *k* en el documento *i*,
   $FREQ_{ik}$

2. Determinar la frecuencia de la colección total $TOTFREQ_k$ para cada palabra sumando las frecuencias de cada término único a través de todos *n* documentos, esto es,

   $TOTFREQ_k = \Sigma_{i=1}^{n} FREQ_{ik}$

3. Arreglar las palabras en orden decreciente de acuerdo a su frecuencia de la colección. Decidir en un cierto valor alto de umbral conveniente y eliminar todas las palabras con una frecuencia de colección sobre el umbral. Tal es el caso de las palabras en la tabla anterior.
4. De la misma forma, eliminar las palabras de frecuencia baja considerable.
5. Las restantes palabras de frecuencia media son las usadas para la asignación del registro como términos índice.

Capítulo 4
Redes Neuronales

4.1. Introducción

Las Redes Neuronales Artificiales (en inglés Artificial Neural Networks ANNs) fueron originalmente una simulación abstracta de los sistemas nerviosos biológicos, formados por un conjunto de unidades llamadas "neuronas" o "nodos" conectadas unas con otras. Estas conexiones tienen una gran semejanza con las dendritas y los axones en los sistemas nerviosos biológicos.

Una primera clasificación de los modelos de ANN podría ser, atendiendo a su similitud con la realidad biológica:

1 Los modelos de tipo biológico. Este comprende las redes que tratan de simular los sistemas neuronales biológicos así como las funciones auditivas o algunas funciones básicas de la visión o el habla.
2 El modelo dirigido a aplicación. Estos modelos no tienen porque guardar similitud con los sistemas biológicos. Su arquitectura está fuertemente ligadas a las necesidades de las aplicaciones para las que son diseñados.

4.2. Redes Neuronales de Tipo Biológico

Se estima que el cerebro humano contiene más de cien mil millones (10^{11}) de neuronas y (10^{14}) sinápsis en el sistema nervioso humano. Estudios sobre la anatomía del cerebro humano concluyen que hay más de 1000 sinápsis a la entrada y a la salida de cada neurona. Es importante notar que aunque el tiempo de conmutación de la neurona (unos pocos milisegundos) es casi un millón de veces menor que en los actuales elementos de las computadoras, ellas tienen una conectividad miles de veces superior que las actuales supercomputadoras.

El objetivo principal de las redes neuronales de tipo biológico es desarrollar un elemento sintético para verificar las hipótesis que conciernen a los sistemas biológicos. Las neuronas y las conexiones entre ellas (sinápsis) constituyen la clave para el procesamiento de la información.

La mayor parte de las neuronas poseen una estructura de árbol llamadas dendritas que reciben las señales de entrada que vienen de otras neuronas a través

de las uniones llamadas sinápsis. Algunas neuronas se comunican solo con las cercanas, mientras que otras se conectan con miles. La neurona esta compuesta de:

1 El cuerpo de la neurona
2 Ramas de extensión llamadas dendritas para recibir las entradas
3 Un axón que lleva la salida de la neurona a las dendritas de otras neuronas.

La forma que dos neuronas interactúan no está totalmente conocida, dependiendo además de cada neurona. En general, una neurona envía su salida a otras neuronas por su axón. El axón lleva la información por medio de diferencias de potencial, u ondas de corriente, que depende del potencial de la neurona. Este proceso es a menudo modelado como una regla de propagación representada por la función de red u(.). La neurona recoge las señales por su sinápsis sumando todas las influencias excitadoras e inhibidoras. Si las influencias excitadoras positivas dominan, entonces la neurona da una señal positiva y manda este mensaje a otras neuronas por sus sinápsis de salida. En este sentido la neurona puede ser modelada como una simple función escalón f(.). Como se muestra en la figura 4.1, la neurona se activa si la fuerza combinada de la señal de entrada es superior a un cierto nivel, en el caso general el valor de activación de la neurona viene dado por una función de activación f(.).

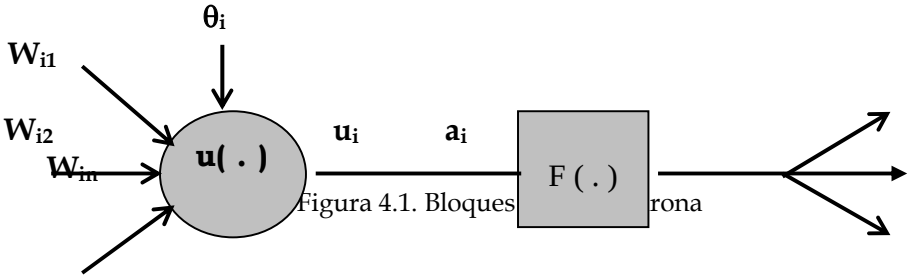


Figura 4.1. Bloques rona

4.3. Red Neuronal Artificial

Una red neuronal es una red de muchos procesos simples cada uno de los cuales tiene memoria local, las unidades están interconectadas por canales de comunicación (conexiones) los cuales lleva un dato numérico a las unidades. Las unidades operan solamente en sus datos locales y en las entradas que reciben de las conexiones con algunas redes vecinas, las ANN son modelos de las redes neuronales biológicas y otras no. Muchas de las redes neuronales surgen del deseo de producir sistemas artificiales capaces de ser semejantes a la inteligencia o entendimiento humano.

Una definición interesante de red neuronal hace uso del concepto matemático de grafo, objeto consistente en un conjunto de nodos (o vértices), más un conjunto de conexiones establecidas entre ellos. En este caso, el grafo describe la arquitectura del sistema y proporciona los canales por los que puede describir su dinámica. Hay diferentes tipos de grafo, por ejemplo, grafo dirigido y no dirigido. En el primer tipo de grafo, las conexiones tienen asignado un sentido, mientras que en el otro son bidireccionales. Existen grafos densos cuando todos los nodos están conectados con casi todos y grafos dispersos cuando son pocas las conexiones entre los nodos. Un grafo puede componerse de diferentes tipos de nodos y diferentes tipos de conexiones.

Una forma de representar el grafo es, como su propio nombre indica, gráficamente, dibujando los nodos como círculos y las conexiones como líneas o flechas, según sean de un solo sentido o bidireccionales. Otra forma común de representación son mediante una matriz de conexiones. En el caso en que el grafo sea no dirigido, la matriz de conexiones será simétrica.

La definición matemática de red neuronal utilizando el concepto de grafo toma la siguiente forma:

1. A cada nodo *i* se asocia una variable de estado Xi
2. A cada conexión (i,j) de los nodos *i* y *j* de asocia un peso wij
3. A cada nodo *i* se asocia un umbral θ j
4. Para cada nodo *i* se define una función f(xj,wij, θ j), que depende de los pesos de sus conexiones, del umbral y de los estados de los nodos *j* a él conectados. Esta función proporciona el nuevo estado del nodo.

En la terminología habitual de las redes neuronales, los nodos son las neuronas y las conexiones son las sinápsis.

4.3.1. Taxonomía de las Redes Neuronales

Existen dos fases en toda aplicación de las redes neuronales: la fase de aprendizaje o entrenamiento y la fase de prueba. En la fase de entrenamiento, se usa un conjunto de datos o patrones de entrenamiento para determinar los pesos (parámetros de diseño) que definen el modelo neuronal. Una vez entrenado este modelo, se usará en la llamada fase de prueba o funcionamiento directo, en la que se procesan los patrones de prueba que constituyen la entrada habitual de la red, analizándose de esta manera las prestaciones definitivas de la red.

1 Fase de Prueba: los parámetros de diseño de la red neuronal se obtienen a partir de unos patrones representativos de las entradas que se denominan patrones de entrenamiento. Los resultados pueden ser tanto calculados de

una vez como adaptados iterativamente, según el tipo de red neuronal, y en función de las ecuaciones dinámicas de prueba. Una vez calculados los pesos de la red, los valores de las neuronas de la última capa, se comparan con la salida deseada para determinar la validez del diseño.

2 Fase de Aprendizaje: una característica de las redes neuronales es su capacidad de aprender. Aprenden por la actualización o cambio de los pesos sinápticos que caracterizan a las conexiones. Los pesos son adaptados de acuerdo a la información extraída de los patrones de entrenamiento nuevos que se van presentando. Normalmente, los pesos óptimos se obtienen optimizando (minimizando o maximizando) alguna "función de energía".

Las aplicaciones del mundo real deben acometer dos tipos diferentes de requisitos en el procesado. En un caso, se requiere la prueba en tiempo real pero el entrenamiento ha de realizarse "fuera de línea". En otras ocasiones, se requieren los dos procesos, el de prueba y el de entrenamiento en tiempo real. Estos dos requisitos implican velocidades de proceso muy diferentes, que afectan a los algoritmos y hardware usados. Atendiendo al tipo de entrenamiento las redes neuronales se clasifican como Supervisadas y No Supervisadas.

4.3.2. Aprendizaje

Aprendizaje es un proceso por el cual los parámetros libres de una red neuronal son adaptados a través de un proceso continuo de simulaciones en el ambiente en la cual la red se encuentra. El tipo de aprendizaje es determinado por la manera en la cual cambian los parámetros. La definición de proceso de aprendizaje implica la siguiente secuencia de eventos:

1 La red neuronal es simulada por medio de un ambiente
2 Los cambios que sufre la neurona son resultado de la estimulación
3 La respuesta de la neurona es una nueva forma de ambiente, debido a los cambios internos que han ocurrido en su interior.

4.3.3. Reglas de Aprendizaje

Por reglas de aprendizaje se entiende el procedimiento para modificar los pesos y el factor de bias de una red neuronal, este procedimiento también es llamado algoritmo de entrenamiento. El propósito de las reglas de aprendizaje es entrenar a la red neuronal para realizar una determinada tarea. Existen varios tipos de reglas de aprendizaje para las redes neuronales, las cuales son supervisadas, y no supervisadas. La adaptación o aprendizaje es uno de los puntos focales de la investigación de redes neuronales.

4.3.4. Reglas de Entrenamiento Supervisado

Las redes de entrenamiento supervisado han sido los modelos de redes más desarrolladas desde inicios de las investigaciones de redes neuronales artificiales. Los datos para el entrenamiento están constituidos por varios pares de patrones de entrenamiento de entrada y de salida. El hecho de conocer la salida implica que el entrenamiento se beneficia por la supervisión de un maestro. Dado un nuevo patrón de entrenamiento.

Las ANNs de entrenamiento supervisado constituyen la línea fundamental de desarrollo en este campo. Algunos ejemplos bien conocidos son red perceptrón, ADALINE/MADALINE, y varias redes multicapa. En el entrenamiento supervisado hay dos fases a realizar: prueba y de entrenamiento.

En el entrenamiento Supervisado, los patrones de entrenamiento se dan en forma de pares de entrada/maestro, donde M es el número de pares de entrenamiento. Dependiendo de la naturaleza de la información del maestro, hay dos aproximaciones al entrenamiento supervisado. Uno se basa en la corrección a partir de una decisión y la otra se basa en la optimización de un criterio de costo. De la última, la aproximación del error cuadrático medio es el más importante. Las formulaciones de decisión y aproximación difieren en la información que tienen los maestros y la forma de usarla.

4.3.5. Reglas de Entrenamiento No Supervisado

Para los modelos de entrenamiento No Supervisado, el conjunto de datos de entrenamiento consiste sólo en los patrones de entrada. Por lo tanto, la red es entrenada sin el beneficio de un maestro.

La red aprende a adaptarse basada en las experiencias recogidas de los patrones de entrenamiento anteriores. Este es un esquema típico de un sistema "No Supervisado".

Los pesos y el factor de bias son modificados en respuesta de las entradas a la red solamente, no existen patrones de salida disponibles, de primera vista este tipo de reglas parecen ser imprácticas. Cómo entrenar una red sino sabemos que se quiere realizar. Muchos de esos algoritmos realizan una operación de agrupamiento, estos aprenden de tal forma que pueden clasificar dentro de un número de clases finitas, este tipo de aprendizaje es especialmente útil en aplicaciones de cuantización vectorial.

En el aprendizaje competitivo como su nombre lo indica las salidas de las

redes neuronales compiten entre ellas mismas para que solo una de ellas sea activada, este tipo de aprendizaje sirve para descubrir las características estadísticas, lo cual sirve para clasificar un conjunto de entradas en clases.

4.4. Arquitectura de una Red Neuronal

Comúnmente una neurona, con múltiples entradas puede ser no suficiente. Se pueden necesitar cinco o diez, operando en paralelo, en este caso se le llama arreglo. Un arreglo de redes de *S* neuronas es mostrado en la figura 4.2, note que cada una de las entradas *R* es conectado a cada neurona y la matriz de pesos tiene *S* renglones.

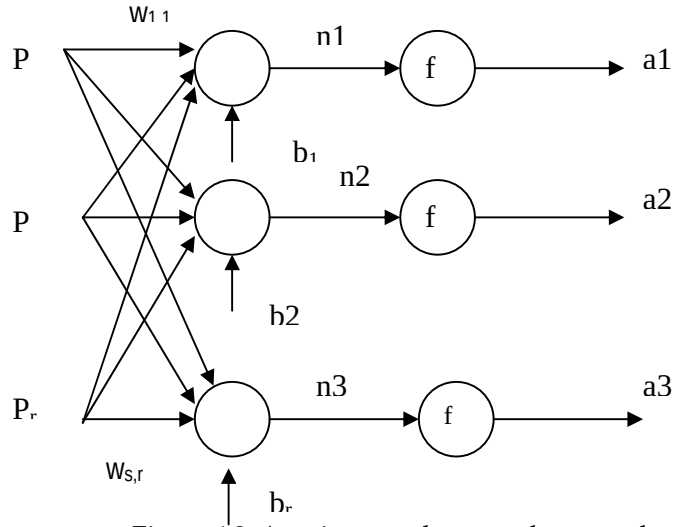


Figura 4.2. Arquitectura de una red neuronal

El arreglo incluye la matriz de pesos, el sumador, el vector de bias, la función de transferencia y el vector de salida a.

Cada elemento del vector de entrada p es conectado a cada neurona a través de la matriz de pesos *W*. Cada neurona tiene un factor de bias, un sumador, una función de transferencia *f* y una salida *a*. Es común que el número de entradas a un arreglo sea diferente del número de neuronas. Además no todas las funciones de transferencia del arreglo deben ser iguales.

La matriz de pesos queda definida como:

$$W = \begin{pmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,r} \\ w_{2,1} & w_{2,2} & \dots & w2,r \\ \dots & & & \\ ws,r & ws,2 & \dots & ws,r \end{pmatrix}$$

Donde los índices de la matrices de los elementos de la matriz indican la neurona destino asociada con su peso, mientras que el índice de la columna indica la entrada fuente, por ejemplo $w_{3,2}$ indica que la conexión es a la tercera neurona de la segunda entrada.

4.5. Representación vectorial

Cada registro, o documento, es representado por un *vector* de términos. Esto es, un documento particular, DOC_i, es identificado por una colección de términos $TERM_{i1}$, $TERM_{i2}$, …, $TERM_{it}$, donde $TERM_{ij}$ está asumiendo la representación del peso, o importancia, del termino j asociado al documento *i*. Por "termino" se entiende a alguna forma de contenido de información, tal como una palabra extraída del texto, una frase o una entrada de un tesauro de términos.

Dada una colección de documentos, pueden ser representados como un arreglo, una matriz, de términos donde cada fila de la matriz representa un documento y cada columna representa la asignación de un término específico de la colección de documentos. Figura 4.3

	$TERM_1$	$TERM_2$	…	$TERM_t$
DOC_1	$TERM_{11}$	$TERM_{12}$	…	$TERM_{1t}$
DOC_2	$TERM_{21}$	$TERM_{22}$	…	$TERM_{2t}$
…				
DOC_n	$TERM_{n1}$	$TERM_{n2}$	….	$TERM_{nt}$

Figura 4.3. Arreglo de términos (n documentos, t términos)

Los pesos positivos del término son elegidos para asignaciones del documento (esto es, TERMij es un número positivo cuando el término *j* ocurre en el documento i); y TERMij es puesto a cero cuando el término *j* no está presente en el documento como un identificador del documento *i*.

Una consulta particular, QUERYj, puede ser similar a un vector QTERMj1, QTERMj2, …, QTERMjt, donde QTERM *jk* representa el peso, o importancia, del termino *k* asociado a la consulta *j*. Así, las recuperaciones de los registros almacenados pueden hacerse dependiendo de la magnitud similar entre vectores de documentos y el vector de la consulta tal como una función de magnitudes. Una medida similar regularmente usada es la medida de coseno definida como,

$$\text{COSINE}(DOC_i, QUERY_j) = \Sigma_{k=1}^{t} (TERM_{ik}\, QTERM_{jk}) / [\Sigma_{k=1}^{t} (TERM_{ik})^2\, \Sigma_{k=1}^{t} (QTERM_{jk})2\,]^{1/2}$$

La correlación del coseno mide el coseno del ángulo entre documentos, o entre consultas y documentos, cuando estos son vistos como el vector en el espacio multidimensional de término de dimensión *t*. En tres dimensiones, cuando sólo tres términos identifican al documento, la situación puede ser representada por la figura 4.4.

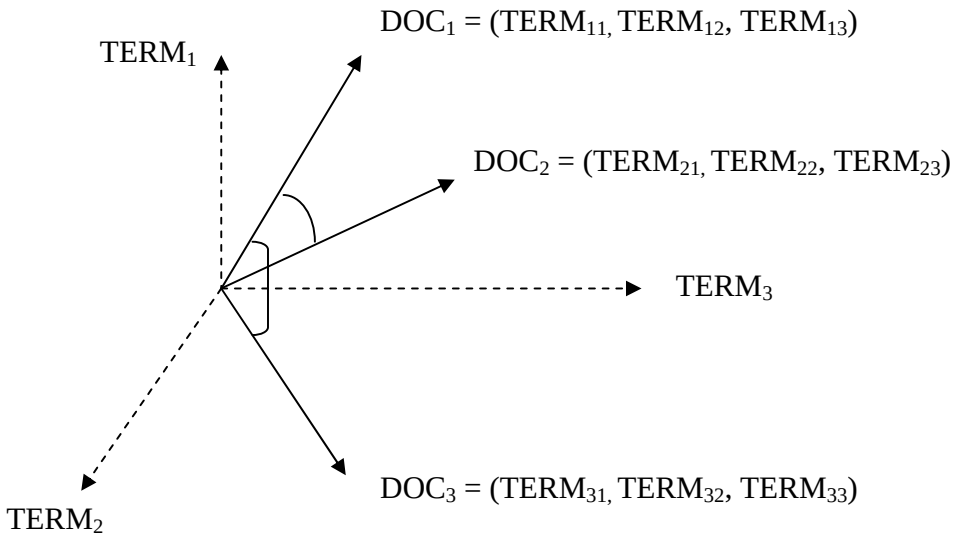


Figura 4. Representación vectorial del espacio de documento

Cada eje corresponde a un diferente término, y la posición de cada vector de documento en el espacio es determinada por la magnitud (peso) de los términos en dicho vector.

La semejanza entre dos vectores se representa como una función inversa relacionada con el ángulo entre ambos. Esto es, cuando dos vectores de documentos son exactamente iguales, el vector correspondiente es superpuesto y el ángulo entre ellos es cero.

El numerador del coeficiente del coseno da la suma de empatar términos entre DOC y QUERY cuando la indexación es binaria, es decir, cuando TERM se asume igual a 1 siempre que el término *k* ocurra realmente en el documento *i*. Cuando la indización no es binaria, el numerador representa la suma de los productos de los pesos del término para los términos de la consulta y del documento. El denominador actúa como un factor de normalización (dividiendo la expresión por el producto de las longitudes de los vectores de la consulta y del

documento). Esto implica que cada elemento es representado por un vector de igual longitud. Si el ángulo entre el vector es pequeño y se utilizan los vectores normalizados, el coseno del ángulo entre los vectores se puede también aproximar por la distancia entre las extremidades de los vectores correspondientes.

Sin embargo, lo anterior puede ser utilizado como método de clasificación de elementos. Dada uno arreglo de datos previamente clasificados, es generado un vector de representación para cada clasificación, así para elementos nuevos es formado un vector que es comparado y el elemento será ingresado dentro de la clase correspondiente.

4.6. Dimensión de la Red

No se pueden dar reglas concretas para determinar el número de neuronas o el número de capas de una red para resolver un problema concreto. Lo mismo ocurre al seleccionar el conjunto de vectores de entrenamiento. En estos casos, lo único que se puede dar son unas cuantas ideas generales deducidas de la experiencia de numerosos autores.

Respecto al número de capas de la red, en general; tres capas son suficientes (entrada-oculta-salida). Sin embargo, hay veces que un problema es más fácil de resolver (la red aprende más deprisa) con más de una capa oculta. El tamaño de las capas, tanto de entrada como de salida, suele venir determinado por la naturaleza de la aplicación. En cambio, decidir cuántas neuronas debe tener la capa oculta no suele ser tan evidente.

El número de neuronas ocultas interviene en la eficiencia de aprendizaje y de generación de la red. No hay ninguna regla que indique el número óptimo, en cada problema se debe ensayar con distintos números de neuronas para organizar la representación interna y escoger el mejor. La idea más utilizada, sobre todo en los sistemas simulados, consiste en tener el menor número posible de neuronas en la capa oculta, porque cada una de ellas supone una mayor carga de procesamiento en el caso de una simulación.

Es posible eliminar neuronas ocultas si la red converge sin problemas, determinando el número final en función del rendimiento global del sistema. Si la red no converge, es posible que sea necesario aumentar este número. Por otro lado, examinando los valores de los pesos de las neuronas ocultas periódicamente en la fase de aprendizaje, se pueden detectar aquellas cuyos pesos cambian muy poco respecto a sus valores iniciales, y reducir por tanto el número de neuronas que apenas participan en el proceso.

4.7. Inicialización y Cambio de Pesos.

Sería ideal, para una rápida adaptación del sistema, inicializar los pesos con una combinación de valores *(W)* muy cercano al punto de mínimo error buscado. Pero es imposible, porque no se conoce a priori dónde está el punto mínimo. Así se parte de un punto cualesquiera del espacio, inicializando los pesos con valores pequeños aleatorios cualesquiera (por ejemplo 0.5), al igual que los términos umbral, que aparecen en las ecuaciones de entrada neta a cada neurona. Este valor umbral se considera como un peso más que está conectado a una neurona ficticia de salida siempre 1. El término umbral es opcional, pues en caso de utilizarse, es tratado exactamente igual que un peso más y participa como tal en el proceso de aprendizaje.

La expresión de la entrada neta a cada neurona se podrá escribir de la siguiente forma:

$$net = \sum_{j=1}^{L} w_{kj} x_{pj} + \theta_k$$

Capítulo 5
Herramienta

5.1. Introducción

En el capítulo 1 se mencionó la importancia del manejador de bases de datos ORACLE, el soporte SQL, su precompilador Pro*C, sus lenguajes procedimentetales principales, PL/SQL y Java, y el soporte de disparadores (o triggers), por fila y por instrucción, además de las vistas; dichas características posibilitan el desarrollo de herramientas de aplicación, tal es el caso del presente trabajo, donde se hace uso de cada una de ellas.

En el presente capítulo se mostraran los proceso para preparar los registros, la forma de indización y recuperación de información para ser presentada e interpretada en el estudio bibliométrico.

En la UNAM es de interés saber cual es la producción temática en las distintas disciplinas, descubrir la corriente de los investigadores, analizar la evolución cronológica de cierta disciplina, en particular de las tesis. Para el presente trabajo la población de estudio son las tesis generadas en la UNAM, registradas y resguardadas por la Dirección General de Bibliotecas. La muestra representativa de las tesis usadas son los registros bibliográficos de las tesis digitales. La propiedad de estudio es la información bibliográfica.

5.2. Requerimientos funcionales

La presente herramienta surge de la necesidad de conocer el comportamiento estadístico de los trabajos recepcionales en la UNAM.

En la actualidad se cuenta con un sistema de administración de tareas de una biblioteca, el sistema ALEPH, sin embargo, y a pesar de que hace uso de una base de datos relacional, dicho sistema no es útil para estudios estadísticos. Debido a ello, el modelado de la herramienta inició con un análisis de la información y recursos existentes, además de indagar sobre las necesidades tangibles que aportarán información.

De lo anterior, se obtienen los casos de uso y requerimientos de la herramienta, que a continuación se mencionan.

Requerimientos funcionales de la herramienta:

- Objetivos

  Desarrollar una herramienta para obtener indicadores bibliométricos y medir factores en la investigación de los trabajos recepcionales que se realizan en la DGB-UNAM.

  Algoritmo general del desarrollo de la herramienta:

  a. Extracción de información bibliográfica de la base de tesis.
  b. Indización de los registros bibliográficos.
  c. Normalización de la información.
  d. Almacenamiento de la información normalizada en tres niveles.
  e. Clasificación automática de los registros bibliográficos de tesis en base a áreas del conocimiento.
  f. Aplicación de modelos matemáticos a la información para la obtención de indicadores bibliométricos.
  g. Presentación de resultados vía Web.

- Características funcionales del sistema

  h. Los datos deberán ser tomados de la tabla Z00 de la base de datos bibliográfica de tesis del sistema ALEPH. Los cuales son:

     - Z00_DOC_NUMBER. Es el número de identificación del registro bibliográfico.
     - Z00_NO_LINES. Es el número de etiquetas de las que está conformado el registro bibliográfico
     - Z00_DATA_LEN. Es la longitud de de la cadena que conforma la información bibliográfica.
     - Z00_DATA. Es la cadena que conforma la información bibliográfica.

  i. Los resultados deberán ser presentados, para su interpretación, mediante una interfaz Web. Los indicadores bibliométricos deberán ser generados en tiempo real, a nivel:
     - Años
     - Asesor
     - Palabra
     - Clasificación

j. La herramienta debe contar con la posibilidad de agregar parámetros para obtener indicadores no considerados inicialmente.

k. Procesar cada registro, en un inicio, todos los que se encuentren en la base y posteriormente cada nuevo registro o cada registro actualizado.

l. Formateo de los registros para su manipulación.

m. Aplicar filtros al registro bibliográfico sobre las etiquetas, para que sean utilizadas aquellas que aporten información para el estudio bibliométrico, las etiquetas pueden variar, entre las cuales están:
   - 245 Título
   - 260 Datos de la publicación
   - 700 Asesor
   - 505 Notas de resumen

n. Aplicar filtros al registro bibliográfico sobre la palabras para descartar aquellas que no aporten información para el estudio bibliométrico, las palabras pueden variar, entre ellas:
   - Las "stop words"

o. Normalizar cada palabra del registro bibliográfico y generación de tokens, considerar:
   - Símbolos de puntuación, acentos
   - Mayúsculas o minúsculas

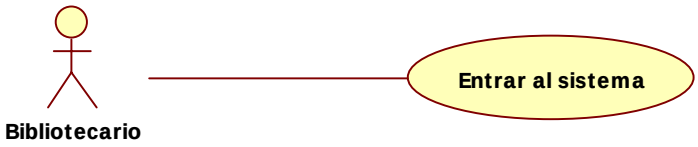
p. La herramienta debe contar con la posibilidad de agregar expresiones regulares que ayuden a la normalización.

q. Para cada nuevo registro aplicar clasificación automática en base a las cuatro área temáticas:
   - Físico matemáticas e ingenierías
   - Ciencias biológicas y de la salud
   - Ciencias sociales
   - Humanidades y Arte

- Casos de uso

El actor de los casos de uso es el bibliotecario que necesita obtener medias cualitativas de la base de datos de registros bibliográficos de tesis producidas por la UNAM.

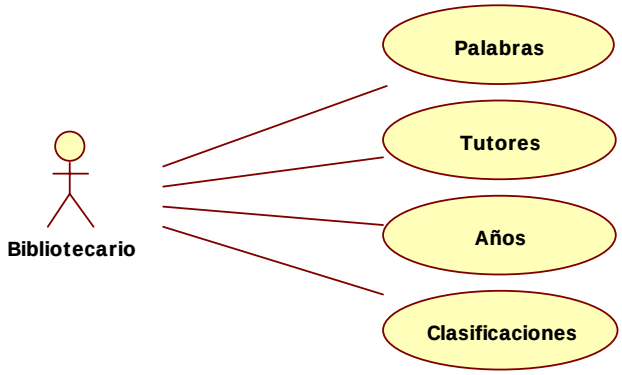
a. Entrar al sistema



	Bibliotecario		**Sistema**	
Entrar al sistema	Ingresa su clave de usuario y contraseña		Muestra la pantalla con el menú de indicadores bibliométricos	E1 E2 E3 E4
E1	No proporciona su clave		Solicita la clave	
E2	No proporciona su contraseña		Solicita la contraseña	
E3	La clave es incorrecta		Solicita verificar sus datos	
E4	La contraseña es incorrecta		Solicita verificar los datos proporcionados	

Tabla 5.1. Caso de uso Entrar al sistema

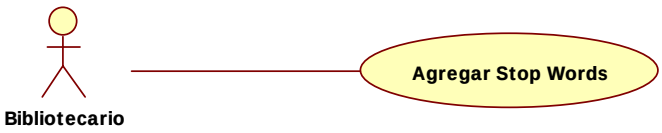
b. Ejecutar indicador bibliométrico. Forma en que el usuario interactúa con el sistema para obtener datos para la interpretación de resultados.



	Bibliotecario		**Sistema**	
Ejecutar indicador bibliométrico	Navega en el menú de indicadores bibliométricos		Muestra los indicadores bibliométricos que se pueden generar	
	Selecciona el indicador bibliométrico		Muestra el resultado	

Tabla 5.2. Ejecutar indicador bibliométrico

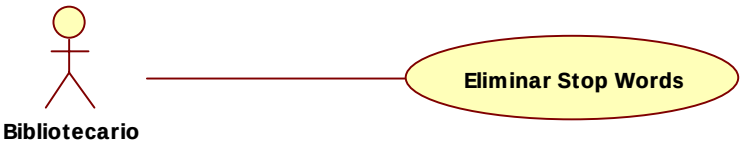
c. Agregar "stop word".
La finalidad de este caso de uso es mostrar la secuencia para agregar una palabra sin significado para el estudio.



	Bibliotecario		**Sistema**	
Agregar "Stop Words"	Selecciona del menú la opción de agregar "stop words"		Muestra la plantilla	E1
	Ingresa la *stop word* y una nota		Actualiza el catálogo de "*stop words*"	
E1	No proporciona la *stop word*		Solicita la palabra	

Tabla 5.3. Caso de uso Agregar

d. Eliminar "Stop Words"
La finalidad de este caso de uso es mostrar la secuencia para eliminar una palabra sin significado.



	Bibliotecario		**Sistema**	
Eliminar "stop words"	Selecciona del menú la opción de eliminar "stop words"		Muestra la lista de *stop words*	
	Selecciona las *stop words* a eliminar y desde un objeto tipo botón manda a eliminar.		Elimina del catálogo las *stop words* seleccionadas	E1
E1	No selecciona *stop words*		Solicita seleccionara al menos una *stop word*	

Tabla 5.4. Caso de uso Eliminar "Stop Words"

e. Agregar etiqueta
La finalidad de este caso de uso es mostrar la secuencia para agregar una etiqueta de contenido.

	Bibliotecario		**Sistema**	
Agregar etiqueta	Selecciona del menú la opción de agregar etiqueta		Muestra la plantilla	E1
	Ingresa la etiqueta y una nota		Actualiza el catálogo de etiquetas	
E1	No proporciona la etiqueta		Solicita la palabra	

Tabla 5.5. Caso de uso Agregar etiqueta

f. Eliminar etiqueta
La finalidad de este caso de uso es mostrar la secuencia para eliminar una etiqueta de contenido

Eliminar etiqueta	Selecciona del menú la opción de eliminar etiquetas		Muestra la lista de etiquetas	
	Selecciona las etiquetas a eliminar y desde un objeto tipo botón manda a eliminar.		Elimina del catálogo las etiquetas seleccionadas	E1
E1	No selecciona etiquetas		Solicita seleccionara al menos una etiqueta	

Tabla 5.6. Caso de uso Eliminar etiqueta

g. Agregar símbolo de puntuación
La finalidad de este caso de uso es mostrar la secuencia para agregar un símbolo de puntuación que se desea no sea considerado en la normalización.

	Bibliotecario		**Sistema**	
Agregar símbolo de puntuación	Selecciona del menú la opción de agregar símbolo de puntuación		Muestra la plantilla	E1
	Ingresa el símbolo de puntuación y una nota		Actualiza el catálogo de símbolos de puntuación	
E1	No proporciona el símbolo de puntuación		Solicita el símbolo de puntuación	

Tabla 5.7. Caso de uso Agregar símbolo de puntuación

h. Eliminar símbolo de puntuación
La finalidad de este caso de uso es mostrar la secuencia para eliminar un símbolo

Eliminar símbolos de puntuación	Selecciona del menú la opción de eliminar símbolos de puntuación		Muestra la lista de símbolos de puntuación	
	Selecciona los símbolos de puntuación a eliminar y desde un objeto tipo botón manda a eliminar.		Elimina del catálogo los símbolos de puntuación	E1
E1	No selecciona símbolos de puntuación		Solicita seleccionara al menos un símbolo de puntuación	

Tabla 5.8. Caso de uso Eliminar símbolos de puntuación

5.3. Análisis

Conforme se vio en el capítulo 3, preparar un registro para su indización requiere de algunos procedimientos, los cuales son analizados a continuación, para el problema plantado en el presente trabajo.

5.3.1 Eliminación de formato

ALEPH en su base bibliográfica tiene una tabla llamada Z00 que almacena registros bibliográficos bajo el estándar de catalogación, el contenido bibliográfico es almacenado en un campo de tipo LOB, ver apéndice A, su estructura se muestra en la figura 5.1:

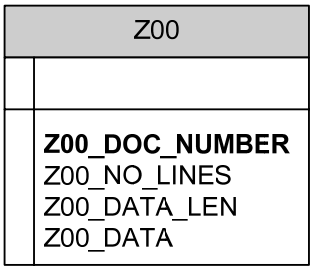


Figura 5.1. Tabla Z00 de la base bibliográfica del ALEPH

Donde

- Z00_DOC_NUMBER es el número de identificación asignado al registro del documento.

- Z00_NO_LINES es la cantidad de líneas que conforman el registro bibliográfico, hablado en términos bibliográficos, es la cantidad de etiquetas que lo conforman.
- Z00_DATA_LEN es la longitud de la cadena que conforma todo el registro bibliográfico.
- Z00_DATA es la cadena de caracteres que representa el contenido del registro. Para cada línea o etiqueta del registro bibliográfico hay una subcadena que se compone de:

  4 dígitos para la longitud de la subcadena, esta se cuenta a partir del primer carácter que representa a la etiqueta.
  3 caracteres que representan a la etiqueta.
  1 dígito para el subcampo uno.
  1 dígito para el subcampo dos.
  1 carácter para el alfabeto.
  Contenido.

Ejemplo; en la siguiente cadena

0008FMT LBK 0030LDR L-----nam--22------a-4500 0021008 LS2003 DGB A ESP 046008.. …

para la subcadena sombreada

La longitud es de 8, esta se cuenta a partir de la F que es el primer carácter que representa a la etiqueta hasta la K que es donde termina la subcadena
La etiqueta es "FMT".
El subcampo uno es un espacio en blanco.
El subcampo dos es un espacio en blanco.
El alfabeto es "L" de Latín.
El contenido es "BK"

En base a la estructura que tiene el registro bibliográfico se obtiene el algoritmo mostrado en la figura 5.2 para obtener la información de forma que sea manipulable, ya que al estar almacenado en un campo de tipo *long* bajo un estándar de catalogación bibliográfica es casi imposible tener acceso a la información, por ello cada elemento de la cadena es extraído.

```
c = cantidad de registros bibliográficos almacenados en la base de datos.
FOR a = 1 TO c
        j=1
        var_z00_doc_number
        var_z00_no_lines
        var_z00_data_len
        var_z00_data
        WHILE i <= var_z00_no_lines
        LOOP
                longitud := SUBSTR(var_z00_data,j,4);
                etiqueta := SUBSTR(var_z00_data,j+4,3);
                indicador1 := SUBSTR(var_z00_data,j+7,1);
                indicador2 := SUBSTR(var_z00_data,j+8,1);
                alfabeto := SUBSTR(var_z00_data,j+9,1);
                contenido := SUBSTR(SUBSTR(var_z00_data,j+10,longitud – 6),1,50);
                k := INSTR(contenido,'$$',1,1)
                IF (k>0) THEN
                l := 1;
                        WHILE k > 0
                        LOOP
                                k := INSTR(contenido, '$$',1,1)
                                l :==
                        END LOOP
                END IF
        END LOOP
NEXT
```

Figura 5.2. Algoritmo de extracción de información de registros bibliográficos

Una vez que son identificados cada uno de los campos que conforman el registro, se manipulan según las necesidades. No todas la etiquetas son relevantes para el estudio, así que en esta parte es donde se van descartando, y considerando aquellas que sí lo son. Las etiquetas a considerar son las 245 para el título, 260 para el año, 505 para las notas y 700 para el asesor, como parte de la etapa de filtración. Para la administración de las etiquetas a ser consideradas se almacenan en una tabla de donde serán leídas, la tabla es TIBETIQUETAS.

Se ha convenido almacenar en tablas la información bajo tres niveles: etiqueta, subcampo y palabra. Para el nivel etiqueta la información almacenada es tal cual se encontró en el registro, para el nivel subcampo la información está desglosada por subcampo para cada etiqueta y para el nivel palabra se almacenan las palabras ya normalizadas, por registro y por etiqueta de interés. En este último nivel está involucrada la etapa de extracción de palabras.

La primer tabla de la base de datos es básicamente para referencia y acceso a la información, se han considerado 7 campos, LON que indica la longitud del contenido, ETI que indica de que etiqueta (o campos del registro bibliográfico) se trata, IND1 es el indicador uno, IND2 el indicador dos, ALFHA el alfabeto donde regularmente encontraremos L de latín, CONTENIDO que es la información que corresponde a la etiqueta y MAT que es la matriz o identificador del registro. Un ejemplo de este nivel se muestra en la tabla 5.9

LON	ETI	I1	I2	A	CONTENIDO	MAT
21	008			L	s2003 dgb a esp	600028
46	008			L	040820s----------r-----000-0-eng-d	600028
19	035			L	$$atdf0600028	600028
26	084			L	$$a001-01132-b0-2003	600028
44	100	2		L	$$aballesteros estrada, silvia Socorro	600028
154	245	1	0	L	$$asistema integral para el control y registro de	600028
36	**260**			**L**	**$$amexico :$$bel autor,$$c2003**	**600028**
23	300			L	$$a104 p. :$$bil.	600028
83	502			L	$$atesis licenciatura (ingeniero en computación)-	600028
10	590			L	$$a1	600028
45	700	2	1	L	$$acontreras barrera, marcial,$$easesor	600028
74	710	2	1	L	$$auniversidad nacional autonoma de mexico.$$b	600028
115	856	4		L	$$uhttp://132.248.9.9:8080/tesdig/ips/entrada.jsp?	600028
43	CAT			L	$$adgb$$b00$$c20051004$$ltdf01$$h0000	600028
32	CAT			L	$$c20060508$$ltdf01$$h1922	600028
32	CAT			L	$$c20060508$$ltdf01$$h1948	600028
32	CAT			L	$$c20060508$$ltdf01$$h2019	600028
49	CAT			L	$$abatch-upd$$b00$$c20060615$$ltdf01$$h201	600028
32	CAT			L	$$c20061011$$ltes01$$h1935	600028

Tabla 5.1. Nivel etiqueta

32	CAT	L	$$c20061011$$ltes01$$h1957	600028
32	CAT	L	$$c20061117$$ltes01$$h1919	600028
32	CAT	L	$$c20061128$$ltes01$$h1508	600028
12	039	L	$$adig	600028
8	FMT	L	Bk	600028
30	LDR	L	-----nam--22------a-4500	600028

Tabla 5.9. (Continuación) Nivel etiqueta

La tabla de la bases de datos con datos a nivel de subcampos está definida por, ETIQUETA que indica de qué etiqueta, CONTENIDO que es la información que corresponde a la etiqueta y subcampo, SUBCAMPO el subcampo correspondiente y MAT que es la matriz o identificador del registro. Un ejemplo de este nivel se muestra en la tabla 5.10

MAT	ETI	SUBC	CONTENIDO
600028	100	a	ballesteros estrada, silvia socorro
600028	245	a	sistema integral para el control y registro de
600028	**260**	**a**	**mexico:**
600028	**260**	**b**	**el autor,**
600028	**260**	**c**	**2003**
600028	300	a	104 p. :
600028	300	b	Il.
600028	502	a	tesis licenciatura (ingeniero en computacion)-u
600028	590	a	1
600028	700	a	Contreras barrera, marcial,
600028	700	e	Asesor
600028	710	a	universidad nacional autonoma de mexico.
600028	710	b	Fac.
600028	856	u	http://132.248.9.9:8080/tesdig/ips/entrada.jsp?
…	…	…	…

Tabla 5.10. Nivel subcampo

Finalmente la tabla de la base de datos con datos a nivel de palabra definida por ETIQUETA que indica a que etiqueta corresponde la información, MATRIZ que es la matriz o identificador del registro, PALABRA tal y como se obtuvo considerando los delimitadores y PALABRAN la palabra normalizada. Un ejemplo de este nivel se muestra en la tabla 5.11

MATRIZ	ETIQUETA	PALABRA	PALABRAN
000074801	505	EvaluaciÃ³n	evaluacion
000074801	505	San	san
000074801	505	CristÃ³bal	cristobal
000074801	505	Chiapas."	chiapas
…			

Tabla 5.11. Nivel palabra

5.3.2. Extracción de palabras

Otro de los problemas encontrados en esta etapa es el de las letras acentuadas, ya que en el formato en el que están almacenadas son tomadas como dos caracteres de la cadena, lo cual afecta al número que indica el tamaño de la longitud, y en consecuencia al algoritmo. Esto es producto del carácter de conversión que es utilizado para almacenar los registros. Tal es el caso del UTF, utilizado por ALEPH.

De tal forma que fue necesario obtener la secuencia de caracteres que representan a cada letra acentuada. La tabla 5.12 muestra dichas secuencias:

L	S1	S2	A1	A2	Ejemplo	L	S1	S2	A1	A2	Ejemplo
à	Ã		-61	-92		À	Ã		-61	-128	
á	Ã	¡	-61	-94	pediÃ¡trico	Á	Ã		-61	-127	SISTEMÃTICA
â	Ã	¢	-61	-95		Â	Ã		-61	-126	
ä	Ã	¤	-61	-96		Ä	Ã		-61	-124	
è	Ã	¨	-61	-85		È	Ã		-61	-120	
é	Ã	©	-61	-86	despuÃ©s	É	Ã		-61	-119	MÃDICA
ê	Ã	ª	-61	-87		Ê	Ã		-61	-118	
ë	Ã	«	-61	-88		Ë	Ã		-61	-117	
ì	Ã	¬	-61	-81		Ì	Ã		-61	-116	
í	Ã	-	-61	-82	CiudadanÃ-a	Í	Ã		-61	-115	
î	Ã	®	-61	-83		Î	Ã		-61	-114	
ï	Ã	¯	-61	-84		Ï	Ã		-61	-113	
ñ	Ã	±	-61	-79	diseÃ±o	Ñ	Ã		-61	-111	
ò	Ã	²	-61	-74		Ò	Ã		-61	-110	
ó	Ã	³	-61	-76	ecolÃ³gicos	Ó	Ã		-61	-109	EDUCACIÃN
ô	Ã	´	-61	-77		Ô	Ã		-61	-108	
ö	Ã	¶	-61	-78		Ö	Ã		-61	-106	
ù	Ã	¹	-61	-68		Ù	Ã		-61	-103	
ú	Ã	º	-61	-69		Ú	Ã		-61	-102	
û	Ã	»	-61	-70		Û	Ã		-61	-101	
ü	Ã	¼	-61	-71	lingÃ¼Ã-sticas	Ü	Ã		-61	-100	

Tabla 5.12 Secuencia de caracteres para letras acentuadas.
L-Letra, S1 Símbolo1, S2 Símbolo 2, A1 ASCII 1, A2 ASCII 2

5.3.3. Filtrado

El registro catalográfico de un registro puede estar conformado por una serie de etiquetas, sin embargo no todas son necesarias para el estudio y no son consideradas para la *tokenización*, por ejemplo la etiqueta que tiene información de catalogador, o la que tiene información de la liga a texto completo, o la que tiene fecha de catalogación, etc. A continuación se muestran algunas:

```
FMT   L BK
LDR   L -----nam--22------a-4500
008   L 960905DGB-A-ESP-19----------------------
035   L $$aTDF0000606
900   L $$aBC$$b000$$x0000000606
CAT   L $$aDGB$$b00$$c20051004$$lTDF01$$h0000
CAT   L $$c20060508$$lTDF01$$h1905
CAT   L $$c20060508$$lTDF01$$h1935
CAT   L $$c20060508$$lTDF01$$h2006
```

CAT L $$c20060508$$lTDF01$$h2031
CAT L $$aBATCH-UPD$$b00$$c20060509$$lTDF01$$h1301
CAT L $$c20060926$$lTDF01$$h1439
CAT L $$c20061009$$lTDF01$$h1351

Otras etiquetas son de importancia por contener información que aporta significado al estudio realizado, entre las que están:

100, Autor. Es el responsable intelectual de una obra, y puede ser una persona o grupo de personas, instituciones, congresos, etc. Esta información también suele ser complementada con datos adicionales con el fin de identificar plenamente una obra, que pueden ser: fechas, títulos nobiliarios, palabras asociadas al nombre, lugares, etc.

245, Título. Por lo general es la denominación con que se imprime la obra; puede presentar variedades como: título propiamente dicho, título uniforme, título paralelo y título clave, entre otros.

260, Pie de imprenta. Contempla el país, la ciudad, la casa editorial donde se imprimió el documento y el año de edición.

505, Notas bibliográficas. Describen ciertas características de las obras; pueden ser de contenido, de encuadernación, de resúmenes, etc.

Para almacenar la información en las tablas de nivel subcampo y nivel palabra se aplican los procesos de limpieza y normalización, cualquier corrección realizada se ve plasmada en dichas tablas, aunque es posible realizar la corrección en el registro bibliográfico y actualizar la tabla Z00 propia del sistema ALEPH esta labor queda a cargo del especialista de la catalogación.

Por ejemplo, para la etiqueta de asesor a nivel subcampo se tienen registros que en el contenido se encuentra más de un asesor plasmado, cada asesor debería de estar en un subcampo, con los procesos de limpieza y normalización se almacenan en las tablas de índices tal como debería de ser.

5.4. Requerimientos no funcionales

- Software para navegar
  - Intenet Explorer o Netscape
- Seguridad
  - Firewall
- Hardware
  - Servidor Sun Fire V890

- 4 procesadores a 1200 Mhz
- RAM 8 GB
- 6 DD de 72 GB

- Software
  - Sistema operativo Solaris 2.9
  - DBMS Oracle Enterprise Edition Release 9.2.0.4.0 - 64bit Production
  - PL-SQL
  - PRO*C
  - Java

5.5 Arquitectura del sistema

Se propone una arquitectura basada en el patrón arquitectónico MVC (Modelo Vista Controlador). La figura 5.3. muestra el patrón tomando en cuenta que la aplicación será desarrollada con tecnología Java.

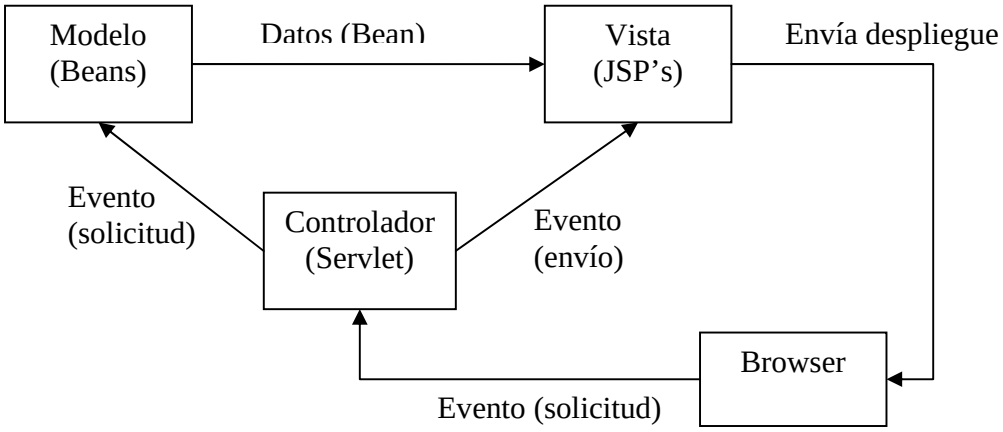


Figura 5.3. Modelo Vista Controlador

5.6. Diseño

Los objetos modelados, y creados en la base de datos son tablas que sirven para el almacenamiento de índices.

Una de las características de ORACLE es el uso de *triggers* (o disparadores) y que en el presente trabajo se hace uso de ello de tal forma que existe un *trigger* cuyo estatus es activo, y que funciona de la siguiente manera:

- Cuando un registro en la tabla Z00 es eliminado todos lo registros correspondientes a él son borrados de las tablas de índices.

- Cuando un registro es agregado o actualizado, en una tabla auxiliar se almacena el número de matriz (sistema o identificador de registro) indicando que dicho registro requiere de indización y permanecerá a la espera de que se lancen los procesos necesarios.

Cuerpo del trigger:

```
create or replace trigger tr_indexa
after INSERT OR UPDATE OR DELETE ON z00
for each row
BEGIN
 if inserting then
  INSERT INTO para_inx VALUES (:new.z00_doc_number)
  ;
 end if;
 if updating then
  INSERT INTO para_inx VALUES (:new.z00_doc_number)
  ;
 end if;
 if deleting then
  DELETE FROM TIBNIVEL1 WHERE MATRIZ = :old. z00_doc_number
  ;
  DELETE FROM TIBNIVEL2 WHERE MATRIZ = :old. z00_doc_number
  ;
  DELETE FROM TIBNIVEL3 WHERE MATRIZ = :old. z00_doc_number
  ;
 end if;
END;
/
QUIT
```

Los procesos a los que se hace referencia en el párrafo anterior son de dos tipos: el de clasificación y el de indización, ambos toman de la tabla temporal los números que van a ser procesados.

Dichos procesos son lanzados cuando en la base de datos no hay actividad mediante un *crontab*. Un *crontab* es un archivo de configuración que está formado por líneas que representan una actividad cuya ejecución se desea programar, las líneas de programación siguen un formato que consta de cinco campos que indican un instante de ejecución y la ruta del archivo, programa o procedimiento a ejecutarse. Así, es posible dejar programada la ejecución en un horario donde la base de encuentre sin actividad.

Ejemplo de contrab:
00 23 * * 1-5 csh -f /exlibris/aleph/a16_1/aleph/proc/indexacion.csh

- Proceso de clasificación

El proceso correspondiente a la clasificación se encarga de tomar todas las palabras para cada una de las clasificaciones establecidas, de tal manera que se forman cuatro vectores de palabras que representan cada clasificación.

El registro de entrada se descompone en palabras y se forma un vector que lo representa. Mediante el uso de la representación vectorial de las redes neuronales se hace la comparación de vectores, según el vector que tenga más similitud con el de entrada será la clasificación otorgada al nuevo registro.

Para formar los vectores necesarios para la clasificación se hizo uso de consultas combinadas o *JOINS*. Con LEFT OUTER JOIN se obtiene los registros de en la tabla que se sitúe a la izquierda de la cláusula JOIN, mientras que con RIGHT OUTER JOIN se obtiene el efecto contrario. Su sintaxis es la siguiente:

SELECT [ALL | DISTINCT] <nombre_campo> [{,<nombre_campo>}]
FROM <nombre_tabla>
[{LEFT|RIGHT OUTER JOIN <nombre_tabla>
ON <condicion_combinacion> }]
[WHERE <condicion> [{ AND|OR <condicion>}]]
[GROUP BY <nombre_campo> [{,<nombre_campo >}]]
[HAVING <condicion>[{ AND|OR <condicion>}]]
[ORDER BY <nombre_campo>|<indice_campo> [ASC | DESC]
[{,<nombre_campo>|<indice_campo> [ASC | DESC] }]]

Así, para la generación de vectores de palabras para la clasificación la tabla de la izquierda es la selección de palabras representativas por área y la otra tabla es la selección de las palabras de cierto registro.

En la tabla 5.13. se muestra el resultado de un *left outer join*, son todos los registros contenidos en TABLA1 aunque no haya correspondencia con TABLA2.

TABLA1	**JOIN**	**TABLA2**
A	A - A	A
B	B – NULL	
		C
D	D – D	D
E	E – E	E
		F

Tabla 5.13. LEFT OUTER JOIN

En la tabla 5.14. se muestra el resultado de un *right outer join*, son todos los registros contenidos en la TABLA2 aunque no haya correspondencia con TABLA1.

TABLA1	JOIN	TABLA2
A	A - A	A
B		
	NULL - C	C
D	D – D	D
E	E - E	E
	NULL - F	F

Tabla 5.14. RIGHT OUTER JOIN

- Proceso de indización

El proceso se encarga de leer el contenido bibliográfico codificado y procesarlo para dejarlo en cierto formato.

Haciendo uso del nivel léxico de procesamiento de lenguaje natural, el proceso lee las tablas que fungen como catálogos de etiquetas y "stop words", formado un vector para cada una de ellas, de esta manera son comparados con la información de entrada, o del registro a procesar, y se descartan o almacenan en las tablas de índices.

Por cada registro bibliográfico la información se descompone en tres niveles, el primero a nivel de etiqueta, el segundo a nivel de subcampo y el tercero a nivel de palabra. Para cada nivel hay una tabla donde se almacenan los datos como parte de la indización y tratamiento posterior.

El nivel subcampo y palabra son los más importantes ya que es de aquí de donde se tomaran los datos para las correspondientes estadísticas.

El resultado estadístico debe ser lo más cercano a la realidad, para cumplir con dicha necesidad se hace uso de una característica del manejador de bases datos, las vistas. De tal forma que teniendo las tablas base se pueden crear tantas vistas como indicadores sean necesarios para el estudio.

A continuación se muestran algunas de las consultas, formadas en función de los modelos de Lotka, vistos en el capitulo 2, para el comportamiento tales como rango de palabras, frecuencias y productividad de tutores, y los resultados para la interpretación.

Frecuencias de palabras

```
CREATE VIEW VIBRANKXPAL AS (
        SELECT ROWNUM RANK
                , PALABRAN
                , FRECUENCIA
                , ROWNUM*FRECUENCIA RF
        FROM (
                SELECT PALABRAN
                        , COUNT(*) FRECUENCIA
                FROM TIBNIVEL3
                GROUP BY PALABRAN
                ORDER BY COUNT(*) DESC
                )
        )
```

RANK	PALABRAN	FRECUENCIA	RF
...			
21	analisis	7027	147567
22	mexico	6624	145728
...			
29	metodos	4700	136300
...			
31	estudio	4671	144801
...			
37	sistema	3559	131683
...			
43	desarrollo	3100	133300
...			
51	diseno	2586	131886
...			

Peso locales por palabras

```
CREATE VIEW VIBPLXPAL AS ( SELECT TIBNIVEL3.MATRIZ D,
TIBNIVEL3.PALABRAN T, COUNT(*) DTF, LOG(10,COUNT(*)) + 1 LOG_DTF,
SUMDTF, (LOG(10,COUNT(*)) + 1) / SUMDTF L
    FROM TIBNIVEL3, (SELECT MATRIZ, SUM(LOG) SUMDTF
            FROM (SELECT MATRIZ MATRIZ
                  PALABRAN
                  , COUNT(*) DTF
                  , LOG(10,COUNT(*)) + 1 LOG
                  FROM TIBNIVEL3
                  GROUP BY MATRIZ, PALABRAN)
            GROUP BY MATRIZ) A
   WHERE TIBNIVEL3.MATRIZ = A.MATRIZ
  GROUP BY TIBNIVEL3.MATRIZ, TIBNIVEL3.PALABRAN, SUMDTF
)
```

D	T	DTF	LOG_DTF	SUMDTF	L
000003287	farmaceutica	5	1.69897	86.0334238	.01974779
000003886	penal	3	1.47712125	49.570063	.029798656
000032270	planta	1	1	46.9199145	.021312912
000032270	estudio	6	1.77815125	46.9199145	.037897581

Pesos globales por año

```
CREATE VIEW VIBPGXANIO as (select CONTENIDO T, count(*) nf, A.N N,
round( log(10,(A.N-count(*))/count(*)),2) G
FROM TIBNIVE2, (SELECT COUNT( DISTINCT MATRIZ) N FROM BAES2
WHERE ETIQUETA = '260' AND SUBCAMPO = 'c' ) A
 WHERE ETIQUETA = '260'
  AND SUBCAMPO = 'c'
 GROUP BY CONTENIDO, A.N)
```

T	NF	N	G
1962	1	29553	4.47058688
1969	2	29553	4.16954219

1979	1	29553	4.47058688
1987	1	29553	4.47058688
1988	1	29553	4.47058688
1996	1	29553	4.47058688
1997	1	29553	4.47058688
1998	11775	29553	.178921986
1999	6924	29553	.514308303
2000	10109	29553	.284077416
2001	720	29553	1.60255734
2002	3	29553	3.99343623
2003	12	29553	3.39124395
2004	3	29553	3.99343623
2005	4	29553	3.8684828

Estadísticas por asesor

```
CREATE VIEW VIBPGXASE AS (SELECT CONTENIDO T, COUNT(*) NF, A.N N,
LOG(10,(A.N-COUNT(*))/COUNT(*)) G
FROM TIBNIVEL2, (SELECT COUNT( DISTINCT MATRIZ) N FROM TIBNIVEL2
WHERE ETIQUETA = '700' AND SUBCAMPO = 'A' ) A
 WHERE ETIQUETA = '700'
  AND SUBCAMPO = 'A'
 GROUP BY CONTENIDO, A.N)
```

T	NF	N	G
Cortes y Huerta, Sergio	71	29368	2.6155648
Avila Ornelas, Roberto	67	29368	2.64080764
Aldape Barrios, Beatriz Catalina	61	29368	2.68164153
Cortes y Huerta, Sergio	71	29368	2.6155648
Avila Ornelas, Roberto	67	29368	2.64080764
Aldape Barrios, Beatriz Catalina	61	29368	2.68164153
Rodriguez Ortiz, Martha	60	29368	2.68883493
Vaquero Cazares, Jose Esteban	59	29368	2.69614899

5.7. Diagrama general de Clases

Una de las actividades importantes en el análisis consiste en identificar y organizar clases en los siguientes tres tipo:

a. Clases de tipo *Boundary.* Son las clases que sirvan como interfaz humana y visual entre el sistema y el usuario.
b. Clases de tipo *Entity.* Son aquellas clases relacionadas con procesos de persistencia o almacenamiento y que generalmente mapean entidades (o tablas) en una base de datos. En el caso de estudio serán mapeadas a vistas.
c. Clases de tipo control. Son aquellas que contienen la lógica del funcionamiento del sistema, así como la del flujo y control de errores.

A continuación se muestra la clasificación de las clases del caso de estudio y el diagrama general de clases en el diagrama 5.1

Clases *Entity*:

1. Conexión
2. Pesog
3. PesogVO
4. Pesol
5. PesolVO
6. Rank
7. RankVO
8. Termino
9. TerminoVO

Clases Boundary:

1. Agregai
2. Borra
3. Index
4. Pglobal
5. Plocal
6. Rank
7. Menu
8. Error

Clases *Control*:

1. ValidaUsuario
2. CrlErrores

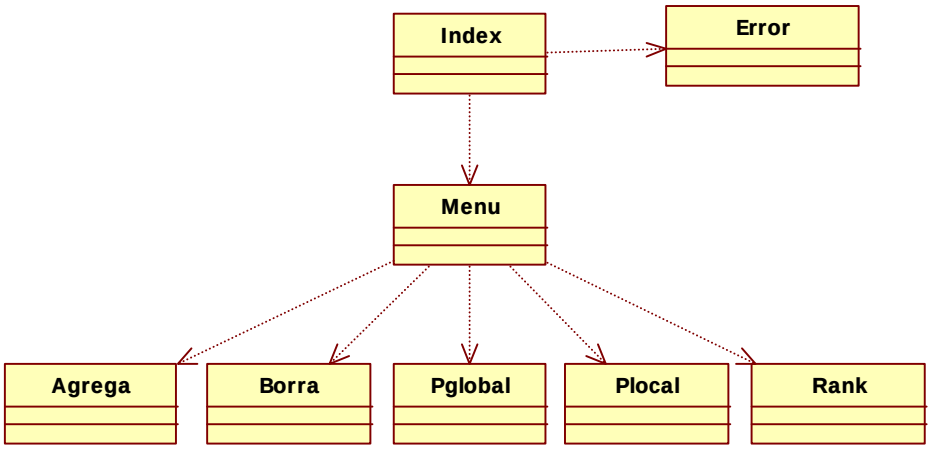


Diagrama 5.1. Diagrama general de clases.

5.8. Diagrama de paquetes

Los paquetes, tabla5.15, están conformados en base al modelo MVC:

Modelo	Vista	Controlador
Conexión Pesog PesogVO Pesol PesolVO Rank RankVO Termino TerminoVO	Agregai Borra Index Pglobal Plocal Rank Menu Error	ValidaUsuario CrtlErrores

Tabla 5.15 Paquetes

5.9. Diagrama de la Base de Datos

En el diagrama 5.2. se muestran las tablas involucradas propias de ALEPH, cuyo nombre inicia con Z, las que fueron creadas para indexar la información normalizada y catálogos, cuyo nombre inicia con TIB, además de las vistas creadas para la recuperación de la información, cuyo nombre inicia con V.

1. Z00. Tesis propiedad de la biblioteca. Los atributos que definen a los objetos de esta clase son:

a. Z00_DOC_NUMBER. Contiene un número único de 9 dígitos que identifica a cada documento.
b. Z00_NO_LINES. Contiene el número de etiquetas que conforman el registro del documento.
c. Z00_DATA_LEN. Contiene la longitud del contenido de la descripción bibliográfica del documento.
d. Z00_DATA. Contiene una cadena de caracteres codificada con la descripción bibliográfica del documento.

2. TIBNIVEL1. Etiqueta contenida en un registro bibliográfico.

    a. LONGITUD. Contiene la longitud del contenido de la etiqueta
    b. ETIQUETA. Contiene el identificador de la etiqueta
    c. CONTENIDO. Contiene la información correspondiente e la etiqueta
    d. MATRIZ. Contiene un número único de 9 dígitos que identifica a cada documento

3. TIBNIVEL2. Subcampo contenido en un registro bibliográfico.

    a. MATRIZ. Contiene un número único de 9 dígitos que identifica a cada documento
    b. ETIQUETA. Contiene el identificador de la etiqueta
    c. SUBCAMPO. Contiene el identificador del subcampo
    d. CONTENIDO. Contiene la información correspondiente a la etiqueta y subcampo

4. TIBNIVEL3. Palabra contenida en un registro bibliográfico.

    a. MATRIZ. Contiene un número único de 9 dígitos que identifica a cada documento
    b. ETIQUETA. Contiene el identificador de la etiqueta
    c. PALABRA. Contiene la palabra tal y como se encontró dentro del contenido bibliográfico
    d. PALABRAN. Contiene la palabra una vez normalizada

5. TIBSTOPW. Palabra que no aporta significado.
    a. PALABRA. Palabra sin significado
    b. NOTA. Nota para la palabra sin significado

6. TIBETIQUETAS. Etiqueta a considerara para el estudio.
    a. ETIQUETA. Cadena de tres caracteres que identifican la etiqueta
    b. NOTA. Nota para la etiqueta

7. TIBSIMBOLOS. Símbolo de puntuación que no es considerado para el estudio
    a. SIMBOLO. Cadena de uno o mas caracteres que contiene el símbolo
    b. NOTA. Nota para el símbolo

8. TIBUSUARIOS. Usuario que hacen uso del sistema.
    a. USUARIO. Clave del usuario
    b. CONTRASENIA. Contraseña del usuario
    c. NOMBRE. Nombre del usuario
    d. NIVEL. Nivel del usuario

9. TIBCLASIFICACIONES. Clasificación a la que pertenece un documento
    a. INSTITUCION. Clave de institución
    b. FACULTAD. Clave de facultad de una institución
    c. CARRERA. Clave de carrera de facultad
    d. AREA. Área temática de la carrera
    e. NOMBRE. Nombre de la carrera

10. TIBPRAINDX. Número de sistema (matriz o identificador de registro) que serán indexados.
    a. MATRIZ

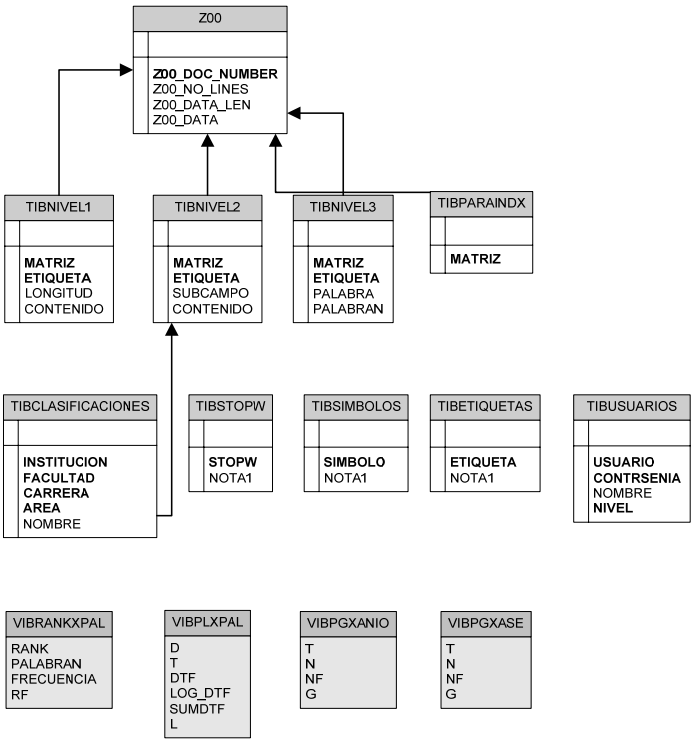


Diagrama 5.2. Diagrama de la base de datos

5.10. Pantallas del sistema

La figura 5.5. muestra la pantalla de acceso al sistema.

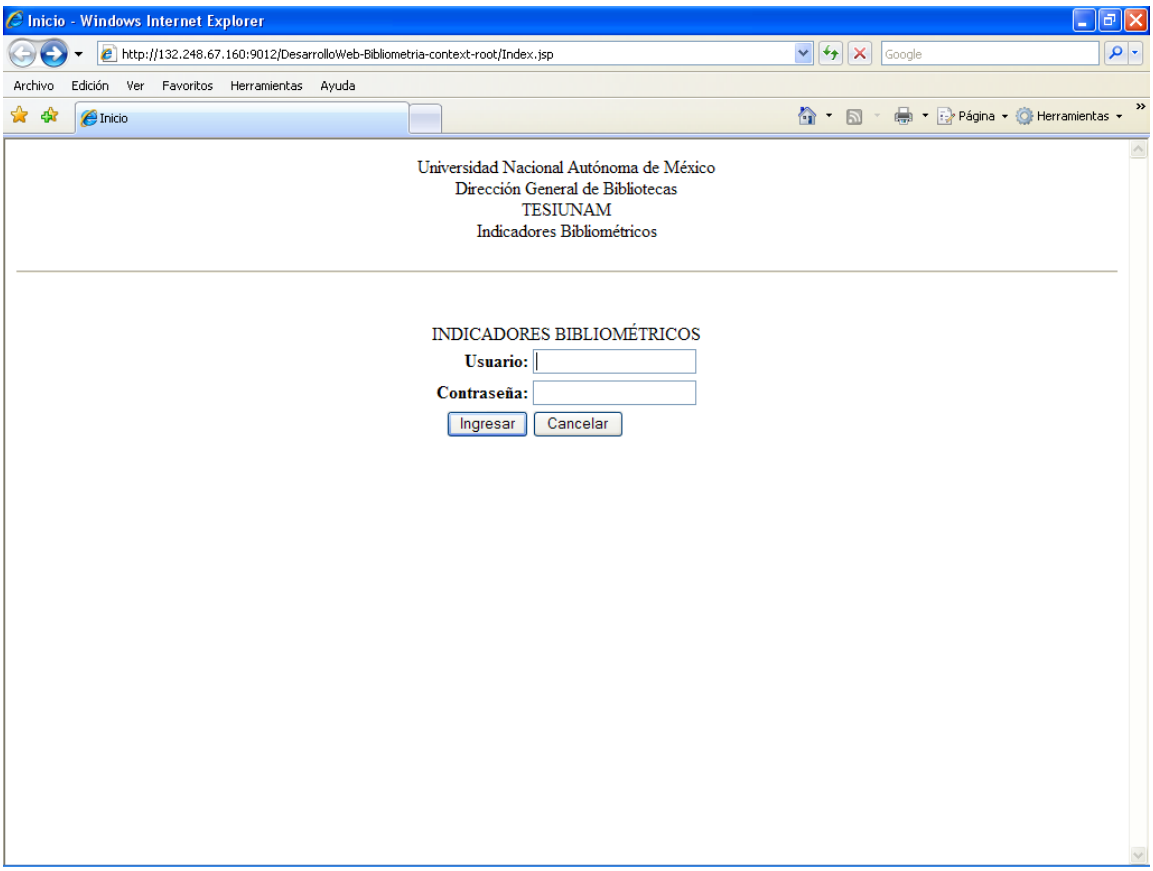


Figura 5.5 Pantalla de acceso la sistema

Una de las pantallas del sistema es la del menú de indicadores bibliométicos, resultado del análisis y aplicación de modelos matemáticos, figura 5.6. Se pueden generar por clasificación temática, años, asesores y palabras, demás de la administración de catálogos como los son de palabras sin sentido (stop words), símbolos de puntuación y etiquetas.

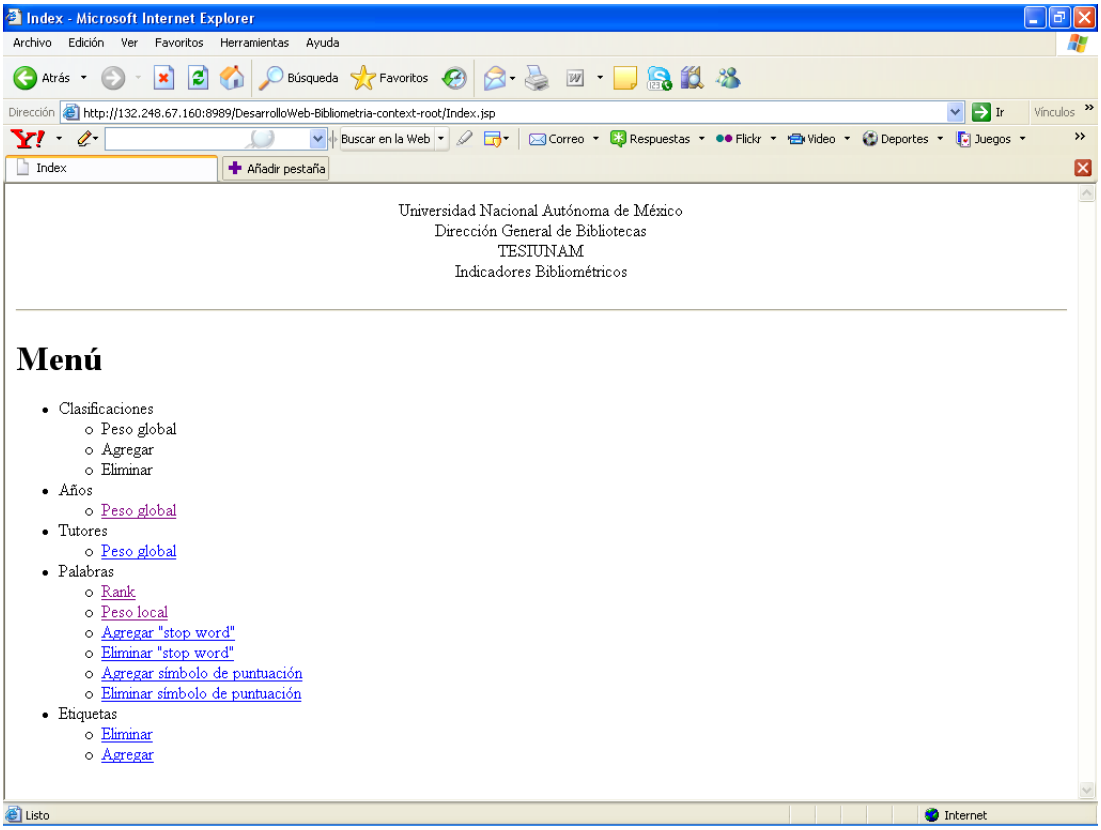
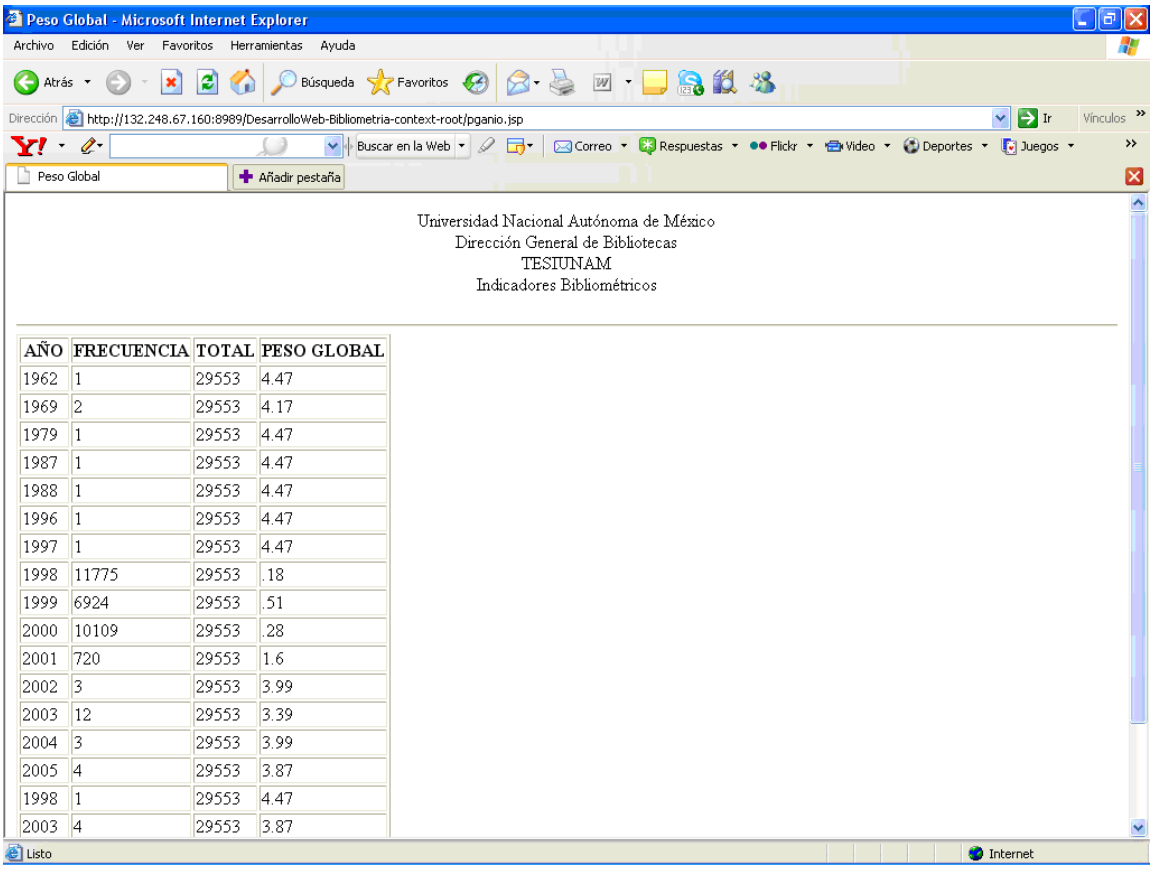


Figura 5.6 Pantalla del menú de indicadores bibliométricos

Los resultados se presentan de forma tabular para su interpretación. En la pantalla 5.7 se observa el resultado del indicador de producción anual de tesis digitales.



AÑO	FRECUENCIA	TOTAL	PESO GLOBAL
1962	1	29553	4.47
1969	2	29553	4.17
1979	1	29553	4.47
1987	1	29553	4.47
1988	1	29553	4.47
1996	1	29553	4.47
1997	1	29553	4.47
1998	11775	29553	.18
1999	6924	29553	.51
2000	10109	29553	.28
2001	720	29553	1.6
2002	3	29553	3.99
2003	12	29553	3.39
2004	3	29553	3.99
2005	4	29553	3.87
1998	1	29553	4.47
2003	4	29553	3.87

Figura 5.7 Pantalla con el resultado de la producción anual de tesis.

Resultados

En la primera etapa desarrollada del proyecto se aplicó el análisis léxico a nivel de palabras, dando como resultado la generación de dos listados, uno de palabras con sentido, es decir, palabras que aportan conocimiento a los estudios bibliométricos, y otro listado de palabras sin sentido (stop word), es decir, palabras que no aportan conocimiento al estudio. Este proceso se realizó analizando la frecuencia de repeticiones de una palabra en el documento; como se expuso, las palabras con mayor y menor frecuencia en los documento son las que menos información aportan en la recuperación de la información, y por tal razón no deben de ser tomadas en cuenta para la indización de los datos.

Se detectaron 121,207 palabras distintas y en total 2,296,899 almacenadas en la base de datos. Se toman como palabras de baja frecuencia aquellas que aparecían menos 10 veces y de altas frecuencias a los artículos, preposiciones, etc. De lo anterior se obtienen los siguientes resultados, 108891 palabras distintas de frecuencias bajas que representan el 89% de las palabras distintas dando un total de 213,232 palabras que representan el 9% de las palabras almacenadas en la base de datos, 22 palabras distintas de frecuencias altas que representan el 0.01% de las palabras distintas dando un total de 713580 palabras que representan el 31% de las palabras almacenadas en la base de datos. Es decir, el 60% de palabras representativas para la aplicación de modelos matemáticos que indiquen el comportamiento en la producción de tesis de la UNAM.

El análisis de la información también sirvió de base en la detección problemas de normalización en el registro de la base de datos, debido a que en el sistema ALEPH 300 la codificación se realizaba sin el uso de acentos, y esto cambió en la versión ALEPH 500.

Una vez realizado el análisis se procedió a formatear la información de acuerdo a las necesidades de desarrollo del sistema. Los datos quedaron almacenados en tres tablas por niveles, en el primer nivel se almacenaron los registros por etiqueta tal cual se extraen del registro bibliográfico. En el segundo nivel se almacenaron por subcampo, en este nivel se aplicó normalización de acuerdo a la optimización de procedimientos de análisis, de almacenamiento y recuperación de información; para los campos, por ejemplo, de asesores y autores en caso de codirección o coautoría se dejó un campo para cada dato referencial. En el tercer nivel se almacenaron por palabra, en este nivel se aplicó indización en base a *tokens* (palabras separadas por un espacio). En los dos últimos niveles se

aplicó filtrado de información en base a la eliminación de símbolos de puntuación y palabras sin sentido.

Un resultado importante en esta etapa es que la información está más uniforme para su manipulación y el desarrollo de la base de datos, otro aspecto importante detectado es la cantidad de errores que comenten los catalogadores al registrar la información en la base de datos.

La aplicación de redes neuronales artificiales en el desarrollo del presente trabajo sirvió para hacer la clasificación de las tesis digitales en áreas temáticas, quedando agrupados los *tokens* por áreas del conocimiento, para su posterior manipulación en la tabla en el nivel 2.

El proceso de indexación, cuyo objetivo es terminar de realizar el procesamiento y la generación de tablas para extracción de información como base del cálculo de frecuencias y determinación de pesos de la palabra en el documento, se realizó para toda la base como referencia de tiempos de procesamiento. Los resultados fueron los siguientes:

Cantidad de registros bibliográficos: 369,090
Cantidad de registros en la tabla en nivel 2 TIBNIVEL2: 22,039,459
Cantidad de registros en la tabla en nivel 3 TIBNIVEL3: 7,483,332
Tiempo de duración del proceso de indexación: 2 días 4 horas 23 minutos 14 segundos

Realizando la indexación en los registros de las tesis digitales, los resultados fueron los siguientes:

Cantidad de registros bibliográficos: 48,089
Cantidad de registros en la tabla TIBNIVEL2 en nivel 2: 3,074,917
Cantidad de registros en la tabla TIBNIVEL3 en nivel 3: 2,296,899
Tiempo de duración del proceso de indexación: 9 horas 33 minutos 50 segundos

De los resultados se observa que el proceso de indexación tardó alrededor de un segundo para cada registro bibliográfico. Con la base de datos en ALEPH 300 la extracción de la información llevaba algunos minutos ya dicha información se encontraba en archivos planos. Así que, considerando que el procedimiento de indización se realizó sólo una vez para todos los registros almacenados, y que para subsecuentes registros la indización es de forma automática una vez que la base de datos está fuera de servicio, la extracción de la información es relativamente más rápida, además que la base de datos se encuentra indexada con los registros bibliográficos catalogados al día.

Una vez terminado el procesamiento de la información se empezó la segunda etapa del proyecto en el cual se desarrolló la interfaz de usuario para ser consultada vía WEB y además se desarrollaron los algoritmos necesarios para realizar los estudios bibliométricos, basados en la teoría expuesta en los capítulos anteriores y la aplicación de los modelos matemáticos.

La aplicación de los modelos tiene como ventajas que en el sistema se identifiquen comportamientos tales como la producción de asesores, años y clasificaciones y uso de palabras.

Se hizo uso de *triggers* para mantener actualizadas las tablas de índices y tomar los datos que forman el contenido de las *vistas* que funcionan como contenedores de información para el estudio bibliométrico.

El uso de vistas hace dinámica la recuperación de información y da la posibilidad de incrementar los indicadores bibliométricos, estos es, si se requiere de un indicador no contemplado basta con crear una vista con la información necesaria de forma que se pueden agregar módulos que resuelvan las necesidades tanto en el caso de nuevas etiquetas como de nuevas consultas.

Por ultimo se realizaron pruebas de funcionamiento y confiabilidad del sistema.

Una vez con el sistema se procedió a la interpretación de datos teniendo los siguientes resultados:

Ejemplos de palabras con frecuencias altas, no representativas (strop words)

Palabra	Frecuencia
a	16107
al	6162
ante	1152
con	9377
de	233025
del	62629
el	43281
el	53839
entre	2226
la	110584
las	18402

lo	772
los	25972
para	19277
por	6493
que	5624
se	1289
su	4935
sus	1864
un	9211
una	8305
y	73054

Ejemplos de palabras más representativas, con valor para estudios bibliométricos.

Palabra	Frecuencia
medio	1,371
estados	1,365
conclusiones	1,354
investigacion	1,342
informacion	1,340
internacional	1,339
ninos	1,319
civil	1,308
recomendaciones	1,307
practico	1,279
problema	1,266
industria	1,259
salud	1,259
federal	1,254
factores	1,067
fiscal	1,063
juicio	1,010
reforma	1,005
humanos	1,003

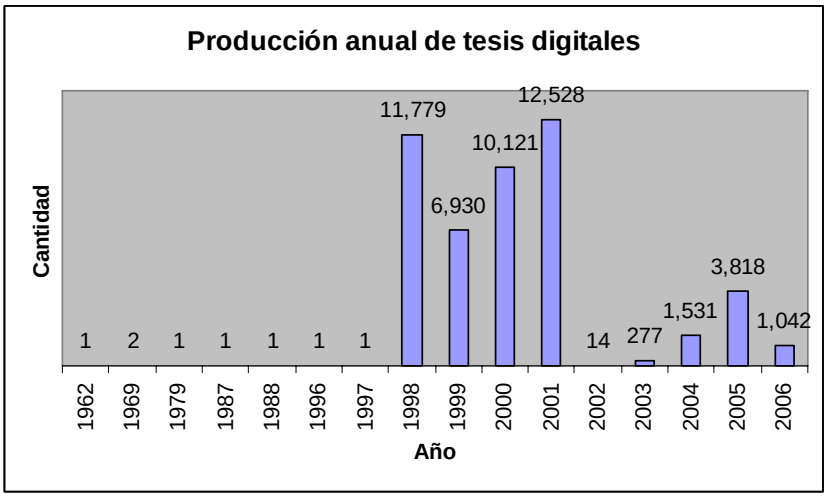
Ejemplos de palabras con frecuencias bajas, no representativas.

Palabra	Frecuencia
lazarsfeld	2
lawler	2
laringotraqueal	2
lacontaminacion	2
juniperus	2
jitomate	2
ixtepeji	2
ixtapaluca	2
acajete-amozoc-tepatlaxco	1
glosotaringeo	1
glicosiltransferasas	1
hikuli	1
nifurtimox	1
reflolux	1
trastocada	1
trashidrogenasa	1
tramadol	1
trajeron	1
tostol	1
termodependencia	1
tenneessee	1

Directores con mayor producción

Director	Frecuencia
Vaquero Cazares, Jose Esteban	137
Cortes y Huerta, Sergio	106
Robles, Francisca	87
Leon Lopez, Maria Graciela	85
Rodriguez Ortiz, Martha	84
Martinez Duran, Maria Elena	83
Aldape Barrios, Beatriz Catalina	82
Avila Ramos, Edy	81
Romero Grande, Gaston	77
Avila Ornelas, Roberto	77
Almazan Alaniz, Jose Antonio	76
Juarez Rojas, Juan Jesus	72

Gráfica de producción de tesis digitales por año



Conclusiones

El crecimiento de la información a nivel nacional e internacional, ha establecido nuevos retos para la correcta administración de la información, áreas del conocimiento como la ciencia de la información, bibliotecología y computación, permiten desarrollar estudios y herramientas para mejorar el registro, la organización, normalización y recuperación de la información.

La carencia de sistemas automatizados para la obtención de indicadores biliométricos en las bases de datos de la Dirección General de Bibliotecas, en particular TESIUNAM lleva al desarrollo del presente trabajo como solución a la necesidad de obtener indicadores bibliométricos, que reflejen la tendencia de las temáticas de las investigaciones que se producen en las tesis de la UNAM.

Uno de los aspectos para el desarrollo del presente trabajo fue conocer la forma en que el sistema ALEPH almacena los registros bibliográficos para su posterior manipulación, y con ello estándares de catalogación para comprender el significado de los registros almacenados. Esto debido a que de forma directa no es posible consultar la información de tales registros y por ello es necesaria su manipulación para tener datos accesibles para estudios bibliométricos.

Uno de los problemas más graves encontrados en la elaboración de los indicadores bibliométricos, fue la falta de normalización de la información, por el crecimiento del volumen de información, la antigüedad de la bases de datos, el registro de la información y la aplicación de diferentes reglas de catalogación, se detectaron errores de normalización y para dar solución a este problema se recomienda seguir las normas de catalogación internacional como el formato MARC 21.

Debido a que el acceso a la información no era de manera directa, el personal del departamento de tesis registraba en bitácora en papel los datos generales de las tesis que ingresan al departamento, de tal forma que si deseaban obtener una estadística, manualmente hacían la búsqueda y el conteo de los datos. Ahora se cuenta con una herramienta de indización y clasificación de registros bibliográficos para determinar indicadores bibliométricos y estudiar el comportamiento de la producción de tesis.

Es de gran ventaja que TESIUNAM haya migrado a la versión 500 del sistema ALEPH ya que se hace uso de una base de datos relacional en plataforma

ORACLE. Una vez conociendo la estructura de los registros bibliográficos y a través del precompilador Pro*C de ORACLE que combina proposiciones SQL con estructuras de un lenguaje de alto nivel como lo es C, se crean aplicaciones eficaces y confiables, como lo es en este caso el procesamiento, recuperación y manipulación de información de la base de datos de tesis, ahorrando tiempo y esfuerzo en el procesamiento de la información bibliográfica y obtención de indicadores bibliométricos.

La forma plantada de indexar los registros bibliográficos y recuperar indicadores bibliométricos permite generar la cantidad de consultas necesarias (y que quizá no hayan sido consideradas inicialmente) que muestren el comportamiento en la producción de tesis ya que no requiere de hacer cambios en las tablas de índices sino que basta con crear una nueva *vista* de las tablas en base a un modelo matemático o en su defecto agregar un nuevo campo de indexación.

La integración de dos áreas del conocimiento como lo es la bibliotecología y la computación, han dado solución a los problemas importantes en el manejo de la información entre ellos la normalización y la recuperación de la información.

Trabajo futuro

Aún hay mucho por hacer para el manejo de la información y nuevos términos que surjan, por tal razón es necesario desarrollar métodos para mejorar el procesamiento de la información y la generación de nuevas expresiones regulares para su validación.

La recepción de documentos de texto completo es un procedimiento que, relativamente va iniciando en DGB, así que aún no se ha trabajado mucho al respecto. Sin embargo la lógica planteada en el presente trabajo es parte medular para el desarrollo de aplicaciones que trabajen bajo este formato, y poder realizar sistemas capaces de hacer búsquedas de información en los contenidos de los documentos, además la aplicación del lenguaje natural en la recuperación de la información es esencial para mejorar los métodos de búsqueda de información en las bases de datos.

Las redes neuronales artificiales, pueden ser utilizadas, para mejorar los esquemas de clasificación y agregar nuevas áreas temáticas, para obtener información más detallada, por ejemplo, dentro del área "Fisicomatemáticas e Ingenierías" podría estar "Bases de Datos" e "Ingeniería de Software" como subáreas.

Apéndice A
Registro Bibliográfico

A continuación se presenta el contenido de un registro bibliográfico con gran cantidad de información, de manera que se puede observar la magnitud y el formato de la información almacenada.

0008FMT LBK0030LDR L-----nam--22------a-45000024008 LS2005 DGB A ESP 00046008 L051024s----------r-----000-0-eng-d0023035 L$$aTDF010006020310026084 L$$a001-00482-G0-200500451002 L$$aGil-Villegas Montiel, Francisco Blas021024510L$$aMax Weber y la guerra de los cien años :$$bAnalisis de la literatura en torno a la polemica centenaria de la tesis de Max Weber sobre :$$bLa etica protestante y el espiritu del capitalismo (1905-2005)0036260 L$$aMexico :$$bEl autor,$$c20050022300 L$$a[13], 1665 p.0104502 L$$aTesis Doctorado (Doctorado en Ciencia Politica)-UNAM, Facultad de Ciencias Politicas y Sociales0049505 L$$aContexto y Objetivos de la Investigacion0089505 L$$aProposiciones Preliminares Para Construir el Criterio de Demarcacion de Analisis0117505 L$$aExposicion y Evaluacion de la Literatura Centenaria en Torno a la Tesis de max Weber Sobre el Protestantismo0059505 L$$aEl Argumento de la Etica Protestante y sus Fuentes0028505 L$$aPasando por Sombart0016505 L$$aScheler0039505 L$$aBrentano y Jaspers (1905-1920)0072505 L$$aF. J. Schmidt en 1905: la Primera Reseña de la Tesis Weberiana0111505 L$$aVon Schultze-Gaevernitz y la Primera Aplicacion en 1906 de la Tesis a un Caso no Contemplado por Weber0052505 L$$aLa Temprana Critica de Karl Fischer en 19070048505 L$$aLa Respuesta de Weber a Fischer en 19080091505 L$$aLa Primera Recepcion de la Tesis Weberiana Fuera de Alemania: P.t: Forsyth en 19100079505 L$$aFelix Rachfahl y la Polemica mas Ruda de la Primera Decada (1909/1910)0099505 L$$aLa Tesis Weber-Troeltsch: una Competencia Entre los dos Expertos de Heidelberg (1911-1912)0052505 L$$aLa Critica de Sombart en el Burgues de 19130053505 L$$aLa Sorprendente Critica en 1915 del Otro Max0062505 L$$aW. Cunningham Adopta en 1914 la Tesis Weber-Troeltsch0046505 L$$aBruno A. Fuchs y su Propuesta de 19140075505 L$$aEmile Doumerge y su Critica de 1917 a la Interpretacion de Calvino0134505 L$$aGeorg von Lukacs (1911) y Thomas Mann (1918): la Presencia de la Tesis Weberiana en la Critica y Produccion Literaria Alemana0070505 L$$aLas Duras Criticas en 1916 del Venerado Maestro Lujo Brentano0075505 L$$aKarl Jaspers y su Psicologia de las Concepciones del Mundo en 19190035505 L$$aOtras Aportaciones de 19190055505 L$$aEl Desconocido Caso de Georg von Below en 19200076505 L$$aLa Cuestion del Papel

de los Judios y los Reduccionismos de Sombart0076505 L$$aErnst Bloch y su Evaluacion del Calvinismo y el Anabaptismo en 19210105505 L$$aMax Weber Como Politico mas que Cientifico en el Historicismo Aleman Tardio de 1922 (Otto Hintze0030505 L$$aTroeltsch y Meinecke)0075505 L$$aVon Schulze-Gaevernitz y el Volumen Conmemorativo de Weber en 19230071505 L$$aLa Tesis de Weber en la Obra Cumbre del Lukacs Marxista (1923)0066505 L$$aLa Critica Marxista Ortodoxa de Karl a. Wittfogel en 19240100505 L$$aEn 1924 Justus Hashagen Cuestiona la Tesis Weberiana Sobre la Base de la Experiencia Renana0077505 L$$aLa Tesis Weberiana en la Sociologia del Saber de Scheler (1922-1924)0129505 L$$aLa Critica de Mannheim a Scheler en 1925 y la Presencia de max Weber en la Edicion Alemana de Ideologia y Utopia de 19290078505 L$$aOtras Aportaciones Alemanas de 1925 (Troeltsch, Wünsch Othmar Spann)0071505 L$$aMax Weber en 1926 Segun las PeRSPectivas de Rickert y Marianne0038505 L$$aLa Critica de Kautsky en 19270069505 L$$aAportaciones de 1928 (Koch, Stoltenberg, Jostock, Schucking)0072505 L$$aLa Interpretacion Existencialista de Siegfried Landshut en 19290049505 L$$aVon Schulze-Gaevernitz Reaparece en 19290095505 L$$aErich Fechner Compara en 1928 y 1929 las Diferenclasy Semejanzas Entre Weber y Sombart0023505 L$$aEscolasticismo0079505 L$$aPuritanismo y Capitalismo en la Critica de 1930 del Jesuita J.B. Kraus0019505 L$$aInglaterra0041505 L$$aFrancia y Estados Unidos: Tawney0050505 L$$aHalbwachs y los Tres Enriques Francofonos0016505 L$$aPirenne0021505 L$$aHauser y See0031505 L$$aPirenne Como Enrique I0065505 L$$aH.G. Wood y la Idea de la Propiedad en la Reforma (1922)0068505 L$$aEl Catolico George O'brien Apoya en 1923 la Tesis Weberiana0120505 L$$aLa Guerra de los Cien Años Llega a Francia Pero sin Juana de Arco (Halbwachs, See, Hauser, Rougier: 1925-1928)0130505 L$$aEl Primer Registro de la Tesis Weberiana en Estados Unidos: H. H. Maurer en el American Journal of Sociology de 1924-19250058505 L$$aLas Tergiversaciones de Pitrim A. Sorokin en 19280018505 L$$aFullerton0037505 L$$aEl Teologo de Harvard (1928)0050505 L$$aLa Deformacion de Frank H. Knight en 19280070505 L$$aLa Tesis Weber-Tawney y el Caso del Puritanismo Ingles (1926)0124505 L$$aLos Ensayos de 1929 de Parsons y la Recepcion Critica de su Traduccion de la Etica Protestante de Max Weber en 19300025505 L$$aLondres y Mexico0106505 L$$aPasando por la Italia de Mussolini: el Problema de la Causalidad y las Relaciones con el Marxismo0048505 L$$aUna Decada Rica y Fructifera Como Pocas0047505 L$$aOtto Hintze en 1931: el Weber Prusiano0068505 L$$aLa Sociologia del Renacimiento de Alfred von Martin en 19320056505 L$$aLa Critica Heideggeriana de Karl Lowith en 19320075505 L$$aLa Critica mal Fundamentada del Sudafricano H.M. Robertson en 19330042505 L$$aTalcott Parsons al Rescate (1935)0074505 L$$aPrejuicios y Pretensiones del Teorico Incurable de Harvard (1937)0065505 L$$aLa Critica Fascista del Catolico Amitore Fanfani en 19340053505 L$$aLa Critica Marxista de los Años 30: Gramsci0017505 L$$aBorkenau0023505 L$$aLaski y Walker0071505 L$$aLa Critica de la Escuela de los Annales: Lucien Febvre en 19340103505 L$$aRobert K. Merton Extiende la

Tesis de Weber al Ambito de la Historia de la Ciencia (1936-1939)0059505 L$$aRaymond Aron Salva el Honor de Francia (1935-1938)0068505 L$$aMexico 1939: Entrada con el pie Izquierdo o Salida en Falso0080505 L$$aItalia 1940: la Magistral Contribucion de la Escuela de Benedetto Croce0042505 L$$aA la Etica Protestante en Noruega0101505 L$$aLas Apropiaciones Marxistas y Funcionalistas y la Critica de los Historiadores de la Iglesia0052505 L$$aErich Fromm y el Miedo a la Libertad (1941)0111505 L$$aJoseph Schumpeter Interpreta en 1942 de Manera Determinista la Tesis Weberiana Sobre el Protestantismo0078505 L$$aMexico de 1943 a 1945: los Traductores Interpretan la Tesis Weberiana0075505 L$$aEphraim Fischoff en 1944: Primera Reseña Historica de la Polemica0076505 L$$aAlbert Salomon en 1945 y el Dialogo Perenne con el Fantasma de Marx0061505 L$$aMaurice Dobb en 1946: una Critica Marxista Adicional0070505 L$$aGerth y Mills en 1946: la Alternativa al Monopolio de Parsons0075505 L$$aNoruega en 1947: Confirmacion de la Tesis o Extrapolacion Invalida0069505 L$$aUna Valiosa Aportacion Desde el Marxismo: leo Kofler en 19480054505 L$$aParsons Insiste en 1948: Weber era Parsoniano0071505 L$$aUn Weber Reclamado Tanto por Marxistas Como por Funcionalistas0054505 L$$aBenjamin Nelson y la Idea de la Usura en 19490068505 L$$aWeber y su Talon de Aquiles: la Critica de la Teologia 19490048505 L$$aAportaciones de Paul Honigsheim en 19500042505 L$$aDos Decadas Ricas en Traducciones0035505 L$$aLefort Samuelsson y Bieler0079505 L$$aAl Debate de los Historiadores de Oxford Sobre el Ascenso de la Gentry0033505 L$$aLa Decada de 1951 a 19600055505 L$$aLa Critica del Historiador Albert Hyma en 19510074505 L$$aLas Insuficienclasde la Critica Marxista de Claude Lefort en 19520067505 L$$aLa Critica de leo Strauss en 1953 al Historicismo de Weber0070505 L$$aNorman Birnbaum Aborda de Nuevo la Relacion de Weber con Marx0057505 L$$aGeorg Lukacs Fragua en 1954 su Asalto a la Razon0116505 L$$aLa Refinada Aportacion de Merleau-Ponty en 1955 a la Historia de la Dialectica y sus Emocionantes Aventuras0108505 L$$aLa Comicidad Involuntaria de un Articulo de 1955 Sobre la Etica Vikinga y el Origen del Capitalismo0138505 L$$aLa Polemica de los Años 50 en Torno a la Tesis de Merton Sobre la Relacion Entre el Protestantismo y el Desarrollo de la Ciencia0132505 L$$aEl Balance de Pietro Rossi en 1956 Sobre el Historicismo Aleman y su Deficiencia en la Interpretacion de la Tesis Weberiana0056505 L$$aLas Tergiversaciones de Recasens Siches en 19560069505 L$$aLimitaciones de la Critica del Sueco Kurt Samuelsson en 19570072505 L$$aLa Critica de Niles Hansen a las Simplificaciones de Samuelsson0060505 L$$aLa Correcta Interpretacion de Stuart Hughes en 19580075505 L$$aLa Critica de los Historiadores Charles y Catherine George en 19580069505 L$$aRaymond Aron Pone Orden con su Critica de 1959 a Leo Strauss0104505 L$$aAndre Bieler Insiste en 1959 en la Correcta Interpretacion del Pensamiento Economico de Calvino0096505 L$$aTenbruck Inicia en 1959 una Reinterpretacion Genetico-Evolutiva de la Obra de Max Weber0091505 L$$aEl Balance de Medio Siglo de Polemica en la Compilacion de Robert W. Green en 19590054505 L$$aSidney A.

Burell en 1960 y el Caso de Escocia0047505 L$$aLa Sintesis de Reinhard Bendix en 19600117505 L$$aEl Debate de los Historiadores de Oxford Sobre las Causas de la Revolucion Inglesa y el Ascenso de la Gentry0057505 L$$aLa Tesis de Tawney Sobre el Ascenso de la Gentry0079505 L$$aEl Salvaje Ataque de Trevor-Roper a las Estadisticas de Lawrence Stone0081505 L$$aTrevor-Roper Ataca a la Escuela de Tawney por Aceptar la Tesis Weberiana0071505 L$$aEl Contrataque de Christopher Hill y Perez Zagorin (1958-1959)0072505 L$$aOxford une Fuerzas Contra la Critica Forastera de Hexter (1958)0096505 L$$aLa Tesis de Weber en la Interpretacion de Christopher Hill Sobre la Revolucion Puritana0079505 L$$aLa Tesis Alternativa de Trevor-Roper a la Explicacion Weberiana (1961)0143505 L$$aLa Victoria de la Escuela de Tawney Equivale a la de Enrique v en la Gloriosa Batalla de Agincourt Durante la Guerra de los Cien Años0088505 L$$aPasando por el uso de la Tesis Weberiana en los Estudios Sobre la Modernizacion0078505 L$$aMcclelland y su Tesis de 1961 Sobre la Sociedad del Logro Adquisitivo0064505 L$$aLa Critica de Macintyre en 1962 a la Explicacion Causal0062505 L$$aLa Extrapolacion Ilogica de Robert E. Kennedy en 19620077505 L$$aEl Caso del Armianismo en Holanda en la Ponencia de Wertheim en 19620118505 L$$aRobert N. Bellah Discute en 1962 la Posibilidad de Extender la Tesis Weberiana Sobre el Protestantismo a Asia0123505 L$$aLa Ilegitima Extrapolacion de la Tesis Weberiana en la Sociologia Empirica de los Estados Unidos Durante los Años0138505 L$$aLa Tesis de Michael Walzer Sobre la Revolucion de los Santos Puritanos y la Deuda Politologica con la Tesis Weberiana (1963-1966)0093505 L$$aCharles H. George Identifica en 1968 a Walzer Como un Weberiano con Falsa Conciencia0059505 L$$aLa Critica de Quentin Skinner a la Tesis de Walzer0026505 L$$aTawney vs. Walzer0070505 L$$aLa Celebracion del Centenario del Nacimiento de Weber en 19640110505 L$$aLa Discusion de la Clasificacion de los Judios Como Pueblo Paria en el Coloquio de Heidelberg de 19640087505 L$$aAbramowski y la Interpretacion Historico-Universal en la Obra de Weber en 19660099505 L$$aEl Pensamiento Historico Universal de Max Weber en la Interpretacion de Wolgang J. Mommsen0039505 L$$aWeber en Alemania Durante 19670077505 L$$aLa Limitada Vision de Habermas del Concepto de Racionalidad en Weber0054505 L$$aLa Sociedad Cortesana (1969) de Norbert Elias0069505 L$$aMommsen Evalua en 1970 la Literatura Sobre Weber en Alemania0049505 L$$aEl Centenario de Weber en Estados Unidos0059505 L$$aBendix y la Paradoja de la Racionalizacion en 19650053505 L$$aBellah y la Teoria de la Evolucion Religiosa0050505 L$$aHelmut Wagner y el Balance de Medio Siglo0057505 L$$aLas Investigaciones de Richard Means (1964-1966)0081505 L$$aLa Delimitacion Historica de la Validez de la Tesis Segun Hansen en 19670060505 L$$aLa Indignacion Catolica de Werner Stark (1964-1968)0057505 L$$aEn 1968 Forcese Intenta Poner Orden en el Debate0064505 L$$aS.N. Eisenstadt y la Teoria de la Modernizacion en 19680014505 L$$aOrden0067505 L$$aReligion y Derecho en la Tesis Doctoral de David Little en0061505 L$$aBenjamin Nelson y la Tesis mas Alla de Weber en 19690062505 L$$aLa Biografia Psicoanalitica de Arthur

Mitzman en 19690079505 L$$aGouldner y la Crisis Orgiastica de la Sociologia Estadunidense en 19700070505 L$$aStephen Warner y la Refraccion de Intereses Materiales (1970)0079505 L$$aEl Centenario de Weber en Mexico y Argentina: Poviña y Sanchez Azcona0056505 L$$aFerraroti en Italia y Moya Valgañon en España0035505 L$$aWeber en Holanda y Polonia0052505 L$$aSuiza y Monaco en la Obra de Herbert Lüthy0070505 L$$aDel Centenario en Francia a la Compilacion de Besnard en 19700074505 L$$aLa Traduccion de j. Chavy en 1964 y la Critica de Mandrou en 19660047505 L$$aEl Gran Libro de Julien Freund en 19660059505 L$$aLos Articulos de Jean-Marie Vincent de 1967 y 19680043505 L$$aLa Version de Raymond Aron en 19670061505 L$$aJean Baechler y su Lectura Marxista de Weber en 19680058505 L$$aLas Confusiones Marxistas de Joseph Gabel en 19690071505 L$$aLa Esplendida Antologia Compilada por Philippe Besnard en 19700063505 L$$aMax Weber en Gran Bretaña Durante la Decada de los 600108505 L$$aAndreski Pone en 1964 a la Tesis Weberiana en el Contexto de la Sociologia de la Religion Comparada0069505 L$$aOtra vez el Debate de los Historiadores Ingleses (1964-1969)0068505 L$$aEn 1969 W.g. Runciman Aprueba a Weber y Reprueba a Durkheim0110505 L$$aThelma Mccomarck Descubre en 1969 la Relacion Entre la Etica Protestante y el Espiritu del Socialismo0067505 L$$aGiddens Replantea en 1970 las Relaciones de Weber con Marx0079505 L$$aEl Acercamiento a Marx y la Discusion de los Casos del Islam y Escocia0096505 L$$aGiddens Niega en su Libro de 1971 que la Tesis Weberiana Proponga una Imputacion Causal0057505 L$$aLa Coleccion de Ensayos de Bendix y Roth en 19710066505 L$$aLenski y la Extrapolacion Ilegitima de la Tesis Weberiana0034505 L$$aEisenstadt en 1971 y 19730053505 L$$aEl Balance de la Polemica Segun Robert Moore0056505 L$$aLa Propuesta Teorica de Jose Luis Reyna en 19710078505 L$$aLa Tesis Weberiana y la Historia del Protestantismo en Estados Unidos0082505 L$$aStephen Berger Resalta en 1971 la Centralidad del Ensayo Sobre las Sectas0077505 L$$aLa Tesis Weberiana Como Explicacion Historica Segun Sprinzak en 19720064505 L$$aJean Cohen y la Dinamica de la Dominacion Racionalizada0046505 L$$aLa Version de J.e.t. Eldridge en 19720066505 L$$aEl Segundo Balance de la Polemica de Robert Green en 19730052505 L$$aLa Comparacion de Weber con Lucien Goldmann0059505 L$$aLas Aportaciones de Benjamin Nelson en 1973 y 19740050505 L$$aWeber y el Islam: Bryan s. Turner en 19740077505 L$$aAfinidades Electivas y Nueva Ideologia Empresarial Segun Alan Winter0045505 L$$aLa Desparsonizacion de Weber en 19750073505 L$$aDos Contrastantes Versiones en 1975 de las Relaciones Weber-Marx0069505 L$$aWeinryb y el Analisis de la Causalidad en la Tesis Weberiana0078505 L$$aLas Contradicciones Culturales del Capitalismo de Daniel Bell en 19760065505 L$$aJapon y el Espiritu del Capitalismo Segun Otsuka en 19760074505 L$$aCriticas Estructuralistas y Naturalistas en Gran Bretaña en 19770077505 L$$aLa Tesis Weberiana y el Patrimonialismo Islamico en mi Tesis de 19770072505 L$$aLa Primera Exposicion Fidedigna de la Tesis Weberiana en Mexico0079505 L$$aUna Exposicion Deficiente en 1980 en la Revista Mexicana de

Sociologia0052505 L$$aDos Aportaciones Filologicas de 1978 y 19790056505 L$$aLa Critica Simplista de Fernand Braudel en 19790081505 L$$aMarshall Defiende en 1979 la Validez de la Tesis Para el Caso de Escocia0068505 L$$aJere Cohen Descubre en 1980 al Capitalismo del Renacimiento0057505 L$$aHolton Refuta a la Critica de Cohen Contra Weber0017505 L$$aSprondel0016505 L$$aMommsen0030505 L$$aTenbruck y Schluchter0059505 L$$aHartmut Lehmann Discute en 1972 la Tesis Weberiana0066505 L$$aGünther dux y su Interpretacion Neoevolucionista de 19710070505 L$$aLa Importancia de la Antologia de 1973 de Seyfarth y Sprondel0080505 L$$aLa Perspectiva Historicista y Antievolucionista de W.j. Mommsen en 19740074505 L$$aWeib y la Fundamentacion Neoevolutiva de la Obra de Weber en 19750027505 L$$aFriedrich Tenbruck0043505 L$$aEl Weberologo Incomodo (1975-1978)0072505 L$$aWolfgang Schluchter y la Paradoja de la Racionalizacion en 19760081505 L$$aEl Gran Libro de Schluchter de 1979 y su Reconstruccion Minimo Evolutiva0076505 L$$aLepsius y el Coloquio de Constanza de 1977 Sobre la Racionalizacion0072505 L$$aMünch y la Anatomia del Racionalismo Occidental en 1978 y 19800074505 L$$aWinckelmann Responde Entre 1979 y 1980 a las Criticas de Tenbruck0078505 L$$aLa Defensa de los Discipulos de Winckelmann y sus Criticas a Tenbruck0056505 L$$aLa Deformada Interpretacion de Habermas en 19810075505 L$$aGordon Marshall y su Busqueda del Espiritu del Capitalismo en 19820083505 L$$aFrank Parkin Encuentra en 1982 dos Tesis en la ep: una Fuerte y Otra Debil0069505 L$$aLos Tres Terminos de la Tesis Segun Gianfranco Poggi en 19830068505 L$$aEl Reduccionismo Neoparsoniano de Jeffrey Alexander en 19830073505 L$$aMax Weber y su Sombra Segun el Balance de Gonzalo Massot en 19830015505 L$$aMexico0065505 L$$aLa ep y la Racionalidad en la Obra de Weber: 1982 y 19840078505 L$$aEl Tema de la Racionalidad y el Renacimiento de Weber en Norteamerica0073505 L$$aTres Volumenes Sobre las Relaciones de Weber con Marx: 1985-19870075505 L$$aRandall Collins Descubre dos Teorlasdel Capitalismo en Weber: 19860069505 L$$aWilhelm Hennis Descubre en 1987 la Tematica Central de Weber0066505 L$$aDesacuerdos Entre Mommsen y Hennis en el Simposio de 19870077505 L$$aDos Balances de 1987 Sobre el Estado de la Polemica en Torno a la ep0063505 L$$aAlan Bloom y la Cerrazon de la Mente Americana en 19870051505 L$$aLa Polemica en la Revista Telos: 1988-19890065505 L$$aPellicani Declara Liquidada en 1988 a la Tesis Weberiana0066505 L$$aGuy Oakes le Responde y Exhibe la Ignorancia de Pellicani0122505 L$$aPiconne Tercia con una Teoria del Complot Para Explicar el Exito de la Tesis Weberiana Sobre la Etica Protestante0049505 L$$aPellicani Mismo Responde a Oakes en 19890067505 L$$aLas Cuatro Preguntas con las que Oakes Reprobo a Pellicani0059505 L$$aWalter Wallace y los dos Espiritus del Capitalismo0075505 L$$aLa Polemica Desatada por la Critica Teologica de Mackinnon en 19880048505 L$$aTres Aportaciones mas Entre 1988 y 19890070505 L$$aLa Oruga y el Aguila: Alan Sica y Lawrence Scaff en 1988-19890077505 L$$aLuis f. Aguilar: la Gran Aportacion a la Idea de la Ciencia en Weber0058505 L$$aLa Ambivalente Critica de Jacques le

Goff en 19900047505 L$$aLepsius y el Guardagujas Falso en 19900066505 L$$aPrimero Nietzsche y Despues Marx: Hartmann Tyrell en 19900033505 L$$aLas Aportaciones de 19910064505 L$$aLa Tesis en 1992 o Sociologia Como Teoria de la Cultura0073505 L$$aMackinnon Hace en 1993 un Insatisfactorio Balance de la Polemica0076505 L$$aLa Etica Catolica y el Espiritu del Capitalismo Segun Novak en 19930035505 L$$aOtras Aportaciones de 19930075505 L$$aLa Critica de Disselkamp en 1994 a las Fuentes Teologicas de Weber0080505 L$$aKalberg Subraya en su Libro de 1994 el Analisis Multifactorial de Weber0057505 L$$aAportaciones Filologicas y Culturalistas de 19940025505 L$$aDe Nuevo Käsler0036505 L$$aHennis y Schluchter en 19950048505 L$$aEl Estudio Introductorio de Dirk Kasler0048505 L$$aLas Reseñas Criticas de Hennis en 19950054505 L$$aSchluchter y la Modernidad Implacable de 19950061505 L$$aLa Pertinente Clarificacion de David Gellner en 19950076505 L$$aRichard Hamilton y su Desconstruccion Social de la Realidad de 19960016505 L$$aKalberg0046505 L$$aGrossein y Ruano de la Fuente en 19960065505 L$$aLa Revaloracion de Kalberg de la ep Como Tratado Teorico0127505 L$$aLa Traduccion Francesa en 1996 de los Escritos Teoricos de la Sociologia de la Religion de Weber por Parte de Grossein0073505 L$$aLa Conciencia Tragica del Ethos Moderno Segun Ruano de la Fuente0063505 L$$aRinger Demuestra en 1997 la Imputacion Causal de la ep0072505 L$$aLa Compilacion del Groupe de Recherche sur la Culture de Weimar0078505 L$$aLa Extrapolacion de la Tesis a Varios Casos no Contemplados por Weber0025505 L$$aLas Estadisticas0049505 L$$aLas Metaforas y la Epistemologia en 19970049505 L$$aOtra vez las Estadisticas de Offenbacher0051505 L$$aOtra vez la Metafora de la Jaula de Hierro0062505 L$$aManuel gil Anton y su Critica Epistemologica de Weber0075505 L$$aGonzalez Leon Aborda en 1998 el Debate Aleman Sobre el Capitalismo0063505 L$$aLa Introduccion de 1998 de Friedhelm Guttandin a la ep0116505 L$$aRenacimiento y Reforma Segun Jorge Velasquez en 1998 y el Articulo de Troeltsch de 1913 Sobre el Mismo Tema0013505 L$$aLowy0036505 L$$aLandes y Schluchter en 19980071505 L$$aLa Sociologia Economica de Weber Segun Swedberg en 1998 y 19990073505 L$$aLa Critica de Grossein en 1999 a la Traduccion de Chavy de la ep0039505 L$$aLas Otras Aportaciones de 19990040505 L$$aLa Edicion Critica de die Stadt0093505 L$$aLa Edicion de Klaus Lichtblau de los Textos de max Scheler Sobre Etica y Capitalismo0066505 L$$aLa Eterna Influencia de la Metafora de la Jaula de Hierro0049505 L$$aEl Desembalsamamiento Finlandes de Weber0043505 L$$aLa Influencia de Jellinek en Weber0070505 L$$aLa Compilacion Postuma en 1999 de los Escritos de f. Tenbruck0064505 L$$aUn Deficiente Balance en 2000 del Estado de la Polemica0068505 L$$aLa Nueva Traduccion Francesa de Isabelle Kalinowsky en 20000081505 L$$aHarrison y Huntington Afirman: la Cultura si Importa y Weber Tenia Razon0080505 L$$aLa Racionalidad y el Desencantamiento en la Tesis de Ramos Lara de 20000061505 L$$aLa Aparicion en 2000 de la Revista max Weber Studies0071505 L$$aGeorge Ritzer y la Tesis de la Macdonaldizacion de la

Sociedad0074505 L$$aLa Tesis Sobre el Protestantismo en los max Weber Studies de 20010074505 L$$aEn España Beriat Señala el Componente Emocional del Capitalismo0130505 L$$aDos Nuevos Tomos de la Edicion Critica Alemana de Economia y Sociedad y la Religionssystematik de Hans Kippenberg en 20010056505 L$$aLas Traducciones de Chalcraft y Abellan en 20010071505 L$$aLa Edicion Critica Britanica de la ep de Baher y Wells en 20020060505 L$$aLa Edicion Estadunidense de Stephen Kalberg en 20020073505 L$$aLa Biografia Intelectual de max Weber por Michael Sukale en 20020058505 L$$aEl Rematch del Desparsonizador Jere Cohen en 20020015505 L$$aRinger0014505 L$$aAtali0109505 L$$aLos Hackers y la Traduccion al Español en 2002 del Antikritisches Schlußwort de Weber en la Rmcpys0072505 L$$aLa Primera Edicion Critica de la ep en Mexico Publicada en 20030031505 L$$aLas Reseñas de Sabido0021505 L$$aPerez Franco0041505 L$$aBallesteros Leiner y Zabludovsky0074505 L$$aLa Traduccion al Ingles en 2003 de la Tesis Doctoral de max Weber0100505 L$$aPhillip Gorski usa la Tesis Weberiana en 2003 Para Explicar la Formacion del Estado Moderno0071505 L$$aEl Paradigma Weber del Coloquio de Heidelberg de Abril de 20030018505 L$$aTucidides0062505 L$$aLa ep y la Guerra del Peloponeso Segun Hennis en 20030065505 L$$aLa Edicion Critica Francesa de Grossein de la ep en 20040071505 L$$aWeber y la Sociedad Posmoderna en el Analisis de Nicholas Gane0058505 L$$aKieran Allen en 2004: un Burdo Retroceso Marxista0077505 L$$aLa Nocion de Profesion en Weber Segun la Tesis de Ballesteros Leiner0094505 L$$aWeber y la Confrontacion con el Mundo Irracional en la Interpretacion de Roger Bartra0073505 L$$aSica y la Tesis Weberiana en el Mundo Periodistico del Siglo XXI0065505 L$$aFritz Ringer y su Biografia Intelectual de Weber en 20040128505 L$$aFerguson Encuentra en 2004 Apoyo Para la Tesis Weberiana en la Superioridad Economica de Estados Unidos Frente a Europa0078505 L$$aLa Conmemoracion en Mexico a Fines de 2004 de los Cien Años de la ep0048505 L$$aLa Reconstruccion de Schluchter en 20050074505 L$$aLa Confrontacion de Weber con la Modernidad Segun Kalberg en 20050025505 L$$aRichard Swedberg0067505 L$$aLa Sociologia Economica y el Diccionario max Weber en 20050073505 L$$aFrancis Fukuyama Celebra en Marzo de 2005 el Centenario de la ep0076505 L$$aTres Articulos de 2005 Discuten la Actualidad de la Tesis Weberiana0104505 L$$aLa Conferencia de Londres del 11 y 12 de Junio de 2005 y el Proximo Numero de max Weber Studies0105505 L$$aEl Centenario de la Tesis Weberiana en Estados Unidos: el Volumen Compilado por Kaelber y Swatos0057505 L$$aLa Decepcionante Contribucion de Hartmut Lehmann0065505 L$$aRiesebrodt y las Tres Dimensiones de la Lectura de la ep0106505 L$$aNielsen Descubre la Cuarta Dimension en la Gran Narrativa Historico Filosofica Subyacente a la ep0118505 L$$aEl Romantico Enfrentmiento de max Weber con Indios y Vaqueros en su Viaje del Otoño de 1904 a Estados Unidos0070505 L$$aKalberg y la Actualizacion de la Tesis de Daniel Bell de 19760115505 L$$aLa Incompleta Relacion de Swatos y Kivisto Sobre la Recepcion de la Tesis Weberiana en el Mundo Anglosajon0105505 L$$aLutz Kaelber y los

Estudios Historicos que Confirman la Validez Cientifica de la Tesis Weberiana0101505 L$$aGorski Complementa la Tesis con su Analisis Estructural de la Pequeña Divergencia en Europa0069505 L$$aLa Renovada Vigencia de la Tesis Weberiana Para el Siglo XXI0019505 L$$aHeidelberg0121505 L$$aTroeltsch y Weber: el Centenario de la Tesis Weberiana en Alemania en el Volumen Compilado por Schluchter y Graf0060505 L$$aLos Contactos Familiares de Weber en Estados Unidos0136505 L$$aLa Tesis de Lehmann Sobre la Transicion de Weber del Protestantismo Cultural al Analisis Cientifico del Protestantismo Ascetico0125505 L$$aLa Interpretacion Causalista y Cientifica de Schluchter Sobre la Manera en que las Ideas Alcanzan Eficacia Historica0039505 L$$aLa Sociologia del Conocimiento0106505 L$$aEl Grupo Eranos y la Primera Exposicion Oral de la Celebre Tesis en Heidelberg en Febrero de 20050016505 L$$aGothein0015505 L$$aSimmel0052505 L$$aSombart y Jellinek Como Modelos Precursores0056505 L$$aSobre la Historik y las Travesuras de Troeltsch0120505 L$$aWeber y Troeltsch: el Conflicto Valorativo del Politico Realista vs. La Sintesis Cultural del Teologo Idealista0075505 L$$aSobre la Historia de la Recepcion de la Tesis Weberiana en Francia0122505 L$$aMexico y la Aportacion mas Completa Sobre el Tema al Cumplirse en 2005 el Primer Centenario de la Tesis Weberiana0024505 L$$aEl Autorretrato0065505 L$$aConclusiones Sobre las Diversas Maneras de Leer el Texto0130505 L$$aEl Misterio de la Equivoca Lectura de la Tesis Weberiana: un Caso Para la Investigacion de la Sociologia del Conocimiento0020505 L$$aQuo Vadimus0047505 L$$aBibliografia Citada en la Introduccion0036505 L$$aBibliografia de 1905 a 19200060505 L$$aBibliografia de la Decada de los Veinte en Alemania0066505 L$$aBibliografia de la Decada de los Veinte Fuera de Alemania0036505 L$$aBibliografia de 1931 a 19400036505 L$$aBibliografia de 1941 a 19500091505 L$$aBibliografia de la Decada de los 50 y de la Polemica de los Historiadores Ingleses0044505 L$$aBibliografia de la Decada de los 600062505 L$$aBibliografia de la Decada de los 70 Fuera de Alemania0053505 L$$aBibliografia de la Decada los 70 en Alemania0044505 L$$aBibliografia de la Decada de los 800044505 L$$aBibliografia de la Decada de los 900036505 L$$aBibliografia de 2000 a 20050010590 L$$a3004470021L$$aSirvent Gutierrez, Carlos,$$easesor009371021L$$aUniversidad Nacional Autonoma de Mexico.$$bFacultad de Ciencias Politicas y Sociales00918564 L$$uhttp://132.248.9.9:8080/tesdig/Procesados_2005/0602031/Index.html$$yTexto completo0043CAT L$$aDGB$$b00$$c20051024$$lTDF01$$h00000032CAT L$$c20060508$$lTDF01$$h19230032CAT L$$c20060508$$lTDF01$$h19490032CAT L$$c20060508$$lTDF01$$h20200032CAT L$$c20060508$$lTDF01$$h20450049CAT L$$aBATCH-UPD$$b00$$c20060615$$lTDF01$$h20220032CAT L$$c20061011$$lTES01$$h19360032CAT L$$c20061011$$lTES01$$h19580032CAT L$$c20061012$$lTES01$$h20530032CAT L$$c20061012$$lTES01$$h20590032CAT L$$c20061117$$lTES01$$h19200032CAT L$$c20061124$$lTES01$$h13470032CAT L$$c20070302$$lTES01$$h23240012039 L$$aDIG

La figura A.1. muestra el registro ejemplo visto desde el módulo de catalogación del sistema ALEPH.

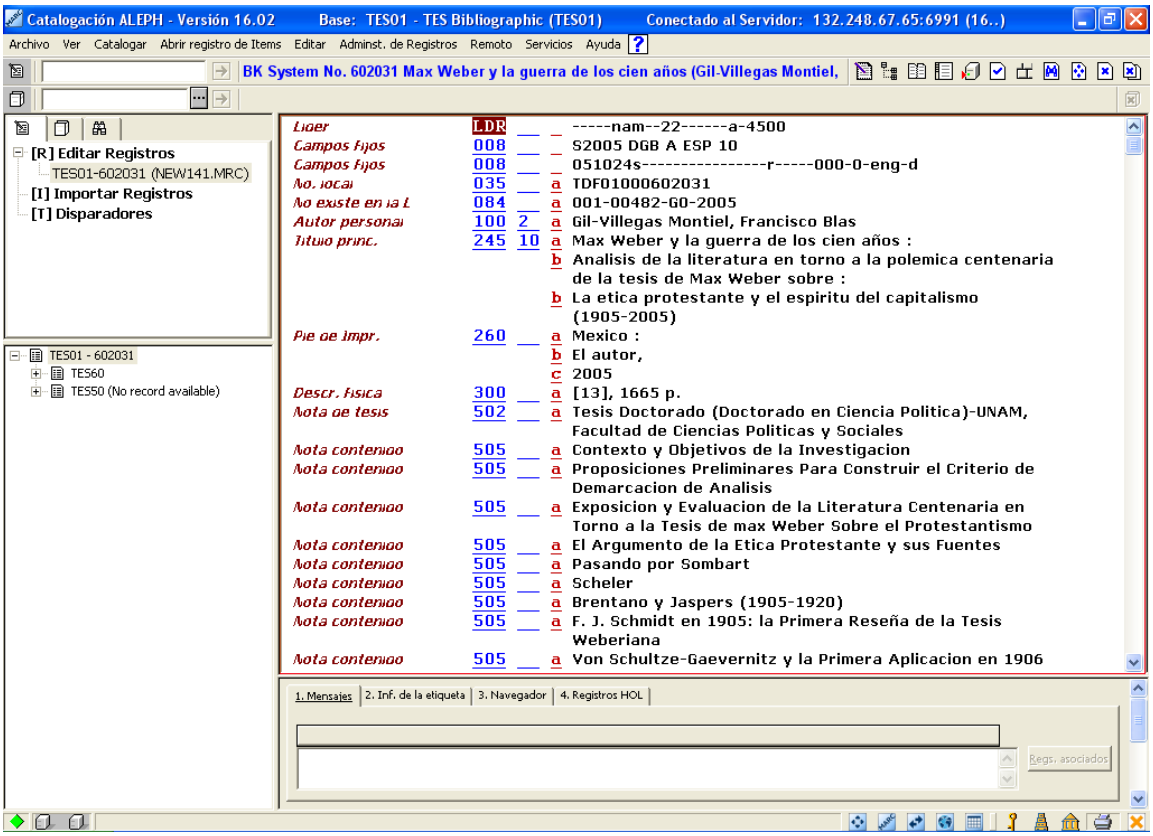


Figura A.1. Módulo de catalogación del sistema ALEPH 500 v16

Glosario

A

ASCII: "Código Estándar Americano para Intercambio de Información". Acuerdo utilizado para representar caracteres, cifras, signos especiales y caracteres de comandos.

Atributo: Es una columna de una tabla.

B

Bibliometría: Es la aplicación de métodos estadísticos y matemáticos dispuestos para definir comportamientos y tendencias de la comunicación escrita mediante técnicas de recuento y análisis de dicha comunicación.

C

Campo: En bases de datos, está representado por las columnas de una tabla.

Campo MARC: Es una unidad lógica de un registro catalográfico. Hay un campo para el autor, uno para el título, etc.

Cardinalidad: Número de tuplas de una tabla.

Char: Un tipo de dato de ORACLE.

Clave primaria: Columna o columnas identificadas para definir e forma única cada fila de una tabla.

D

Disparador: Son bloques nominados de PL/SQL con secciones declarativa, ejecutable y de tratamiento de excepciones.

Dominio: Conjunto de valores de una tabla.

E

Etiqueta MARC: Es un número de tres dígitos asociado a cada campo MARC.

G

Grado: Número de atributos de una tabla.

I

Independencia física de los datos: Es la inmunidad de las aplicaciones a cambios en la representación física y la técnica de acceso.

Indicador bibliométrico: Son datos estadísticos para medir resultados de la actividad científica.

Índice: Clasificación lógica que acelera el tiempo de acceso a los datos de una tabla.

M

MARC: *Machine Readable Cataloging,* es un formato estándar para la representación y comunicación de información bibliográfica

P

PL/SQL: lenguaje de "procedimientos" en SQL que permite construir bloques de programación o procedimientos.

R

Registro catalográfico: información que tradicionalmente se presenta en una ficha de catálogo de biblioteca.

Registro de una tabla: Representa elementos individuales de la tabla.

S

SQL: *Structured Query Language,* es el lenguaje estándar para trabajar con bases de datos relacionales y es soportado prácticamente por todos los productos del mercado.

Stop Word: Palabra que no aporta significado a estudios, tal como, artículos, adjetivos, etc.

Subcampo MARC: Es un tipo de dato dentro de un campo MARC.

T

Tecnologías de información: hacen referencia a la utilización de medios informáticos para almacenar, procesar y difundir todo tipo de información o proceso de formación educativa.

Término: Alguna forma de contenido de información.

Tesis digitales: Documento en formato electrónico.

Tupla: Fila de una tabla.

U

UML: Lenguaje estándar para el modelado de software.

V

Vista: Representación lógica de una tabla.

Bibliografía

[1] *Fundamentos de bases de datos*. Silberschatz Abraham. McGraw-Hill

[2] *Introducción a los sistemas de bases de datos*. C. J. Date. Pearson Educación. 2000.

[3] *Introduction to modern information retrieval*. Gerard Salton. McGraw-Hill.

[4] *Oracle 8i programación avanzada con PL/SQL*. Scout Urman. Mcgraw-Hill

[5] *Building Oracle XML Applications*. Steve Muench. O'Reilly,2000

[6] *The evolution of the interdisciplinarity of information science. A bibliometric study*. Al-Sabbagh, Imad *A.* Ann arbor, michigan : University Microfilms International, 1992

[7] *Clustering for data mining a data recovery approach.* Boris Mirkin. Chapman & Hall/CRC. 2005

[8] Josefa Marín Fernández. *Estadística aplicada a las ciencias de la documentación*. Diego Marin, librero editor. Murcia 2000

[9] *Los indicadores bibliométricos. Fundamentos y aplicación al análisis de la ciencia*. Bruno Maltrás Barba. Ediciones Trea, S. L. España 2003.

[10] *El modelo matemático de Lotka: Su aplicación a la producción científica latinoamericana en ciencias bibliotecológica y de la información*. Gorbea Portal, Salvador**.** México: UNAM, Centro Universitario de Investigaciones Bibliotecológicas, 2005

[11] *Modelo bibliográfico basado en formatos de intercambio y en normas internacionales orientado al control bibliográfico universal*. Roberto Garduño Vera. CUIB-UNAM, 1996.

[12] *Modelo teórico para el estudio métrico de la información documental*. Salvador Gorbea Portal, ediciones TREA, España 2005.

[13] *Measuring information: an information services perspective.* Tague-Sutcliffe, Jean, San Diego : Academic, 1995.

[14] Contreras Barrera, Marcial. *Reconocimiento del parlante utilizando cuantización vectorial y redes neuronales de tipo LVQ*. Tesis de Maestría (Maestría en Ingeniería Eléctrica) – UNAM, Facultad de Ingeniería. México, 2002.

[15] Sierra Flores, Maria Magdalena. *Identificación y estudio de los principales grupos de investigación de campo de la física de la UNAM a través de indicadores bibliométricos***.** Tesis de Maestría (Maestría en Bibliotecología y Estudios de la Información) - UNAM, Facultad de Filosofía y Letras. México, 2005.

[16] Madera Jaramillo, Maria de Jesús. *Una aplicación Web para la obtención de indicadores bibliométricos en ciencia y tecnología.* Tesis de Maestría (Maestría

en Ciencias de la Computación) - UNAM, Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas. México, 2003.

[17] Arellano Sandoval, Gustavo Adolfo. *Integración del contexto técnico y tecnológico al proceso de desarrollo para la generación de Software con calidad*. Tesis de Maestría (Maestría en Ciencias de la Computación) - UNAM, Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas México, 2003.

[18] Language Independent Morphological Analysis, Tatsuo Yamashita, Yuji Matsumoto, April 2000, Proceedings of the sixth conference on Applied natural language processing, Publisher**:** Morgan Kaufmann Publishers Inc.

[19] *One tokenization per source*, Jin Guo, August 1998, Proceedings of the 36th annual meeting on Association for Computational Linguistics - Volume 1, Proceedings of the 17th international conference on Computational linguistics - Volume 1, Publisher: Association for Computational Linguistics.

[20] *Term-Specific Smoothing for the Language Modeling Approach to Information Retrieval: The Importance of a Query Term*. Djoerd Hiemstra. University of Twente.

[21] *Natural language techniques for intelligent information retrieval*. Paul S. Jacobs, Lisa F. Rau, NY

[22] *Conceptual information retrieval*. Roger C. Schank, Janet L. Kolonder, Gerald DeJong

[23] *Information retrieval and situation theory.* T.W.C Huibers, M. Lalmas, C.J. Rijsbergen

[24] *"Maximal-Munch" Tokenization in Linear time*. Thomas Reps. University of Wisconsin

[25] *Tokenization as the initial phase in NLP*. Jonathan J, Webster, Chunyu Hit.

[26] *Language indeoendent morphological analysis*. Yamashita, Tatsuo; Matsumoto, Yuji

[27] *Bibliometría ¿para qué?*. Revista de Biblioteca universitaria. Nueva Época, vol.5 numero 1, enero-junio de 2002.

[28] http://www.loc.gov/marc/

[29] http://www.cs.umbc.edu/help/oracle8/server.815/a68022/toc.htm

[30] http://www.miislita.com/term-vector/term-vector-1.html