FACULTAD 10

METODOLOGÍA PARA LA EVALUACIÓN DE SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN WEB EN LA UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS.

AUTOR (A)

MIRELDIS GARCÍA DEL VALLE

TUTOR

ING. EDUARDO MANUEL MACÍAS SOTOLONGO

Ciudad de la Habana, Cuba, junio de 2009

"Año del 50 Aniversario del Triunfo de la Revolución"

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter no exclusivo.

Para que así conste firmo la presente a los _____ días del mes de _____ del año 2009.

MIRELDIS GARCÍA DEL VALLE

_____

Firma del Autor.

EDUARDO MANUEL MACÍAS SOTOLONGO

_____

Firma del Tutor(a).

AGRADECIMIENTOS

DEDICATORIA

RESUMEN

Los Sistemas de Recuperación de Información (SRI) constituyen una herramienta muy útil para los usuarios en Internet o en alguna subred específica, ya que localizan la información deseada de una manera rápida y relativamente sencilla. Dada la gran utilidad de estos sistemas, los usuarios anhelan altos niveles de eficiencia, calidad y rapidez en los mismos. Para lograr altos niveles de aceptación de los usuarios, estos sistemas deben ser exhaustivamente evaluados, aplicando métricas y modelos de evaluación que garanticen su buen funcionamiento, por lo cual, el objetivo fundamental de la investigación es proponer una metodología para evaluar el funcionamiento y la eficiencia de los SRI.

En la presente investigación se realiza un estudio sobre el surgimiento y evolución de Internet y con esta la creación y desarrollo de los primeros SRI, sus características y funcionamiento. Se profundiza en el estado del arte actual sobre la Recuperación de Información (RI) y los SRI, así como las características de los principales modelos de RI. Posteriormente se realiza un análisis sobre las principales tendencias y modelos para la evaluación de SRI, se exponen las características de varios SRI exitosos de Internet, uno cubano y dos de la Universidad de las Ciencias Informáticas.

Para cumplir con los propósitos de la investigación se propone una metodología general integradora para la evaluación de los SRI en la Universidad de las Ciencias Informáticas, dicha metodología está conformada por un modelo de evaluación, una propuesta de métricas técnicas, de calidad y orientadas al usuario, una guía de pasos a seguir para realizar la evaluación de los SRI y por último una serie de características documentales que deben cumplir los SRI.

Palabras clave: Sistema de Recuperación de Información, modelo, métricas, metodología.

ÍNDICE

INTRODUCCIÓN

Con el surgimiento de Internet se dio un paso trascendente en materia de comunicación y gestión de información. Las dimensiones crecientes del contenido en Internet y las facilidades que posibilita el formato digital hacen que la información crezca aceleradamente a cada instante y por tanto se hace más difícil su almacenamiento y recuperación. De esta necesidad surgieron herramientas que localizan la información publicada en las páginas web, llamadas Sistemas de Recuperación de Información (SRI).

En la Universidad de las Ciencias Informáticas (UCI), creada con una amplia base tecnológica y un perfil docente-productivo, aumentan a diario la cantidad de sitios web publicados, abarcando todas las esferas que en ella se desarrollan. El crecimiento acelerado de este fenómeno durante los 6 años de existencia de la universidad, provocó el surgimiento de SRI, no obstante, éstos no son lo suficientemente eficientes para satisfacer las necesidades de los usuarios. Algunos resultados no deseados que presentan hoy los SRI de la UCI son: Baja calidad de los primeros resultados; abundan los enlaces duplicados o muertos; muchas de las páginas web de la UCI no han sido almacenadas en sus bases de datos y no cuentan con una estrategia eficaz para esto, dado que en la universidad no todos los portales web cuentan con otros portales que apunten a ellos, es decir, no están conectados; muchos solo recuperan la información de un portal específico; sus algoritmos de posicionamiento y cálculo de relevancia de la información son ineficientes; los modelos para representar la información usados en muchos casos no son los más

actuales y por tanto los resultados de una búsqueda no son los más esperados; no prestan muchos servicios adicionales relevantes para la comunidad universitaria; entre otros.

Adicionalmente se ha manifestado en los últimos tiempos una marcada tendencia al uso de herramientas colaborativas en Internet, los SRI no están exentos de este fenómeno y desde hace algún tiempo se ha venido aplicando en estas herramientas un enfoque colaborativo, estrategia que intenta fomentar el trabajo en equipo, permitiéndole a los usuarios comunicarse y trabajar conjuntamente sin importar que no estén en un mismo lugar físico. Por tanto es evidente la necesidad de implementar en la UCI un SRI más flexible, donde los usuarios jueguen un papel más activo e interactivo con los criterios de búsqueda y posicionamiento, lo cual podría contribuir a mejorar el rendimiento, la calidad y eficiencia de los resultados.

Los SRI que existen en la UCI no aplican estrategias colaborativas o al menos de retroalimentación, elemento que podría desarrollar nuevas técnicas de recuperación de información como producto de la colaboración entre los usuarios y potenciar otras funcionalidades que ya existan, además de permitirle a los mismos administrar de cierta forma la información de los resultados que se muestran en las búsquedas, mediante opciones de “sugerir”, “editar”, entre otras.

Podría pensarse que con los exitosos SRI de internet como Google, Yahoo!, Wikia Searsh, Cuil, Live Search (antiguamente MSN), entre otros, el problema está resuelto, sin embargo se deben hacer cuatro acotaciones al respecto:

Primero: Estos sistemas consumen cuota de Internet y ancho de banda.
Segundo: La información que se obtiene como resultado al usar estos en muchas ocasiones no está acorde con los principios de la Revolución Cubana.
Tercero: Se tiene la imposibilidad por parte de los estudiantes de 1er y 2do año de la UCI de acceder a los mismos pues no cuentan con navegación plena.
Cuarto: Estos sistemas no tienen indexados en sus bases de datos los documentos y páginas publicadas en el dominio uci.cu.

Por tales motivos es evidente la necesidad de optimizar los SRI existentes en la universidad, así como los que se estén desarrollando para corregir algunas de estas deficiencias e introducir nuevas tendencias que en el mundo de la Recuperación de Información (RI) están surgiendo y de esta manera dar respuesta a

las necesidades de la comunidad universitaria en materia de localización de la información presente en las páginas de la UCI.

Con este objetivo, la necesidad de realizar una evaluación exhaustiva de los SRI es evidente y por tanto se hace necesario diseñar una metodología de evaluación eficiente que tenga en cuenta las características específicas de la UCI y proponiendo las métricas para realizar la evaluación así como la guía de pasos a seguir.

Teniendo en cuenta lo expuesto anteriormente queda definido como **problema científico** de la investigación: ¿Cómo evaluar el funcionamiento y la eficiencia de los Sistemas de Recuperación de Información Web en la UCI?

En correspondencia con la problemática planteada se define como **objeto de estudio**: Los procesos de búsqueda y recuperación de información web.

Delimitando así el **campo de acción** a: Los procesos de evaluación de los Sistemas de Recuperación de Información Web en la UCI.

Es **objetivo general** de la investigación: Proponer una metodología para evaluar el funcionamiento y la eficiencia de los Sistemas de Recuperación de Información Web en la UCI.

Para guiar el desarrollo de la investigación se formularon las siguientes **preguntas científicas:**

- ¿Son eficientes los Sistemas de Recuperación de Información Web que hoy existen en la UCI?
- ¿Qué parámetros se deben tener en cuenta para evaluar el funcionamiento y eficiencia de los Sistemas de Recuperación de Información Web en la UCI?
- ¿Qué modelo seguir para evaluar la eficiencia de los Sistemas de Recuperación de Información Web en la UCI?

Se han enumerado los siguientes **objetivos específicos**:

- Analizar el estado del arte relacionado con los Sistemas de Recuperación de Información Web.

- Analizar los modelos de evaluación de los Sistemas de Recuperación de Información Web más reconocidos.
- Realizar un análisis crítico de los Sistemas de Recuperación de Información Web en internet y en la UCI.
- Definir las métricas y su relevancia, para evaluar el funcionamiento y eficiencia de los Sistemas de Recuperación de Información Web en la UCI.
- Elaborar una guía de pasos para realizar la evaluación a los Sistemas de Recuperación de Información Web en la UCI de manera eficiente.

Para cumplir exitosamente con los objetivos de la investigación se han definido las siguientes **tareas:**

- Levantamiento bibliográfico y webgráfico para estudiar los principales elementos teóricos y conceptos que permitan analizar la historia de los Sistemas de Recuperación de Información Web y elaborar un marco teórico-conceptual.
- Análisis de las características, ventajas y desventajas de Sistemas de Recuperación de Información Web existentes en Internet, Cuba y la UCI así como de los modelos de Recuperación de Información usado por estos.
- Análisis de las métricas más usadas en la evaluación de Sistemas de Recuperación de Información Web.
- Análisis de guías y metodologías de evaluación de Sistemas de Recuperación de Información Web existentes.

La investigación deberá sentar las bases para mejorar cualitativamente los Buscadores Web presentes en la UCI, así como los que se desarrollen en el futuro y se esperan obtener importantes **resultados** como:

- Investigación sobre el estado del arte de los Sistemas de Recuperación de Información Web.
- Investigación del estado actual en la UCI de los Sistemas de Recuperación de Información Web.
- Guía de métricas a tener en cuenta para evaluar el funcionamiento y la eficiencia de los Sistemas de Recuperación de Información en la UCI.
- Metodología para evaluar el funcionamiento y la eficiencia de los Sistemas de Recuperación de Información Web en la UCI.

Métodos de investigación:

Teóricos:

Analítico-Sintético: Se utiliza este método centrándose en el análisis de las teorías y documentos, permitiendo la selección de los elementos más importantes de manera que se elabore correctamente la información.

Histórico-Lógico: Para estudiar el origen, la evolución, las tendencias y las perspectivas de los Sistemas de Búsqueda y Recuperación de Información Web.

Empíricos:

Observación Científica: Permite conocer los elementos relacionados con los temas abordados en la investigación.

Encuesta: Empleando cuestionarios con preguntas predominantemente cerradas y categorizadas que permitan recopilar el criterio de estudiantes y profesores de la universidad.

El presente trabajo consta de 3 capítulos estructurados de la siguiente forma:

Capítulo 1: Estado del arte y fundamentación teórica: Este capítulo trata los principales elementos relacionados con el surgimiento y evolución de internet, así como la evolución de los SRI Web. Están presentes también conceptos importantes que serán de ayuda para un mejor entendimiento de la investigación, como los lenguajes de consulta que utilizan operadores para formular de una manera más optima las preguntas y por último recoge un estudio de los modelos para la recuperación de la información más utilizados.

Capítulo 2: Análisis de tendencias para la evaluación de SRI: Analiza algunas metodologías de evaluación de SRI que se han realizado, entre las que se encuentran varias que han sido pioneras de las evaluaciones de los SRI y se realiza un análisis de algunos de los buscadores más exitosos de internet, uno cubano y algunos de los existentes hoy en la UCI.

Capítulo 3: Metodología para evaluar los SRI en la UCI: En este capítulo se propone un modelo de evaluación para los SRI y las principales métricas a tener en cuenta para la evaluación de estos en la universidad. Se propone además una guía para realizar dicha evaluación y por último algunas de las características documentales que deben cumplir los SRI.

CAPÍTULO 1

ESTADO DEL ARTE Y FUNDAMENTACIÓN TEÓRICA

1.1 SURGIMIENTO Y EVOLUCIÓN DE INTERNET

Internet remonta sus orígenes a 1969 cuando se establece la primera conexión de computadoras llamada ARPANET, la misma fue creada por encargo del Departamento de Defensa de los Estados Unidos como medio de comunicación para los diferentes organismos del país; un poco más adelante se realiza en 1973 la primera conexión ARPANET fuera de EEUU y se hizo con NORSAR[1] en Noruega justo antes de las conexiones con Gran Bretaña.

Este avance en la tecnología fue un gran impacto para la humanidad y revolucionó todas las grandes mentes de esa época. Posteriormente otro momento de gran significado para la historia de la informática y las comunicaciones fue la creación de la World Wide Web (WWW, o "la Web") por Tim Berners-Lee en 1989. La WWW es un conjunto de protocolos que permite, de forma sencilla, la consulta remota de archivos de hipertexto y utiliza a Internet como medio de transmisión. Sin duda alguna este fue uno de los

[1] ***NORSAR:*** *Mediante esta red se establece la primera conexión entre Inglaterra y Estados Unidos donde la University Collage of London se conecta a ARPANET.*

servicios que más éxito proporcionó y aun continúa proporcionando a los cibernautas, contribuyendo en gran medida a convertir a Internet en la gran autopista de información y conocimiento que hoy es.

Con el paso de los años, el volumen de información digital presente en Internet se hace mayor, formando así una gran Red de Redes donde millones de personas a través de todo el mundo se encuentran interconectadas. No ha habido un día en que Internet haya dejado de crecer, convirtiéndose con el transcurso del tiempo en una enorme biblioteca virtual contenedora de una buena parte del saber humano, evolucionando hacia lo que podría considerarse un dinámico almacén donde se albergan informaciones muy diversas en contenidos, relevancia y utilidad.

Como se puede apreciar, buscar y localizar la gran cantidad y variedad de información existente en los sitios web de forma manual es prácticamente imposible, dado el gran flujo de información que circula hoy en día y la velocidad a la que sufren modificaciones. Poder localizar cualquier tipo de información que necesite un usuario cualquiera en Internet, de manera rápida y eficiente sería perfecto. Aquí es donde entran a jugar un papel importante los Sistemas de Recuperación de Información en la Web (SRI), los cuales surgen como la muy necesaria solución a tal problema. En general pueden ser usados para localizar no sólo información textual, sino también imágenes, sonidos y videos.

1.2 EVOLUCIÓN DE LOS SRI WEB

Los SRI son herramientas que posibilitan localizar la información digital presente en internet o en una subred determinada. Un SRI permite la RI, previamente almacenada, por medio de la realización de una serie de consultas (“queries”) a los documentos contenidos en la base de datos. Esta serie de preguntas o interrogaciones se conceptúan como *sentencias formales de expresión de necesidades de información,* y suelen venir expresadas por medio de un lenguaje de interrogación.

Esta información que se recupera puede ser de tipo texto (Word, PDF, páginas Web), y con el avance de la tecnología se pueden incluir ya multimedia, incorporándose fotografías, ilustraciones gráficas, vídeo animado y audio.

Estas herramientas encargadas de la RI pueden ser fundamentalmente clasificado en dos grupos: directorios web o índices temáticos y los motores de búsqueda o buscadores web aunque se podrían mencionar otros, pero estos son los más utilizados por los cibernautas.

En la actualidad los más exitosos son los buscadores pues realizan de manera automática la búsqueda e incluso los directorios utilizan motores de búsqueda. Hoy en día es muy usada esta tendencia de herramientas hibridas, muy pocos de los directorios existentes son puramente directorios y ocurre de la misma manera con los buscadores. Ej. Yahoo! Es un directorio y tiene su propio motor de búsqueda, ya que de esta manera se hace más factible y menos engorrosa la búsqueda para el usuario y además presenta varios servicios adicionales tratando de que el usuario tenga todo lo que necesite con solo dar un clic. Por este motivo la investigación se centra fundamentalmente en los SRI con características de buscadores.

Se puede definir a un **directorio o índice temático** como un sistema de búsqueda por temas o categorías jerarquizados (aunque también suelen incluir sistemas de búsqueda por palabras clave). Se trata de bases de datos de direcciones Web elaboradas "manualmente", es decir, hay personas que se encargan de asignar cada página web a una categoría o tema determinado. [1] Ejemplos de directorios: Yahoo!, Terra (Antiguo Olé), entre otros. Ahora, ambos utilizan tecnología de búsqueda jerárquica, y Yahoo! conserva su directorio.

Por su parte los **motores de búsqueda o buscadores web** son herramientas encargadas de presentar al usuario donde encontrar documentos relevantes de un tema buscado, son sistemas de búsqueda por palabras claves, utilizan programas automáticos, generalmente llamados crawlers o spiders, que rastrean la web recolectando páginas o parte de ellas, y las almacenan en bases de datos documentales, proporcionando una herramienta para hacer búsquedas sobre estas bases de datos. Es decir está compuesto por tres partes fundamentales:

Robot, Araña (Spider) o Crawler: Es un programa que inspecciona las páginas del World Wide Web de forma metódica y automatizada. Su funcionamiento más común es que se le da al programa un grupo de direcciones iniciales, la araña descarga estas direcciones, analiza las páginas y busca enlaces a páginas nuevas. Luego descarga estas páginas nuevas, analiza sus enlaces, y así sucesivamente. Al finalizar el proceso la araña debe haber sido capaz de descargar una copia de las páginas visitadas para algún directorio que se le especifique.

Indexador: Consiste en un programa capaz de descomponer los documentos descargados por la araña hasta obtener las listas de palabras que los forman. Este proceso puede ser muy sencillo, como en los archivos de texto plano, que tienen todas las palabras separadas por espacios u otros caracteres, o muy complejo, como un documento PDF que debe ser decodificado, separando el formato y las imágenes, extrayendo solamente el texto plano.

Base de Datos Documental: Una vez procesados los documentos por el indexador, se almacenan en la base de datos documental para ser recuperados posteriormente cuando se desee.

Motor de búsqueda: Recupera y jerarquiza contenidos de la base de datos en función de términos y criterios de búsqueda entrados por los usuarios para dar los resultados.

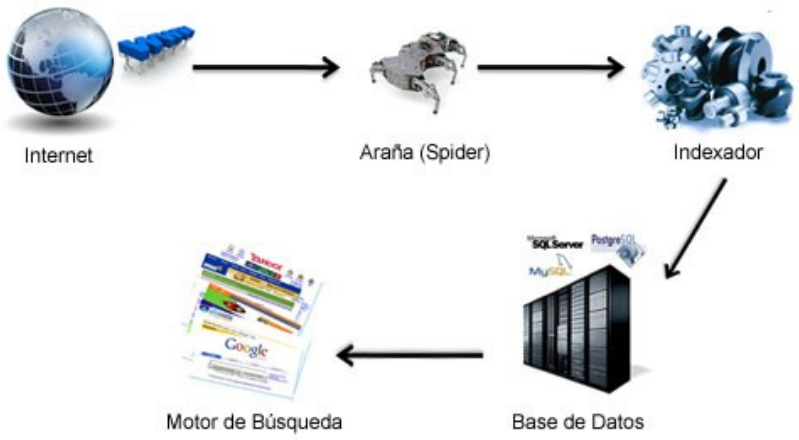

Figura 1.1: *Elementos que componen un buscador web.*

	Directorios	**Motores de Búsqueda**
Descubrimiento de	De forma manual.	Principalmente de forma

recursos		automática por medio de robots (spiders)
Representación del contenido	Clasificación manual	Indización automática
Representación de la consulta	Implícita (Navegación por categorías)	Explícita (palabras clave, operadores, etc.)
Representación de los resultados	Páginas creadas antes de la consulta. Poco exhaustivos, muy precisos.	Páginas creadas dinámicamente en cada consulta. Muy exhaustivos, poco precisos.
Almacenamiento de información	Almacena la información mediante directorios, clasificados en categorías.	Almacena la información mediante una base de datos propia.
Organización de los resultados	La presentación de los resultados se lleva a cabo mediante un listado de todos los documentos correspondientes en la categoría, sin ningún criterio de presentación.	La presentación de los resultados se establece por orden de relevancia según unos criterios establecidos en la ecuación de búsqueda.
Búsqueda de información	La búsqueda se realiza jerárquicamente según las categorías establecidas.	La búsqueda se realiza en la base de datos mediante la ecuación de búsqueda.

Tabla 1.1: *Comparación entre motores de búsqueda y directorios.*

Luego del gran avance que introdujo la WWW, estas herramientas encargadas de la RI han tenido que sofisticarse a través de los años y han ido evolucionando desde la década de los años 50, cuando el objetivo era manejar información bibliográfica. En un principio los ficheros informáticos se dejaban en un servidor al que había que conectarse mediante una interfaz de interrogación mediante comandos. Luego se desarrolla Archie el primer motor de búsqueda, diseñado para indexar archivos FTP, permitiendo a las personas encontrar archivos específicos. La implementación original se escribió en 1990 por Alan Emtage, Bill Heelan, y Peter J. Deutsch, en aquel entonces estudiantes de la Universidad McGill de Montreal (Canadá); posteriormente le suceden un sin número de aplicaciones informáticas de este tipo, pero esto no era suficiente para el gran cúmulo de información que se movía ya diariamente en Internet, por lo que han ido evolucionando con el transcurso del los años, dada la gran necesidad de los usuarios de localizar la información y la gran exigencia de los mismos en términos de calidad de los resultados que estas

herramientas devuelven, la rapidez con que lo hacen y las interfaces más o menos amigables que puedan brindan.

A continuación se muestra un orden cronológico de algunos de los SRI más significativos que constituyeron pasos trascendentales para llegar al avance tecnológico que se tiene hoy con estas herramientas:

En 1991, Paul Lindner y Mark P. McCahill de la Universidad de Minnesota desarrollan Gopher consistente en el acceso a la información a través de menús; el mismo incorporó el uso de hipertexto, para tratar de clasificar toda la información mediante menús y submenús, lo cual facilitó enormemente la clasificación y localización de la información.

En 1992, la Universidad de Nevada lanza Verónica, una herramienta de búsqueda para Gopher.

En 1993, se crea Excite un portal de Internet y una de las primeras "punto.com"[2] de la década de 1990, uno de los más reconocidos dominios en internet en esta época, ya que fue una gran revolución del momento. Se crea además el primer buscador llamado "Wandex", un índice (ahora desaparecido) realizado por la World Wide Web Wanderer[3] y un robot desarrollado por Mattew Gray en el Instituto de Tecnología de Massachusetts (MIT). Se crea también en este año Mosaic, primer navegador gráfico. En este año surge también otro buscador llamado Aliweb (http://www.aliweb.com/), el cual todavía se encuentra en funcionamiento.

En 1994, se crea WebCrawler, el primer motor de búsqueda de texto completo, otro importante avance de esta época. Infoseek es otro de los buscadores lanzados en este año, fue un motor de búsqueda muy popular creado por Steve Kirsch. Más adelante es comprado por The Walt Disney Company[4] en 1998, y la

---

[2] ***.com*** *(del inglés commercial, comercial) es un dominio de internet genérico que forma parte del sistema de dominios de internet. En los años 1990, .com se convirtió en el dominio más frecuentemente utilizado para sitios web, especialmente los de uso comercial. Ejemplo (para comprender la palabra dominio): El dominio para las páginas de Cuba es .cu, el de españa .es.*

[3] ***World Wide Web Wanderer:*** *Utilizado para medir el tamaño de la World Wide Web en aquellos momentos y para ello utilizaba un índice denominado Wandex, proporcionando el primer motor de búsqueda en la web.*

[4] ***The Walt Disney Company:*** *(también conocida como Disney Enterprises, Inc. o simplemente Disney) es la segunda compañía de entretenimiento en el mundo, después de Time Warner.*

tecnología se fusionó con el de la Disney y Starwave[5] para formar el Go.com (Figura 1.2) el cual sigue en funcionamiento desde entonces y ha sido sustituido por las búsquedas de Yahoo!.

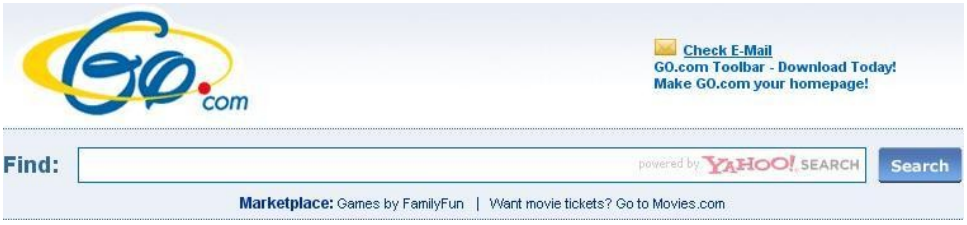

Figura 1.2: *Formulario de búsqueda de Go.com.*

También en este año David Filo y Jerry Yang lanzan Yahoo! el primer directorio categorizado para búsquedas en Internet.

En 1995, la compañía Digital lanza AltaVista y se pueden mencionar otros ejemplos obtenidos de motores de búsqueda populares en este año como son World Wide Web Worm, Lycos y World Wide Web Home Pages; mientras que entre los índices se pueden encontrar a Galaxy y Yahoo! como los más usados.

En 1996, es creado Inktomi un motor de búsqueda desarrollado específicamente para dar servicio a otros portales.

En 1997, se lanza Ask (Figura 1.3) (http://www.ask.com/) el cual es un motor de búsqueda muy utilizado hoy en día.

Figura 1.3: *Formulario de búsqueda de Ask.com.*

En 1998, se lanza Google creado por Larry Page y Sergey Brin (dos estudiantes de doctorado en Ciencias de la Computación de la Universidad de Stanford), su más preciada creación fue el PageRank el cual analiza los links entre páginas web, cuanto mayor fuera la WWW, mejores serían los resultados y esta es una de las formulas que permite un mejor posicionamiento de las páginas que muestra Google al realizar

[5] ***Starwave:*** *Corporación realizadora de video juegos.*

una consulta. Este hecho inspiró a que Page y Brin bautizasen definitivamente su buscador con el nombre de Google, en alusión a la palabra 'googol' (el número representado por un '1' seguido de 100 ceros). Anteriormente, en agosto de 1996 lanzaron públicamente la primera versión de Google bajo el dominio 'google.stanford.edu'. Ahora se encuentra inaccesible (fue sustituido por 'google.com', http://www.google.com/)

En 1999, AlltheWeb (Figura 1.4) (http://www.alltheweb.com/) es creado en este año. Ya en estos momentos utiliza los servicios de Yahoo! y no es muy utilizado por los cibernautas.

Figura 1.4: *Formulario de búsqueda de AlltheWeb.*

En 2000, Yahoo! utilizaba los servicios de búsqueda de Google y en este año cambia el buscador de Google por Inktomi que como se planteó anteriormente el mismo no tiene su propia interface de búsqueda pero entrega los resultados de su rastreador a muchos otros buscadores.

En 2003, Overture[6] compra AltaVista y Fast/AlltheWeb. Por otro lado Microsoft[7] anuncia su propio buscador y Yahoo! compra Overture.

En 2004, Yahoo! lanzó su propio buscador basado en una combinación de tecnologías de sus adquisiciones y proporcionando un servicio en el que ya prevalecía la búsqueda en la Web sobre el directorio.

En 2005, Noxtrum fue el primer motor de búsqueda global diseñado por una empresa española, Telefónica Publicidad e Información, S. A. (TPI), lanzado al mercado en versión Beta el 1 de diciembre de este año. Realizaba búsquedas en internet del mundo hispanohablante. (Ya no se encuentra en funcionamiento).

En febrero de este mismo año Microsoft lanza su propio buscador Live Search (Figura 1.5) (http://www.live.com/)

Figura 1.5: *Formulario de búsqueda de Live Search.*

En 2005 también Yahoo! compra Flickr (un buscador de fotos muy exitoso y útil http://flickr.com/).

En 2006, Google compra YouTuve (sitio web que permite a los usuarios compartir vídeos digitales a través de Internet e incluso, permite a los músicos novatos y experimentados dar a conocer sus vídeos al mundo.)

---

[6] ***Overture:*** *Sucesor del motor World Wide Web Worm. Representa la primera herramienta que implementa el "pago por clic".*

[7] ***Microsoft Corporation:*** *Es una empresa multinacional estadounidense, fundada en 1975 por Bill Gates y Paul Allen. Dedicada al sector de la informática, con sede en Redmond, Washington, Estados Unidos. Microsoft desarrolla, fabrica, licencia y produce software y equipos electrónicos.*

En 2008, Ex empleados de Google lanzan en Estados Unidos un nuevo buscador de Internet llamado Cuil (Figura 1.6) (http://www.cuil.com/) que, según los expertos, es una prometedora página con posibilidades de convertirse en un serio competidor del gigante de las búsquedas en la red.

Figura 1.6: *Formulario de búsqueda de Cuil.*

En este mismo año también se lanza Wikia Search (Figura 1.7), el motor de búsqueda en Internet creado por Jimmy Wales, el fundador de la enciclopedia online Wikipedia[8], y su objetivo es el de convertirse en un buscador transparente y abierto a todos los usuarios.

El creador plantea que el futuro de las búsquedas en Internet debe basarse en cuatro principios organizativos (Four Organizing Principles, TCQP en inglés):

1. **Transparencia:** que el público sepa cómo operan los sistemas y algoritmos de búsqueda, ya sea con licencias de código abierto, como con contenidos abiertos + APIs[9].
2. **Comunidad:** cualquier persona puede colaborar de alguna forma (de forma individual o mediante una organización), con un fuerte acento social y comunitario.
3. **Calidad:** mejorar significativamente la relevancia y exactitud de los resultados y la experiencia de búsqueda.

---

[8] ***Wikipedia:*** *Es un proyecto de la Fundación Wikimedia (organización sin ánimo de lucro) para construir una enciclopedia libre. Los más de 12 millones de artículos de Wikipedia han sido redactados conjuntamente por voluntarios de todo el mundo, y prácticamente todos pueden ser editados por cualquier persona que pueda acceder a Wikipedia.*

[9] *Un* ***API*** *o interfaz de programación de aplicaciones (del inglés Application Programming Interface) es el conjunto de funciones y procedimientos (o métodos, si se refiere a programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción.*

4. **Privacidad:** habrá una protección de los datos de identificación de los usuarios y de sus preferencias de búsqueda. [2]

Figura 1.7: *Formulario de búsqueda de Wikia Search.*

En la actualidad la tecnología ha avanzado de un modo deslumbrante y se estima que existen en la red alrededor de 5 300 buscadores, de los cuales 5 000 son internacionales y unos 300 son hispanos y entre ellos tiene lugar una gran carrera para ganar la preferencia de los navegantes. [3] Por esto, dichas herramientas se perfeccionan continuamente, poseen una interfaz cada vez más amigable, y se han adaptado a nuevas exigencias, especialmente en el campo de la recuperación de información. Uno de los buscadores más exitosos hoy en día es Google y dentro de los directorios se puede mencionar a Yahoo!. Catalogamos a Yahoo! como un directorio porque inicialmente fue así y su concepción desde el inicio era la de un directorio pero en la actualidad no se pueden ver separados estos dos conceptos, muchos buscadores presentan servicios de directorios y viceversa, entre ellos están los antes mencionados.

En la actualidad se ha podido constatar que los SRI más exitosos en Internet son los buscadores web (atendiendo a la cantidad de usuarios que hacen uso de los mismos y al porciento de tráfico en la red que se genera a través de estos) pues realizan de manera automática la búsqueda y son por este motivo los más utilizados por los usuario. Se puede apreciar incluso que muchos directorios en la actualidad utilizan motores de búsqueda para hacer más fácil a los usuarios localizar determinada información que estos deseen, tal es el caso de Yahoo! y otros directorios que a la vez son buscadores.

El mercado está en estos momentos dominado por SRI como Google, Yahoo!, Live Search y otros pocos en menor medida. El resto de grandes buscadores tienden a ser portales que muestran los resultados de otros buscadores y ofrecen, además, otro tipo de contenidos que tienen mayor o menor importancia en la página como hace el propio Yahoo!, el cual en un tiempo utilizó el servicio de búsqueda de Google pero luego como se pudo apreciar en la cronología antes descrita en el año 2002 compra Inktomi y utiliza sus servicios y de esta manera comienza una competencia de popularidad entre los dos: Google y Yahoo!.

En la actualidad las herramientas colaborativas y la filosofía de Software Libre han alcanzado un gran auge en internet por las facilidades y los beneficios que estas aportan a las comunidades de cibernautas. Con este principio algunos SRI están implementando modelos de recuperación de información interactivos como el caso de Wikia Search (http://search.wikia.com/), donde el usuario participa de una manera activa en los criterios de posicionamiento y búsqueda, optimizando los resultados teniendo como base la retroalimentación.

La arquitectura de la información y su organización son temas sumamente importantes cuando los volúmenes de esta aumentan exponencialmente, los niveles investigativos actuales de las empresas e instituciones necesitan no sólo formas de organizar bien la información, además de ello necesitan herramientas para encontrarlas fácilmente; empresas como Google o Yahoo! brindan dichos servicios pero para esto se debe de entregar la documentación a ellos y además de esto pagar por el servicio. El mundo del Software Libre ofrece alternativas, libres y además gratis que permiten indexar la información de un servidor web y encontrar la misma de una manera rápida. Existen en la actualidad algunas de estas herramientas libres que funcionan como directorios o buscadores ya sea internos de una página en específico o busca la información solicitada en varias páginas que le son definidas, se tienen como ejemplo de ello los programas que se mencionan a continuación, por mencionar solo algunos:

TSEP - The Search Engine Project

TSEP es un motor de búsqueda para un sitio web, creado en PHP[10]. Se puede poner un botón "Buscar en este Sitio" y permitir que las personas encuentren fácilmente lo que están buscando. TSEP es gratis y de código abierto. Está diseñado para ser muy fácil de instalar y usar. Desde la versión 0.913 lo suministran con instalador. Puede mejorar en su diseño, usando CSS[11] para aplicar estilos. [4]

PhpDig

---

[10] ***PHP (PHP Hypertext Pre-processor)*** *es un lenguaje de programación interpretado, diseñado originalmente para la creación de páginas web dinámicas. Es usado principalmente en interpretación del lado del servidor (server-side scripting) pero actualmente puede ser utilizado desde una interfaz de línea de comandos o en la creación de otros tipos de programas incluyendo aplicaciones con interfaz gráfica.*

[11]*Las* ***hojas de estilo en cascada*** *(Cascading Style Sheets, CSS) son un lenguaje formal usado para definir la presentación de un documento estructurado escrito en HTML o XML.*

Spider o robot y motor de búsqueda creado en PHP y con base de datos MySQL[12]. Crea un glosario con palabras encontradas en las páginas indexadas. En una búsqueda muestra resultados que tienen las palabras clave, ordenados por ocurrencias de esas palabras. [5]

Blasten blt-SEARCH

Un buscador adaptado a usted y sus visitantes. Utilizando un sistema escaneador de enlaces, basado en la tecnología PageRank con un poder de penetración superior a los 10 puntos, logra solucionar y automatizar la gran necesidad de encontrar contenidos de forma rápida y precisa en una página web, para de este modo aumentar la navegabilidad y la comodidad de todos sus visitantes. Operando con un Motor de búsqueda, que permite ordenar de manera instantánea todos los resultados calculando la similitud entre el criterio buscado y los contenidos indexados, basándose en la importancia de la página. Además de asignarle una posición adecuada acorde a la cantidad de enlaces referidos durante el proceso de indexación. [6]

Lucene

Es un API para recuperación de información de código abierto, originalmente implementada en Java. Es útil para cualquier aplicación que requiera indexado y búsqueda a texto completo. Lucene ha sido ampliamente usado por su utilidad en la implementación de motores de búsquedas, lo cual ha llevado a la falsa idea de que Lucene es un motor de búsquedas con funciones de "crawling" y análisis de documentos en HTML[13] incorporadas. [7]

El centro de la arquitectura lógica de Lucene se encuentra en el concepto de Documento (Document) que contiene Campos (Fields) de texto. Esta flexibilidad permite a Lucene ser independiente del formato del

[12] ***MySQL*** *es un sistema de gestión de base de datos relacional, multihilo y multiusuario con más de seis millones de instalaciones.*

[13] ***HTML****, siglas de HyperText Markup Language (Lenguaje de Marcas de Hipertexto), es el lenguaje de marcado predominante para la construcción de páginas web.*

fichero. Textos que se encuentran en PDF[14], páginas HTML, documentos de Microsoft Word (.doc), así como muchos otros pueden ser indexados mientras que se pueda extraer información de ellos.

Nutch

Es un robot y motor de búsqueda basado en Lucene. Nutch ofrece una solución transparente, pues al ser una tecnología de código abierto es posible conocer como organiza el ranking de resultados de las búsquedas. Está desarrollado en Java[15], y basa su arquitectura en la plataforma Hadoop[16] de desarrollo de sistemas distribuidos.

Algunas de las características del buscador son:

- No distingue entre mayúsculas y minúsculas.
- Usando comillas (") al principio y al final de un grupo de palabras o frase realiza la búsqueda de ese texto exacto.
- Añadiendo el signo más (+) delante de una palabra fuerza la búsqueda de palabras no habituales.
- Añadiendo el signo menos (-) delante de una palabra realiza la búsqueda excluyendo esa palabra.

En los resultados se puede encontrar diversa información:

- En caché: muestra la versión de la página visitada por Nutch.
- Explicar: muestra una explicación de cómo Nutch otorgó la puntuación a esa página.
- Anchors: muestra una lista con el texto que aparece en enlaces que apuntan a esa página.

Nutch es un software que, sobre la base aportada por Lucene, integra todo lo que hace falta para completar un motor de búsqueda de páginas web. [7]

1.3 CONCEPTOS BÁSICOS

[14] ***PDF*** *(acrónimo del inglés Portable Document Format, formato de documento portátil) es un formato de almacenamiento de documentos, desarrollado por la empresa Adobe Systems.*

[15] ***Java:*** *Lenguaje de programación orientado a objetos. Fue desarrollado por James Gosling y sus compañeros de Sun Microsystems al principio de la década de los 90.*

[16] ***Hadoop****: Es una plataforma que permite desarrollar software para el proceso de cantidades enormes de datos mediante la utilización de clusters de ordenadores.*

En el campo de la RI se deben tener en cuenta varios conceptos fundamentales para comprender los procesos básicos de un SRI. A continuación se abordarán los aspectos teóricos o conceptos que soportan la investigación y además contribuirán a un mejor entendimiento de la misma.

1.3.1 Lenguaje de consulta

Es un lenguaje informático que se puede utilizar para hacer consultas en Bases de Datos (lenguajes de consulta de bases de datos) y sistemas de información (lenguajes de consulta de RI). Este lenguaje ayudará a realizar la búsqueda que se necesita mucho más efectiva. En esencia provee al usuario de un mecanismo de expresión de sus necesidades informativas.

Enfocados en el tema que se trata en la investigación se puede plantear que el lenguaje de consulta de recuperación de información le hará al usuario más factible y eficiente su búsqueda.

Un lenguaje de consulta básico está formado por una sucesión de términos o palabras claves que el usuario específica para buscar la información a fin con dichos términos. Los SRI en la web suelen presentar dos interfaces de búsqueda, una simple (Anexo 1) donde el usuario solo inserta las palabras claves y otra avanzada (Anexo 2) donde se utilizan operadores que brindan facilidades para realizar búsquedas más complejas y específicas. Los SRI se han sofisticando a través de los años y ya la interfaz de la búsqueda avanzada es mucho más amigable, así los usuarios no tiene que memorizar los operadores ni la sintaxis y siguen construyendo expresiones bastante complejas.

La búsqueda avanzada suele ofrecer la posibilidad de utilizar distintos tipos de operadores y estos se incluyen en 5 grupos:

1. Operadores booleanos.
2. Operadores posicionales.
3. Operadores de existencia.
4. Operadores de truncamiento.
5. Operadores de límite o comparación.

Operadores booleanos

Los operadores booleanos (Figura 1.8) son muy utilizados en los SRI. Los tres operadores básicos son:

1. Operador suma/unión (**AND**)
2. Operador producto/intersección (**OR**)
3. Operador resta/negación **(NOT**)

Dichos operadores pueden combinarse, dando como resultado operaciones más complejas.

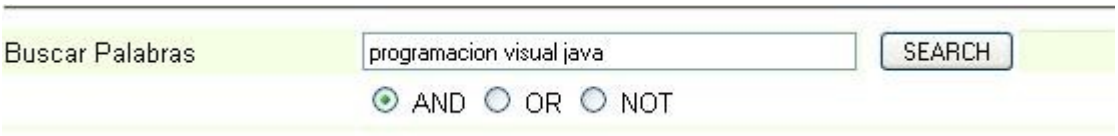

Figura 1.8: *Búsqueda utilizando operadores booleanos.*

Por ejemplo: Al utilizar el operador OR la búsqueda se hace más general pues si se plantea: *juego* OR *entretenimiento* OR *descanso*; los documentos que se obtendrán podrán contener al menos uno de los 3 términos.

Para simplificar su uso a los usuarios no familiarizados con este tipo de operadores, algunos buscadores utilizan frases que son mucho más sencillas y el resultado se convierte en el mismo; ejemplo: *"con todas las palabras"* en lugar del operador AND (se puede usar el operador **&** como alternativo) o *"con alguna de las palabras"* en lugar del operador OR. Como ejemplo de esto se tiene la búsqueda avanzada de Google (Figura 1.9).

Figura 1.9: *Búsqueda utilizando operadores booleanos.*

Operadores posicionales

Los operadores posicionales tienen por objetivo superar determinadas limitaciones de los operadores booleanos. Ellos toman como punto de partida la valoración del término dentro del contexto en el que se encuentre; se dividen en dos grupos:

Posicionales absolutos: Son operadores que permiten buscar un término en un lugar específico del documento como por ejemplo, el título, la url, en el contenido de la pagina, en los enlaces hacia esta página, entre otros.

Posicionales relativos o de proximidad: Son operadores para establecer la posición de un término respecto a otro. Se pueden buscar palabras que estén juntas, separadas por varias palabras o caracteres, que se encuentren en la misma frase o párrafo e incluso si se debe o no respetar el orden en que se han introducido los términos. Un operador posicional muy común es **NEAR**, también se pueden mencionar el **WITH**, **ADJ.** A continuación se explica con más detalle, los que se consideran más importantes:

- Se utiliza el operador **SAME** para localizar registros en los que el campo de registro bibliográfico contiene todos los términos especificados. Todos los términos de búsqueda están localizados dentro del mismo campo del registro, aunque no necesariamente en la misma frase. Por ejemplo, si se busca "Cuba SAME historia", sólo se recuperarán los registros que contengan tanto "Cuba" como "historia" dentro del mismo campo bibliográfico.

- Se utiliza el operador **WITH** para localizar registros en los que un campo contiene una frase con todos los términos especificados. Por ejemplo, si se buscar "Cuba WITH historia", sólo se recuperarán los registros que contengan tanto "Cuba" como "historia" en la misma frase del campo bibliográfico.

- Se utiliza el operador **NEAR** para localizar registros en los que un campo contiene todos los términos de búsqueda juntos; sin embargo, el orden de los términos no tiene que coincidir con el orden en el que se introdujeron. Por ejemplo, si se busca "Cuba NEAR historia", sólo se recuperarían los registros con los términos "Cuba " e "historia" juntos dentro del mismo campo bibliográfico. "Cuba" o "historia" podrían aparecer al principio del campo.

- Se utiliza el operador **ADJ** para localizar registros en los que un campo contiene todos los términos de búsqueda juntos y en el orden en el que se introdujeron. Por ejemplo, si se buscar "Cuba ADJ historia", sólo se recuperarían los registros con los términos "Cuba " e "historia" juntos dentro del mismo campo bibliográfico y con "Cuba" delante.

- Además, se puede añadir un número a los operadores posicionales **NEAR** y **ADJ** para limitar o ampliar la proximidad entre palabras. Por ejemplo, "DE ADJ1 AQUI ADJ3 ETERNIDAD" muestra cómo buscar el título "De aquí a la eternidad". ADJ3 significar que las palabras pueden encontrarse a dos palabras buscables la una de la otra, pero en el orden en el que se hayan introducido. [8]

Operadores de existencia

Estos operadores forman lenguajes de consulta básicos, pues tienen como fin forzar la presencia o no de determinados términos o palabras en los documentos recuperados, por lo general se usa el operador (+) delante de una palabra si se desea que esta esté presente en el documento recuperado, de igual manera si una palabra es precedida por el operador (-), esta será excluida de los resultados.

Una regla a tener en cuenta para usar estos operadores es que se deben poner delante de la palabra clave sobre la que actúan sin dejar espacio en blanco.

Ejemplo: +olímpico +baloncesto +fútbol +voleibol

Se puede representar también de la siguiente manera: +(olímpico baloncesto fútbol voleibol)

Una vez hecha la consulta, el SRI devolverá los documentos que contengan las palabras olímpico, baloncesto, fútbol y voleibol.

Se pueden utilizar algunos trucos y combinarlo con otros operadores.
Ejemplo: +voleibol "deporte escolar"

Ahora, si lo que se va a utilizar es el signo (-) habitualmente se indica añadiéndolo al inicio de la palabra clave.

Ejemplo: juventud –racismo

El mismo resultado se puede obtener de otra manera a través de estos dos operadores lógicos.

Ejemplo: juventud **AND NOT** racismo

Operadores de truncamiento

Pueden darse situaciones en las cuales no sea necesario utilizar un término simple, sino también sus derivados, es decir, cuando se realiza una búsqueda se utiliza una palabra o frase para la consulta, pero sería adecuado en determinado caso que se tuviese en cuenta las palabras derivadas de la palabras clave, es decir prefijos o sufijos, variantes léxicas mínimas, etc.

Los operadores de truncamiento facilitan este tipo de búsqueda. Se trata de operadores, normalmente símbolos como: (*), ($), (?) y (!), cuya presencia puede sustituir a un carácter o a un conjunto de caracteres, situados a la izquierda, dentro o a la derecha del término especificado.

Ejemplo:

ARREND* Recuperará documentos que contengan las palabras arrendamiento, arrendar, arrendatario, arrendador...

INFORM* Recuperará documentos que contengan las palabras informe, informes, informar, información...

CAS? Recuperá documentos que contengan las palabras caso, casa, casi..., pero no las palabras casos, casas...

PAGA Recuperá documentos que contengan las palabras impagado(s), pagaré, pagador(es)...

Los operadores de truncamiento se usan tanto para la búsqueda de referencias numéricas como alfanuméricas. Hay que tener en cuenta que cuanto más a la izquierda de la palabra se coloquen estos signos, más lenta será la respuesta a la consulta.

Operadores de límite o comparación

Los operadores de límite/comparación especifican el rango de búsqueda fijando las cotas para la misma, cotas que pueden ser tanto numéricas como alfabéticas, dichos operadores suelen ser del tipo: (< **menor**, > **mayor**, = **igual a**, <>**diferente de**, <= **menor o igual que**, >= **mayor o igual que**) o combinaciones de éstos. Estos operadores se utilizan principalmente en documentos contenedores de información numérica. [9]

Se puede llegar entonces a una conclusión en este sentido y es que dependiendo del SRI que sea se puede utilizar uno u otro de estos operadores en dependencia de cual o cuales de estos tenga implementado.

Se puede apreciar que el uso de los paréntesis para agrupar, también se utilizan en varios buscadores y así poder hacer más precisa y a la vez más compleja la búsqueda. De esta manera además podrían combinarse los operadores, ejemplo de esto sería:

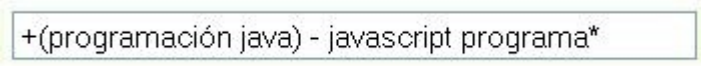

Figura 1.10: *Consulta utilizando combinación de operadores.*

En el caso de la consulta mostrada en la figura 1.10 se utilizan los operadores de existencia (+ y -) y el operador de truncamiento (*). El resultado de dicha consulta serían los documentos donde aparezcan las palabras “programación” y “java”, no aparezca la palabra “javascript” y esté presente además la palabra programa y todas sus derivaciones (programadores, programación, programando, entre otras).

Para un mejor entendimiento de la investigación es necesario además tener en cuenta algunos conceptos relacionados con Usabilidad, Calidad y Evaluación, entre otros ya que estos términos serán utilizados en la misma.

1.3.2 Usabilidad

Como la física estudia los fenómenos naturales y la filosofía el pensamiento humano, la *usabilidad* es la que estudia la interacción usuario - software. Es quien busca que los usuarios se sientan cómodos al usar un software determinado, que el trabajo en este no se vuelva difícil. Si el software es capaz de atraer al usuario entonces este tiene calidad y por consiguiente una técnica de usabilidad correctamente aplicada. El reto de la usabilidad es entender como los usuarios ven el software.

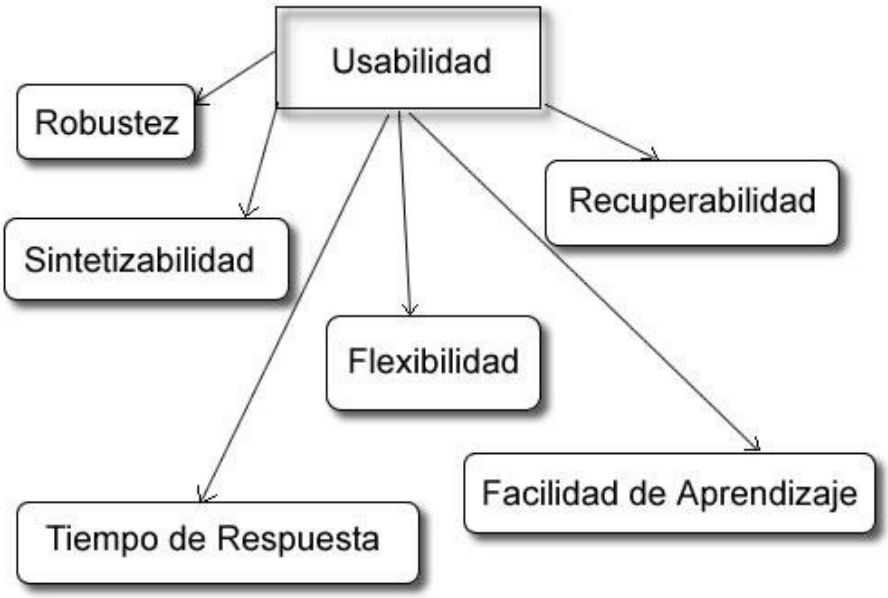

Figura 1.11: *Principales factores que definen Usabilidad.*

Facilidad de Aprendizaje

Necesidad de minimizar el tiempo que el usuario emplea en aprender a utilizar correctamente el software.

Tiempo de Respuesta

Capacidad del software para dar respuesta a las peticiones que le hace el usuario. Este factor es muy variable ya que depende de las características que tenga la PC (Personal Computer, computadora personal en español) donde se encuentre el usuario.

Flexibilidad

Diversidad de formas de intercambiar el usuario con el sistema la información. Aportar flexibilidad al sistema implica brindar control al usuario, capacidad de sustitución y capacidad de adaptación.

Robustez

Caracteriza la necesidad de que el usuario cumpla con sus objetivos.

Recuperabilidad

Facilidad que brinda el software o la aplicación al usuario para corregir alguna operación que previamente se haya hecho pero que se ha reconocido el error.

Sitetizabilidad

Este factor se caracteriza porque el usuario sea capaz de captar cuando ocurra algún cambio de operación en el sistema.

1.3.3 Calidad

La palabra calidad tiene múltiples significados:

- Propiedad o conjunto de propiedades inherentes a algo, que permiten juzgar su valor.
- Conjunto de propiedades inherentes a un objeto que le confieren capacidad para satisfacer o no necesidades implícitas o explícitas.
- Es la percepción que el cliente tiene del mismo, es una fijación mental del consumidor que asume conformidad con dicho producto o servicio y la capacidad del mismo para satisfacer sus necesidades. [10]

1.3.4 *Evaluación*

El concepto de evaluación se refiere a la acción y efecto de evaluar. Es un verbo cuya etimología[17] se remonta al francés *évaluer* y que permite señalar, estimar, apreciar o calcular el valor de algo.

En el lenguaje cotidiano, el concepto de evaluación es polisémico[18] porque éste se impone o no en la práctica según las necesidades mismas de la evaluación y en función de las diferentes formas de concebirla. En efecto, puede significar tanto estimar y calcular como valorar o apreciar. Así pues, la evaluación, en términos generales, supone una instancia de valoración.

1.3.5 Relevancia

El término de relevancia queda definido como calidad o condición de relevante, importancia, significación, y el término "relevante" se dice de algo importante o significativo.

En el contexto que trata en la investigación, un documento relevante seria aquel en que el contenido del mismo posea alguna significación o importancia con motivo de la consulta realizada por el usuario, es decir, con su necesidad de información.

Un mismo documento puede ser considerado relevante, o no relevante, por dos personas distintas en función de los motivos que producen la necesidad de información o del grado de conocimiento que sobre la materia posean ambos. Llegados a un caso extremo, un mismo documento puede parecer relevante o no a la misma persona en momentos diferentes de tiempo.

1.3.6 Posicionamiento web

El posicionamiento web es algo trabajoso y continuo, consiste en *"…aplicar diversas técnicas tendentes a lograr que los buscadores de Internet encuadren nuestra página web en una posición y categoría deseada dentro de su página de resultados para determinados conceptos clave de búsqueda."* [11]

Existen dos tipos de posicionamiento web, en dependencia del resultado que generan los motores de búsqueda, se puede clasificar en:

[17] *Se denomina **etimología** al estudio del origen de las palabras, cuándo son incorporadas a un idioma, de qué fuente y cómo su forma y significado han cambiado.*

[18] *La **polisemia** se presenta cuando una misma palabra tiene significados distintos.*

- Posicionamiento en enlaces patrocinados SEM *(del inglés: Search Engine Marketing*).
- Posicionamiento natural SEO *(del inglés: Search Engine Optimization, optimización para motores de búsqueda en español)*.

El primero se refiere al basado en los resultados patrocinados *"(…) cuya clasificación depende del dinero que se invierta en los anuncios".* [11]

Mientras que el segundo se refiere al cimentado en los resultados naturales u orgánicos, que "*(…) están basados en el algoritmo imparcial de los buscadores (…)".* [11]

1.3.7 Métricas

Las métricas pueden ser vistas de diferentes maneras y en todos los campos no significan lo mismo, la definición que más se asemeja al contenido tratado en la investigación es:

En el campo de la ingeniería del software una **métrica** es cualquier medida o conjunto de medidas destinadas a conocer o estimar el tamaño u otra característica de un software o un sistema de información, generalmente para realizar comparativas o para la planificación de proyectos de desarrollo. Un ejemplo ampliamente usado es la llamada métrica de punto función[19].

Métricas técnicas

Se centran en las características de software por ejemplo: la complejidad lógica, el grado de modularidad. Mide la estructura del sistema, el cómo está hecho.

Métricas de calidad

Proporcionan una indicación de cómo se ajusta el software a los requisitos implícitos y explícitos del cliente. Es decir cómo se va a medir para que el sistema se adapte a los requisitos que pide el cliente.

[19] *La **métrica del punto función** es un método utilizado en ingeniería del software para medir el tamaño del software.*

Métricas orientadas a la persona

Proporcionan medidas e información sobre la forma que la gente desarrolla el software de computadoras y sobre todo el punto de vista humano de la efectividad de las herramientas y métodos. [12]

1.4 MODELOS CONCEPTUALES DE RI

Una de las problemáticas actuales es lograr que las herramientas de búsqueda realicen una recuperación de la información lo más eficiente posible, dada la creciente cantidad de usuarios que hacen uso de las mismas para localizar la información deseada, así como la exigencia creciente de los mismos en cuanto a la calidad de los resultados que estos sistemas devuelven, el tiempo que demoran y la eficiencia de manera general.

En este punto cabría preguntarse: ¿Cómo estas herramientas realizan su trabajo y recuperan la información? Precisamente la respuesta se halla en los Modelos de Recuperación de Información, mediante los cuales se define la forma en que se van a realizar las comparaciones entre una consulta determinada y los documentos candidatos a convertirse en resultado de dicha consulta, estos modelos solo son aplicables a información de contenido textual y su funcionamiento consiste en la creación de un índice determinado en función del contenido de dicho documento a recuperar. Para la creación de estos índices se tienen en cuenta factores como por ejemplo la frecuencia con la cual aparece la palabra en el documento.

Se puede plantear que la mayor dificultad de los SRI es predecir que documentos son los más relevantes para una consulta determinada. En dependencia de las premisas que se adopten se implementará un determinado Modelo de Recuperación y en algunas ocasiones se puede optar por la combinación de algunos de ellos.

Luego de lo anterior, se puede afirmar que, conceptualmente un modelo de recuperación de información es una tupla < *D, Q, F, R* (q_i, d_j) >, donde:

- *D*: Representación lógica de los documentos.
- *Q*: Representación lógica de los requerimientos de información (*queries* o consultas).

- *F*: Marco para modelar documentos, consultas y sus relaciones. (modelo de recuperación).
- *R* (q_i, d_j): Función de Ranking.

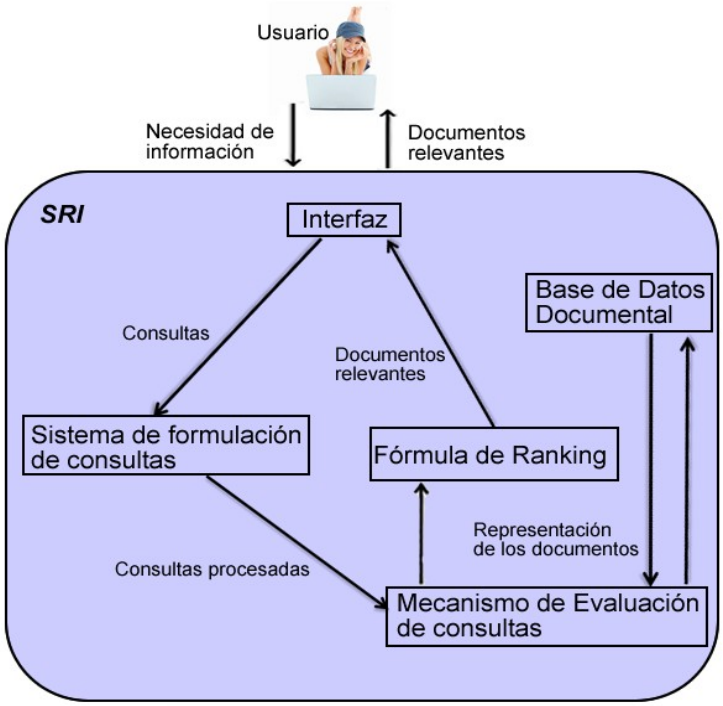

Figura 1.12: *Procesos básicos y funcionamiento de un SRI.*

En la tabla 1.2 se representan los conjuntos en los que se agrupan los principales modelos de recuperación de información hasta ahora definidos, siendo los clásicos los más utilizados por los SRI más exitosos en la actualidad.

Modelo	Descripción
Clásicos	Booleanos, Probabilísticos y basados en el Espacio Vectorial.
Alternativos	Basados en la Lógica Fuzzy.
Lógicos	Basados en la Lógica Formal.
Basados en la interactividad	Posibilidades de expansión del alcance de la búsqueda y uso de retroalimentación por relevancia.
Basados en la Inteligencia Artificial	Redes neuronales, bases de conocimiento, algoritmos genéticos y procesamiento del lenguaje natural.

Tabla 1.2: *Modelos conceptuales de recuperación de información.*

1.4.1 Clásicos

Modelo booleano

De todos los modelos que se analizarán, el booleano fue el primero que se empleó para la recuperación de información. Este modelo se basa en la teoría del Álgebra de Boole[20] (ya que utiliza los símbolos del algebra de Boole AND, OR, NOT, IF...THEN).

Un SRI booleano puro divide en dos categorías los posibles resultados de una búsqueda efectuada: satisface o no satisface. Es decir, la idea principal de este modelo es que una palabra clave puede estar ausente o presente en un documento y por tanto serán relevantes solo aquellos documentos que contengan las palabras clave especificadas en la consulta. Al considerar presente o ausente las palabras claves los pesos de estas en los documentos serán de (0 y 1) y formarán el conjunto de documentos recuperados aquellos que tengan un valor igual a 1.

Una de las principales desventajas que presenta este modelo es que no se devolverán documentos que podrían ser relevantes a pesar de que no coincidan exactamente con la consulta. Además cuando se encuentren muchos documentos que satisfagan la consulta, el sistema no tiene un modo eficiente de decidir cuáles de ellos son más relevantes y por tanto mostrarlos en un mejor orden por relevancia.

[20] ***Álgebra de Boole:*** *Boole definió un álgebra aplicable a los razonamientos sobre proposiciones lógicas: una proposición puede ser cierta o falsa y esto se anota con un 0 o con un 1. Una variable lógica o booleana es una variable binaria que toma los valores que anotamos convencionalmente con los símbolos 0 y 1. Una función lógica o booleana es una función de n variables lógicas que toman valores del conjunto {0, 1}.*

Modelo de espacio vectorial

Se observaba claramente en aquel momento, que el modelo booleano tenía deficiencias y no satisfacía las exigencias de los usuarios por lo que se comienzan a generar modelos alternativos como el Modelo Vectorial. Salton[21] fue el primero en proponer los SRI basados en Espacio Vectorial (SRI-EV) a finales de los 60, dentro del marco del proyecto SMART.

La idea básica de este modelo de recuperación vectorial reside en la construcción de una matriz (*podría llamarse tabla*) de términos o alfabeto inicial y documentos, donde las filas fueran estos últimos y las columnas correspondieran a los términos incluidos en el alfabeto. El alfabeto se obtiene inicialmente extrayendo todos los términos de la colección de documentos y luego aplicando algoritmos de normalización, extracción de palabras poco relevantes y sin valor así como eliminando las repeticiones de los mismos términos entre otras técnicas utilizadas.

De esta manera las columnas de esta matriz estarían representadas por una palabra o término determinado del alfabeto y las filas (que en términos algebraicos se denominan *vectores*) serían equivalentes a los documentos que se expresarían en función de las apariciones (*frecuencia*) de cada término.

Finalmente, los documentos podrían expresarse de la siguiente manera:

Suponiendo que los textos de los distintos documentos sean:

d1 → “Hola mundo”

d2 → “El mundo de la informática es el mundo del mañana”

d3 → “Universidad de las Ciencias Informáticas”

d (i-ésimo) → “…..”

	d1	*d2*	*d3*	*dn*
hola	1	0	0	Valor1
mundo	1	2	0	Valor2
informática	0	1	1	Valor3

[21] ***Gerard Salton*** *(Nuremberg (Alemania), 8 de marzo de 1927 - Nueva York (EUA), 28 de agosto de 1995) fue un informático y documentalista científico estadounidense de origen alemán. Especialista en RI y en procesamiento del lenguaje natural.*

mañana	0	1	0	Valor4
universidad	0	0	1	Valor5
ciencias	0	0	1	Valor6

Tabla 1.3: *Representación de la matriz de vectores.*

Si se obtienen las filas de esta matriz se estarían representando los documentos en forma de un vector de la siguiente manera.

d1 = (1; 1; 0; 0; 0; 0)

d2 = (0; 2; 1; 1; 0; 0)

d3 = (0; 0; 1; 0; 1; 1)

d (i-ésimo) = (Valor1; Valor2; Valor3; Valor4; Valor5; Valor6)

Siendo cada uno de estos valores el número de veces que aparece cada término del alfabeto en el documento. La longitud del vector de los documentos sería igual al total de términos del alfabeto (el número de columnas).

Partiendo de que se pueden representar los documentos como vectores de términos, estos podrán situarse en un espacio vectorial de *n* dimensiones, es decir, con tantas dimensiones como elementos tenga el vector. Situado en ese espacio vectorial, cada documento cae entonces en un lugar determinado por sus coordenadas, al igual que en un espacio de tres dimensiones cada objeto queda bien ubicado si se especifican sus tres coordenadas espaciales. Se crean así grupos de documentos que quedan próximos entre sí a causa de las características de sus vectores (clúster de documentos).

De igual manera la consulta cuando es formulada por el usuario, también se representa en un vector de la misma manera que los documentos y se deja caer en este espacio vectorial y, así, aquellos documentos que queden más próximos a ella serán, en teoría, los más relevantes para la misma. La representación de los documentos y las consultas se realiza mediante la asociación de un vector de pesos no binarios (un peso por cada término de índice).

En una base de datos documental organizada de esta manera, resulta muy rápido calcular la relevancia de un documento a una consulta, y siendo muy rápida también la ordenación por relevancia, ya que, de forma natural, los documentos ya están agrupados por su grado de semejanza.

Modelo probabilístico

El modelo probabilístico está compuesto por conjuntos de variables, operaciones con probabilidades y el Teorema de Bayes[22]. Está basado en el que se ha traducido como el *Principio de la ordenación por probabilidad,* este principio, formulado por Robertson[23], asegura que el rendimiento óptimo de la recuperación se consigue ordenando los documentos según sus probabilidades de ser juzgados relevantes con respecto a una consulta, siendo estas probabilidades calculadas de la forma más precisa posible a partir de la información disponible. Es decir él sugiere que cuantas más pruebas o evidencias se tengan sobre la consulta, sobre los documentos y sobre las relaciones entre ellos, mayores serán las probabilidades de que los resultados se adecuen a la necesidad informativa del usuario.

En este modelo, los documentos y las consultas se representan por un vector binario. Así, un documento cualquiera tiene la siguiente forma:

$$d_j = (t_1, t_2, \ldots, t_n)$$

Donde *ti* = 0 ó 1 indica la ausencia o presencia del término i-ésimo, respectivamente, y *n* el número de términos de la colección.

Existen dos eventos mutuamente excluyentes:

[22] ***El teorema de Bayes****, enunciado por Thomas Bayes, en la teoría de la probabilidad, es el resultado que da la distribución de probabilidad condicional de una variable aleatoria A dada B en términos de la distribución de probabilidad condicional de la variable B dada A y la distribución de probabilidad marginal de sólo A. Es válido en todas las aplicaciones de la teoría de la probabilidad.*

[23] ***Stephen E. Robertson*** *es un investigador de Microsoft Research Laboratory en Cambridge, Reino Unido. Mantiene un profesorado a tiempo parcial en el Departamento de Ciencias de la Información, que forma parte de la Escuela de Informática en la Universidad de la Ciudad. Sus principales intereses de investigación se encuentran en las teorías y modelos de recuperación de información, en particular modelos probabilísticos, el diseño y evaluación de sistemas de infrarrojos, los métodos de evaluación y optimización. Autor de varias publicaciones referidas al tema.*

w1, que representa el hecho de que un documento sea relevante, y *w2*, que indica que no lo sea. Este modelo asume que se conocen, o por lo menos se suponen, el conjunto de documentos relevantes (*R*) y no relevantes (*r*) de una consulta dada.

El objetivo que se persigue es calcular *p (w1|dj)* y *p (w2|dj)*, es decir, la probabilidad de que el documento *dj* sea relevante o no relevante, respectivamente, dada una consulta *q* y desarrollar una función que ofrezca un valor de relevancia para así poder ordenar los documentos según ella. En este caso, esa función tendrá la forma:

$$Sim(d_j, q) = \frac{p(\omega_1|d_j)}{p(\omega_2|d_j)}$$

Dentro de la recuperación probabilística, se utiliza el **modelo de recuperación probabilístico de independencia de términos binarios** donde: "*La probabilidad de los términos es independiente (un término es independiente de los otros)" y "los pesos asignados a los términos son binarios".*

La equiparación probabilística se basa en que, dado un documento y una consulta, es posible calcular la probabilidad de que ese documento sea relevante para esa consulta.

Si un documento es seleccionado aleatoriamente de la base de datos hay cierta probabilidad de que sea relevante a la pregunta. Si una base de datos contiene N documentos, n de ellos son relevantes, entonces la probabilidad se estima en:

$$P\ (rel) = n/\ N$$

En concordancia con la teoría de la probabilidad, el valor de que un documento no sea relevante a una consulta realizada viene expresado por la siguiente fórmula:

$$P\ (\downarrow rel) = 1 - P\ (rel) = N - n/\ N$$

Obviamente, los documentos no son elegidos aleatoriamente, sino que se eligen sobre la base de la equiparación con la consulta, basado en el análisis de los términos contenidos en ambos. Así, la idea de relevancia está relacionada con los términos de la pregunta que aparecen en el documento. Dividiendo la colección de documentos en dos conjuntos: los que responden a la consulta y los que no.

1.4.2 Alternativos

Modelo de Lógica Difusa (Fuzzy)

El modelo de lógica difusa es similar al probabilístico, pero cambia la necesidad de estimar la probabilidad por una necesidad de primar la creencia sobre la relevancia de un documento dado. Ésta es una regla estricta que gobierna el uso de las probabilidades no aplicadas.

En sentido real, un documento *fuzzy* no existe. Es decir, los autores no asignan grados de pertenencia a los términos o a los conceptos en sus documentos. De cualquier forma, puede hacerse un "juicio difuso" sobre cuándo un documento debería estar en el conjunto de coincidentes con la pregunta. Ésta es la base del conjunto de términos que describen un documento o los términos usados en él. En la equiparación probabilística, el cálculo último devuelto sobre la probabilidad de que los términos de los documentos sean potencialmente relevantes a una pregunta, está contenida en los documentos relevantes y en los no relevantes. En el modelo de lógica difusa, el cálculo se define basándose en el grado de pertenencia de los términos. La cuestión llega a ser tal, que el grado de confianza de que un documento contenga un término dado es relevante. Si esto se usa para definir el grado de pertenencia, entonces este grado con respecto al conjunto de documentos relevantes, puede ser computado para cualquiera de los documentos.

El concepto de lógica difusa ha tenido diferentes enfoques. Si se consideran términos relacionados semánticamente, entonces se podrá considerar en qué grado un término relacionado es equiparable con un término dado. Por ejemplo, en la pregunta sobre el "cocker spaniel", un documento que contiene el término "springer spaniel" no será comparable, pero las dos razas de perros están relacionadas estrechamente, y en grado suficiente para que el documento pueda contener información útil. Otro documento sobre canes en general también puede contener información útil, pero quizás menos que el anterior. Así que, todo depende de cómo de específica sea la consulta sobre el "cocker spaniel", o el juicio *fuzzy*.

Otro enfoque se basa en la búsqueda de descriptores que ofrezcan alguna indicación del valor de la información en el documento. Y éstos pueden ser tanto cualitativos como cuantitativos. Los indicadores cualitativos pueden ser adjetivos calificativos como “pequeño”, “grande”, etc., que puedan indicar una escala cercana a la numérica u otros que no como, por ejemplo, “bonito”, “feo”, “colorido”, etc. Los descriptores que pueden considerarse “cuantitativos” incluyen palabras como “pocos”, “la mayoría de”, etc. El uso de un término en un documento también puede ser descrito de manera *fuzzy*: tales como “muy significativo”, “muy importante”, y los resultados de la recuperación descrita como altamente relevantes o parcialmente relevantes. El problema, entonces, es decidir cómo cada término se traduce o transforma en una función de pertenencia asociada con la recuperación difusa. Un proceso de comparación difusa puede aparecer combinado con otros procesos como, por ejemplo, el enfoque booleano extendido.

1.4.3 Basados en la interactividad.

Modelo Interactivo basado en el uso de retroalimentación por relevancia

Este modelo pretende obtener el mayor número de documentos relevantes tras establecer varias estrategias de búsqueda. La idea es que, tras determinar unos criterios de búsqueda y observar los documentos recuperados se vuelva a repetir nuevamente la consulta pero esta vez con los elementos interesantes, seleccionados de los documentos primeramente recuperados. El Algoritmo Genético[24] es el que se ha utilizado para llevar a cabo este tipo de técnicas de recuperación.

Además se puede plantear que una retroalimentación por relevancia también tiene como significado que los usuarios confiables de una comunidad actúan en conjunto para mejorar la búsqueda del SRI de una manera pública, transparente y abierta, mejorando de esta manera la herramienta en cuestión. Ejemplo: Wikia Search.

---

[24] ***Algoritmo Genético:*** *Son llamados así porque se inspiran en la evolución biológica y su base genético-molecular. Un algoritmo genético es un método de búsqueda dirigida basada en probabilidad. Bajo una condición muy débil (que el algoritmo mantenga elitismo, es decir, guarde siempre al mejor elemento de la población sin hacerle ningún cambio) se puede demostrar que el algoritmo converge en probabilidad al óptimo. En otras palabras, al aumentar el número de iteraciones, la probabilidad de tener el óptimo en la población tiende a 1 (uno).*

Se identificó a partir de un estudio realizado diferentes tipos de retroalimentación con necesidades informativas reales:

Retroalimentación por relevancia de contenido: Se trata de la reformulación de una consulta teniendo en cuenta los juicios de relevancia que el usuario emite, como consecuencia de la respuesta que el sistema ofrece a una consulta anterior de éste. Este tipo de retroalimentación puede ser negativa o positiva, dependiendo de que según el juicio del usuario, los documentos sean relevantes o no.

Retroalimentación por relevancia de los términos: Los usuarios reformulan la consulta incorporando nuevos términos. Éstos se obtienen de la observación y valoración de los resultados obtenidos en la consulta anterior.

Retroalimentación por magnitud de la respuesta: Este tipo de retroalimentación la realiza el usuario en función del número de documentos que el sistema le ofrece como respuesta a su consulta. Puede ser tanto positiva, como negativa, dependiendo del exceso o defecto en el número de resultados.

Retroalimentación por revisión de consultas anteriores: Muchas veces, a raíz de largas interacciones con el sistema, como consecuencia de una necesidad informativa poco clara o de un SRI poco eficiente, el usuario se encuentra perdido y no sabe qué estrategias de búsqueda ha utilizado. Como solución mira en el historial[25] de su pantalla de búsqueda, en el caso de que la interfaz disponga de él.

Retroalimentación por revisión de términos utilizados en consultas anteriores: Se trata del mismo caso que el anterior, con la diferencia de que aquí se buscan términos utilizados en consultas anteriores en lugar de estrategias. Su utilización en las interacciones es prácticamente despreciable.

De los modelos vistos anteriormente los más flexibles y eficientes para la RI son el modelo vectorial y el modelo probabilístico (aunque este último utiliza formulas de probabilidad con determinado grado de complejidad, lo cual puede perjudicar el tiempo de respuesta del SRI), puesto que el booleano es muy simple y se vasa solo en si el documento cumple o no con la consulta formulada, careciendo de flexibilidad en la recuperación de información. Los restantes modelos de RI vistos presentan algunos problemas,

[25] *En el **historial** se guardan los lugares a donde ha accedido un usuario, es como un rastro que se va quedando guardado en la PC de las cosas que ha hecho un usuario, por ejemplo: en el historial de un navegador se guardan las páginas a dónde se ha entrado.*

fundamentalmente por la complejidad de su implementación práctica y el tiempo de respuesta que presentarían estos luego de realizar tareas complejas de lógica difusa o inteligencia artificial.

CAPÍTULO 2

ANÁLISIS DE TENDENCIAS PARA LA EVALUACIÓN DE SRI

La evaluación de Sistemas de Recuperación de Información se encuentra en un momento crucial en el que se suceden cambios, se está realizando constantes aportes y se emprenden rigurosas investigaciones. Sus antecedentes, los nuevos proyectos, prototipos y experimentos de evaluación que actualmente se llevan a cabo, se remontan a las décadas de los años 50 y 60, desde entonces se pueden detectar varias aproximaciones a la evaluación de los SRI. La necesidad crítica de evaluación que poseen los sistemas de recuperación de información, propicia que las evaluaciones de estos nuevos sistemas surjan de forma casi simultánea a su puesta en marcha, y que paralelamente, comience la comunidad de

usuarios a plantearse dudas sobre cuál de ellos es el que mejor responde a sus necesidades y posee la mayor porción de documentos de la web.

Es decir los SRI como cualquier otro sistema son sometidos a evaluación con el fin de que sus usuarios se sientan en condiciones de valorar su efectividad y de esta manera adquieran confianza en los mismos. Las evaluaciones de los SRI se encuentran estrechamente vinculadas con la investigación y el desarrollo de la recuperación de la información, de forma paralela al desarrollo de su tecnología, ha surgido un amplio campo de trabajo dedicado específicamente al establecimiento de medidas que permitan valorar su efectividad.

Una serie de trabajos especializados permiten identificar varios grupos de evaluaciones las cuales serán tratadas en este capítulo.

2.1 PRINCIPALES ESTUDIOS SOBRE EVALUACIÓN DE SRI

2.1.1 Proyectos Cranfield

Se consideran los primeros estudios significativos, que proporcionaron una nueva dimensión de la investigación de los SRI. Estos estudios se llevaron a cabo en el Instituto Cranfield de Tecnología y se consideran además el punto de partida de las investigaciones empíricas y experimentales sobre la recuperación de la información. Originalmente, este estudio proyectaba evaluar el funcionamiento de varios sistemas de indización y el rendimiento de los SRI basados en ellos, este trabajo además trajo consigo nuevas métricas para la evaluación como son la Precisión, Exhaustividad, Tasa de fallo, entre otras.

Son 2 los estudios Cranfield más importantes. El primero de ellos, dirigido por Cleverdon[26], comenzó en 1957 y tenía como objetivo comparar la efectividad de cuatro sistemas de indización: un catálogo alfabético de materias basado en una lista de encabezamientos; una clasificación decimal (concretamente la Clasificación Decimal Universal o CDU); un catálogo basado en una clasificación por facetas y

[26] ***Cyril W. Cleverdon*** *(Bristol, 1914 - Cranfield, 1997), documentalista científico inglés, pionero de la disciplina Recuperación de información en sistemas documentales. Proporcionó un objeto de estudio, una metodología de investigación y un lenguaje terminológico, siendo el comienzo de investigaciones empíricas y de corte experimental. Además, estableció las directrices apropiadas en la indización automatizada.*

finalmente, un catálogo compilado de un índice coordinado de unitérminos[27] (concretamente el modelo UNITERM).

El test probó que el rendimiento de un sistema no depende de la experiencia del indizador; en segundo lugar, mostró que los sistemas donde los documentos se organizan por medio de una clasificación facetada rendían menos que los basados en un índice alfabético.

El segundo proyecto Cranfield consistió en un experimento controlado destinado a fijar los efectos de los componentes de los lenguajes de indización en la ejecución de los SRI.

Los resultados que proporcionó este nuevo experimento fueron contradictorios, principalmente a la hora de seleccionar los términos más adecuados para representar los conceptos contenidos en los documentos, ya que los sistemas de indización libre (no controlados) ofrecieron mejor rendimiento que los controlados, obteniéndose mejores resultados con lenguajes de indización basados en los títulos de los artículos que en los basados en los resúmenes.

2.1.2 Conferencias TREC

Las conferencias TREC (Text REtrieval Conferences) se han convertido en el foro de intercambio científico más prestigioso del campo de la recuperación de información.

TREC reúne a creadores de diferentes sistemas y compara los resultados que éstos obtienen en diferentes pruebas, previamente estandarizadas y acordadas por todos. Este foro se viene celebrando anualmente desde 1991. Nace con la idea de resolver uno de los mayores problemas de las evaluaciones de los SRI: las mismas suelen llevarse a cabo sobre pequeñas colecciones de documentos, y sus resultados resultan de difícil extrapolación a la totalidad de la colección almacenada.

---

[27] *Sistema de almacenamiento y recuperación de la información (y no sistema de clasificación, como erróneamente es definido en muchas fuentes) creado por el investigador Mortimer Taube en 1951, que consiste en un método de indización por palabras únicas o asuntos simples, generalmente tomadas del lenguaje natural, recuperables mediante la postcoordinación de conceptos.*

La primera conferencia, TREC-1 (1992), ofreció como resultado principal el hecho de la existencia de una amplia similitud entre los SRI que hacen uso de técnicas basadas en lenguaje natural y los basados en los modelos probabilístico y los basados en el modelo del vector.

En la conferencia TREC-2 (1993), se detectó una significativa mejoría de la recuperación de información, con respecto a la anterior. Las siguientes conferencias aportaron nuevas prestaciones a los experimentos: localización de información en varias bases de datos de forma simultánea, presencia de errores ortográficos con el fin de valorar el comportamiento de los SRI ante ellos y recuperación de información en idiomas distintos del Inglés (se eligieron el Español y el Chino) para valorar los posibles cambios de comportamiento de los SRI. Estas conferencias han aportado la evaluación de variadas modalidades de recuperación de información (desde el clásico modelo booleano a la búsqueda por cadenas de texto o las búsquedas basadas en diccionarios), y han demostrado hasta qué punto pueden alcanzarse resultados significativos de investigación a través de la cooperación entre investigadores en el ámbito mundial.

2.1.3 Oppenheim (2000)

Los autores de este trabajo, Oppenheim, Morris y McKnight llevan a cabo un estudio exhaustivo de las metodologías aplicadas en la evaluación de los distintos motores de búsqueda por varios autores anteriores y a partir de la síntesis de estos procedimientos, exponen una serie de parámetros que les permiten formular una metodología de evaluación de los motores de búsqueda. En dicha investigación se expone que los métodos más empleados para la evaluación de los motores de búsqueda pueden agruparse en cuatro categorías:

1- Evaluaciones a pequeña escala.
2- Evaluaciones basadas en los test Cranfield.
3- Evaluaciones basadas en los test Cranfield con estimación del tamaño del motor.
4- Evaluaciones que eluden la exhaustividad.

El primer método lo componen estudios que han llevado a cabo un conjunto de evaluaciones a pequeña escala. En estos casos, los experimentadores realizan un escaso número de preguntas sobre un grupo reducido de motores de búsqueda. A continuación, analizan todos los resultados y determinan la precisión, el solapamiento entre los motores de búsqueda y la exhaustividad relativa de cada motor.

El segundo método lo componen las evaluaciones basadas en los test Cranfield. Es usual que en la práctica la mayoría de las evaluaciones realizadas calculen la precisión y la exhaustividad de cada motor o alguna medida derivada de las anteriores, tal como es el caso frecuente de la exhaustividad relativa. Generalmente, estas medidas se calculan a partir de una muestra de documentos recuperados, muestra cuyo tamaño es variable según el estudio, aunque, por regla general, su valor oscila entre los veinte y treinta primeros documentos recuperados por cada motor.

Un tercer método, variante del anterior, lo conforman las evaluaciones basadas en los test Cranfield con estimación del tamaño del índice del motor. Estos estudios suelen ser de bastante extensión ya que realizan cientos de preguntas a distintos motores de búsqueda y manejan colecciones de resultados de gran magnitud. Para calcular los tamaños de los índices se siguen diversas técnicas. Una de ellas, la más rudimentaria, es identificar el motor que más resultados ofrece sobre una pregunta, contar cuántas páginas relevantes devuelve y luego calcular hasta qué porcentaje de este valor alcanza el resto, extrapolando que, esos porcentajes de tamaño relativo van a mantenerse. Esta técnica tan simple adolece de varios problemas, porque ¿cómo se puede estar seguro de que el motor que más devuelve puede representar toda la web? y ¿qué pasa con el resto de los documentos relevantes que un motor no tiene en común con el motor de mayor volumen?

El último método consiste en evaluaciones que eluden la exhaustividad, es decir que evitan hacer uso de esta medida. Varios estudios la obvian con el objeto de no basar sus juicios en una medida algo imprecisa, sino solamente ser inferida a partir de unas estimaciones. No por ello estos estudios dejan de ser completos, ya que la mayoría analiza muchas variables de los motores de búsqueda, tales como: precisión, características de la interface, herramientas para la formulación de preguntas, presentación de los resultados, sintaxis empleada y tipos de agentes inteligentes usados en la recopilación de información.

Para los autores de esta investigación las evaluaciones de los motores de búsqueda deberían incluir, como mínimo, los siguientes criterios:

- Precisión.
- Exhaustividad relativa.
- Velocidad de respuesta, analizada varias veces al día y calculada en términos de promedio.
- Consistencia de resultados a lo largo de un determinado período tiempo.

- Proporción de enlaces fallidos.
- Proporción de duplicados.
- Calidad promedio de resultados.
- Evaluación de la amigabilidad de la interface.
- Calidad de la ayuda.
- Opciones para la visualización de los resultados.
- Presencia de avisos en la pantalla.
- Cobertura.
- Longitud esperada de búsqueda.
- Longitud y legibilidad del resumen.
- Efectividad del motor de búsqueda.

Asimismo, los autores consideran que los ensayos deben realizarse haciendo uso de tres tipos de búsqueda: de referencia simple, frases en lenguaje natural y expresiones booleanas.

2.1.4 Savoy y Picard (2001)

Este estudio aborda la evaluación de los SRI en la web desde una perspectiva diferente, en tanto que no analiza a unos motores de búsquedas específicos, sino que evalúa la efectividad de los modelos de recuperación de información en los que se encuentran basados los algoritmos sobre los que se construyen estos SRI, “investigando si las técnicas usadas en los SRI mejoran la efectividad en la recuperación de información cuando se aplican a una colección de documentos web”.

Este estudio pone en duda la aplicación de técnicas tradicionales tan comunes como los procedimientos de aislamiento de la base de las palabras, la presencia de una lista de palabras vacías o la tradicionalmente aceptada importancia de la aparición de las palabras en el título de la página.

Aunque este estudio analiza la realización de los distintos modelos y no ofrece como resultado la designación de un motor de búsqueda como el mejor de los evaluados, sí que resulta interesante conocer qué parámetros emplea para comparar la efectividad de los distintos modelos y qué resultados ofrecen. Además, presenta una serie de datos muy interesantes procedentes de las conferencias TREC sobre el cálculo de la precisión en términos de promedio tras analizar los primeros mil documentos devueltos.

Este estudio indica que el añadir palabras en las preguntas puede mejorar significativamente la precisión promedio y que asignar una mayor importancia (o peso) a las palabras que aparecen en el título o en los encabezados de las páginas, no posee efectos significativos sobre la precisión. Casi de forma paralela a la realización de este estudio, estos mismos autores plantearon un análisis de la efectividad mostrada por tres modelos de recuperación de información en la web: el clásico (representado por el TRECeval software y que siguen la mayoría de los motores de búsqueda del mercado), el de extensión de enlaces (el usado por Google) y uno basado en el modelo probabilístico.

Las conclusiones resultan bastante clarificadoras, los resultados que ofrece el modelo probabilístico son, al menos, igual de buenos que los ofrecidos por un SRI en la web cuyo algoritmo de alineamiento hace un uso extensivo de los enlaces y ambos mejoran ligeramente (alrededor de un cinco por ciento), los resultados de precisión media frente a los sistemas basados en el modelo clásico del vector.

2.1.5 Schlichting y Nilsen (1997)

Para los autores de este trabajo, las primeras evaluaciones que se realizaron de los motores de búsqueda "poseían un escasísimo nivel científico". Schlichting y Nilsen indican que los trabajos de Winship y de Leighton[28] en 1995 en cuanto a la evaluación de SRI constituyen un avance importantísimo en este campo de trabajo. En estos trabajos, los autores se percataron del problema de la presencia de documentos duplicados en las respuestas de los motores de búsqueda, de hecho, Leighton suprimió los documentos duplicados en su análisis. No obstante, tal como opinan Schlichting y Nilsen, "con la puesta en funcionamiento de motores de búsqueda de gran tamaño (como Alta Vista), cuyos índices contienen más de treinta millones de páginas, medir el número de aciertos que proporcionan a una pregunta, no resulta a la larga una medida efectiva, la calidad de los resultados es mucho más importante que la cantidad de documentos que se entreguen como resultado".

[28] *Estos trabajos fundamentalmente se basaban en crear metodologías cuantitativas para evaluar la eficacia de los SRI utilizando un número reducido de preguntas, en otro de los estudios se plantearon ocho consultas de diferente dificultad en Infoseek, Lycos, WebCrawler y WWWWorm. Los mejores resultados en cuanto a precisión y tiempo de respuesta fueron los de Lycos e Infoseek. Posteriormente, Leighton y Srivastava mejoraron y ampliaron este estudio, comparando Altavista, Excite, Hotbot, Infoseek y Lycos.*

El objetivo de este trabajo es mostrar una metodología que pueda emplearse para comparar motores de búsqueda y otro tipo de posibles futuros sistemas inteligentes de recuperación de información. Esta metodología debía encontrar alguna manera de medir la calidad de los resultados ofrecidos por un motor de búsqueda. Los autores proponen hacer uso de la metodología SDA[29] (Signal Detection Analisis), que proporciona dos medidas: d' que mide la sensibilidad del motor de búsqueda en hallar información útil y ß que mide cómo de conservador (o de liberal) es el comportamiento del motor de búsqueda a la hora de determinar qué páginas deben formar parte de la respuesta (mide el grado de flexibilidad del motor a la hora de considerar relevante un nuevo documento analizado).

En esta metodología se le asigna una categoría a cada uno de los resultados devueltos. Estas categorías se fundamentan en la asignación de relevancia de cada uno de estos sujetos, y son cuatro:

- **Acierto:** resultado relevante localizado en el experimento.
- **Falsa alarma:** resultado irrelevante localizado en el experimento.
- **Perdido:** documento relevante no localizado en el experimento.
- **Rechazado correctamente:** resultado irrelevante debidamente localizado en el experimento.

El más claro resultado de este estudio es que el rendimiento de los motores de búsqueda está muy lejos de ser considerado ideal. Obviando los pobres resultados de este estudio en particular, la gran contribución de este trabajo es la sugerencia de un método objetivo para la evaluación de la efectividad de la recuperación de información.

2.1.6 Johnson, Griffiths y Hartley (2001)

Este estudio se desarrolló en la Universidad Metropolitana de Manchester (Reino Unido) y estuvo destinado a establecer un marco global para la evaluación de los motores de búsqueda en Internet. La estructura de este informe se divide en cuatro capítulos.

El primero de ellos proporciona una introducción general a los motores de búsqueda en Internet y a su evaluación. El segundo capítulo muestra de forma gráfica la evolución de estos ingenios en la búsqueda de la mejora de su ejecución. El tercer capítulo revisa las metodologías empleadas en la evaluación de

[29] *La **metodología SDA** proporciona una visión más detallada del rendimiento de los motores de búsqueda incorporando más información dentro de una visión global.*

estos sistemas, partiendo desde una perspectiva sistémica hacia un punto de vista centrado en el usuario. En último lugar, en el cuarto capítulo de este informe se construye un marco global para la evaluación de los SRI en la web, basado en criterios de satisfacción del usuario. En la revisión de las metodologías empleadas hasta entonces para la evaluación de los motores de búsqueda, los autores identifican tres grupos:

- Evaluaciones basadas en los test Cranfield.
- Evaluaciones de sistemas interactivos centradas en el usuario.
- Evaluaciones basadas en la satisfacción del usuario.

El método de las evaluaciones basadas en los test Cranfield es el más empleado a lo largo de las evaluaciones realizadas. Las medidas más significativas empleadas dentro de este método son la exhaustividad y la precisión, a pesar de sus problemas de determinación exacta.

Resultan especialmente interesantes las recomendaciones de procurar evitar, en lo máximo, favorecer a un motor frente al resto de los analizados cuando se lleva a cabo un proceso de evaluación. También recomiendan contemplar la naturaleza dinámica de la web y llevar a cabo los ensayos en períodos breves de tiempo para que los resultados puedan ser comparables. Por último, hacen referencia a las posibles dimensiones en las que se puede medir la relevancia.

El segundo grupo de métodos es el que los autores denominan evaluaciones de sistemas interactivos centradas en el usuario. Johnson, Griffiths y Hartley comentan que muchos autores reniegan del trascendental protagonismo que el método anterior le confiere a los juicios de relevancia, considerando a esta trascendencia como la culpable de los bajos niveles de efectividad que proporcionar normalmente estos estudios.

Este grupo lo conforman aproximaciones alternativas a la evaluación de los SRI que eluden hacer uso de los juicios de relevancia. Algunas de estas medidas alternativas propuestas son: la utilidad de los documentos devueltos, la usabilidad, una medida que “pretende involucrar al usuario aún más en la evaluación” ya que se ha estudiado desde muchos puntos de vista: exactitud, tasa de error, número de comandos empleados, número de descriptores utilizados, pantallas necesarias para la recuperación de información y percepciones del usuario.

El tercer grupo de métodos empleados para la evaluación de los motores de búsquedas, es el de las evaluaciones basadas en la satisfacción del usuario. Una experiencia piloto en este campo agrupó quince medidas en cinco categorías: relevancia, eficiencia, utilidad, satisfacción del usuario y conectividad de la página. Un total de once evaluadores participaron en la experiencia realizando diversas cuestiones sobre cuatro motores de búsqueda, cada uno de ellos llevó a cabo sus juicios de relevancia (con base en las categorías anteriores).

Con lo cual, este estudio, vino a representar de forma empírica, las principales necesidades de un usuario de un SRI, el usuario necesita información de valor que cubra sus necesidades de información y la misma le debe ser entregada en un espacio de tiempo pequeño.

Como se evidenció anteriormente los proyectos Cranfield constituyeron la base fundamental y el inicio de las evaluaciones de los SRI. Sus aportes a la recuperación de información fueron fundamentales y dieron origen a toda una serie de conceptos y métricas de evaluación de SRI que en muchos casos, a través de los años se han mantenido vigentes. Por su parte, las Conferencias TREC se han constituido como un espacio propicio para el intercambio en materia de recuperación de información, dando lugar a importantes aportes e investigaciones en este campo. El tercero de los trabajos analizados, el de Oppenheim, presenta una sugerencia de criterios mínimos necesarios a tener en cuenta en el diseño de una metodología de evaluación, fruto de una exhaustiva síntesis de las medidas empleadas en otros trabajos de evaluación de estos sistemas. A este trabajo le sigue un interesante estudio realizado por Saboy y Picard (2001), en el que analizan la efectividad de los distintos modelos sobre los que se basan los SRI en la web. Este estudio no analiza el comportamiento de un motor específico frente a otro, sino que estudia la reacción producida al trasladar un modelo diseñado en su origen para el entorno de los SRI tradicionales al nuevo contexto de la web. El siguiente trabajo analizado, expone la necesidad de encontrar una metodología ajena a los juicios de relevancia, basada en parámetros de sensibilidad y utilidad de los documentos. Por último, el estudio propuesto por Johnson, Griffiths y Hartley, presenta una propuesta global de evaluación de los SRI, elaborada desde el punto de vista del usuario final.

Del análisis de los estudios anteriores se desprende la necesidad de aplicar un enfoque multidimensional a la evaluación de la efectividad de la recuperación de información en la web, donde exista un cierto grado de relación entre las dimensiones analizadas. Estas metodologías resultan de difícil aplicación por dos razones: la complejidad del cálculo de las medidas propuestas y, en segundo lugar, el considerable nivel de abstracción de algunas de estas medidas, que las alejan del usuario convencional de la web.

2.2 ANÁLISIS DE ALGUNOS SRI

El amplio conjunto de herramientas que han surgido con el desarrollo de Internet y que ayudan a la recuperación organizada de la información, hacen más difícil seleccionar las que mejor se adaptan a las necesidades de recuperación de información de cada persona.

A medida que pasan los años el cúmulo de información en Internet se hace mayor y esto trae consigo que los usuarios se hagan más dependiente de estas herramientas de búsqueda, las cuales a su vez tienen el enorme reto de ser más eficientes, agradables y útiles a quienes las usan, aunque aún les queda un gran camino por recorrer estas herramientas hacen su mayor esfuerzo por satisfacer al usuario y tener cada día más popularidad entre los mismos.

Se podrían mencionar algunos de los problemas que afectan hoy estas herramientas, por ejemplo: uno de los grandes problemas de los buscadores tradicionales es que para búsquedas muy comunes o de gran actividad comercial, los primeros resultados están copados por páginas que muchas veces no aportan gran interés real a la búsqueda. Esto se produce debido al gran trabajo de posicionamiento en buscadores realizado por las empresas mediante las técnicas de posicionamiento SEO (Search Engine Optimization, optimización para motores de búsqueda en español), y genera mucha insatisfacción entre los usuarios que quieren informarse con rapidez.

De hecho, este comienza a ser uno de los principales problemas de Google; a medida que la Red se hace más popular, y por tanto comercial, el buscador pierde eficiencia real por causa de las estrategias de optimización.

Hoy en día los buscadores más utilizados son Google y Yahoo! y lo demuestran claramente las cifras, se realizan 61 billones de búsquedas cada mes y el 80% de ellas se hacen solamente en Google y Yahoo!

- Más del 90% de los usuarios de Internet usan los buscadores para encontrar artículos o servicios.
- Más del 75% de los usuarios de Internet, su principal actividad es hacer búsquedas.
- Más del 70% de las transacciones por Internet se hacen a partir de una búsqueda.
- Entre el 85% y 90% del tráfico que recibe un sitio web viene de una búsqueda. [13]

A continuación se realiza un análisis de algunos de los buscadores más exitosos en Internet y algunos de los existentes en la UCI, tratando de demostrar así la necesidad de realizar evaluaciones más exhaustivas para lograr desarrollar un SRI más eficiente en la universidad.

2.2.1 AltaVista

AltaVista que significa "una visión desde las alturas" fue desarrollado por Digital Equipment Corporation en 1995 en laboratorios de investigación de Palo Alto. Es un motor de búsqueda para localizar todos los documentos publicados en la WWW. Otras características que se pueden mencionar de este buscador es que realizó las primeras búsquedas multilingües en Internet, desarrolla también el Babel Fish que fue el primer servicio de Internet (en la web) que traduce palabras, frases, y sitios enteros de la Web en línea en diversos idiomas como el español, alemán, portugués e italiano, lanzo además recientemente un buscador de fotografías con gran tecnología de búsqueda de la imagen y un filtro que reduce el ruido de los resultados de la búsqueda.

Este buscador es utilizado en el mundo entero y presenta varios aspectos destacados por él mismo en su página, algunos ya se mencionaron anteriormente otros no y se muestran a continuación:

- Ofrece el primer índice de la Web de Internet (1995).
- Primeras capacidades de búsqueda multilingüe en Internet.
- Primer motor de búsqueda de Internet en lanzar capacidades de búsqueda de imágenes, audio y video.
- Funciones y capacidades más avanzadas de búsqueda en Internet: búsqueda multimedia, traducción y reconocimiento de idiomas, búsqueda especializada.
- Ha obtenido 61 patentes de búsqueda, más que ninguna otra empresa de búsquedas a través de Internet. [14]

Para adentrarse un poco más en esta herramienta se pueden mencionar varias características de la interfaz de este buscador. Presenta posibilidades de una búsqueda simple y una avanzada siendo las dos de fácil entendimiento para los usuarios, es capaz de realizar una búsqueda en lenguaje natural pues puede reconocer cuales de las palabras que introduce el usuario son de un real significado y es capaz de suprimir las palabras sin significado como: artículos, preposiciones, adverbios entre otras. Se puede

apreciar un servicio de ayuda al usuario donde se encuentra información sobre los distintos tipos de búsqueda que se pueden realizar entre otras opciones que presenta la herramienta.

El diseño de la página es sencillo, solo con una imagen en la parte superior y presenta servicios adicionales que son de gran interés para el usuario como un servicio de noticias y directorio, traducción de páginas como se había descrito anteriormente y el usuario además puede descargar una barra de herramientas para su navegador. Se debe además tener en cuenta que esta herramienta pertenece a las empresas Overture, compradas hace unos años por Yahoo!, por tanto hay algunos servicios que presenta este buscador donde se redirecciona a alguna página de Yahoo!.

Los resultados de la búsqueda pueden ser mostrados por un ranking o un orden específico de los resultados, basando su ranking, más o menos, en los siguientes factores:

- Las páginas largas con mucho texto significativo.
- Páginas con un buen sistema de navegación, con una buena cantidad de vínculos a páginas con contenido relacionado.
- La conectividad de las páginas, incluyendo no sólo cuantos vínculos hay hacia una página sino también desde dónde vienen los vínculos; el número de distintos dominios y la "calidad" de esos sitios desde los que apuntan los vínculos. Un sitio o página es "bueno" si muchas páginas apuntan a ella y especialmente si muchos buenos "sitios" apuntan a ella.
- El nivel de directorio donde se encuentra la página. Los más altos son considerados como más importantes. Si una página está muy al fondo, el Spider no irá tan abajo y nunca la encontrará. Estos factores estáticos son recalculados una vez a la semana, y según vaya mejorando la página irá subiendo en el ranking.

El índice de AltaVista se construye enviando "Spiders" (programas robot) que capturan texto y lo almacenan. En este proceso no interviene ninguna acción humana ni juicio. Lo que se ve es lo que almacenan. El principal Spider, "Scooter", recoge miles de peticiones HTTP simultáneamente, almacenándolo y enviándolo a las máquinas indexadoras para que el texto pueda ser clasificado. "Scooter" tiene otros Spiders que lo ayudan a realizar tareas específicas para ayudar a mantener el índice actualizado, cómo, por ejemplo, comprobar vínculos rotos - páginas que se han movido o borrado y no serán indexadas. ¿Cómo sabe Scooter dónde tiene que ir? Sigue los vínculos que se encuentra en las páginas que visita. Cuando una página es capturada, los vínculos desde esa página se almacenan en una

lista. En teoría, no es necesario describir a AltaVista su sitio: el resto del sitio se encontrará automáticamente. En un día normal, Scooter y los otros robots visitan más de 10 millones de páginas. [15]

En la búsqueda avanzada, permite especificar fechas, idioma y muestra en primer lugar frases o palabras que se indiquen en las opciones especificadas. Al realizar la búsqueda muestra el titulo y el primer párrafo de la página, la URL y la posibilidad de traducción, ofrece términos relacionados y otras páginas que pertenecen a la misma URL. Al final de la página sugiere fuentes alternativas de búsqueda donde se pone de manifiesto la retroalimentación (en algunos casos son links que llevan a páginas de Yahoo! para realizar las búsquedas desde allí).

Presenta además la posibilidad de elegir el idioma del buscador (despliega una ventana con 25 idiomas), se pueden realizar búsquedas de imágenes audio y video, en la pantalla principal te da la posibilidad de seleccionar las opciones. Se puede disfrutar además de una búsqueda por campos, Host, URL y por Links.

A continuación se presenta una tabla con algunos criterios de búsqueda que se pueden utilizar en esta herramienta.

domain:*domainname*	Encuentra páginas dentro del dominio especificado. Se utiliza domain:uk para encontrar páginas del Reino Unido, o domain:com para encontrar páginas de sitios comerciales.
host:*hostname*	Encuentra páginas en un ordenador específico. La búsqueda host: www.shopping.com encontrará páginas que se hallen en el ordenador Shopping.com, y host: dilbert.unitedmedia.com encontrará páginas en el ordenador llamado "dilbert" dentro de unitedmedia.com.
link:*URLtext*	Encuentra páginas con un vínculo a una página con el texto de URL especificado. Se utiliza link: www.myway.com para encontrar todas las páginas con vínculos a myway.com.
title:*text*	Encuentra páginas que contienen la palabra o frase especificada en el título de la página (que aparece en la barra de título de la mayor parte de los navegadores). La búsqueda title:puesta de sol encontrará las páginas que contienen en el título la frase "puesta de sol".

inurl:*text*	Encuentra páginas con una palabra o frase específicas en la URL. Se utiliza por ejemplo inurl:jardín para encontrar todas las páginas de todos los servidores que tengan la palabra *jardín* en cualquier parte del nombre del host, la ruta, o el nombre del archivo.

Tabla 2.2: *Consultas reservadas de AltaVista.*

La posibilidad de realizar búsquedas con los operadores booleanos no está excluida de este buscador, presenta una interfaz para expresar condiciones de búsquedas poderosas, aunque un poco complicada. Incluye soporte para los operadores AND, NOT, OR, NEAR, truncamiento, uso de paréntesis y búsquedas textuales. En la búsqueda simple el operador por defecto es el or.

También permite filtrar los resultados por idioma, tipo de recurso o fecha. El formato de salida por defecto incluye título, descripción, URL y palabras claves, permitiendo opcionalmente incluir el idioma, como parte de la personalización de la interfaz. El ranking se basa en la frecuencia en que los términos ingresados están presentes en los documentos y en la cercanía de los distintos términos entre sí.

POSIBILIDADES	***ALTA VISTA***
Truncamiento (*)	Permite truncamientos en el medio y al final de la palabra y reconoce como mínimo tres caracteres. Puede reemplazar de 0 a 5 caracteres.
Frases exactas (“..”)	Utiliza las comillas.
AND	Utiliza el operador AND y el signo “+”.
OR	Utiliza el operador OR.
NOT	Utiliza el operador NOT o el signo “-“.
NEAR	Utiliza el operador NEAR o “~”, aproximando 10 palabras en búsqueda avanzada y 8 en búsqueda simple.
Reconoce signos lingüísticos (sensibles e insensibles)	Es insensible, no considera: comas (“,”); puntos (“.”); guión (“-“); subrayado (“_”); barras (“/”), acentos, minúsculas. Es sensible a las palabras escritas con mayúscula o con palabras que utilizan ambas.

Tabla 2.3: *Algunos operadores reconocidos por AltaVista.*

Por último se puede mencionar que ofrece otras herramientas como filtro familiar (el cual es utilizado para reducir el contenido no deseado en sus resultados de búsqueda.), servicio de páginas amarillas, buscador

de personas y comparador de precios de diferentes productos a través de Dealtime (http://altavista.dealtime.com).

2.2.2 Google

El motor de búsqueda de Google es uno de los sistemas de recuperación en la web más utilizados (solo comparable en estos momentos con Yahoo!) y esto se debe no solo por la eficiencia en la búsqueda de información sino también por el diseño de su arquitectura, este motor de búsqueda que fue desarrollado en la Universidad de Stanford en California, utiliza el Modelo del Espacio Vectorial para el proceso de almacenamiento y recuperación de la información explicado anteriormente en la investigación. El nombre de esta herramienta proviene de un juego de palabras con el término "googol", acuñado por Milton Sirotta, sobrino del matemático norteamericano Edward Kasner, para referirse al número representado por un 1 seguido de 100 ceros. El uso del término refleja la misión de la compañía de organizar la inmensa cantidad de información disponible en la web y en el mundo.

Esta herramienta se encuentra en constante cambio y todos estos cambios son determinados por los clientes los cuales a través del paso de los años se hacen más exigentes en cuanto a la calidad de los servicios. Google al igual que otros buscadores incorpora una búsqueda simple y otra avanzada y de igual manera es de fácil entendimiento para los usuarios, su interfaz es sencilla al igual que su nombre y ha tenido éxito entre los usuarios de una manera explosiva y es fácil hasta para los niños. Presenta disímiles servicios para que en una sola página se encuentren todos los servicios que se puedan necesitar y estos servicios que brinda también han sido un paso decisivo para que los internautas lo prefieran por encima de todos los buscadores existentes y lo cataloguen como el mejor.

Figura 2.1: *Algunos servicios que brinda Google.*

Incluso su intención es la de hacer el directorio más grande la web utilizando técnicas colaborativas, es decir donde los usuarios puedan participar.

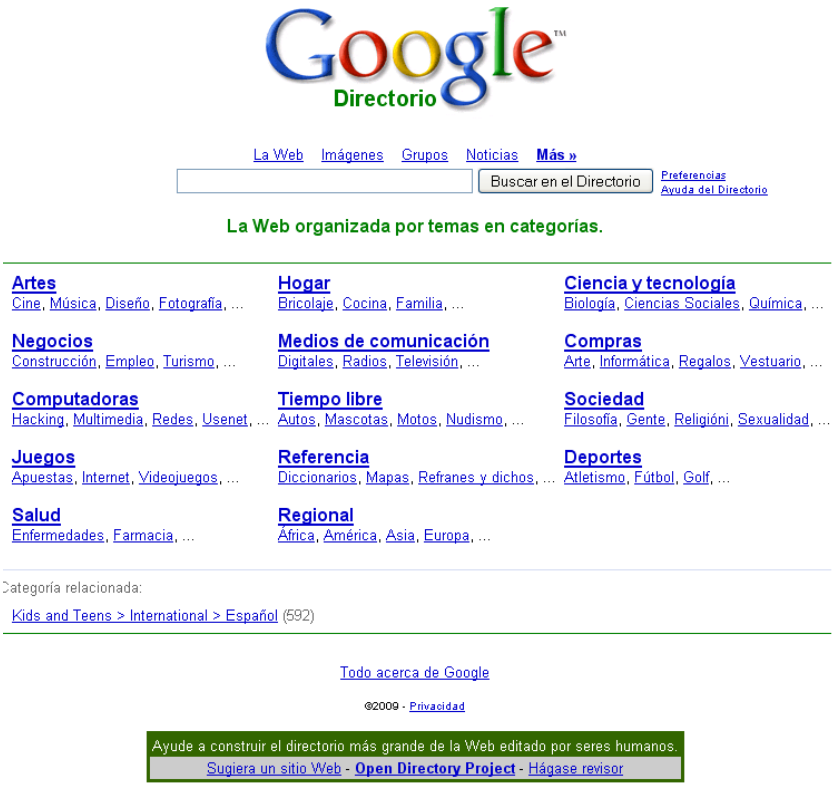

Figura 2.2: *Directorio de Google.*

Se pueden mencionar varios servicios que brinda Google:

- Buscador de imágenes. (http://images.google.com/).
- Buscador dentro de los grupos. Los grupos se encuentran organizados por categorías y allí puedes encontrar un foro de debate o un grupo sobre el tema que busques. (http://groups.google.com/).
- Presenta (como se apreciaba en la imagen anterior) servicio de directorio. (http://directory.google.-com/).
- Buscador de noticias, en más de 4.000 medios de comunicación de Internet (http://news.google.-com/).
- Este es un buscador de información de productos online. Google no es el que vende, simplemente se aprovechan de su motor de búsqueda para reconocer los sitios web que ofrecen productos online y crear una base de datos con sus datos y sus URLs. (http://www.google.com/products).
- Servidor de pruebas de Google, donde se pueden hacer búsquedas de significado de palabras, conjuntos de términos, búsqueda por voz, y búsqueda avanzada por el teclado. Además Google Labs expone algunas ideas que aún no están preparadas para integrarlas en Google y se exponen

allí los productos para que los usuarios las prueben y dejen sus comentarios o las mejoren. (http://labs.google.com/).

- Integración de una barra de búsqueda de Google dentro del navegador web. (http://www.google.-com/tools/firefox/toolbar/FT5/intl/es/index.html).
- Busca en los servidores cuyos dominios son .gov, .mil, ó .us es decir busca en los servidores del gobierno de Estados Unidos. (http://www.google.com/unclesam).
- Busca términos relacionados con el Sistema Operativo Linux. (http://www.google.com/linux).
- Busca dentro de todo lo relacionado con los sistemas operativos de Apple y Macintosh. (http://www.google.com/mac).
- Busca términos relacionados con la compañía Microsoft. (http://www.google.com/microsoft).
- Puedes realizar búsquedas mediante Google en las Universidades de todo el mundo. (http://www.-google.com/options/universities.html).
- Es una tienda online donde puedes comprar de todos los tipos de productos Camisetas, bolígrafos, gorras,... Todo relacionado con Google (no aparece Cuba). (http://www.googlestore.com/).
- Traduce páginas webs, palabras o textos en varios idiomas (inglés, español, alemán, francés,...). (http://www.google.com.cu/language_tools?hl=es).
- Knol es el nombre en clave de un proyecto anunciado por Google el 13 de diciembre de 2007, y que pretende convertirse en una colección de artículos, escritos por los propios usuarios, y que cubrirán aspectos relacionados con la Ciencia, información médica, Geografía e Historia, entretenimiento, manuales, información sobre productos, etc. Pretenda hacerle la competencia a la Wikipedia (servicio que se ha convertido en la referencia de consulta y aportación de información por excelencia de la WWW). (http://www.google.com/help/knol_screenshot.html).
- Tiene un servicio de correo utilizado por casi todos los internautas de la Red de Redes (http://www.gmail.com).
- Otro de los servicios más usados es el Google Earth el cual te permite volar a cualquier parte de la Tierra para ver imágenes de satélite, mapas, relieves, edificios en 3D... desde galaxias del espacio exterior hasta cañones en los océanos. Podrás explorar un rico contenido geográfico, guardar los lugares que visites y compartirlos con otras personas. (http://earth.google.es/).

Estos entre otros son los servicios que hacen de Google una compañía de excelencia para los usuarios. Otras de las claves importantes del éxito mundial de este buscador “híbrido” (ya que presta muchos servicios) es su algoritmo de posicionamiento llamado PageRank, es un sistema de clasificación que es el encargado de establecer un orden de relevancia entre las páginas web. Este algoritmo usa los enlaces

existentes entre las páginas como base para calcular el valor o relevancia de ellas, Google interpreta un vínculo desde la página A hacia la página B como un voto de A por la página B, a su vez analiza la relevancia de la página que emite el voto. Los votos emitidos por páginas que son en sí mismas "importantes" pesan más y ayudan a convertir a otras páginas también en "importantes". El PageRank mide objetivamente la importancia de las páginas web y la calcula resolviendo una ecuación de alta complejidad con grandes números. Los complejos mecanismos automáticos de búsqueda de Google permiten prescindir de la interferencia humana. Está estructurado de manera que nadie puede comprar un lugar privilegiado en la lista ni alterar los resultados con fines comerciales (por ejemplo: nadie puede comprar un PageRank más elevado, además la compañía solo explica cómo funciona el algoritmo de manera muy superficial pero la fórmula del PageRank nunca es publicada, la real porque hay algunas que se asemeja a la que debe hacer la que ellos utilizan), esto no tiene nada que ver con los enlaces patrocinados que si están presentes en Google donde las empresas pagan porque sus sitios web aparezcan en la primera página del buscador, pero se colocan aparte de los resultados mostrados por posicionamiento natural . Su robot de búsqueda es Googlebot, programado en Python[30] e implementado en plataformas de Software Libre. Para poder dar abasto a tal cantidad de información, Google tiene un sistema distribuido de rastreo. Un servidor general emite a cada rastreador (suelen ejecutar hasta 3 al mismo tiempo) una lista de páginas, de manera que cada araña rastreadora mantiene unas 300 conexiones abiertas, Googlebot colecciona documentos de la WWW para construir una base de datos para el motor de búsqueda Google y no solamente indexa páginas web (HTML), sino que también extrae información de ficheros PDF, PS, XLS, DOC y algunos otros, una de las desventajas de Googlebot es que se encarga de rastrear la red mensualmente y las noticias no salen de manera momentánea a su trasmisión por eso los creadores se dieron la tarea de crear dos versiones de Googlebot y estas son: Deepbot y Freshbot.

Freshbot es el encargado de recorrer la red frecuentemente, de esta manera los sitios web como los de las mayores cadenas de noticias del mundo (CNN, Reuters, BBC), o páginas que Google considera que actualizan frecuentemente sus contenidos, son rastreados por Freshbot cada cierto tiempo y las noticias son trasmitidas al usuario casi instantáneas a la hora de su publicación. La frecuencia de Freshbot es variable, aunque generalmente suele realizar su rastreo diariamente. [16]

[30] ***Python*** *se utiliza como lenguaje de programación interpretado, lo que ahorra un tiempo considerable en el desarrollo del programa, pues no es necesario compilar ni enlazar.*

El llamado Deepbot es el que hace el trabajo más detallado, el que busca en todos lados y trata de seguir cualquier enlace, el que pone las páginas en la caché y las deja allí para que Google las procese. Se dice que este trabajo es completado en un mes y luego, comienza otra vez. Claro que el encargado de determinar con qué frecuencia visite el Deepbot o el Freshbot tu página web depende en gran medida del PageRank que tenga la misma.

2.2.3 Live Search

Primeramente conocido como MSN Search, Live Search es una versión más simplificada que parece estar bastante inspirada en Google (su máximo rival) con el cambio de nombres también aparecieron varios cambios como por ejemplo: ha llevado a cabo esta modificación para mejorar su sistema de búsqueda e incluir en él varios enlaces a sus principales servicios, como Hotmail, Messenger, Noticias, búsqueda de imágenes entre algunas otras que se mencionaran a continuación:

- **Live Search Cashback:** La compañía de Microsoft, mediante Live Search Cashback, recompensa con descuentos a los usuarios que utilicen el buscador para hacer compras. Microsoft Live Search Cashback funciona acumulando el dinero de los descuentos ofrecidos en una cuenta. Cuando la cuenta alcanza los 5 dólares, puedes pedir el reembolso. (http://search.live.com/cashback/).
- **Live Search Club:** Este servicio de Live Search es otra de las estrategias para que los usuarios hagan búsquedas en su buscador y así poder hacer la competencia en Internet entre los buscadores más exitosos. El cómo ayudan estos juegos a subir los porcentajes de búsquedas es muy simple: todos están enfocados a las palabras. Por ejemplo, en uno de ellos tienes que formar una palabra a partir de letras que se dan, una vez que está formada la palabra, el juego automáticamente la busca en Live Search. Si a alguien le gusta jugar y sigue por varias rondas, terminará haciendo entre 30 y 50 búsquedas por partida. Además, para que esto sea constante, precisan que los usuarios continúen jugando y no tienen mejor idea que el dar disponibilidad de premios. En cada partida, los jugadores ganan puntos y al acumularlos los pueden cambiar por premios. (http://club.live.com/Pages/Home/HomePage.aspx)
- **Live Search Farecast:** Farecast es una "agencia de viajes online" premiada internacionalmente y que adquirió la compañía de Microsoft en 2008 y que integraron como un servicio de Live Search, aunque por el momento está en fase de prueba, y le permite al usuario buscar viajes, hoteles y organizar sus propias rutas. Live Search Farecast permite buscar vuelos de ida y vuelta, sólo ida, o

viajes con escalas, especificar la cantidad de adultos que desean viajar, el tipo de clase (Económica, Business, etc), las fechas de los vuelos, etc. La base de datos a la que recurre Farecast es bastante completa, incluye vuelos de todo el mundo, además permite comparar los resultados del buscador con los de otros servicios como Expedia y Hotwire, e incluso la posibilidad de agregar ciudades y aeropuertos que no se encuentren almacenados. También el usuario puede comprar los boletos y hacer las reservaciones de hoteles sin salir de Farecast. Gracias a un sistema de procesamiento estadístico, Farecast es capaz de decir si los valores subirán o bajarán en los próximos días, y en base a esto si conviene comprar o no los boletos y hacer las reservaciones ahora, o si es preferible esperar a que precios bajen. Lamentablemente, la información sobre hoteles comprende solo ciudades en los EE.UU. (http://farecast.live.com/)

- **Live Search Images:** Live Search al igual que Google y otros buscadores incluye un servicio de búsqueda de imágenes. Tienes según los usuarios que lo utilizan ventajas sobre google.
- **Live Search Local:** Es una herramienta que, aunque aún sólo está disponible para USA, permite dar de alta tu negocio o localización en el buscador de forma gratuita. La idea de esta herramienta es la de poder encontrar tu empresa o negocio en los mapas del buscador una vez dada de alta.
- **Live Search Macros:** Este servicio es una manera de personalizar las búsquedas del usuario es como la opción de favoritos en los navegadores donde guarda la páginas que se accede a menudo o no se quiere olvidar porque era del agrado del usuario o le llamó la atención en algún sentido. Entonces se puede resumir: lo que hacen las macros es facilitarle la búsqueda al usuario, es decir ya tiene en su buscador la búsquedas personalizadas, podría ser de algo que busca a diario como noticias por ejemplo y el puede sugerir páginas e incluírsela a esa búsqueda personalizada, no se enfoca en la cantidad de páginas indexadas para una búsqueda como Google sino que le trata de dar al usuario lo que de verdad le sirve en una búsqueda.
- **Live Search Maps:** Es una potente herramienta parecida al Google Maps.
- Live Search News: Es una herramienta muy útil para el usuario pues lo mantiene actualizado de la ultimas noticias que acontecen en el mundo (http://search.live.com/news)
- **Live Search Products:** Live Search products, es una facilidad más que brinda Live Search a sus clientes, es un buscador de productos para comprar, en los que puedes ver el precio y la valoración de otros usuarios, entre otras facilidades que brinda. (http://search.msn.com.br/products)

Este buscador fue lanzado por la compañía de Microsoft y previamente dependía de otros para listar sus búsquedas. En 2004 debutó una versión beta con sus propios resultados, impulsada por su propio robot (llamado MSNBot). Al principio de 2005 comenzó la versión definitiva.

MSNBot tiene un funcionamiento similar a GoogleBot. Rastrea la web a través de los links establecidos de una página a otra y a partir de esta información realizará una clasificación a partir de los enlaces recibidos y la importancia de las páginas que los enlazan. Presenta una interfaz agradable al usuario pero es muy pesada a la hora de cargar la página y se tarda mucho en aparecer la página principal.

Como se había mencionado anteriormente esta herramienta está muy inspirada en el funcionamiento de Google, aunque la mayoría de los buscadores más exitosos le brindan la mayor cantidad de facilidades al usuario, como cambiar de idioma (tiene 41 idiomas para escoger), una búsqueda simple y otra avanzada o una ayuda donde el usuario puede aprender a realizar las búsquedas que por lo general se realizan de la misma manera, en el caso de Live Search también utiliza los símbolos:

- + Encuentra páginas web que contengan todos los términos que van precedidos por el símbolo +. Además, permite incluir términos que normalmente se omiten.
- “” Encuentra las palabras exactas de una frase.
- () Encuentra o excluye páginas web que contengan un grupo de palabras.
- AND o & Encuentra páginas web que contengan todos los términos o frases.
- NOT o - Excluye aquellas páginas web que contengan un término o frase.
- OR o | Encuentra páginas web que contengan alguno de los términos o frases. [17]

Entre otros servicios y estilos de búsqueda que lo hacen uno de los buscadores más exitosos pero sin opacar aun el brillo de Google o Yahoo! y las cifras que se muestran al inicio corroboran este planteamiento.

2.2.4 Wikia Search

Es una herramienta de búsqueda con nuevas tendencias lanzada por los creadores de la popular Wikipedia (www.wikipedia.org) la primera versión se lanza el 7 de enero de 2008 y la cual dentro de unos años podría ser una de las competencias más duras que enfrentarán los buscadores más exitosos del momento, este buscador tiene estrategias nuevas para triunfar entre los usuarios, el mismo es una

herramienta web de código abierto, transparente y colaborativa, utiliza la tecnología abierta de Nutch, una implementación de la API de Lucene para indexar y realizar búsquedas dentro de los documentos web que rastrean mediante la tecnología de Grub, el proyecto pretende recolectar toda la información rastreada dentro de un mismo repositorio, que estaría a disposición pública mediante la licencia GFDL (licencia de software libre).

Wikia Search utiliza además Hadoop, la plataforma libre que permite ejecutar Nutch (y otras aplicaciones de software) en grandes clústeres, servidores construidos con hardware a partir de componentes clónicos, y que es una implementación libre del famoso MapReduce de Google que le permite a éste disponer de su sistema de almacenamiento. Para construir este clúster, se utilizaron casi mil servidores. Lucene y Hadoop son dos sub-proyectos del proyecto Apache.

Por su parte el buscador Nutch permite crear tus propios algoritmos de búsqueda, y Wikia Search pretende modificarlos para ofrecerles un componente social que otros buscadores no ofrecen directamente. De esta manera, se intentará que los usuarios de este buscador puedan tanto votar positiva o negativamente por cada URL, como modificar manualmente los resultados de la clasificación realizada por el algoritmo.

Wikia Search pretende cambiar la forma en que los usuarios se relacionan con los buscadores, haciendo que estos sean más transparentes y con un nivel mayor de calidad.

A la hora de realizar una búsqueda, Wikia Search ofrece tres índices de ordenación de resultados para la misma búsqueda: Whitelist, que es el índice por defecto y está basado en la indización de enlaces destacados. Además están Smaller Test y Visvo, ambos índices basados en la tecnología de Nutch, un método de búsqueda que tiene en cuenta las categorizaciones de los usuarios.

La herramienta en cuestión incluye una descripción o “mini artículos” en las búsquedas más comunes, que pueden ser añadidos por los usuarios registrados y que suelen incluir enlaces a la Wikipedia y a otros productos colaborativos de la factoría de Jimmy Walles.

Para ello, se pide a los usuarios que realicen su labor de una forma objetiva, abierta y responsable. Aunque si alguien realiza alguna acción de vandalismo, ésta es rápidamente subsanada por la comunidad.

Los usuarios de Wikia Search pueden controlar la privacidad de sus actividades en el sitio web, y eligen si muestran o no sus preferencias de búsquedas a otros usuarios, así como otras opciones de seguridad.

Otra de las opciones de Wikia Search es la de mostrar qué usuarios registrados han realizado la misma búsqueda, para ayudar así a crear un grupo de intereses dentro de la red social que lo conforma. [18]

Se podrían mencionar otras de las funcionalidades que ofrece:

- Poder editar los elementos de cualquier resultado.
- Añadir resultados de forma inmediata.
- Eliminar u ocultar resultados.
- Puntuar los resultados, que hará mejorar la calidad a lo largo del tiempo.
- Sugerir búsquedas relacionadas.
- Publicar comentarios en un resultado.

2.2.5 Yahoo!

Yahoo! significa “Yet Another Hierarchical Officious Oracle” y fue creada por David Filo y Jerry Yang, estudiantes de doctorado de Ingeniería Eléctrica de la Universidad de Stanford. [19] Esta herramienta tiene una cobertura internacional con nodos locales que sirven a sus respectivos ámbitos geográficos, los cuales incluyen 8 países europeos y a otros continentes (por ejemplo: hay una versión para Japón), para ibero-América hay una versión española, una mexicana y una brasileña. Es uno de los servicios de directorio más antiguos y grandes de la web. Está organizado en forma de índice jerárquico de temas, poseyendo una herramienta de búsqueda dentro del directorio (para búsqueda en toda la web).

Como se había mencionado anteriormente Yahoo! en sus inicios se beneficiaba de las búsquedas de Google pero luego se separan para hacerle la competencia entonces lanza su propio buscador basado en una combinación de tecnologías de sus adquisiciones y proporcionando un servicio en el que ya prevalecía la búsqueda en la web sobre el directorio.

En la actualidad utiliza un motor de búsqueda más intuitivo bautizado "Search Assist" (asistente de búsqueda), incluye una función que permite ver sugerencias mientras el usuario se decide entre las

opciones encontradas por el buscador. También permite incluir varios tipos de búsquedas en una misma página, combinando textos, fotos, videos o audio.

Yahoo! Slurp es el robot rastreador o spider de Yahoo! para el indexado de páginas web, el mismo recopila documentos de la WWW para construir un índice rastreable para servicios de búsqueda que usan el motor de búsqueda de Yahoo!. Estos documentos son descubiertos y rastreados porque otros sitios web contienen enlaces que dirigen hacia ellos. Como parte del sistema de rastreo, Yahoo! Slurp tomará en cuenta los estándares robots.txt para asegurarse de que no se rastrean e indexan las páginas que no se quiere que aparezcan en resultados de búsqueda a través de Yahoo! Search Technology. Si una página está protegida por un fichero robot.txt no será considerada para inclusión ni indexación en la base de datos de Yahoo!. Yahoo! Slurp, rastrea la web más rápidamente que su anterior versión, de manera que, tal y como comentan sus creadores, los propietarios de los sitios notarán en un 25% la reducción de peticiones de descarga y el ancho de banda utilizado por este rastreador. Almacena toda la información que recoge durante el proceso de rastreo y luego la pasa a la base de datos.

Diariamente rastrea la web para mantener sus páginas actualizadas y dos veces por semana se encamina en busca de nuevos contenidos.

Esta herramienta también presta servicios adicionales al cliente al igual que Google.

2.2.6 2x3

El primer buscador cubano en internet (http://www.2x3.cu/) fue presentado en La Habana por la Oficina para la Informatización (INFOSOC), en el pabellón que representa a Cuba en la Exposición Internacional Informática 2007, como uno de los proyectos más avanzados en los que trabaja dicha entidad para facilitar el uso masivo, ordenado y eficiente de las Tecnologías de la Información y las Comunicaciones (TICs) a escala nacional.

La función principal de esta herramienta es facilitar la búsqueda, revisión y consulta de los más diversos contenidos sobre Cuba, publicados en páginas y sitios web nacionales. El mismo contiene indexadas en su base de datos más de 100 mil direcciones de sitios cubanos en internet correspondientes al dominio .cu, pero próximamente se irá ampliando esta cobertura a todos los sitios de entidades cubanas o mixtas

relacionadas con el país. El robot de búsqueda diariamente indexa aquellos sitios que poseen mayor nivel de actualización, como son los medios de prensa, el resto de los sitios se revisan con menor frecuencia proporcionando una actualización semanal a su base de datos.

El sistema cuenta con la opción de introducir manualmente nuevos sitios por los usuarios para luego ser recorridos por su robot, aunque para ser incorporados tienen que cumplir con ciertos criterios, el primero es que el sitio exista, o sea, que esté en funcionamiento. Y después, que esté en el dominio .cu o que cumpla con las condiciones de pertenecer a entidades de cubanas o mixtas que radiquen en el territorio nacional, con independencia de la ubicación de los servidores donde están hospedadas.

Además, en todos los casos se revisará que su contenido no sea ofensivo de las normas de conductas y educación aceptadas. Entre sus principales funcionalidades ofrece la búsqueda por palabras o por frases, el sistema no busca por todas las palabras introducidas ya que el mismo posee un listado de palabras tales como preposiciones, conjunciones, artículos y otras que no aportan significado por sí mismas y que son conocidas como “palabras vacías” o “stopwords”, las cuales no se tienen en cuenta para la búsqueda. También permite realizar búsquedas especiales en sitios de medios de prensa así como en los discursos del compañero Fidel. Además ofrece la búsqueda de imágenes, en la misma se puede seleccionar diferentes tamaños para las imágenes. En criterio de ordenamiento para devolver los resultados de las búsquedas aparecen primero las páginas que contienen las palabras tecleadas en la dirección URL, a continuación cuando aparecen en el título de la página, luego las que aparecen en los metadatos y por último, las que aparecen en el cuerpo o contenido principal de la página.

En la portada presenta categorías o temas que agrupan la información disponible, cuenta con una búsqueda avanzada sencilla y de fácil comprensión para los usuarios, pero carente de algunas opciones. Además cuenta con varios servicios de búsqueda especializada como el de Imágenes, Prensa, Discursos de Fidel, Archivo (Word y PDF), Multimedia y más (Tiempo, Noticias, Diccionarios). Uno de los principales problemas que tiene es que en los primeros resultados de una búsqueda muestra noticias muy desactualizadas.

2.2.7 Buscador GPI++ de la UCI

El Grupo de Procesamiento de Imágenes (GPI), es uno de los proyectos que se desarrolla en la Universidad de Ciencias Informáticas (UCI) con el objetivo de proveer con productos de software de alta calidad y de elevado valor agregado; por su carácter científico, en el tema de Procesamiento Digital de Imágenes y Señales, al Sistema Nacional de Salud y a otros Centros e Instituciones.

En la actualidad GPI constituye uno de los proyectos líderes de la UCI, su más conocida identificación es el portal GPI basado en el CMS[31] Joomla[32] que entre sus objetivos promueve una continua publicación de noticias sobre tecnologías desarrolladas en el mundo y un servicio de búsqueda de información en la intranet de la UCI el cual se basa en un módulo de este CMS.

El funcionamiento del mismo realiza procesos irregulares de búsqueda debido a la información desactualizadas en sus bases de datos y la estructura de sus índices no permite la eliminación de redundancia para los criterios de búsqueda establecidos por los usuarios. Dicho buscador esta fuera de servicio desde hace varios meses.

2.2.8 *Buscador de la biblioteca UCI*

La herramienta utilizada para hacer este servicio fue el WebLIS[33], utilizado por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) como gestor de bases de dato documentales. La interfaz WEB ha sido traducida al español y personalizada por la Dirección de Informatización del Ministerio de Educacion Superior (MES) a partir de la versión en inglés desarrollada por el Instituto de Computación e Ingeniería de la Información ubicado en Polonia (ICIE en ingles Institute for Computer and Information Engineering).

---

[31] *Un Sistema de gestión de contenidos (Content Management System en inglés, abreviado **CMS**) es un programa que permite crear una estructura de soporte (framework) para la creación y administración de contenidos, principalmente en páginas web, por parte de los participantes.*

[32] ***Joomla!** es un sistema de gestión de contenidos (CMS) de código abierto construido con PHP bajo una licencia GPL. Este administrador de contenidos se usa para publicar en Internet e intranets utilizando una base de datos MySQL.*

[33] *Este sistema ha sido desarrollado por el Instituto de Computación e Ingeniería de la Información (ICIE, Polonia) sobre la base de su experiencia en la construcción de sistemas de bibliotecas para las organizaciones internacionales. Corre a través del motor WWW-ISIS, también desarrollado por el ICIE.*

Presenta una búsqueda simple y otra avanzada pero es muy complejo su entendimiento, aún después de leer las instrucciones. Es poco intuitivo ya que los operadores booleanos están en su forma simple, es decir no están sustituidos por oraciones que le faciliten el entendimiento al usuario. Además la información indizada en su base de datos es interna al sitio y no realiza recorridos en la web de la universidad recopilando la gran cantidad de documentos y la información presente en las mismas.

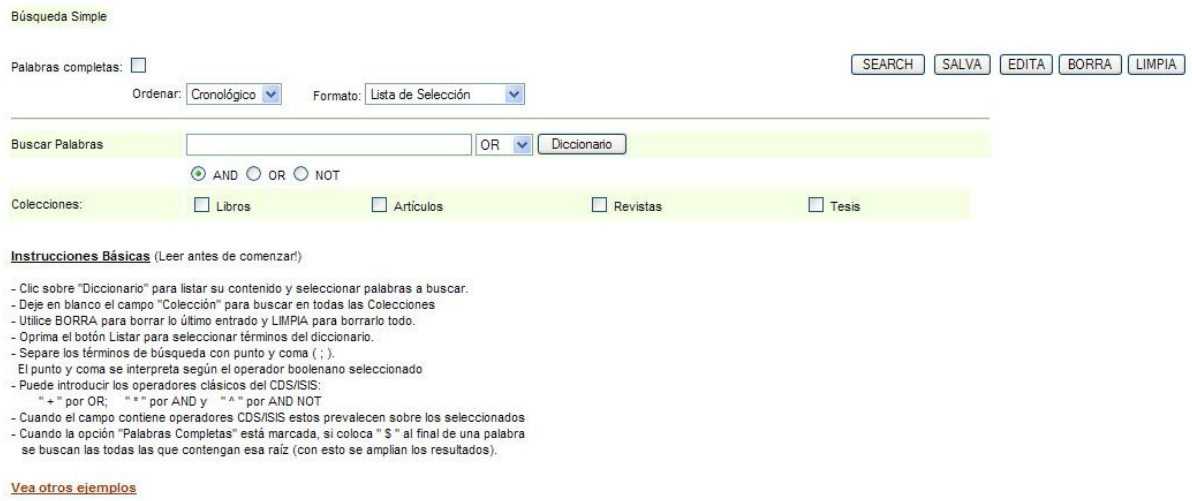

Figura 2.3: *Interfaz de búsqueda simple en el buscador de la Biblioteca UCI.*

A medida que la tecnología evoluciona, las necesidades de los usuarios van cambiando y por este motivo se vuelven más críticos con los servicios que se brindan en la red, los SRI no están exentos de esto, por tanto aumenta la necesidad de elevar la calidad y eficiencia de estos sistemas mediante evaluaciones para determinar que tan eficientes son y el grado de aceptación que manifiestan los usuarios sobre dichos sistemas. Se han planteado en investigaciones realizadas por especialistas y estudiosos del tema varias estrategias y parámetros para la evaluación de los SRI bajo condiciones específicas, distintas a la realidad presente en la Universidad de las Ciencias Informáticas, por tal motivo surge la necesidad de definir una metodología de evaluación de SRI apropiada para la UCI. El objetivo principal de realizar una metodología para evaluar los SRI existentes en la universidad, es realizar un análisis crítico sobre los sistemas de búsqueda que se tienen hoy y demostrar que estos tienen muchos aspectos que mejorar. En la actualidad, la web universitaria está nutrida de mucha información útil para la comunidad, pero esta se encuentra dispersa, por tal motivo surge la necesidad de mejorar los SRI existentes o desarrollar uno capaz de cubrir las necesidades de los usuarios en materia de localización de información.

CAPÍTULO 3

METODOLOGÍA PARA EVALUAR LOS SRI EN LA UCI.

3.1 PROPUESTA DE UN MODELO DE EVALUACIÓN DE SRI EN LA UCI.

Desde la década del 50 paralelamente con el surgimiento de los SRI aparece la necesidad de evaluar los mismos para determinar que tan eficientes eran y la aceptación que estos tenían por los usuarios. Desde entonces se han planteado varios modelos y tendencias para enfrentar la evaluación de SRI.

Se considera un modelo al esquema teórico de un sistema, una investigación o un proceso que se elabora para facilitar su comprensión y el estudio de su comportamiento. En el caso particular de esta investigación se abordarán los modelos de evaluación de SRI.

En 1986 Dervin y Nilan plantearon dos tendencias o modelos de investigación en torno a la evaluación de SRI: la centrada en sistemas frente a la orientada a usuarios, provocando un creciente interés por incorporar a éstos en el proceso de evaluación. Por su parte, varios años después, Ingwersen (1992) sintetiza las líneas de estudio en tres corrientes: la clásica o algorítmica, la orientada a usuarios y la cognitiva, que pueden resumirse en dos: la aproximación tradicional frente al modelo cognitivo.

Luego de analizar los principales modelos de RI y tendencias fundamentales que se han desarrollado en los últimos años, en la investigación se propondrá hacer uso de un modelo general integrador que agrupe aspectos y criterios tanto de la tendencia algorítmica o tradicional como de la tendencia cognitiva. Adicionalmente se propondrá hacer uso de una tendencia por dominios, algo de especial significado dadas las características del entorno y la comunidad universitaria en la UCI.

A continuación se pasarán a detallar cada una de las tendencias propuestas y sus principales fundamentos teóricos.

3.1.1 Tendencia algorítmica o tradicional

La tendencia tradicional o algorítmica debe centrar la evaluación en los algoritmos y estructuras de datos necesarios para optimizar la eficacia de las búsquedas que pueden realizarse en bases de datos textuales, así como el modelo de RI utilizado. La evaluación de los SRI desde esta perspectiva es de mucha importancia para garantizar su adecuado funcionamiento, es decir, la recuperación de información pertinente y una correcta adaptación a las necesidades de los usuarios: facilidad de uso, respuestas rápidas y coste razonable.

Un SRI, sin embargo, en su forma más simple, puede verse como una "caja negra" que acepta *inputs* y produce *outputs,* durante este proceso realiza actividades que incluyen: el reconocimiento de la estrategia de búsqueda planteada, la aplicación de diferentes algoritmos de recuperación y de ordenación de los resultados según su relevancia o utilidad probable para el usuario, la selección de los documentos o su representación, etc. Los SRI intentan localizar los documentos y recuperarlos tan veloz y económicamente como sea posible, por lo que su valor depende de su capacidad para identificar rápida y correctamente la información útil, de su facilidad para rechazar los documentos irrelevantes y de la versatilidad de los métodos que emplean.

La aproximación algorítmica se utilizó en varios estudios relacionados con teorías de clasificación, de indización y de cuestiones vinculadas a los lenguajes controlados y con la representación del lenguaje natural, así como para analizar métodos alternativos a la recuperación booleana, la ordenación de resultados por relevancia, los sistemas probabilísticos o vectoriales de recuperación, entre otros.

En la actualidad, los buscadores de la WWW, que cuentan con enormes y dinámicas bases de datos a texto completo, han llevado al límite este problema y la evaluación de grandes SRI en funcionamiento. Dada la ausencia de lenguajes documentales para la recuperación en internet y el enorme volumen de datos existente, más que exhaustividad los usuarios anhelan una alta tasa de precisión en sus consultas, si bien es éste un aspecto que las herramientas de búsqueda deben mejorar.

3.1.2 Tendencia cognitiva

La tendencia cognitiva propone centrar el análisis en el usuario y las fuentes de conocimiento implicadas en la recuperación de información. La tendencia actual es que el estudio del fenómeno informacional se realice mediante enfoques centrados en el usuario o individuo, lo que ha llevado a concebirla como algo subjetivo, individualizado, que forma parte del proceso continuo que sigue cualquier persona en su relación con el entorno que le rodea.

El término usuario, igualmente, es bastante ambiguo. De forma genérica pueden establecerse varios tipos: el potencial, el previsto y el beneficiario. En el contexto de una organización los primeros son los que aún no disfrutan de un acceso al servicio de información. Los previstos son los que sí lo tienen y planean hacer uso del mismo. En cuanto a los últimos, son los que ya han obtenido algún beneficio de los datos recuperados. Sin embargo, el éxito o el fracaso de la búsqueda pueden depender de características personales muy dispares (experiencia previa en búsquedas, edad, personalidad, estatus académico y el tipo de usuario). En consecuencia, la forma de enfrentarse a los procesos y sistemas de recuperación se ven afectadas por numerosos factores.

Por otra parte, la interacción cobra un protagonismo fundamental. Puede definirse como aquellos procedimientos de comunicación en los que intervienen todos los agentes importantes: el usuario, el intermediario y el propio sistema, donde este último se entiende como la suma de la información potencial, sobre todo en forma de texto o de su representación y de sus propias características, tales como la estructura de la base de datos y las técnicas o modelos de Recuperación de Información que utilice el sistema.

El análisis del comportamiento de los usuarios es de gran importancia y debe tenerse muy en cuenta en el diseño de los sistemas. Por ello su análisis ha de reflejar su capacidad para satisfacer al usuario. En el caso de los SRI interactivos, una evaluación realista debe ser multi-dimensional, porque estos modelos han creado la necesidad de nuevas medidas que contemplen todas las facetas de la interacción con el usuario.

La perspectiva de la RI orientada a usuarios se centra en la representación de los documentos y de los problemas de información, el comportamiento en las búsquedas y los componentes humanos de los sistemas en situaciones reales. Se nutre principalmente de la psicología cognitiva y emplea métodos de las ciencias sociales.

La aproximación orientada al usuario dio paso al acercamiento cognitivo propiamente dicho, cuya principal finalidad es mejorar la representación documental y diseñar y construir sistemas que le sean más cercanos. Asimismo, se centra en las actividades mentales cognitivas, emocionales y de motivación en relación con todos los componentes del proceso. Por tanto, en este enfoque cobran especial importancia la semántica del texto y el estudio del lenguaje natural en el entorno de las necesidades de información del usuario. Sin embargo, el eje de esta tendencia es el análisis de sus estructuras cognitivas.

Ambas tendencias de evaluación (Algorítmica y Cognitiva) deben complementarse, ya que por separado ninguna sería capaz de dar una evaluación completa de un SRI, por lo que en la presente investigación se propone un modelo integrador que haga uso de ambas en función de garantizar no solo la eficiencia del sistema sino también tener en cuenta los usuarios y dominios de usuarios.

Por las características presentes en la UCI, más que a los usuarios por separado se debe tener en cuenta a la hora de evaluar un SRI la comunidad tanto de estudiantes, profesores y trabajadores que harán uso de la herramienta. Por tanto se propone tener especial atención en una perspectiva o tendencia por dominios para realizar la evaluación.

3.1.3 Tendencia por dominios

La tendencia por dominios, que se ha querido proponer como una tercera vía o tendencia afirma que el horizonte más fructífero para las ciencias de la información es estudiar dominios de conocimiento como comunidades de pensamiento o de discurso, que son parte integrante de la división del trabajo en la sociedad. A pesar de sus enormes aciertos, sobre todo en sus críticas a modelos anteriores, la propuesta no hace sino desplazar el problema de la abstracción del usuario al cuerpo social. Algo que sería especialmente importante en la comunidad universitaria pues permitiría explotar servicios y áreas temáticas a fines con la Ingeniería Informática, elementos que tienen gran valor dada las características de los usuarios presentes en la universidad.

De esta manera haciendo uso de las características propias del medio se podrían hacer uso de modelos interactivos basados en la retroalimentación de los usuarios y el trabajo colaborativo, donde cada estudiantes y trabajador de la universidad podría aportar esfuerzo y conocimiento para mejorar el sistema y el proceso de RI en general.

La premisa, de todo estudio evaluativo ha de tener en cuenta la combinación de los tres elementos fundamentales que intervienen en la recuperación de información: el sistema, el usuario (dominio de usuarios) y la información. El primero debe estudiarse para verificar su idoneidad respecto a un usuario (o un grupo) y un tema (o conjunto de ellos) o tipos de información. La información, su calidad y su misma naturaleza son relevantes porque no todos los temas permiten un tratamiento similar y su exclusión puede distorsionar los resultados de la evaluación.

3.2 MÉTRICAS PROPUESTAS PARA EVALUAR LOS SRI EN LA UCI.

Para realizar de manera efectiva y concreta la evaluación de los SRI, esta debe tener como base una serie de métricas que justifiquen cualquier juicio a emitir y además puedan servir para realizar comparativas entre varios SRI. Para cumplir con el modelo de evaluación propuesto anteriormente donde se tengan en cuenta las características del sistema, los usuarios y el dominio o la comunidad universitaria se proponen tres grupos de métricas a tener en cuenta: las métricas técnicas, de calidad y orientadas a la persona.

3.2.1 Métricas técnicas

Composición de los índices

La composición de los índices afecta de forma muy directa a la calidad de la recuperación de información. En este aspecto se destacan tres componentes importantes: *Tamaño del índice, Frecuencia de Actualización y Porción de página web indexada* (título, primeros párrafos, página completa, etiquetas meta). Las magnitudes de cada motor dependerán del hardware y el software dedicado.

Tamaño del índice del motor de búsqueda (Cubrimiento del SRI)

En un contexto ideal, si un motor de búsqueda recopila en su índice la totalidad de los documentos de la Web (o un porcentaje cercano), sin duda alguna, ese motor seria el predilecto de todos los usuarios de internet, otorgando a este parámetro un valor prioritario por encima de otros, quizás mucho más importante que la presencia de documentos duplicados o la inclusión de enlaces erróneos en el índice.

La realidad es bien distinta, el constante cambio y expansión de la web, provoca que ninguno de los motores de búsqueda pueda indexar la totalidad de sus documentos. Por tanto a la hora de realizar la evaluación de algún SRI es necesario tener en cuenta la cantidad de documentos indizados y comparar esta cantidad con la cantidad total probable en existencia y de aquí sacar posibles conclusiones. Por lo general podría considerarse un SRI con buen cubrimiento aquel que haya indizado más del 70 % de las páginas posibles a recopilar y mientras más elevado sea este porciento mejores resultados presentaría el SRI.

$$Cubrimiento = \frac{Páginas\ indizadas}{Total\ de\ páginas}$$

Frecuencia de Actualización

La información presente en la web es en gran medida cambiante y se actualiza constantemente, por tal motivo es importante que los SRI se actualicen regularmente. Esta métrica está relacionada con el número de enlaces muertos que pueda presentar un SRI pues si la actualización de los índices no es muy frecuente entonces aumentarán los enlaces muertos. Esta frecuencia de actualización se debe determinar en dependencia de las características del sistema y el contexto en el que se implante, lo mejor sería actualizar contantemente la base de datos pero esto trae grandes sacrificios en recursos y los SRI en ocasiones no soportan actualizaciones excesivamente frecuentes, por lo tanto se debe definir un punto medio donde el índice se mantenga correctamente actualizado. Una alternativa que han aplicado algunos SRI es definir un grupo de páginas que se actualizan regularmente y que por su importancia es necesario actualizar de manera más frecuente y realizar actualizaciones más periódicas con estas, mientras que las restantes se actualizarían en periodos de tiempo más distantes.

Porción de página web indexada

Según las características de cada SRI se definen las porciones de páginas que sus indexadores almacenan para luego comparar con las consultas de los usuarios y en base a la semejanza entre las consultas y estos fragmentos de páginas almacenados en la base de datos entonces devolver el listado de documentos pertinentes para dicha consulta. De ahí se puede deducir entonces que mientras mayor sea

la porción de página indexada, mayor será la probabilidad de encontrar documentos pertinentes para una consulta realizada por el usuario. Por esta razón en la actualidad la mayoría de los SRI en Internet tiene lo que se ha dado en llamar bases de datos documentales a texto completo y en las mismas almacenan la totalidad de los documentos encontrados. Esto lógicamente supone grandes gastos en recursos y sobre todo en capacidad de almacenamiento, así como los tiempos de respuestas pues la cantidad de palabras por documentos a comparar con la consulta es mucho mayor.

Eficacia en la ejecución o tiempo de respuesta

Es medida por el tiempo que se toma un sistema o una parte de un sistema para realizar una operación, es decir, el tiempo entre el pedido y la respuesta. Este parámetro ha sido siempre una de las preocupaciones principales en un SRI, un largo tiempo en obtener la respuesta interfiere con la utilidad del sistema, y llega a alejar a los usuarios del mismo. Esta métrica es en ocasiones difícil de evaluar debido a que está afectada por las características de hardware y conectividad presentes en el momento de la prueba, así el tiempo de respuesta en una PC de gran velocidad de procesamiento no será el mismo que en una PC de más bajo procesamiento, de igual manera en un momento de mayor congestión en la red el tiempo de respuesta se vería afectado. Por tal motivo se recomienda realizar estas pruebas en distintos horarios y días en función de llegar a calcular el tiempo promedio de respuesta de un SRI.

Capacidades y sintaxis de las consultas

Un motor ha de poseer operadores booleanos, búsquedas por expresiones literales, truncamiento de los términos y facilidades de acotar una búsqueda en un determinado campo. De hecho, este conjunto de prestaciones comienzan a considerarse básicas.

Especialización en materias

En ocasiones es deseable que un SRI brinde la posibilidad de realizar búsquedas específicas en una materia determinada. Se podría mencionar, por ejemplo, realizar búsquedas de información académica, política, técnica, entre otras.

Eficiencia del almacenamiento

Es medida por el número de bytes que se precisan para almacenar los datos. Una medida típica para medir esta eficiencia es el tamaño del índice de los ficheros unido al tamaño de los archivos del documento y dividido entre el tamaño de los archivos del documento. Esta métrica está muy relacionada con el **número de páginas cubiertas por un servidor,** pues mientras más eficiente sea el sistema en su almacenamiento, mayor será el número de páginas cubiertas por el servidor y por tanto será mayor el porciento de que se devuelva al usuario lo que busca.

$$Exceso\ de\ espacio = \frac{Tam.\ índic. + Tam.\ doc.}{Tam.\ doc.}$$

Interfaz y accesibilidad al buscador

Se trata de un factor indirecto, pero resulta importante para los usuarios a la hora de decantarse por uno u otro buscador. En general el acceso se realiza mediante una página Web, pero también resulta común ver programas instalables como barras de búsqueda que se integran con los navegadores o el propio sistema operativo para realizar búsquedas de forma supuestamente más cómoda. También resulta importante en el caso del acceso mediante la Web la sobriedad y apariencia de la página en la que está alojado el buscador, pues una página muy recargada puede desviar la atención del usuario y repercutir en un aumento del tiempo de carga de la misma.

Servicios adicionales

El usuario también valora en un SRI la existencia de otras funcionalidades o servicios que en un determinado momento le pueden resultar útiles. Por ejemplo, contadores de enlaces de visitas, funciones de búsqueda avanzada y monitorización de sitios Web. Además, servicios como las macros serían muy útiles para el usuario, pues su objetivo fundamental es el de personalizar las búsquedas que realiza el mismo y es semejante a la opción de favoritos en los navegadores, donde quedan guardas la páginas a las que se accede con mayor frecuencia y no se quiere olvidar o cualquier otra que sea de interés, luego de tener estas páginas agrupadas con un solo clic se podría acceder a ellas sin necesidad de realizar una búsqueda. Estas y otros muchos servicios se deben tener en cuenta para lograr SRI con mayores prestaciones y utilidad para los usuarios.

3.2.2 Métricas de calidad

Relevancia Vs Pertinencia

Ya se conoce que un documento será relevante cuando el contenido del mismo posea alguna significación o importancia en relación con la pregunta realizada por el usuario, es decir, con su necesidad de información. Pero no se puede afirmar con exactitud cuando un documento es relevante o no, pues un mismo documento puede ser considerado relevante, o no relevante, por dos personas distintas en función de su necesidad de información o su grado de conocimiento de la materia. Llegados a un caso extremo, un mismo documento puede parecer relevante o no a la misma persona en momentos diferentes de tiempo, entonces:

- Resulta difícil definir, a priori, criterios para determinar cuándo un documento es relevante e incluso resulta complicado explicitarlo de forma clara y concisa, siendo más fácil proceder a la determinación de la *relevancia* que explicar cómo la misma se lleva a cabo.

- Es muy aventurado calificar categóricamente un documento como relevante o no relevante con un tema, en la realidad lo normal es encontrarnos con documentos relevantes con una materia determinada en alguno de sus apartados, pero no en el resto de sus contenidos. Para subsanar esta circunstancia, algunos autores introducen el concepto de *relevancia parcial*.

Estas objeciones condicionan, en cierto grado, la viabilidad de la *relevancia* para constituirse en un criterio de evaluación de la recuperación de la información. Cooper aporta la idea de “utilidad de un documento” o *pertinencia*, considerando que es mejor definir a la *relevancia* en términos de la percepción que un usuario posee sobre la utilidad de un documento recuperado, es decir, *si el mismo le va a ser útil o no*. Este nuevo punto de vista supera alguna de las limitaciones anteriores, ya que un usuario tendrá problemas a la hora de definir qué es relevante y qué no lo es, pero tendrá pocos problemas a la hora de decidir si el documento le parece o no útil.

Frants plantea otra acepción de *relevancia* muy similar a la anterior, en términos de *eficiencia funcional*. Así, *relevancia* queda asociada con el concepto de la relación existente entre los contenidos de un documento con una temática determinada y *pertinencia* se restringe a la *relación de utilidad* existente

entre un documento recuperado y una necesidad de información individual. Por otro lado, si bien es considerable el número de problemas que presenta la *relevancia*, no se ha encontrado sustituto práctico para el concepto de *relevancia* como criterio de medida de la efectividad de los SRI.

Calidad de los primeros resultados mostrados

Cuando un Buscador Web ofrece los resultados tras haber realizado una consulta, proporciona al usuario miles de resultados. El usuario medio suele consultar solamente los diez primeros enlaces. De ahí, la importancia de considerar un Buscador Web de buena o mala calidad según este parámetro. Esta métrica puede evaluarse a partir de dos criterios más específicos (*Número de enlaces relevantes y Número de enlaces duplicados o muertos*).

Número de enlaces relevantes (pertinentes)

Indica el número de páginas realmente relacionadas con el tema buscado y que son útiles en un momento dado que aparecen en las primeras posiciones.

Número de enlaces duplicados o muertos

Indica el número de enlaces rotos (la página no lleva a ningún lado) o duplicados, que no aportan ninguna información al usuario.

En ambos casos anteriores las páginas pueden ser evaluadas otorgándoles un peso según la relevancia de las mismas (por ejemplo clasificándolas de 0 a 3 según su contenido), sumándose finalmente los pesos de todos los resultados para la evaluación del SRI.

Como el orden en el que aparecen las páginas en las primeras posiciones de la consulta resulta muy importante, los pesos anteriores serán multiplicados por el inverso de la posición que ocupan (1/posición), exceptuando las que reciban cero como puntuación. Según lo anterior, si la primera página coincide con lo esperado aportarán 2 puntos (2/1), mientras que si la cuarta página ofrece parte de la información buscada contará a la evaluación con 0.25 puntos (1/4).

RELEVANCIA	**0:** Enlaces **duplicados, inactivos e irrelevantes** (que no satisface la pregunta ni recoge los términos de la ecuación de búsqueda)
	1: Enlaces técnicamente **adecuados pero no útiles** (que recogen en el HTML las diferentes partes de la pregunta pero no en el contexto adecuado o mencionan el tema en el contexto adecuado pero sólo contienen un mínimo de información relevante.
	2: Enlaces potencialmente **útiles** que no abordan el tema en profundidad o se centran en algún aspecto específico del mismo, o páginas con al menos un enlace a otra página a la que se le asignan 3 puntos)
	3: Enlaces probablemente **más útiles** (que tratan el tema extensamente, contienen enlaces a otros documentos que tratan el tema, ofrecen una bibliografía de página web o "webibliografía")

Figura 3.1: *Asignación de pesos según la relevancia de los primeros resultados.*

Calidad de los resúmenes

Aspecto que parecía relegado a un papel residual dentro de la evaluación de los motores. Sin embargo en los últimos años se le he prestado especial interés a este aspecto pues un buen resumen que recopile lo esencial de un documento en relación con la consulta realizada puede ser de mucha ayuda a la hora de decidirse un usuario por un resultado en concreto que devuelva el SRI.

Efectividad de la recuperación de información

Desde casi el inicio de las evaluaciones de los SRI (que surgen a partir de los experimentos desarrollados por Cleverdon en el Instituto Cranfield y son casi contemporáneos a la implementación de los primeros sistemas), se ha conferido mucha importancia a **la efectividad de la recuperación de información**, normalmente basada en la relevancia de los documentos recuperados (lo cual representa un serio problema, ya que valorar esa relevancia es un proceso altamente subjetivo y sin confianza, ya que diferentes juicios personales asignarán siempre diferentes valores de relevancia a un documento recuperado en respuesta a una petición).

A pesar de la seriedad del problema y al alto número de autores que critican el uso de la relevancia como criterio para calcular la efectividad (el debate no sólo no ha concluido, sino que está plenamente vigente), muchos investigadores consideran que la subjetividad no es suficiente argumento para invalidar el sistema, máxime cuando no se han aportado medidas alternativas verdaderamente aplicables.

Así, para calcular la efectividad (en juicios basados en la relevancia), las medidas más comúnmente usadas son la **exhaustividad** y la **precisión**.

Exhaustividad o Recall

La exhaustividad es el ratio de documentos relevantes recuperados en una búsqueda dada, sobre el número de documentos relevantes para esa búsqueda contenidos en la base de datos.

$$Exhaustividad = \frac{Doc.\ relev.\ recup.}{Doc.\ relev.}$$

En la práctica ocurre que, excepto para test realizados sobre pequeñas colecciones, este denominador es generalmente desconocido y debe ser estimado por muestreo o por otros métodos, lo que le proporciona otra dosis de incertidumbre a la valoración de la efectividad.

Por ejemplo, suponiendo que en la base de datos existen 40 documentos relevantes para la consulta de un usuario y que el sistema de recuperación obtiene 20 documentos relevantes, por lo tanto la exhaustividad es de 20/40, es decir 50%.

Precisión

La precisión es el ratio del número de documentos relevantes recuperados, sobre el número total de documentos recuperados. Esta segunda medida es mucho más fácil, cierta y simple de calcular, no presentando tantos problemas como la anterior.

Por ejemplo, suponiendo que un SRI contiene 40 documentos relevantes que satisfacen una consulta dada, y el sistema de recuperación solamente obtiene 30 documentos, de los cuales sólo 20 son relevantes; entonces la precisión del sistema es de 20/30, es decir 67%.

$$Precisión = \frac{Doc.\ relev.\ recup.}{Doc.\ recup.}$$

Es importante tener en cuenta que estos parámetros (precisión y exhaustividad) necesitan estar compensados ya que un sistema con una exhaustividad muy alta pero con baja precisión y viceversa no será adecuado.

La tasa de fallo

Refleja el porcentaje de documentos recuperados no relevantes sobre el total de documentos no relevantes de la base de datos. Esta medida cobra especial importancia cuando se considera que la precisión se encuentra muy sujeta a posibles variaciones en el contenido de la base de datos y se observa que la tasa de fallo no adolece tanto de esta dependencia: los cambios en la Generalidad de una colección afectan menos a la tasa de fallos que a la precisión, que resulta más sensible.

$$Tasa\ de\ fallo = \frac{Doc.\ recup.\ no\ relev.}{Doc.\ no\ relev.}$$

Métricas complementarias para la precisión y la exhaustividad

Complemento del ratio de precisión

También se le denomina "factor de ruido". Consiste en los documentos no relevantes recuperados partido por todos los documentos recuperados (relevantes y no relevantes).

$$Comple.\ ratio\ precisión = \frac{Doc.\ no\ relev.\ recup.}{Doc.\ recup.}$$

Complemento del ratio de exhaustividad

Su ecuación se calcula dividiendo los documentos relevantes no recuperados entre el total de los documentos relevantes.

$$Comple.\ ratio\ exha. = \frac{Doc.\ relev.\ no\ recup.}{Doc.\ relev.}$$

El primero en formularlo fue Swets 1963 que lo denominó probabilidad condicional de una pérdida. En 1964 Fairthorne lo denominó ratio del esnobismo (“snobbery ratio”).

Índice de irrelevancia

Este índice se obtiene de dividir los documentos recuperados no relevantes a la pregunta entre el total de los documentos contenidos en la colección. Como muchas de las medidas anteriores fue formulada en primer lugar por Swets en 1963, que se refirió a él como *probabilidad condicional de bajada falsa* (*conditional probability of false drop*). Cleverdom, Mills and Keen la llamaron posteriormente *fallout*. También ha sido denominada "*desechado*" (*discard*).

$$Índi.\ irrelev. = \frac{Doc.\ recup.\ no\ relev.}{Total\ doc.}$$

Según Kowalski con esta medida se puede establecer con qué efectividad está actuando un sistema de recuperación. Esta medida es el inverso de la exhaustividad y nunca se encontrará un resultado de 0/0, a menos que todos los documentos sean relevantes para la búsqueda.

Complemento del índice de irrelevancia

Swets en 1963, lo denominó "*probabilidad condicional de una correcta respuesta negativa*" (condictional probability of a correct rejection). Goffman y Newill la llamaron *"especificidad".* Se calcula dividiendo los documentos no relevantes no recuperados entre el total de los documentos no relevantes.

$$Comple.\ índi.\ irrelev. = \frac{Doc.\ no\ relev.\ no\ recup.}{Doc.\ no\ relev.}$$

Generalidad

La generalidad sirve para calcular la densidad de documentos relevantes. Se calcula dividiendo los documentos relevantes entre el total de los documentos de la base de datos.

$$Generalidad = \frac{Doc.\ relev.}{Total\ doc.}$$

Salton se refiere a la generalidad como: el grado de documentos relevantes contenidos en una colección. Una colección con un alto grado de generalidad es una colección donde los documentos relevantes son mayoría.

La precisión, la exhaustividad, el índice de irrelevancia y la generalidad se relacionan mediante la siguiente ecuación:

$$\frac{P}{Ir} = \frac{P/(1 - P)}{G/(1 - G)}$$

Donde P/(1-P) es el ratio de los documentos relevantes recuperados entre los no relevantes recuperados. G/(1-G) es el ratio de los documentos relevantes en la colección entre los documentos no relevantes en la colección.

P/ir es la ejecución de la recuperación en los documentos relevantes entre la ejecución de la recuperación en los documentos no relevantes. Es deseable tener el primero de los dos altos.

3.2.3 Métricas orientadas a la persona

Medidas relacionadas con el usuario

La precisión y la exhaustividad se basan en que el conjunto de documentos recuperados para unas preguntas es el mismo, independientemente del usuario. Sin embargo, lo habitual es que la valoración de la respuesta obtenida, varíe de unos usuarios a otros, o incluso en un mismo usuario dependiendo del

momento de la recuperación, por este motivo son necesarias las medidas orientadas a los usuarios ya que ellos son la razón de ser de la existencia del sistema.

La efectividad de un sistema es una medida ajena al propio sistema que relaciona la satisfacción del usuario con la salida que el sistema proporciona. Medir la satisfacción del usuario resulta muy importante, pero es complicado y es menos objetivo que las medidas vistas anteriormente.

Se pueden definir las siguientes medidas orientadas al usuario:

Ratio de cobertura

Es la proporción de documentos relevantes conocidos por el usuario que son actualmente recuperados.

$$Ratio\ cobert. = \frac{Doc.\ relev.\ conocidos\ recup.}{Doc.\ relev.\ conocidos}$$

Suponiendo que el usuario conoce 15 documentos relevantes, y el sistema recupera 10 relevantes, incluyendo 4 documentos que son conocidos por el usuario. El ratio de cobertura sería 4/15 es decir 26,6%. De aquí el usuario puede inferir que hay aproximadamente 38 documentos relevantes, aproximadamente cuatro veces el número de documentos recuperados. Si el usuario ha visto 6 nuevos documentos relevantes añadidos a esos 15 previamente conocidos, se puede estimar que la base de datos contiene 16 ó 17 documentos relevantes que él nunca ha visto y a partir de aquí puede intentar recuperarlos, modificando, si lo considera oportuno, su estrategia de búsqueda.

Ratio de novedad

Proporción de documentos relevantes recuperados que previamente no son conocidos por el usuario.

$$Ratio\ noved. = \frac{Doc.\ relev.\ recup.\ no\ conocidos}{Total\ doc.\ relev.\ recup.}$$

Siguiendo con el ejemplo, el ratio de novedad sería 6/10. Un ratio de cobertura alto, podría dar al usuario alguna confianza en que los sistemas localicen todos los documentos relevantes. También sugiere que el sistema es efectivo en la localización de documentos desconocidos para el usuario. Del ejemplo anterior, el usuario puede inferir que aproximadamente el 60% de algún grupo de documentos relevantes recuperados para esta pregunta y esta base de datos, en particular, no será previamente conocida. Por supuesto, al usuario no le interesa saber que puede recuperar, aquellos documentos que él ya conoce, por lo tanto, es deseable que el ratio de novedad sea alto.

Exhaustividad relativa

Ratio de documentos relevantes recuperados, examinados por el usuario, partido por el número de documentos que el usuario quiere examinar.

$$Exha.\ relat. = \frac{Doc.\ relev.\ recup.\ examinados}{Doc.\ que\ el\ usuario\ quiere\ examinar}$$

En cuanto a la exhaustividad relativa, puede referirse más directamente a la cuestión de cómo el usuario quiere algunos documentos. Suponiendo que el sistema presenta 20 documentos al usuario y que éste quiere 5 documentos relevantes.

Si solo hay 3 documentos relevantes entre los 20, la exhaustividad relativa será 3/5, el usuario solo obtiene 3 de los 5 que busca. Si por el contrario, hay 5 o más documentos relevantes entre los 20, entonces, presumiblemente el usuario podrá abandonar después de encontrar los 5 deseados con una exhaustividad relativa de 5/5 es decir de 1. Si la exhaustividad relativa es de 1, la medida falla al referirse los esfuerzos a localizar los documentos.

Podría ser que el usuario encuentre los documentos entre los primeros 5 ó 6 examinados o podría ser que necesitara examinar los 20, por lo tanto esto da pie para definir una nueva medida: *Esfuerzo de exhaustividad.*

Esfuerzo de exhaustividad

Es el ratio del número de documentos relevantes deseados partido por el número de documentos examinados para encontrar el número de documentos relevantes deseados. Esta medida asume que la colección contiene el número de documentos relevantes deseado y que el sistema de recuperación permite al usuario localizarlos todos, lo cual aunque es deseable no siempre es posible. Este ratio puede ser 1, si los documentos relevantes deseados son los primeros documentos examinados por él, y más próximo a 0, si el usuario necesita examinar un gran número de documentos para encontrar los pocos que desea.

Otras medidas relacionadas con el usuario son la utilidad y satisfacción. De las medidas vistas hasta ahora, éstas son las más subjetivas, por lo que habrá que valorarlas con mucho cuidado. La satisfacción pone énfasis en la coincidencia entre lo que el usuario quiere y lo que el usuario recibe.

3.3 GUÍA GENERAL PARA EVALUAR LOS SBRI EN LA UCI.

Con el propósito de aplicar la evaluación siguiendo el modelo y las métricas propuestas anteriormente se ha elaborado la siguiente guía general que describirá los principales pasos a seguir para realizar la evaluación de un SRI.

Las principales etapas de la evaluación propuestas son:

- Definición de las necesidades de información de los usuarios y elaboración de los enunciados de búsqueda o consultas.
- Realización de las consultas.
- Aplicación de las métricas.
- Análisis de los resultados.

El proceso de evaluación se inicia con la elaboración de las ecuaciones de búsqueda o consultas mediante las sintaxis correspondientes a partir de las necesidades de información más comunes en la universidad.

3.3.1 Definición de necesidades de información y elaboración de las consultas.

Una vez que se analicen las principales demandas de información, se debe pasar a seleccionar las preguntas con sus respectivas sintaxis de búsqueda. La selección de las preguntas es un aspecto clave, pues de ello depende el éxito o fracaso de la recuperación. Las preguntas ofrecen el punto de partida para realizar las consultas, controlar el proceso de búsqueda y también para valorar los resultados que ofrece el sistema.

Las preguntas deberían presentar las siguientes características:

- Preguntas sobre las que deba haber información en el SRI en cuestión.
- Que constituyan una combinación de preguntas “fáciles” y “difíciles”, en relación con la cantidad de información que sobre ellas pueden encontrarse.
- Que se trate de preguntas heterogéneas, relacionadas con distintos temas.

En algunos casos puede ser oportuno plantear dos (o más) opciones de sintaxis de búsqueda para asegurar los resultados y conseguir de esta manera que la pregunta se formule de la manera más correcta para recuperar los documentos realmente relevantes, pues no hay una única manera de plantear la consulta, ya que para elaborar la expresión de búsqueda, hay que decidir cuántos y qué términos de la pregunta incluir, además de elegir si se formula la pregunta en lenguaje natural o usando la lógica booleana. Esto da lugar a expresiones de búsqueda de diversos tipos:

- Algunas utilizan términos generales y otros más específicos.
- Algunas usan la lógica booleana.
- Algunas se plantean como búsquedas de frase y otras como búsquedas en lenguaje natural.
- Algunas utilizan nombres de persona.
- En algunos casos se utilizan la mayúscula y el truncamiento, entre otros operadores.

En el mejor de los casos es recomendable hacer uso de la mayor cantidad posible de tipos de expresiones de consulta (siempre que el SRI lo permita) con el fin de comprobar si el SRI soporta gran variedad de sintaxis de consultas y además comparar la calidad de los resultados utilizando distintas formas de

formular la misma consulta. Se recomienda realizar como mínimo 10 consultas para la evaluación de los SRI.

3.3.2 Realización de las consultas

Se recomienda utilizar para la realización de las consultas tanto la interfaz de búsqueda simple como la interfaz de búsqueda avanzada (en caso de que el SRI evaluado cuente con una), de esta manera se podrían evaluar ambas interfaces.

En caso de estar realizándose la evaluación de manera simultánea en dos o más SRI para comparar los mismos, se debe tener en cuenta formular las mismas preguntas en los SRI que se están evaluando sin que transcurra demasiado tiempo entre el uso de los distintos motores pues el retraso podría aumentar las probabilidades de cambio o eliminación de localización de las páginas recuperadas y llevar a un análisis menos confiable.

3.3.3 Aplicación de las métricas

Algunas métricas deben calcularse paralelamente a la realización de las consultas, tal es el caso de la eficacia en la ejecución o tiempo de respuesta y las capacidades y sintaxis de las consultas. El resto puede analizarse luego de realizar cada consulta.

Anteriormente se plantearon un gran número de métricas a tener en cuenta para la evaluación de SRI, de las cuales se considera que al menos no deben faltar en la realización de una evaluación las siguientes:

- Calidad de los primeros resultados mostrados.
    - Número de enlaces relevantes (pertinentes).
    - Número de enlaces duplicados o muertos.
- Eficacia en la ejecución o tiempo de respuesta.
- Precisión.
- Exhaustividad.
- Interfaz y accesibilidad al buscador.
- Frecuencia de Actualización.

- Tamaño del índice del motor de búsqueda (Cubrimiento del SRI).
- Calidad de los resúmenes.
- Ratio de novedad.
- Esfuerzo del usuario, medido a partir de la exhaustividad relativa y el esfuerzo de exhaustividad.

3.3.4 Análisis de los resultados

Luego de determinar las métricas para el SRI evaluado, estas deben analizarse pormenorizadamente en función de determinar los aspectos negativos detectados y sugerir mejoras al sistema en función de estos. El análisis puede resumirse en gráficas y tablas que ayuden a la comprensión y presentación de los resultados.

Generalmente un proceso de evaluación involucra a dos o más SRI con el objetivo de establecer comparaciones entre estos. Las comparaciones son además la base para guiar el juicio de evaluación y en muchas ocasiones se tiende a tomar un SRI probadamente exitoso para comparar con el sistema que interesa evaluar y en fusión de esto determinar que tan bien funciona.

3.4 CARACTERÍSTICAS DOCUMENTALES QUE DEBEN CUMPLIR LOS SRI.

Los SRI son las herramientas más importantes para localizar información en Internet. Tanto los índices temáticos como los motores de búsqueda utilizan bases de datos documentales.

Las bases de datos documentales deben realizar unas funciones básicas y cumplir una serie de características, que a los buscadores también se les debe exigir. A continuación se describen dichas funciones y características agrupadas en tres apartados: *recogida y análisis de información, búsqueda y resultados*.

3.4.1 Recogida y análisis de información

En documentación, en términos generales, son más aconsejables los sistemas que son capaces de analizar más pormenorizadamente las características formales y de contenido de cada uno de los registros que gestionan. En el caso de los SRI, no se deben contraponer de forma general índices y

motores de búsqueda, sino más bien hay que indagar las características particulares tanto de unos como de otros.

Consecuentemente con lo anterior, si hubiera que establecer un ranking de categorías de SRI desde el punto de vista de sistema de recogida y análisis de direcciones el orden sería el siguiente: 1) Índice temático con al menos título, URL, clasificación, descripción de cada una de las páginas. 2) Buscador con reconocimiento y análisis de etiquetas Meta. 3) Buscador sin reconocimiento y uso de las etiquetas Meta.

3.4.2 Búsqueda

Un sistema de búsqueda es mejor cuanto más flexible es y cuantas más posibilidades de recuperación ofrece. Las características que se han considerado básicas desde el punto de vista de la recuperación de la información que debe cumplir un buscador, ya sea un índice o un motor de búsqueda, son las que a continuación se citan:

- Formularios de búsqueda. Posibilidad de elegir entre un formulario simple y otro más detallado (búsqueda avanzada). El simple da la posibilidad de acercarse a los buscadores a aquellas personas que no están muy familiarizadas con el uso de las bases de datos. Los formularios más completos dan la oportunidad de realizar búsquedas más complejas a los usuarios expertos.
- Operadores de búsqueda. Posibilidad de uso del truncado de palabras, de los operadores booleanos: AND, OR y el NOT y del paréntesis. También es recomendable la localización de términos compuestos (utilizando comillas o buscando frases). El contar con todas estas facilidades permite realizar búsquedas que definen muy ajustadamente la información que se requiere.
- Clasificación temática. Existencia de un índice general, pues facilita mucho la localización de información a aquellos que no saben concretar su tema de búsqueda y prefieren explorar los apartados temáticos propuestos por el sistema.
- Campos de búsqueda. La posibilidad de dirigir las búsquedas a campos determinados aumenta notablemente la pertinencia en la recuperación de información. Las búsquedas en texto libre son más exhaustivas pero menos pertinentes. La búsqueda en campos determinados está muy relacionada con la segmentación de datos en la recogida de datos. Cuanto más sectorializada esté la información extraída de un recurso de información y más se pueda concentrar la búsqueda a un tipo de dato concreto más pertinente será la recuperación. Como en el caso de la toma de datos se consi-

deran campos imprescindibles: título, URL, descripción; y campos recomendables: palabras-clave, localización, idioma, tipo de información y tipo de propietario.

- Control del vocabulario. Uno de los problemas fundamentales en la recuperación de información es el control de vocabulario. Eliminar las sinonimias[34] y las polisemias es una labor a realizar en todo sistema documental serio, pues disminuye considerablemente el ruido y el silencio.
- Detección de novedades. En todo sistema documental es imprescindible poder detectar los nuevos registros incorporados. Existen varios sistemas a través de los que detectar las novedades: creación de ficheros aparte, uso de etiquetas especiales, acotar por fecha de alta y ordenar los recursos por fecha de alta.

3.4.3 Resultados

Es aconsejable que los usuarios puedan elegir entre diferentes formatos de presentación de resultados o incluso diseñarlos a la medida. También es recomendable que todos los documentos que componen un sistema documental vayan acompañados de una descripción de su contenido. En cuanto al orden de presentación de los resultados, es importante que se pueda seleccionar entre varios criterios de ordenación. En el caso de los motores de búsqueda, ya que gestionan URL, es recomendable que puedan presentar las páginas recuperadas agrupadas por clusters, ya sea de formatos, temáticas o relevancia, entre otras.

[34] ***Sinonimia**: igualdad, semejanza, analogía, equivalencia, paralelismo.*

CONCLUCIONES

En la investigación se ha realizado un estudio sobre las características y el funcionamiento de los Sistemas de Recuperación de Información, así como de los modelos fundamentales de Recuperación de Información, llegando a la conclusión de que el modelo de espacio vectorial debe ofrecer mejores resultados, unido al modelo interactivo que permitirá que la comunidad universitaria aporte esfuerzo y conocimiento para mejorar el sistema y el proceso de RI.

El modelo de evaluación propuesto contempla tres tendencias que permitirán realizar una evaluación integral y garantizar la calidad y eficiencia del sistema. La tendencia Algorítmica o tradicional permitirá realizar la evaluación desde el punto de vista de la efectividad del sistema, prestando especial atención a la implementación del SRI, los algoritmos y el modelo de RI utilizados. Por otra parte la tendencia Cognitiva centra la evaluación en el usuario y las fuentes de conocimiento implicadas en la RI, elemento de gran importancia para lograr una alta aceptación del sistema por los usuarios, así como que a estos les resulte útil y cómodo utilizar el SRI. Finalmente se propone la tendencia por Dominio, dada las características de la comunidad universitaria donde podría pensarse en incorporar servicios y áreas temáticas a fines con la Ingeniería Informática, lo cual sería un valor añadido para el sistema.

Por último se llegó a la conclusión de que era necesario utilizar métricas técnicas orientadas a medir de manera concreta y precisa la efectividad del sistema, métricas de calidad para evaluar la calidad de los resultados devueltos así como del proceso de RI y finalmente métricas orientadas al usuario para poder medir la aceptación del sistema y que tan usable es este para los usuarios.

RECOMENDACIONES

Por la importancia que tiene recuperar la información disponible en la web universitaria de manera eficiente y brindar un servicio de localización de la misma con calidad a los usuarios, se recomienda:

- Aplicar la presente metodología de evaluación en los SRI presentes en la universidad con el fin de optimizar su funcionamiento y eficiencia.
- Tener en cuenta los elementos planteados en la investigación a la hora de implementar un SRI para que este cumplan con creces las expectativas de los usuarios.
- Continuar el desarrollo de investigaciones relacionadas con la RI, los SRI y la evaluación de los mismos, con el fin de profundizar más en el tema.

REFERENCIAS BIBLIOGRÁFICAS

[1] ¿qué es motor de búsqueda?, disponible en: http://es.answers.yahoo.com/question/index?qid=20090412174400AALEPpF consultado el 12 de enero del 2009.

[2] Wikipedia lanza su buscador Wikia Search, disponible en: http://www.seccperu.org/?q=node/556 consultado el 30 de enero 2009.

[3] Motores de búsqueda sobre salud en Internet, disponible en: http://bvs.sld.cu/revistas/aci/vol11_5_03/aci02503.htm consultado el 3 de febrero del 2009.

[4] TSEP - The Search Engine Project, disponible en: http://www.desarrolloweb.com/scripts/step-motor-busqueda-php.html consultado el 3 de febrero del 2009.

[5] PhpDig, disponible en: http://www.desarrolloweb.com/scripts/phpdig-motod-busqueda-php.html consultado el 15 de febrero del 2009.

[6] Blasten blt-SEARCH, disponible en: http://www.desarrolloweb.com/scripts/blasten-motor-busqueda.html consultado el 15 de febrero del 2009.

[7] Lucene, disponible en: http://es.wikipedia.org/wiki/Lucene consultado el 17 de febrero del 2009.

[8] Operadores, disponible en: http://biblio.uah.es/iBistro_helps/Castellano/htip7102.html consultado el 1 de marzo del 2009.

[9] CRUZ ALMAGUER, J. A.; Buscador Web; Universidad de La Habana, Facultad de Matemática-Computación; Ciudad de La Habana; 2004; 47 páginas.

[10] Calidad, disponible en: http://es.wikipedia.org/wiki/Calidad consultado el 3 de marzo del 2009.

[11] Definiendo algunos términos SEO; disponible en: http://www.bloginformatico.com/definiendo-algunos-terminos-seo.php consultado 4 de marzo del 2009

[12] Aplicación Métricas Para Evaluación Diseño; disponible en: http://www.mitecnologico.com/Main/AplicacionMetricasParaEvaluacionDise%F1o consultado 4 de marzo del 2009.

[13] Optimización de Páginas Web y Posicionamiento en Buscadores, disponible en: http://www.weblifeclub.com/ consultado el 5 de marzo del 2009.

[14] Breve presentación de AltaVista, disponible en: http://www.altavista.com/about/default consultado el 5 de marzo del 2009.

[15] BRITO SALAZAR C., MACÍAS ROMERO E.; Sisweb; Universidad de las Ciencias Informáticas, Facultad de Entornos Virtuales; Ciudad de La Habana, 2008; 98 páginas.

[16] FreshBot, disponible en: http://google.dirson.com/freshbot.php consultado el 10 de marzo del 2009.

[17] Búsquedas eficaces, disponible en: http://help.live.com/help.aspx?project=wl_searchv1&market=es-ES&querytype=keyword&query=egapemoh&domain=www.live.com:80, consultado el 13 de marzo del 2009.

[18] Wikia Search: el buscador colaborativo, disponible en: http://www.consumer.es/web/es/tecnologia/internet/2008/04/21/176022.php consultado el 15 de marzo del 2009.

[19] La diferencia entre Google y Yahoo!; disponible en. http://www.taringa.net/posts/info/2027308/La-diferencia-entre-Google-y-Yahoo!.html consultado el 1 de abril del 2009.

BIBLIOGRAFÍA

OLVERA LOBO M. D.; Evaluación de sistemas de recuperación de información: aproximaciones y nuevas tendencias; 1999; disponible en: http://www.elprofesionaldelainformacion.com/contenidos/1999/noviembre/evaluacion_de_sistemas_de_recuperacion_de_informacion_aproximaciones_y_nuevas_tendencias.html

GUTIÉRREZ SÁEZ A.; Evaluación de buscadores Web; disponible en: http://es.geocities.com/evaluacionbuscadoresweb/aspectosEvaluablesBuscadoresWeb.htm

ANONÍMO; Criterios de Evaluación para los Buscadores Web; disponible en: http://csanzc.en.eresmas.net/EvaluacionBuscadoresWeb/paginas/criteriosdeevaluacion.htm

ANONÍMO; Aspectos a evaluar en los sistemas de recuperación de información; 2004; disponible en: http://irsweb.blogspot.com/2004/10/aspectos-evaluar-en-los-sistemas-de.html

MARTÍNEZ MÉNDEZ F. J.; Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en internet; España, Universidad de Murcia; 2002; 301 páginas.

ANEXOS

ANEXO 1: Búsqueda simple en Google.

ANEXO 2: Búsqueda avanzada en Google.

GLOSARIO DE TÉRMINOS

Eficiencia: *(Del lat. efficientĭa):* Según el Diccionario de la Real Academia Española: *Es la capacidad de disponer de alguien o de algo para conseguir un efecto determinado*. Según el diccionario de la lengua española eficiencia es la: capacidad para lograr un fin empleando los mejores medios posibles.

Free Software Foundation: La Fundación para el Software Libre (FSF) es una organización creada en Octubre de 1985 por Richard Matthew Stallman y otros entusiastas del Software Libre con el propósito de difundir este movimiento, está dedicada a eliminar las restricciones sobre la copia, redistribución, entendimiento, y modificación de programas de computadoras. Con este objeto, promociona el desarrollo y uso del software libre en todas las áreas de la computación, pero muy particularmente, ayudando a desarrollar el sistema operativo GNU.

GPL: Es una licencia creada por la Free Software Foundation a mediados de los 80, y está orientada principalmente a proteger la libre distribución, modificación y uso de software.

Hipertexto: En informática, es el nombre que recibe el texto que en la pantalla de una computadora conduce a su usuario a otro texto relacionado. La forma más habitual de hipertexto en documentos es la de hipervínculos o referencias cruzadas automáticas que van a otros documentos. Si el usuario selecciona un hipervínculo, hace que el programa de la computadora muestre inmediatamente el documento enlazado.

Indexar: Indizar.

Indizar: Hacer índices. Registrar ordenadamente datos e informaciones para elaborar su índice.

Licencia de Software: Es un contrato entre el titular del derecho de autor (propietario) y el usuario del programa informático (usuario final), para utilizar éste en una forma determinada y de conformidad con unas condiciones convenidas.

Las licencias de software pueden establecer entre otras cosas: la cesión de determinados derechos del propietario al usuario final sobre una o varias copias del programa informático, los límites en la responsabilidad por fallos, el plazo de cesión de los derechos, el ámbito geográfico de validez del contrato e incluso pueden establecer determinados compromisos del usuario final hacia el propietario, tales como

la no cesión del programa a terceros o la no reinstalación del programa en equipos distintos al que se instaló originalmente.

Para que una documentación o un software sean libres tiene que ser publicada con una licencia de documentación libre. Generalmente se utiliza la Licencia de Documentación Libre de GNU (GNU FDL), aunque en ocasiones también se utilizan otras licencias de documentación libre.

Metabuscador: Clase de buscador que carece de base de datos propia y, en su lugar, usa las de otros buscadores y muestra una combinación de las mejores páginas que ha devuelto cada buscador. Un buscador normal recopilan la información de las páginas mediante su indexación, como Google o bien mantiene un amplio directorio temático, como Yahoo!. La definición simplista sería que un metabuscador es un buscador de buscadores.

Motores de búsqueda de Primera y Segunda Generación: Con el surgimiento de internet se creó la necesidad de herramientas que ordenaran o catalogaran la información disponible para posibilitar su fácil acceso. Se produce así la "primera generación" de buscadores, la cual se caracteriza por catalogar los documentos en forma manual o en base a la información interna de los mismos, utilizando técnicas provenientes de Recuperación de Información (Information Retrieval) tradicional.

Debido a las limitaciones intrínsecas de este tipo de buscadores, al crecimiento continuo y exponencial de la web (un orden de magnitud en un año luego de que EUA anunció que dejaría de subsidiar a Internet a partir de1995, comenzando así el proceso de privatización) y los cambios cualitativos debido a la explosión de los sitios ".com" (pasaron a cubrir más del 60% de los dominios registrados), estas soluciones empezaron a resquebrajarse en este mismo año, dando lugar al origen de una nueva gama de herramientas, llamadas la "segunda generación" de herramientas de búsqueda. Se comienzan a utilizar además nuevas técnicas de ranking. La lucha por ser el mejor de los buscadores llevo a las personas a realizar acciones anti éticas como: repetir un número grande de veces las palabras claves con el mismo color del fondo, para que los clientes no los vean pero los robots sí, y aumenten el ranking de la página; incluir en las etiquetas meta palabras claves muy buscadas, (como sex, free, o nombres de empresas como Microsoft), aunque la página no tenga nada que ver con estos temas, entre otras. La red se vuelve imposible de rastrear por un solo buscador y se crean los metabuscadores.

Ranking: Tabla o lista en que se clasifican una serie de elementos por orden de mayor a menor categoría o puntuación: Ejemplo: esta película encabeza el ranking de las películas más premiadas de la historia del cine; el número uno del ranking mundial de tenis ha sido eliminado por un jugador que ocupa el número 90 de la clasificación.

Software Libre (SWL): El SWL, no se trata de software gratis, sino que las empresas o comunidades que los desarrollan, son libres de pedir por adquirirlos, incentivos financieros, siempre que esto no afecte las libertades para estudiarlos, usarlos, modificarlos y redistribuirlos. Este movimiento busca preservar el patrimonio intelectual informático accesible para las futuras generaciones.

Un programa es Software Libre si:

- **Libertad 0:** Usted tiene libertad para usar el programa, con cualquier propósito.
- **Libertad 1:** Usted tiene la libertad de estudiar cómo funciona el programa, y adaptarlo a sus necesidades (Para que esta libertad sea efectiva en la práctica, usted debe tener acceso al código fuente, porque modificar un programa sin disponer del código fuente es extraordinariamente dificultoso).
- **Libertad 2:** Usted tiene la libertad para redistribuir copias, tanto gratis como por un costo.
- **Libertad 3:** Usted tiene la libertad para distribuir versiones modificadas del programa, de tal manera que la comunidad pueda beneficiarse con sus mejoras.