Capítulo 2
Máquinas de Búsqueda

2.1. Comienzos de la World Wide Web

La Web se inicia en marzo de 1989 propuesto por el investigador Tim Berners-Lee [10], como un proyecto de desarrollo de un sistema de hypertexto, es decir un sistema de creación y distribución de documentos que permita compartir información desarrollada en diferentes aplicaciones, de forma sencilla y eficiente, entre equipos de investigadores ubicados en distintos lugares geográficos. Luego se abocaron a desarrollar una solución que permitiera cubrir los siguientes requerimientos:

- Obtener una interfaz consistente, es decir, el sistema debería permitir una conexión que al menos asegurara una transferencia de datos consistente.
- Permitir incorporar un alto rango de tecnologías y distintos tipos de documentos.
- Proveer una herramienta que permita leer los documentos desde cualquier lugar y por cualquier individuo que esté navegando dentro de este sistema de información, y deberá permitir que cualquier documento sea accesible en forma paralela por dos o mas personas en forma sencilla.

Luego de esta primera etapa, comenzó a desarrollarse lo que se conoce como buscadores o $browser$ para poder acceder a la información desde cualquier lugar en forma rápida. El primer browser construido en 1992, fue llamado WWW

y estaba orientado al trabajo vía FTP. Paralelamente se encontraba disponible para sistemas X Windows, el browser Viola que fue líder en ese campo, ya que permitió los primeros vistazos a gráficos y a un sistema de hypertexto basado en mouse.

En esta época los sitios Web eran creados manualmente, página por página, comenzando con una única página principal y luego realizando a mano los enlaces a las diferentes secciones [16]. Personas interesadas en colocar sus organizaciones en la Web, crearon estos primeros sitios Web o *websites*. Estos esfuerzos eran bien intencionados, pero en general era un trabajo que requería mucho tiempo, y la mayoría de los sitios no soportaba muy bien la incorporación de nuevas páginas. Luego, las organizaciones comenzaron a aportar mas dinero y esfuerzo en sus sitios web, con la esperanza de atraer mas clientes.

A principios de 1993 surgió el browser Mosaic, que cumplía con todos los requerimientos que se buscaban (funcionamiento en diversas plataformas, poseer una interfaz gráfica y ser fácil de usar). Después aparecieron Netscape e Internet Explorer. Finalmente en 1995, se formó el consorcio World Wide Web o W3C que esta bajo la dirección del fundador de la Web.

Hoy en día la mayoría de los sitios Web son realizados con lenguajes de programación e intentan hacer las páginas “dinámicas” (por ejemplo php [40], asp [39], Phyton [41], ColdFusion [37], etc.). Pero en la mayoría de los casos, las presentaciones no se encuentran separadas de las publicaciones lógicas. Usualmente se utilizan mezclas de SQL [21] y HTML [43].

2.2. La Web

La Web se ha convertido en un repositorio de conocimiento y cultura humana que ha permitido compartir ideas e información en una escala sin precedentes, nunca

vista anteriormente. Su éxito se basa en el concepto de una interfaz de usuario estándar que siempre es la misma, sin importar el ambiente computacional utilizado para ejecutar la interfaz. Como resultado de ésto, al usuario se le ocultan todos los detalles de comunicación, ubicación de máquinas, y operaciones del sistema. Además, cada usuario puede crear sus propios documentos Web y hacerlos apuntar a cualquier otro documento Web sin restricciones. Esto es un aspecto fundamental, porque transforma la Web en un nuevo medio de publicación accesible por cualquiera. Como una consecuencia inmediata, cualquier usuario de la Web puede colocar su agenda con poco esfuerzo y casi sin costo alguno. Este universo sin fronteras ha atraído la atención de millones de personas desde el principio, y ha causando una revolución en la manera en que las personas utilizan las computadoras y realizan sus tareas diarias [6].

El crecimiento indiscriminado de la Web, la replicación de información y la falta de una organización apropiada la han llevado, a pesar de su utilidad, a introducir nuevos problemas. La tarea de encontrar información útil en la Web, es frecuentemente tediosa y difícil. Por ejemplo, para que el usuario satisfaga su necesidad de información debe navegar todo el espacio de vínculos o $links$ en la Web buscando información interesante. Pero como el hyperespacio es vasto y desconocido, dicha tarea de navegación es usualmente ineficiente. Para los usuarios nuevos, el problema es mas difícil aún, con lo cual se pueden frustrar todos sus esfuerzos. El principal obstáculo es la falta de un modelo de datos bien definido para la Web, lo cual implica que la definición de información y estructuras es frecuentemente de una baja calidad [7].

2.2.1. Caracterizando la Web

La Web se puede considerar como un gigantesco texto distribuido en una red de baja calidad, con contenido probablemente escrito, no focalizado, sin una buena organización y consultado por los usuarios mas inexpertos.

Los principales desafíos son [7]:

- Gigantesco volumen de texto.
- Información distribuida y conectada por una red de calidad variable.
- Texto altamente volátil (el 40 % cambia cada mes).
- Información mal estructurada y redundante.
- Información de mala calidad, sin revisión editorial de forma ni contenido.
- Información heterogénea: tipos de datos, formatos, idiomas, alfabetos.

2.3. Arquitectura de los Motores de Búsqueda

Los motores o máquinas de Búsqueda (Figura 2.1) son sistemas que indexan en forma automática una porción de los documentos residentes en la totalidad de la Web, y permiten localizar información a través de la formulación de una pregunta. Los motores de búsqueda manejan también grandes bases de datos de referencias a páginas Web, que han sido creadas a través de un proceso automático, sin intervención humana y generalmente de mayor tamaño [24].

Un motor de búsqueda no cuenta con subcategorías como los servicios de directorios, sino con avanzados algoritmos de búsqueda que analizan las páginas que tienen en su memoria y con ello proporcionan el resultado mas adecuado a una búsqueda (por ejemplo Google).

Los puntos a considerar a la hora de evaluar la calidad de un buscador son [15]:

1. El número de documentos de Internet que almacena en su base de datos.
2. La flexibilidad y calidad del lenguaje de consulta.

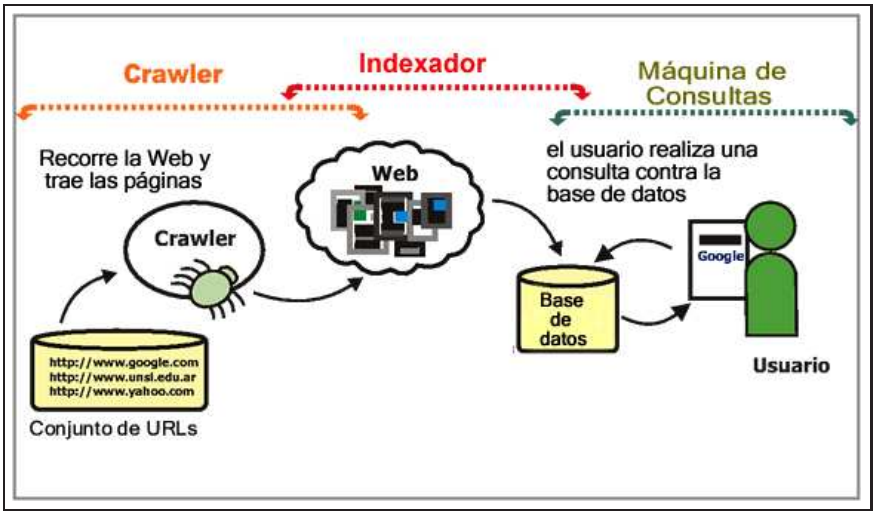

Figura 2.1: Motor de Búsqueda

3. La pertinencia de los resultados (ruido y silencio).

4. Los servicios de valor añadido que incorporan.

5. La periodicidad de actualización de la base de datos.

6. La velocidad en la recuperación y las dificultades de conexión.

Un motor de búsqueda convencional está formado por cuatro elementos principales como se muestra en la Figura 2.2.

Crawler: recupera información de la Web.

Indexador: organiza la información de la Web.

Máquina de consultas: realiza las búsquedas en el índice.

Interfaz: interactúa con el usuario.

La eficiencia y escalabilidad de los motores de búsqueda están directamente relacionado con los “crawlers”, los cuales mantienen actualizado el índice sobre el cual trabaja la máquina de búsqueda.

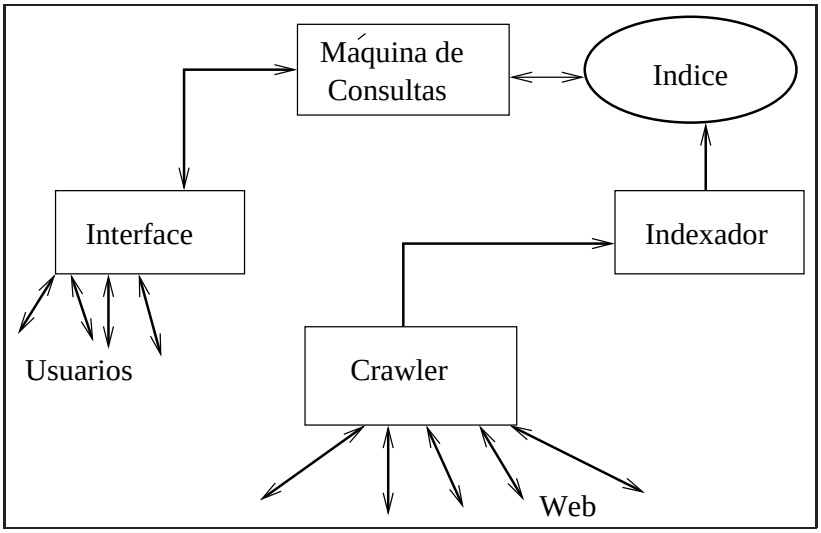

Figura 2.2: Arquitectura centralizada de un motor de búsqueda

Este trabajo está enfocado en una arquitectura de motor de búsqueda centralizado (ej. AltaVista), donde el crawler (robot, *spider*) corre localmente en la máquina de búsqueda, atravesando la Web y trayendo los textos de las páginas Web que va encontrando. El indexador corre localmente y mantiene un índice sobre las páginas que le trae el crawler.

La máquina de consultas también corre localmente y realiza las búsquedas en el índice, retornando básicamente las URLs rankeadas.

La interfaz corre en el cliente (en cualquier parte de la Web) y se encarga de enviar la consulta, de mostrar los resultados, de la retroalimentación, etc.

2.4. Como Trabaja una Máquina de Búsqueda en Internet

Un aspecto positivo sobre Internet y su componente mas visible, la World Wide Web, es que existen cientos de millones de páginas disponibles, para mostrar información sobre una gran variedad de tópicos. Un aspecto negativo es que

existen cientos de millones de páginas rotuladas de acuerdo a los deseos del autor, ubicadas en servidores con nombres secretos. Cuando se desea buscar información sobre un tema en particular, se visitan las máquinas de búsqueda de Internet [46].

Las máquinas de búsqueda son sitios Web diseñados para ayudar a las personas a encontrar información almacenada en otros sitios. Existen diferencias en la forma en que varias máquinas de búsqueda trabajan, pero todas realizan tres operaciones básicas [20]:

- Buscan en Internet en base a palabras relevantes.
- Mantienen un índice de las palabra encontradas y de dónde fueron encontradas.
- Permiten que los usuarios realicen búsquedas de palabras o conjunto de palabras.

Las primeras máquinas de búsqueda mantenían un índice de unos pocos cientos de miles de documentos y páginas, y recibían quizás mil o dos mil requerimientos por día. Hoy en día, una de las mejores máquina de búsqueda (como Google o Yahoo) indexa cientos de millones de páginas, y responde a cientos de millones de consultas por día.

2.4.1. Buscando en la Web

Cuando la mayoría de las personas habla de máquinas de búsqueda en la Web, en realidad se refieren a máquinas de búsqueda en la World Wide Web. Antes que la Web se convirtiera en la parte mas visible de Internet, ya existían buscadores para ayudar a las personas a buscar información en la Red. Programas con nombres como “golpher” y “Archie” mantenían índices de archivos almacenados en servidores conectados a Internet [20].

Antes que un motor de búsqueda pueda indicar dónde se encuentra un archivo o un documento, éste debe ser encontrado. Para encontrar información el buscador emplea robots de software especiales llamados arañas (*spides*), para construir la lista de palabras encontradas en los sitios Web. El proceso de construcción de las listas realizado por las arañas, es denominado *Web crawling*. Para poder construir una lista eficiente de palabras, las arañas de la máquina de búsqueda deben recorrer muchas páginas.

Para que las arañas comiencen a viajar por la Web, generalmente se utiliza una lista de servidores y páginas muy populares. Las arañas comenzarán con un sitio popular, indexa las palabras y sigue cada enlace encontrado en cada sitio.

Luego que estas arañas recuperan los documentos, éstos últimos deben ser procesados para obtener la información de importancia. El procesamiento de los documentos prepara, procesa e ingresa los documentos, páginas o sitios que el usuario busca. El procesamiento de documentos consiste de algunos o todos de los siguientes pasos [46]:

1. Normalizar los documentos a un formato predefinido.
2. Romper el documento en unidades recuperables.
3. Identificar en los documentos potenciales elementos a ser indexados.
4. Eliminar las $stopwords$ o palabras vacías, y extrae las entradas para el índice.
5. Calcular los pesos.
6. Crear y actualizar el índice invertido.

Los dos primeros pasos son denominados “pre-procesamiento” del documento. Aquí simplemente se estandarizan los diferentes formatos encontrados. Estos

pasos sirven para unir los datos en una estructura de datos consistente.

En el tercer paso, se identifican los elementos potencialmente indexables. En el diseño del sistema se debe definir la palabra término. En el sistema implementado para este trabajo de investigación, se definió como un conjuntos de caracteres alfa-numéricos separados por espacios en blancos o signos de puntuación.

En el cuarto paso, se eliminan los términos que tienen poco valor al momento de resolver una consulta. Una lista de palabras vacías típicamente consiste de aquellas clases de palabras que tienen poco significado sustantivo, tales como los artículos, conjunciones, preposiciones, etc. Para eliminar las stopwords, un algoritmo compara los términos candidatos a ser indexados contra una lista predefinida de stopwords. En el quinto paso, se asigna pesos a los términos del índice invertido. A medida que los motores de búsqueda son mas sofisticados los esquemas de pesos se vuelven mas complejos. Finalmente, en el último paso se construye el índice o la lista invertida que es la estructura de datos que almacena la información indexada y que será buscada por cada consulta.

Por otro lado, para el procesamiento de las consultas se requieren de las siguientes etapas. Primero la consulta enviada por el usuario a la máquina de búsqueda, es separada en fragmentos entendibles, generalmente en términos. Luego se parsea la consulta, debido a que los usuarios pueden utilizar símbolos especiales como los booleanos, comillas, etc. Una vez que se obtienen los términos en la consulta, se comparan con las stopwords, para eliminar las palabras de menor relevancia.

El paso final en el procesamiento de las consultas consiste en calcular los pesos de los términos de la consulta. Algunas veces los usuarios pueden controlar este paso, indicando qué término o concepto es mas relevante en la consulta. Dejar que los usuarios asignen los pesos a los términos es poco común, porque

los investigadores han mostrado que los usuarios no son particularmente buenos en determinar la importancia relativa de los términos. Ellos no pueden hacer esta determinación porque no conocen los documentos existentes en la base de datos, y porque buscan información de un tema poco conocido por lo que pueden desconocer la terminología correcta.

2.5. La Máquina de Búsqueda Google

Google es uno de los buscadores mas populares de nuestros días. Está compuesto por una arquitectura bien diseñada y varias capas de software de aplicación para realizar las tareas de búsqueda, crawling e indexación.

En Google, el proceso de crawling es realizado por varios crawlers distribuidos. Existe un servidor de URL que envía listas de URLs para ser buscadas por los crawlers. Las página web recuperadas, son enviadas al *store server* o servidor de almacenamiento. El *store server* comprime y almacena las páginas en un repositorio. Cada página tiene un identificador denominado $docID$. La función de indexación es realizada por el indexador y el $sorter$ u ordenador. El indexador lee el repositorio, descomprime los documentos, y los parsea. Cada documento es convertido en un conjunto de palabras y frecuencias, llamadas $hits$. En este hit se almacena la palabra, su posición en el documento, una aproximación del tamaño de la letra, y las mayúsculas. El indexador distribuye estos $hits$ en un conjunto de barriles. El indexador también recupera todos los enlaces de las páginas web y almacena información importante sobre ellos en un archivo llamado $anchor$. Este archivo posee suficiente información para determinar hacia dónde apunta cada enlace, desde donde es apuntado, y el texto del enlace.

El *URL resolver* lee el archivo $anchor$ y convierte las URLs relativas en absolutas. También genera una base de datos de enlaces que contiene pares de docIDs. La base de datos de enlaces es utilizada para computar el ranking de los

documentos.

El ordenador toma los barriles, que están ordenados por identificador de documentos, y los reordena de acuerdo a los identificadores de palabras para generar el índice invertido. El ordenador también produce una lista de identificadores de palabras y desplazamientos en el índice invertido. Un programa denominado $DumpLexicon$ toma esta lista junto con el léxico producido por el indexador, y genera un nuevo léxico que será usado por el buscador ($searcher$).

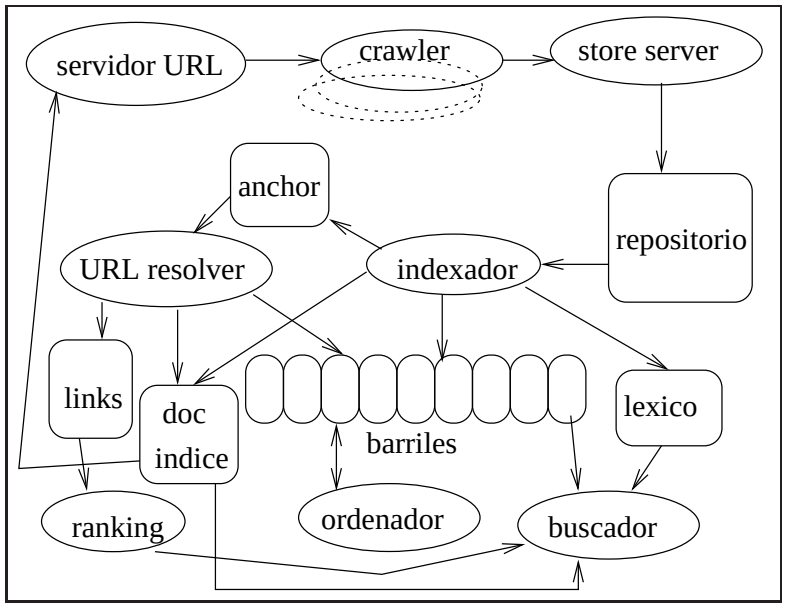

Figura 2.3: Arquitectura de Google.

La operación de crawling es la mas difícil debido a que debe interactuar con cientos de servidores web, y varios servidores de nombres que están mas allá del control del sistema. Para llegar a las páginas web, Google tiene un sistema de crawling distribuido. Un único servidor de URLs envía las listas de URL a un número de crawlers. Cada crawler mantiene unas 300 conexiones abiertas a la vez, y cada una de las conexiones puede estar en diferentes estados: buscando el

DNS, conectándose a un $host$, enviando requerimientos o recibiendo respuestas. El crawler utiliza entrada/salida asincrónica para manejar eventos y un número de colas para mover las páginas buscadas de un estado a otro [11].

La operación de indexación de la Web consiste de tres etapas.

- $Parsing$: cualquier parser que este diseñado para recorrer la Web completa, debe manejar una gran cantidad de posibles errores. Estos errores van desde tipos incorrectos de rótulos HTML, hasta largas cadenas de ceros en la mitad de un tag.

- *Indexar los documentos en los barriles:* luego que los documentos son parseados, se almacenan en un conjunto de barriles. Se convierte cada palabra en un $wordID$ utilizando una tabla de hash en memoria.

- *Ordenamiento:* Para generar el índice invertido, el ordenador toma cada uno de los barriles y los ordena de acuerdo a los $wordID$. Esta operación se realiza de a un barril por vez y se paraleliza la etapa de ordenamiento en distintas máquinas.

En este sistema, diferentes consultas pueden ser ejecutadas en distintos procesadores, y el índice es particionado de modo tal que una consulta pueda utilizar varios procesadores [9].

Para responder a una consulta (de la forma www.googlr.com/search?q=ieee+sociedad), el $browser$ del usuario primero realiza una búsqueda en el sistema de dominio de nombre (DNS), para mapear www.google.com a una dirección IP en particular [36]. El sistema de Google consiste en varios cluster distribuidos por todo el mundo. Cada cluster tiene cientos de máquinas, y una distribución geográfica de las máquinas protege al sistema de catástrofes naturales (terremotos, fallas de energía a larga escala). Un sistema de carga de balance basado en DNS selecciona el cluster que se encuentra

mas cerca físicamente del usuario.

Luego el browser del usuario envía una consulta, y de allí en adelante el procesamiento de la consulta es totalmente local a ese cluster. En cada cluster, un balanceador de carga basado en hardware monitorea el conjunto de Servidores Web Google (SWG) disponibles, y realiza un balance local de requerimientos. Luego de recibir una consulta, una máquina de SGW coordina la ejecución de la consulta y obtiene el resultado en formato HTML (la Figura 2.4 ilustra estos pasos).

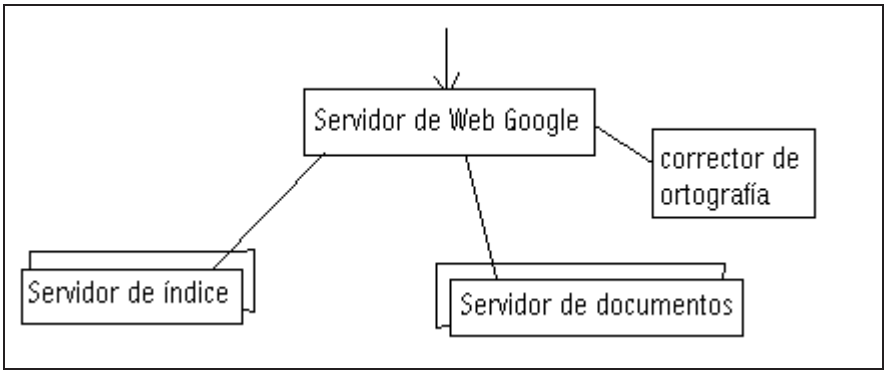

Figura 2.4: Arquitectura para el procesamiento de una consulta.

La ejecución de una consulta consiste de dos etapas principales. En la primera etapa, el servidor de índice busca en el índice invertido cada palabra de la consulta y las mapea a la lista de documentos hit. Luego, el servidor de índice determina un conjunto de documentos relevantes.

La salida de la primera etapa consiste de una lista ordenada de identificadores de documentos ($docID$). La segunda etapa consiste de tomar esta lista, y buscar cada documento en el disco para recuperar el título y las palabras de importancia. Esta operación es realizada por el *Servidor de Documentos*. La estrategia consiste en particionar el procesamiento de los documentos de la

siguiente manera:

- Distribuyendo en forma aleatoria los documentos en pequeñas piezas.
- Teniendo réplicas de múltiples servidores para manejar cada pieza.
- Ruteando los requerimientos a través de un balanceador de carga.

Google también proporciona otras funciones como el sistema de corrector de ortografía y un sistema de servicio de avisos, para generar avisos importantes (si es que los hay). Cuando todas las etapas están completas, genera el HTML apropiado para la página de salida, y se le entrega al browser del usuario.

La mayoría de los accesos al índice y otras estructuras involucradas en la resolución de las consultas son de sólo lectura, debido a que los creadores de Google aseguran que las actualizaciones son relativamente poco frecuentes, y casi siempre se pueden realizar en forma segura derivando las consultas a otras réplicas de servidores durante la actualización.

También en este sistema se explota el paralelismo inherente que existe en las aplicaciones. Por ejemplo la búsqueda de documentos en el gran índice se transforma en varias búsquedas sobre conjuntos de índices mas pequeños, seguido por un proceso de unión (o $merge$).

2.6. Discusión

Este capítulo se ha dedicado por una parte a la Web, a sus comienzos y características. Este gran repositorio de información es un tópico importante para el desarrollo de esta tesis, debido a que es necesario conocer el comportamiento y las particularidades del sistema donde serán ejecutados los algoritmos implementados.

Otro tema que se ha abordado en este capítulo, son los motores o máquinas de búsqueda que permiten facilitar la búsqueda de información en el vasto espacio de la Web. Se ha descripto la arquitectura de estas máquinas y sus componentes mas importantes: el crawler, el indexador y el buscador.

El crawler es el encargado de recolectar las páginas de la Web, el indexador se encarga de procesar las páginas obtenidas por el crawler, para luego actualizar los índices (cuyo fin es acelerar las operaciones de consultas sobre grandes bases de datos de textos); y finalmente el buscador es el reponsable de resolver los requerimientos de los usuarios.

Un buscador es un software que realiza búsquedas de archivos almacenados en los ordenadores. Cuando se les pide información sobre algún tema, las búsquedas se hacen con palabras clave y/o con árboles jerárquicos por temas; el resultado de la búsqueda es un listado de direcciones Web en los que se mencionan temas relacionados con las palabras clave buscadas. Existen distintas clases de buscadores:

- Los Spiders: La mayoría de grandes buscadores internacionales que todos usan y conocen son de este tipo. Requieren muchos recursos para su funcionamiento, y no están al alcance de cualquiera. Recorren las páginas recopilando información sobre los contenidos de las páginas. Cuando se busca una información en los motores, ellos consultan su base de datos, y la presentan clasificados por su relevancia. Cada cierto tiempo, los motores revisan la Web, para actualizar los contenidos de su base de datos, por lo que no es infrecuente, que los resultados de la búsqueda no estén actualizados. Algunos ejemplos de Spiders son: Google, Altavista, Hotbot, Lycos

- Los Directorios: Una barata tecnología, que es ampliamente utilizada por programas scripts en el mercado. No se requieren muchos recursos de informática. En cambio, se requiere más soporte humano y mantenimiento.

Los algoritmos son mucho mas sencillos, presentando la información como una colección de directorios. No recorren la Web ni almacenan su contenido. Sólo registran algunos de los datos de algunas páginas. Los resultados son presentados haciendo referencia a los contenidos y temática de la Web. Algunos ejemplos de directorios son: Antiguos directorios, Yahoo, Terra (Antíguo Olé). Ahora, ambos utilizan tecnología spider, y Yahoo, conserva su directorio.

- Los sistemas mixtos Buscador - Directorio: Además de tener características de buscadores, presentan la Web registrada en catálogos sobre contenidos. Informática, cultura, sociedad. Que a su vez se dividen en subsecciones.

- Metabuscadores: En realidad, no son buscadores. Lo que hacen, es realizar búsquedas en auténticos buscadores, analizan los resultados de la página, y presentan sus propios resultados. No suelen ser bien vistos por los buscadores. Para utilizar los servicios gratuitos de un buscador de esta forma, es necesario pedir permiso. El motivo es el siguiente: El Buscador, pone el dinero para opera el servicio, los contenidos que utilizará el metabuscador, y no percibe nada a cambio. Al eliminar la publicidad, no se obtienen ingresos. Solo gasto y pérdida de visitantes que utilicen este servicio de búsqueda.

- Multibuscadores: Permite lanzar varias búsquedas en motores seleccionados respetando el formato original de los buscadores.

- FFA Enlaces gratuitos para todos: FFA, página de enlaces gratuitos para todos. Cualquiera puede inscribir su página durante un tiempo limitado en estos pequeños directorios. Los enlaces, no son permanentes.

- Buscadores de Portal: Bajo este título, se engloban los buscadores específicos de sitio. Aquellos que buscan información solo en su portal o sitio web.

El buscador es el módulo estudiado en este trabajo, y su importancia se debe a que es la parte mas visible del motor de búsqueda, ya que es el responsable de entregar las respuestas en un tiempo considerable, como así de la efectividad de dichas respuestas. El software que implementa a un buscador, utiliza distintas técnicas que permiten mejorar las búsquedas y optimizar el uso de los recursos disponibles.