Universidad Nacional de Mar del Plata
Facultad de Humanidades
Departamento de Documentación

Acceso y recuperación de información en la World Wide Web

Análisis de motores de búsqueda y metabuscadores

por

Cristian Merlino Santesteban

Director: Oscar A. Fernández

Marzo 2001
Mar del Plata, ARGENTINA

Tesina presentada para acceder al grado académico de Licenciado en Bibliotecología y Documentación.

Comité evaluador:

Lic. Gustavo Liberatore
CC. César Archuby
Lic. Oscar A. Fernández

A
mis padres,
mi hermano,
Zimba y Thor.

Agradecimientos a:

A. Ariel Barrios
Víctor F. Herrero Solana
Jorge A. Ruocco
M. Laura Montagna

Tabla de Contenido

1. Introducción

El almacenamiento, localización y recuperación de información eficiente ha siempre requerido del desarrollo de mecanismos de acceso, ya sea, para uso de los profesionales que organizan y gestionan la información, o uso de los usuarios finales.

Con el desarrollo de Internet, la recuperación de información eficiente ha ganado nueva importancia. Como un sistema de comunicación entre sistemas de computadoras abierto a la comunidad mundial, Internet, brinda a cientos de miles de organizaciones y personas la posibilidad de alojar información sin ningún tipo de control estructural u organizativo, o medio para su acceso; lo que provoca confusión y frustración al momento de ser localizada y recuperada por los usuarios. Por ello, desde su inicio se han ido desarrollando diversos programas de búsqueda para usos específicos: entre ellos, Archie para identificar archivos potencialmente relevantes vía *File Transfer Protocol* (FTP); Telnet para buscar en catálogos y/o directorios en línea como HYTELNET y LIBS; *Very Easy Rodent-Oriented Net-wide Index to Computerized Archives* (VERONICA) y Judhead para reconocer ítems de menúes potencialmente relevantes en Gopher; y *Wide Area Information Server* (WAIS) para acceder a documentos en texto completo usando diferentes bases de datos[1].

El auge de la World Wide Web (WWW, W3 o Web), que por un lado, está revolucionando la manera en que la sociedad accede a la información, y por otro, está creando nuevos desafíos para el campo de la Recuperación de Información, determinó que los sistemas de búsqueda mencionados fueran sólo el preludio de las actuales herramientas de recuperación de información.

[1] WAIS era un sistema en que múltiples bases de datos especializadas se distribuían en servidores dispersos controlados por un directorio, y cuyos contenidos eran accesibles y recuperables mediante el empleo de programas cliente. Los usuarios obtenían una lista de las bases de datos y, en repuesta a una expresión de búsqueda dirigida a una base de datos seleccionada, se accedía a los servidores que la contenían. Como resultado se obtenía una descripción de los textos y la posibilidad de obtener completos los documentos.

Los usuarios de la Web tienen básicamente dos maneras de encontrar la información que están buscado: pueden usar un sistema de búsqueda[2], o pueden navegar los nodos por medio de sus enlaces hipertextuales. Este último modo ha quedado casi en desuso puesto que cuando el tamaño de la Web creció más allá de unos pocos sitios y un reducido número de documentos quedo claro que un *browsing* manual a través de una porción significativa de la estructura hipertextual no es posible.

En febrero de 1999 investigadores del NEC Research Institute estimaron el tamaño de la World Wide Web en 800 millones de páginas *indizables* distribuidas aproximadamente en 3 millones de servidores (Lawrence y Giles, 1999). Al año siguiente en el mes julio un estudio conducido por Cyveillance Inc. determinó el tamaño de la Web en 2,1 billones de páginas, con un crecimiento diario de 7,3 millones de páginas (Moore, Murray y Brian, 2000). Estas cifras muestran claramente la necesidad de conocer cada vez más sistemas que permitan (o pretendan) hacer posible una rápida, efectiva y eficiente recuperación de información.

Antes de elaborar un juicio formal acerca de la efectividad de los servicios de búsqueda es necesario conocer las capacidades de la herramienta, la conducta del usuario y el entorno en que opera.

Este trabajo pretende aunar (aunque someramente) todos estos factores haciendo hincapié en dos tipos de servicios de búsqueda: motores de búsqueda y metabuscadores. En la primer parte se analizan y describen a fin de comprender su estructura y funcionamiento; luego nos centramos en los métodos de recuperación de información que utilizan dichos sistemas y finalmente se evalúa su rendimiento.

[2] De acuerdo a la décima encuesta del Graphics, Visualization, and Usability Centre (GVU) llevada acabo a finales 1998, el 84% de los "navegantes" usan servicios de búsqueda para localizar información.

1.1. Problema

¿Cuál es el grado de efectividad de los motores de búsqueda y metabuscadores de la World Wide Web para recuperar documentos relevantes y ordenarlos por relevancia?

1.2. Objetivos

Objetivo general:

- Analizar las características de los sistemas de recuperación de información (SRI) que ofrecen los motores de búsqueda y los metabuscadores.

Objetivos específicos:

- Detallar mecanismos de funcionamiento.
- Comparar la metodología de búsqueda utilizada por los sistemas.
- Describir métodos de recuperación de información.

Objetivo general:

- Evaluar el rendimiento de varios SRI.

Objetivos específicos:

- Examinar en qué forma los motores de búsqueda de mayor cobertura rankean los primeros documentos recuperados.
- Calcular la precisión, exhaustividad y cobertura de los buscadores y los metabuscadores.
- Establecer el grado de similitud de los mismos.

1.3. Estado de la Cuestión

Desde los inicios de Internet, el estudio de las herramientas de búsqueda ha despertado gran interés en la comunidad mundial. Con la rápida popularidad adquirida por la WWW, los nuevos SRI, se convirtieron en el centro de atracción tanto de la literatura de divulgación como de la académico-científica. Los trabajos publicados presentan enfoques cualitativos y en un menor número enfoques

cuantitativos. Habitualmente los estudios de divulgación son pobres evaluaciones de carácter meramente descriptivo (o publicitario). Las investigaciones realizadas por profesionales de la información e ingeniería en sistemas muestran más profundidad en sus aspectos comparativos mediante el uso de metodologías (aún no consolidadas) cualitativas y cuantitativas. En esta revisión de la literatura nos centraremos en aquellos trabajos que hayan evaluado la recuperación de información.

Mayormente los estudios científicos no han evaluado a las herramientas de búsqueda con las dos medidas más usadas en el campo de la Recuperación de Información: la precisión y la exhaustividad. Han optado principalmente por la primer medida. La exhaustividad muchas veces ha sido deliberadamente omitida, o bien por la incapacidad para determinar cuantos ítems relevantes hay para una búsqueda en particular en la base de datos del SRI, o por la aplicación de un método inadecuado para dicho testeo.

Leighton (1995) evaluó la relevancia y precisión de los resultados obtenidos de InfoSeek, Lycos, WebCrawler, y WWWWorm mediante ocho preguntas con diferentes grados de dificultad.

Zorn y otros (1996) usaron tres expresiones booleanas complejas en AltaVista, InfoSeek, Lycos y OpenText para mostrar el comportamiento de la búsqueda avanzada en la recogida de registros.

Chu y Rosenthal (1996) testearon AltaVista, Lycos y Excite usando diez consultas fáciles y difíciles derivadas de preguntas de referencia reales. Calcularon la precisión de los primeros diez resultados de cada herramienta y determinaron sus capacidades de búsqueda.

Ding y Marchioni (1996) compararon la precisión, duplicación y validación de vínculos, y solapamiento de InfoSeek, Lycos y OpenText. Cinco búsquedas complejas fueron ejecutadas y se juzgaron los primeros veinte resultados.

Gauch y Wang (1996) equipararon la precisión del metabuscador ProFusion con los metabuscadores MetaCrawler, Savvy, y los servicios AltaVista, Excite, InfoSeek, Lycos, OpenText y WebCrawler por medio de doce consultas y trabajando con los veinte primeros aciertos.

Schlichting y Nilsen (1996) analizaron AltaVista, Lycos, InfoSeek y Excite empleando de 4 a 6 palabras clave. La relevancia fue determinada para los diez primeros resultados usando *signal detection analysis.*

Leighton y Srivastava (1997) computaron la precisión de AltaVista, Excite, HotBot e InfoSeek comparando los primeros veinte resultados retornados por quince búsquedas con diferentes juicios de relevancia.

Hou (1998) analizó Yahoo, AltaVista y Lycos en la facilidad de uso y calidad de la información recuperada. A través de ocho preguntas calculó la precisión para los primeros cinco y diez ítems.

Ljosland (1999) comparó AltaVista, Fast y Google, por medio de veinte búsquedas y evaluó sus primeros diez documentos.

Courtois y Berry (1999) expusieron en su trabajo otro punto de vista para testear AltaVista, Excite, HotBot, InfoSeek y Lycos. A partir de su experiencia personal determinaron que la metodología empleada para hallar la precisión, a pesar de ser efectiva en muchos casos, no es la misma manera como los usuarios consideran sus listas de resultados.

Olvera Lobo (2000) publicó un estudio realizado en 1997 donde evaluó la precisión y exhaustividad de Excite, InfoSeek, HotBot, AltaVista, Magellan, Lycos, WebCrawler, OpenText y Yahoo usando veinte preguntas con diferentes grados de dificultad. Implementó para ello la metodología propuesta por Salton y McGill (1983).

Nicholson (2000) replicó el trabajo de Ding y Marchioni diez veces con una semana de separación entre cada replicación y determinó la precisión de los primeros diez y veinte aciertos para dos categorías de relevancia.

Gran parte de los estudios reseñados presentan diversas falencias desde un punto de vista crítico, ya que en general no detallan la metodología empleada, usan un número insuficiente de búsquedas o servicios de recuperación, no listan las consultas o no brindan el análisis estadístico correspondiente. Otra gran falencia de todos los estudios realizados sobre el rendimiento de herramientas de búsqueda web de acceso libre, e incluso el presente, es la imposibilidad de

conocer el algoritmo real de ordenación de resultados. Siempre se ha trabajado con la información escasa de carácter general que brindan las empresas comerciales o sitios web dedicados a esta temática. Aún hoy la competitividad del mercado ha determinado que el algoritmo sea un secreto celosamente guardado.

2. Herramientas de recuperación de información web

De la variedad de herramientas de recuperación disponibles en la W3 nos centraremos en dos de los servicios que tienen mayor impacto en la comunidad mundial, los motores de búsqueda y metabuscadores.

2.1. Motores de búsqueda

Los motores de búsqueda web tienen sus raíces en los sistemas de recuperación de información que fueron desarrollados originalmente para gestionar el fondo documental de las unidades de información.

Estos sistemas de búsqueda están alojados en servidores web basados en la arquitectura cliente/servidor que caracteriza a Internet y son accesibles a través del *HyperText Transfer Protocol* (HTTP). Constan, generalmente, de 4 (cuatro) partes: un programa para localizar e indizar documentos (robot), una base de datos para gestionar y almacenar la información extraída de ellos, un sistema de consulta e interrogación (incluye los métodos de búsqueda y recuperación, y un programa pasarela que sirve de unión entre el servidor que gestiona las peticiones de información del programa cliente y la base de datos), y finalmente una interfaz de usuario que permite interrogar al sistema y recibir los resultados obtenidos; como muestra la figura 1.

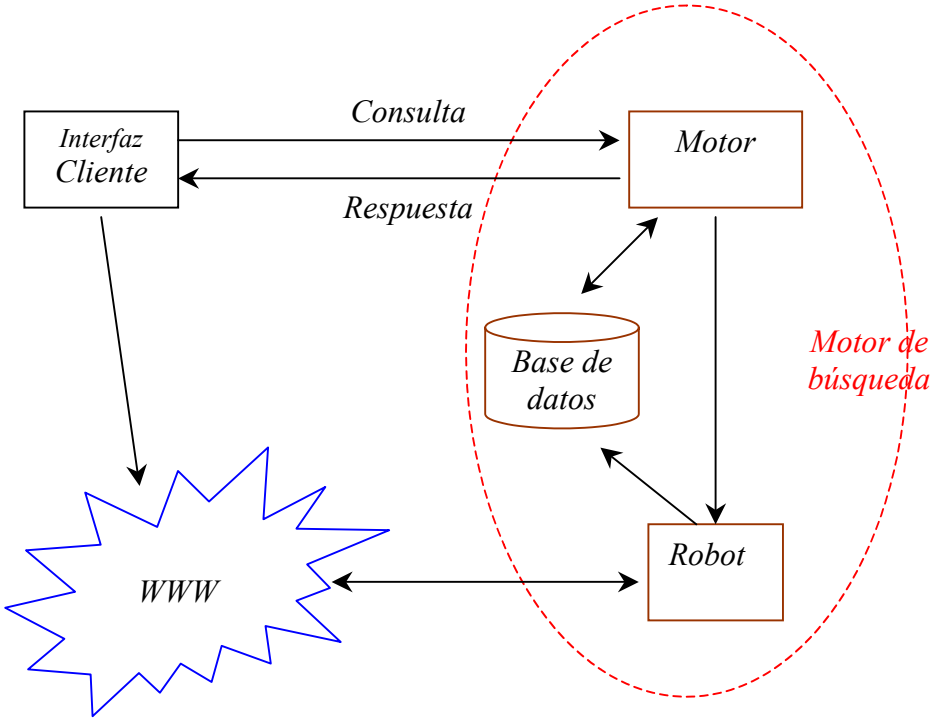


Figura 1 - Estructura general de un motor de búsqueda

2.1.1. Robots

El motor de búsqueda antes de aceptar consultas del usuario necesita determinar la colección de documentos contra la cual las consultas serán equiparadas. Usualmente esa recogida de documentos se realiza usando un robot.

Un robot es un programa que goza de cierta autonomía de acción y de capacidad de adaptación a su entorno, que recorre automáticamente la estructura hipertextual de la Web en forma recurrente recuperando todos los documentos que coinciden con un criterio específico (Koster, 1995).

La terminología usada en idioma inglés para designar a los robots (*spiders, wanderers, worms* o *crawlers*) da la impresión de que estos en realidad viajan a través de toda la WWW, lo cual es una imprecisión pues estos programas son estáticos. No tienen la propiedad de moverse de una computadora a otra pero pueden hacer uso de los recursos de red para accesar recursos remotos. Casi todos los robots de búsqueda convencionales recuperan documentos de

diferentes localizaciones en la Web a la computadora donde residen usando el protocolo HTTP.

2.1.1.1. Método de recolección de documentos

Los robots usan para recuperar los documentos en la Web el siguiente método:

a) El algoritmo usa una lista inicial de *Universal Resource Locators* (URLs) que ha sido compilada a mano por el administrador del robot. La lista debe contener al menos un URL.
b) Un URL es tomado de la lista (usando un método heurístico que es diferente para cada robot), y el documento referenciado es recuperado de la Web.
c) El documento es copiado en forma parcial o completa para recuperar la información que enviará a la base de datos y que además extraerá los anclajes hacia otros documentos.
d) Los URLs encontrados en el documento se agregan a la lista de URLs iniciales (el orden y posición en que se agregan URLs a la lista difiere entre los robots).
e) Si la lista está vacía o algún límite es excedido (número de documentos recuperados, tamaño del índice de la base de datos, etc.) el algoritmo se detiene, de lo contrario el algoritmo continúa con el paso b.

Cabe aclarar que un solo punto de arranque no es suficiente para encontrar la Web entera. Por lo tanto, la composición de la lista inicial de URLs es un paso muy importante para encontrar la mayor cantidad de documentos como sea posible. Una manera consistiría en determinar la lista en base a la popularidad de los sitios web. Intuitivamente podemos esperar que sitios muy populares contengan muchos anclajes que apunten a los recursos de información más frecuentes en el servidor local o servidores externos.
También, al usar un robot para encontrar información sobre un tema específico o zona geográfica, una lista inicial de URLs pertinentes es un paso clave.
Otra manera complementaria de localizar nuevos sitios web consiste en permitirle a los proveedores de información adicionar sus propios URLs, ya sea desde el mismo motor o por medio de un programa ad hoc. Esta red de trabajo juega un rol

clave en el éxito (desde el punto de vista de la cobertura) de un buscador en particular. Usualmente, un SRI con una colección grande de URLs atrae más usuarios y los hace más visibles. Consecuentemente, más proveedores de información harán el esfuerzo de remitir sus URLs al motor y así el motor tendrá una colección más completa.

2.1.1.2. Estrategia de navegación

El proceso de toma y agregando de nuevos URLs a la lista determina la estrategia de navegación del robot. Si los anclajes encontrados recientemente se agregan del mismo lado de la lista donde se escogen los URLs próximos a ser recuperados, el robot navega primero en profundidad (*depth-first*). Si los nuevos anclajes se agregan en un extremo de la lista y los URLs son escogidos del otro extremo, el robot navega primero en anchura (*breadth-first*).

La pluralidad de los robots combina ambas estrategias para beneficiarse de sus ventajas sin padecer sus inconvenientes.

La estrategia *depth-first* da la mejor distribución global de URLs en la Web, lo cual es importante cuando solo una parte relativamente pequeña de la WWW es recuperada (De Vocht, 1994). Hay que tener en cuenta que un documento web no consiste en un conjunto de páginas yuxtapuestas, sino en una red de nodos con n niveles de profundidad. Esta navegación genera el peligro de incurrir en una repetición infinita de servidores que generen documentos "en vuelo" (*on the fly*). Muchos documentos contienen enlaces a ellos mismos u otros documentos que son creados usando el mismo URL.

Cuando se usa, por ejemplo, el registro oficial de servidores como lista inicial, la estrategia *breadth-first* da resultados excelentes porque alcanza muchos servidores diferentes. Sin embargo esta estrategia de navegación es menos eficaz al profundizar a la Web. También, porque los vínculos de la lista se toman en el mismo orden en que fueron recolectados de los documentos. En muchos casos, estos anclajes apuntan a documentos en el mismo servidor, lo que ocasiona sobrecargas en la conexión.

Los robots tienen el potencial de hacer tantas solicitudes que pueden fácilmente ocupar todo el ancho de banda de la red y hacer uso de todo el poder de cómputo de los servidores. Incluso varios motores de búsqueda utilizan múltiples robots en paralelo para obtener un mejor rendimiento. Esto no solo acelera el proceso de localización y recolección de documentos, sino también evita bloqueos cuando un robot encuentra enlaces o servidores muy lentos.

2.1.1.3. Métodos de indización automática

La indización automática consiste en el reconocimiento (por parte del robot) de los términos e/o imágenes que figuran dentro del documento; y el almacenamiento tal cual de las palabras, o bien después de transfórmalas en otros términos equivalentes o conceptualmente próximos en el archivo invertido, ó previa conversión de los términos y las imágenes en patrones de bits para luego utilizarlos como elementos de recuperación[3].
Los robots utilizan distintos métodos para indizar las páginas web que incorporan a la base de datos de los motores de búsqueda. La indización puede ser por técnicas estadísticas, semilingüísticas, lingüísticas y reconocimiento de patrones binarios. Por ahora, las dos últimas técnicas están disponibles en sistemas de uso restringido.

La indización por técnicas estadísticas (simples) consiste en una indización por extracción, donde las palabras (cadenas de caracteres separadas por un espacio en blanco) que aparecen en un documento son extraídas y utilizadas para representar el contenido del documento como un todo. Mediante este método se crea un archivo invertido de palabras clave, ubicación y frecuencia de aparición. Este enfoque basa la recuperación de información en la similitud formal de las palabras. Algunos robots obtienen las palabras clave de determinados campos (título, encabezamientos, metaetiquetas HTML), pero la mayoría indiza el texto completo de las páginas, incluyendo o no las palabras vacías de significado y eliminando a veces las más frecuentes.

La indización por técnicas semilingüísticas implementa análisis estadísticos más complejos y procesamientos lingüísticos básicos para extraer, además de palabras, frases que ocurran con significativa frecuencia en el texto y seleccionar raíces de palabras. De este modo, por ejemplo, la raíz metal es selecciona y almacenada en lugar de la variante metálico y metalmecánica. Este tipo de técnica también permite la indización por asignación a través de análisis de co-ocurrencia, esto es detectar relaciones entre términos y/o frases a fin de mejorar la eficacia de las búsquedas. Cuanto más frecuente dos términos ocurran juntos, más probable será que ellos traten de un tópico similar. Por ejemplo, si la palabra u ocurre muy frecuentemente con g y la palabra g aparece muy frecuentemente con u, los dos términos tienen un alto grado de interdependencia. Una relación indirecta entre términos podría ser que la palabra h casi nunca ocurra sin i en una base de datos y que el término j casi siempre co-ocurra con i, sin embargo h y j nunca co-ocurren juntos en los textos. Esto indicaría alguna relación entre h y j, debido al hecho de que ambos términos co-ocurren fuertemente con i. A partir de estos análisis y procesamientos el robot determina qué palabras y/o frases aparecen juntas o relacionadas en textos que se centran en un tema concreto. De esta manera un documento que trata de un tema dado podrá ser recuperado, incluso aunque las palabras y frases incluidas en la página web no coincidan formalmente con las de la pregunta.

La indización por reconocimiento de patrones (nivel submorfológico) consiste en la clasificación de un objeto dentro de alguna categoría definida, en base a las características del objeto que son comunes a los elementos de una misma clase. La clasificación, desde el punto de vista matemático, consiste en la partición del espacio n-dimensional definido por las características de un objeto, en varias regiones, donde cada región corresponde a una clase.

[3] Todavía la indización automática de imágenes (en la WWW) se encuentra en su primera fase de desarrollo.

Algunos sistemas de recuperación han incorporado la indización y búsqueda de texto, imagen y sonido por medio del reconocimiento de patrones[4] de bits (*bits patterns*), es decir, usando la misma forma de representación. El reconocimiento de patrones opera por cálculo de probabilidad de ver x si vemos y, y luego cual es la probabilidad de ver z si ambas x e y están presentes, y así sucesivamente (Wiley, 1998). Esta técnica ofrece un método muy flexible y exacto para la recuperación.

La herramienta comercial de búsqueda Excalibur Visual RetrievalWare[5] ofrece un proceso de reconocimiento adaptativo de patrones basado en redes neuronales. El sistema es capaz de indexar y recuperar información textual e icónica introduciendo patrones binarios en ella.

Por medio de la indización a nivel binario, se consiguen importantes ventajas: a) es independiente del lenguaje, ya que toda información independientemente del idioma es almacenada por la computadora en código binario b) y asimismo es independiente del tipo de dato por el mismo motivo, y puede ser aplicado para la recuperación completa del contenido de textos, imágenes y videos.

Los métodos lingüísticos se basan en el *Natural Language Processing* (NLP)[6]. El procesamiento del lenguaje natural se efectúa por medio de diferentes niveles de análisis: morfológico, sintáctico, semántico, discursivo, pragmático y fonético. El nivel morfológico persigue la segmentación de la palabra ortográfica con el fin de obtener la gramatical y determinar su estructura y propiedades; el nivel sintáctico detecta las relaciones sintácticas entre las palabras de una frase gramatical; el nivel semántico determina el posible significado de las oraciones en su contexto; el nivel discursivo interpreta la estructura y significado del texto más allá de una oración; el nivel pragmático trata de entender el uso del lenguaje en situaciones, particularmente aquellos aspectos donde se requiere una base de conocimiento previa; y el nivel fonético interpreta la manera de pronunciar las palabras.

---

[4] Modernamente podemos distinguir varias corrientes o "enfoques" en el reconocimiento de patrones: el enfoque estadístico, el numérico, el sintáctico, el lógico-combinatorio y el de redes neuronales entre otros.

[5] URL: http://www.excalibur.com

[6] El tratamiento NLP manipula los documentos como objetos lingüísticos.

El sistema *Document Retrieval using LINguistic Knowledge* (DR-LINK)[7], hace un completo procesamiento del documento y la consulta (necesidad de información) en todos los niveles aludidos.

Efectividad de la indización

La efectividad de un sistema de indización es controlada por dos parámetros principales: exhaustividad y especificidad (Gudivada y otros, 1997). La exhaustividad de la indización refleja el grado en que todos los conceptos realmente significativos manifestados en el documento son reconocidos por el sistema de indización.

Cuando el sistema de indización es demasiado exhaustivo (genera un largo número de términos para reflejar todos los aspectos del tema tratado) hará que se recuperen documentos que no contengan información pertinente sobre los conceptos de la consulta; disminuye la precisión y aumenta el ruido. Si la exhaustividad es demasiado reducida (genera unos pocos términos correspondientes a los temas principales que caracterizan el contenido del texto) hará que no se recuperen documentos pertinentes, aumenta el silencio y disminuye la tasa de llamada.

La especificidad se refiere a la precisión de los términos usados para la indización. Los términos generales recuperan muchos documentos útiles pero con un número significativo de documentos irrelevantes, disminuye la precisión y aumenta la tasa de llamada; y los términos relacionados recuperan pocos documentos e incluyen algunos irrelevantes, disminuye la precisión.

[7] URL: http://drlink.mnis.net

2.1.1.4. Otros usos

Los robots, ciertamente, pueden realizar otras tareas a parte de localizar e indizar documentos, entre ellas podemos mencionar (Heinonen, Hätönen y Klemettinen, 1996):

Mirroring: esta aplicación consiste en copiar una colección de documentos de un sitio a otro, con frecuencia en otro país o continente. Esto evita la conexión innecesaria entre sitios remotos geográficamente y reduce sustancialmente los tiempos de acceso. También es útil cuando el sitio original no está disponible o se encuentra sobrecargo por el número de conexiones recibidas. Un aspecto negativo del *mirroring* es la generación de una gran cantidad de contenido duplicado. Se estima que alrededor del 30% de la WWW está repetida (Bharat y Broder, 1999).

Mantenimiento de URLs: una de las principales dificultades de una estructura hipertextual dinámica y volátil, como es la Web, es que las referencias a otros documentos se conviertan en enlaces inactivos cuando el nodo enlazado ha sido borrado o movido a otra localización. Por consiguiente, el robot realiza periódicamente tareas de mantenimiento de la lista de direcciones, actualizando las modificaciones que se puedan producir en un nodo, cambios de URLs, etc. El último estudio de NEC proveyó evidencia de una demora de meses en la renovación de páginas (Lawrence y Giles, 1999).

Análisis estadístico: los primeros desarrollos de robots estuvieron enfocados a este tipo de tarea, por ejemplo, contar el número de servidores web, determinar el número promedio de documentos por servidor, la proporción de ciertos tipos de archivos, el tamaño promedio de una página web, el grado de interconectividad, etc.

2.1.1.5. Problemas de uso

El uso de robots automáticos, a pesar de su gran utilidad, presenta desventajas al ambiente de la Web. Podemos identificar los siguientes problemas:

- El ancho de banda de la Red es en la actualidad bastante limitado debido a la cantidad de información que circula a través de ella, por ello el uso los robots de manera indiscriminada produce un incremento notable del tráfico y una sobredemanda en los servidores.
- Su funcionamiento es demasiado simple, por lo cual, el tipo de datos recogidos no es muy útil.
- No pueden determinar automáticamente si una página web debe o no debe ser incluida en el archivo invertido (por ejemplo, páginas temporales y *mirrors*).
- Los sistemas de indización aún no están lo suficientemente desarrollados como para representar adecuadamente el contenido de los documentos.
- Los documentos disponibles en la Web son muy heterogéneos (textos, imágenes, sonidos, animaciones, etc.); pero la mayor parte de los robots son incapaces de indizar documentos no textuales.
- La frecuencia de actualización del índice no se corresponde con el dinamismo de la Web, haciendo casi imposible reflejar su estado en tiempo real.

2.1.2. Base de Datos

La base de datos es básicamente similar a la de un programa de gestión documental convencional; recibe como entrada la información indizada de los documentos localizados por el robot y produce como salida un archivo invertido.

El índice o archivo invertido, en la casi totalidad de los sistemas, se genera por medio de la asociación de cada palabra con su posición exacta dentro del documento. En otros casos la información posicional no es tan precisa y se consigna solo la frase, el párrafo o incluso solo el documento al que pertenece la palabra en cuestión (Moya Anegón, 1995).

En una primera fase el sistema lee todos los términos indizados del documento y almacena de forma temporal cada uno de los que considera claves y su ubicación;

en la segunda fase comienza la ordenación alfabética del archivo generado en la fase anterior, de tal forma que los términos repetidos aparecen juntos, o bien, las raíces de las palabras, ya que, con frecuencia en vez de palabras completas el archivo invertido almacena las raíces de las palabras[8]. El número de palabras distintas no crece en forma proporcional al texto, sino que crece en forma sublineal. Esto se debe a que el vocabulario es finito y entonces muchas palabras se repiten. Por otra parte, la frecuencia de las palabras sigue una variante de la Ley de Zipf que caracteriza la ocurrencia de palabras en un texto[9].

También suelen excluirse del índice las palabras que no contienen información semántica (preposiciones, conjunciones, etc.), tales como *"y", "a"*, *"the"* y *"de"*, o no cumplan con un umbral preestablecido. Estas palabras son colocadas en una colección de palabras llamada lista de palabras vacías (*stop list*). Ordenada esa información se actualiza el índice que contiene todos los términos de indización diferentes aparecidos en los distintos documentos, junto con el número de documentos en que aparece y el total de apariciones. Y por último, en la tercer fase se completa la información del invertido con otro archivo que contendrá las informaciones posicionales de cada una de las claves del diccionario.

Estas bases de datos no contienen los documentos originales en texto completo, sino únicamente las direcciones de los documentos. Pero hay excepciones, como por ejemplo, la base de datos del motor de búsqueda Google que posee un repositorio que contiene cada página web en forma completa (no incluye imágenes).

2.1.3. Sistema de interrogación e interfaz de búsqueda

El sistema de interrogación toma las preguntas del usuario (a veces casi en lenguaje natural), elimina asiduamente las palabras no significativas comparando

[8] Los vocablos con la misma raíz se tratan como semejantes basándose en la premisa que términos similares morfológicamente también lo son semánticamente. Si embargo, si el programa no aplica esta prestación adecuadamente puede incurrir en resultados incorrectos, incrementado la tasa de ruido en la recuperación.

las ocurrencias de la consulta contra la lista de palabras vacías, y recorre el archivo invertido de la base de datos para seleccionar las entradas relevantes.
El lenguaje de interrogación[10] es la parte más importante del sistema de consulta, mediante el cual el usuario puede expresar su necesidad de información, redefinir la búsqueda, acotar el número de aciertos, aplicar diversos filtros, etc.
La interfaz típica de un motor de búsqueda presenta una casilla para el ingreso de la ecuación de búsqueda y un botón para ejecutarla, otras veces además incorpora una casilla para optar como serán procesados los términos de la consulta (por ejemplo *any of the words/all the words/the exact phrase*). Pero la mayor parte de las opciones del sistema serán visibles cuando se haga uso de la búsqueda avanzada.

Las opciones de búsqueda avanzada incluyen:

- búsqueda booleana y paréntesis
- especificación de los términos que deben estar o no presentes
- truncamiento (manual o automático)
- frase exacta
- búsqueda por proximidad
- búsqueda por dominio, enlace y sitio web
- búsqueda por campos (título, cuerpo del documento, etiquetas meta, etc.)
- filtrado por fecha, dominio, idioma, tipo de caracteres (normas ISO) o tipo de archivo (basándose en el nombre de la extensión)
- búsqueda por retroalimentación
- búsqueda sensible a letras mayúsculas

La mayoría de los motores, como respuesta a una consulta, presentan 10 (diez) o más resultados al mismo tiempo con un formato de visualización por defecto mostrando el título y algo de texto.

---

[9] Esta Ley indica que la j-ésima palabra más frecuente aparece una cantidad de veces proporcional al inverso de j. Es decir, hay un conjunto pequeño de palabras muy frecuentes y muchas que aparecen muy pocas veces o sólo una vez.
[10] Generalmente está basado en el álgebra relacional y el cálculo relacional.

El formato de visualización (varia en corto, mediano y largo) puede incluir alguno de los siguientes elementos:

- título
- grado de relevancia (expresado en diferentes escalas)
- sumario (los sumarios pueden ser resúmenes preparados, líneas extraídas de las etiquetas de los encabezamientos, las primeras palabras, palabras más frecuentes, alguna representación construida automáticamente, etc.)
- fecha del archivo
- tamaño del archivo en bytes
- URL
- idioma
- categoría (si el servicio de búsqueda ofrece también un sistema clasificatorio)
- términos de búsqueda resaltados
- fecha de indización

A su vez, pueden ser incluidos otros elementos en la visualización de los resultados:

- traducción
- páginas similares (búsqueda por retroalimentación)
- páginas relacionadas (páginas referenciadas por el recurso)
- resultados del sitio

2.2. Metabuscadores

A pesar los motores de búsqueda de ser herramientas útiles y populares poseen una gran limitación: su cobertura. Se estima que ninguno de ellos indiza más del 16% de la Web (Lawrence y Giles, 1999).

Como cada servicio de búsqueda ofrece una recogida incompleta de recursos, los usuarios, se ven forzados a utilizar y reutilizar sus consultas en diferentes motores hasta encontrar la respuesta adecuada a su necesidad de información. El proceso de emplear múltiples motores es tedioso puesto que hay que lidiar con interfaces y sistemas de interrogación diferentes. La solución a este inconveniente fue el diseño de metabuscadores.

El metabuscador es una herramienta de recuperación de información que está constituida únicamente por una interfaz de interrogación y el algoritmo de recuperación. Cuando el metabuscador recibe la consulta del usuario, al no mantener su propia base de datos, la remite automáticamente en paralelo a más de un motor de búsqueda y/o directorio temático y luego recoge (a veces, reorganiza) los resultados obtenidos (fig. 2).

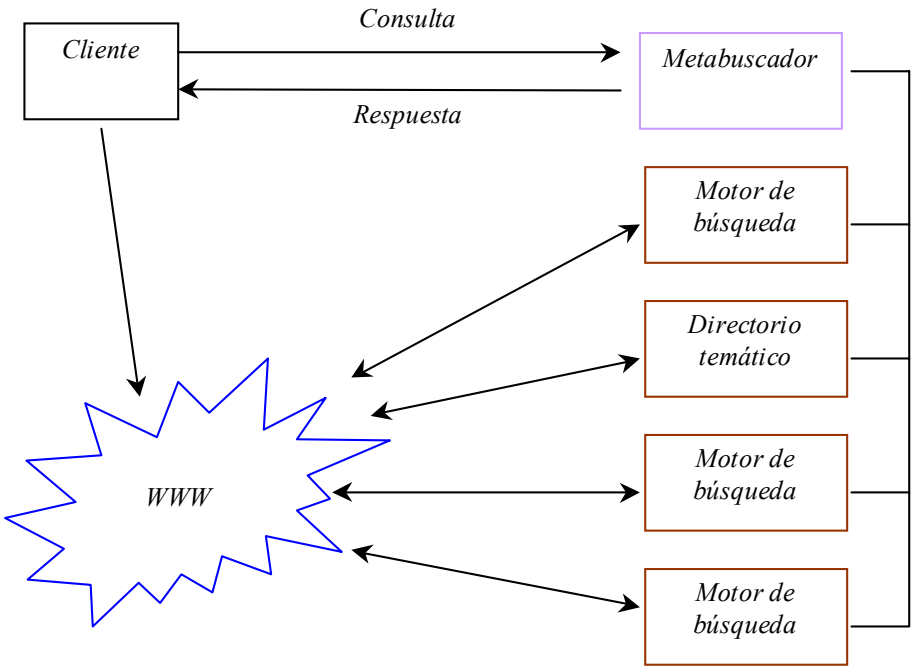


Figura 2 - Estructura general de un metabuscador

Las ventajas principales de los metabuscadores actuales son la habilidad para combinar los resultados de diversos motores de búsqueda y la habilidad para proveer una interfaz de usuario consistente para buscar en estos motores.
La idea de consultar e intercalar resultados de múltiples bases de datos no es nueva. Compañías como Dialog, Verity, Lexis-Nexis y Personal Library Software (PLS) llevan un tiempo largo creando sistemas que integren los resultados de varias bases de datos heterogéneas (Selberg y Etzioni, 1995).

Desde el punto de vista del usuario un metabuscador luce como cualquier buscador. Sin embargo la gran diferencia entre ellos está dada en su funcionamiento interno.
Para Sander-Beuermann y Schomburg (1998) la caracterización de un real metabuscador va más allá de ofrecer una única interfaz para interrogar diferentes motores. Se requiere además el cumplimiento de ciertos requisitos:

- Las búsquedas deben ser realizadas en paralelo.
- Los resultados de los diferentes motores de búsqueda deben ser combinados, el metabuscador debe hacer más que un listado simple de un acierto detrás del otro.
- Los aciertos idénticos encontrados de los diferentes buscadores (duplicados) deben ser eliminados.
- Los operadores booleanos deben estar disponibles.
- La descripción de los aciertos, si es presentada en el motor, debe ser transferida al metabuscador. La información que presenta el metabuscador no debe ser menor a la que ofrece el buscador.
- Las especificaciones fundamentales de los servicios de búsqueda utilizados deben ser invisibles al usuario. El usuario no debe necesitar conocer el funcionamiento específico de cada sistema de recuperación.
- El metabuscador debe permitir una búsqueda completa, es decir, debe presentar los aciertos en la medida que cada motor de búsqueda sea capaz de ofrecer sus resultados.

2.2.1. Funcionamiento

El principio de funcionamiento de un metabuscador puede ser descripto fácilmente en los siguientes pasos:

a) Acepta la consulta del usuario.
El usuario tipea la consulta en el formulario de búsqueda simple o avanzada y realiza su envío.
b) Convierte la consulta en la sintaxis correcta para cada servicio de búsqueda.
La expresión de búsqueda es traducida por el algoritmo a la sintaxis apropiada para cada motor de búsqueda o directorio temático que abarca.
c) Remite la consulta en sus múltiples sintaxis a los diversos SRI.
Finalizada la conversión transmite las diferentes ecuaciones de búsqueda a sus respectivos servicios de búsqueda.
d) Espera por las respuestas.
Espera un tiempo prudente para recoger la totalidad de las respuestas brindadas por las herramientas de recuperación interrogadas.
e) Analiza los resultados, elimina duplicados.
Captura los resultados y elimina los URLs duplicados exactos (no parciales).
f) Combina los resultados.
Todos los aciertos son mezclados conformado una única lista de ítems.
g) Elabora el ranking.
Ordena la lista por relevancia de los documentos.
h) Entrega los resultados postprocesados al usuario.
Presenta la lista final al usuario como respuesta a su consulta.

2.2.2.Tipos de metabuscadores

Al no requerir de una base de datos propia, fueron desarrollados dos tipos de metabuscadores. Los que se ejecutan desde el servidor (*server-side*) o aquellos que se instalan y ejecutan desde la computadora cliente (*client-side*) sin necesidad de conectarse previamente a ningún sitio.
Los metabuscadores como aplicación cliente sufren de dos inconvenientes (Sander-Beuermann y Schomburg, 1998):

- problema de último tramo de la conexión
- problema de actualización

El primer problema se debe a que el último tramo de la conexión entre el proveedor y el usuario es la parte con el menor ancho de banda, por otro lado cada metabuscador crea un gran flujo de datos desde los servicios de búsqueda. De estos flujos de datos cerca del 50% o más es remitido al metabuscador para su procesamiento sobreexigiendo la computadora y el sistema del usuario.
El problema de actualización resulta del hecho de que los administradores de los servicios de búsqueda tienden a cambiar sus formatos de salida con frecuencia. Con cada cambio en el formato el programa de postprocesamiento del metabuscador debe ser actualizado. La experiencia ha demostrado que al menos una vez por mes, sin tener en cuenta las propias modificaciones del programa cliente; lo cual es impráctico para el usuario.
En contrapartida a los inconvenientes señalados el programa cliente permite al usuario manipular los resultados devueltos de múltiples formas, grabarlos en diferentes formatos, reordenarlos o mostrarlos en función de su estado (descargado, nuevo, seleccionado, etc.) y mayor capacidad de filtrado.

2.2.3. Interfaz y opciones de búsqueda

La interfaz de un metabuscador es muy similar a la de un buscador. El usuario inexperto ni siquiera notará la diferencia. Normalmente exhibe una casilla para ingresar la consulta, un botón para ejecutarla y, a su vez (aunque no en todos lo casos) una caja de verificación donde se indican los servicios de búsqueda utilizados[11].

La generalidad de los metabuscadores permite formular la expresión de búsqueda de manera similar a los motores populares e incorporan la lógica booleana. En su búsqueda avanzada incorporan distintos filtros (tiempo de espera, tamaño del formato, etc.), búsqueda por host, enlace, título, URL, búsqueda por ostención, búsqueda por proximidad y truncamiento.

Una de las cosas más importantes a saber antes de usar la herramienta es si realmente traduce la sintaxis de búsqueda a la sintaxis propia de cada servicio de búsqueda antes de transmitirla, o si la transmite como está (tal como se ingreso). Si ocurre esto último, es muy probable que los resultados obtenidos sean de baja relevancia (ruido documental) puesto que los buscadores y directorios van a interpretar la interrogación de forma incorrecta o no haciendo uso de todo su potencial de búsqueda.

2.2.4. Beneficios e inconvenientes

Como toda herramienta de recuperación de información, los servicios de búsqueda múltiple, tienen sus virtudes y defectos; sobretodo en un ambiente tan inestable como es la W3. Podemos indicar los siguientes:

- Los usuarios necesitan aprender una interfaz sola para acceder a un conjunto de servicios de búsqueda. Cada motor de búsqueda tiene un único formulario para expresar la consulta y un único formato para presentar sus resultados. Estas interfaces no son difíciles de aprender pero se requiere de un cierto

[11] En otros casos el nombre y la cantidad de servicios de búsqueda utilizados está semioculta en menúes de ayuda o algunas veces no son descriptos por completo.

tiempo para familiarizarse con ellas, y a su vez, el cambio de una a otra puede confundir.

- Los usuarios de los metabuscadores no participan de la localización de nuevos documentos, en cambio en casi todos los buscadores disponen de una opción para adicionar manualmente una nueva dirección al robot. En lugar de ello depende de los administradores encontrar e integrar nuevas herramientas de búsqueda. Un beneficio adicional de un metabuscador es que nos alertan de las herramientas subyacentes.
- Es mucho más eficiente y conveniente consultar en paralelo con el metabuscador que realizar la misma tarea secuencialmente con cada servicio de manera individual. Antes que tomarse el tiempo de encontrar un buscador que no esté sobrecargado y ofrezca resultados útiles para su consulta
- Pueden ofrecer una búsqueda más exhaustiva. En contraposición, la recogida de los buscadores es mucho menor, no obstante algunos están especializados en ciertos tópicos en particular, con lo cual el metabuscador recupera más documentos relevantes para una consulta en particular.
- Los metabuscadores consumen más recursos de red que los motores convencionales - específicamente recursos de ancho de banda y servidor -. La generación de múltiples consultas y la recepción de sus múltiples respuestas de manera simultánea aumenta en forma considerable el tráfico de la Red.
- Las interfaces de usuario son engañosas. Algunos motores proveen elaboradas interfaces diseñadas especialmente para que la herramienta sea vista a través de ellas. Los usuarios de los metabuscadores son aislados de estas interfaces. Además, algunos recursos experimentales, nuevos o actualizados tienen características que no son generalizadas en los formatos de metabuscadores, por lo tanto deben ser accedidas desde la interfaz nativa. Los buscadores son distanciados de los usuarios humanos; no pudiendo distinguir quien hace la consulta.
- Una demanda alta y continua dificulta el rendimiento del sistema. Por cada interrogación de usuario el sistema abre diversas conexiones. Los

metabuscadores de hoy tiende a centralizar el servicio, o sea, los usuarios acceden a una computadora central en la cual procesan sus búsquedas.

- La combinación del potencial de muchos buscadores y directorios temáticos no necesariamente provee mejores resultados. Porque cada base de datos ofrece diferentes facilidades, como búsqueda por proximidad, ponderación de términos, equiparación exacta o parcial. Las herramientas que distribuyen sus consultas sobre dichas bases de datos no pueden aprovechar la fortaleza de todos esos sistemas.
- Los metabuscadores pueden introducir sus propias deficiencias, por ejemplo, pueden tener dificultades para ordenar por relevancia las listas de resultados. Si un SRI devuelve muchos documentos poco relevantes hará más difícil la tarea de encontrar páginas relevantes en la lista.

3. Recuperación de información

La recuperación de información en la WWW se basa en tres elementos: el sistema, el usuario y el corpus documental.

El sistema es un programa informático (*software*) en el cual el usuario puede ejecutar una consulta comprimida en unas pocas palabras (representación formal del requerimiento de información) y recibir referencias o documentos previamente evaluados por el algoritmo de recuperación (similitud pregunta-documento).

El usuario es el individuo que interactúa con el sistema por medio de la interfaz de búsqueda al efecto de satisfacer su necesidad de información. Existen distintas tipologías de usuarios determinadas por características personales y aspectos cognitivos que no vamos aludir en este trabajo.

El corpus documental conforma la base de datos del sistema. A diferencia de sistemas tradicionales la Web es un repositorio de información heterogénea, mal estructurada desde el punto de vista de la semántica del lenguaje de marcado y desorganizada.

3.1. El sistema

Los sistemas de recuperación de información de la Web han incorporado una diversidad de métodos de recuperación que varían entre simples (lógica booleana, ponderación de términos) y de cierta sofisticación (modelo de espacio vectorial, análisis de enlaces). Aunque en realidad integren características de más de uno de estos métodos.

Baeza-Yates y Ribeiro-Neto (1999) estiman que la mayoría de los buscadores usan técnicas de ponderación de términos, equiparación (y sus variaciones) o modelo de espacio vectorial.

3.1.1. Equiparación exacta y lógica booleana

La equiparación exacta demanda que la representación del documento contenga la representación exacta de la consulta. Si la representación consiste en la indización de los términos presentes, la equiparación exacta debe chequear cada palabra de la búsqueda contra los términos indizados del documento.

Usualmente, la equiparación exacta es implementada como el álgebra booleana (AND, OR y NOT). Los documentos recuperados son aquellos donde los términos de indización están o no están presentes. Todos los documentos obtenidos por este método serán considerados iguales, es decir, no habrá ningún tipo de ordenación por relevancia entre ellos.

Este modelo binario de recuperación ha recibido muchas críticas por los especialistas del área. Moya Anegón (1995) recopila las siguientes:

- La lógica booleana no se adquiere por intuición sino que requiere formación por parte de los usuarios.
- Los operadores booleanos son unas veces demasiados restrictivos (AND) o demasiado inclusivos (OR).
- El álgebra de Boole es demasiado rígida en cuanto a las posibilidades de entrada y salida que ofrece a los usuarios.

- Parecen no existir la incertidumbre y parcialidad como características inherentes a los procesos de indización y recuperación.
- Dificultad de transformar las necesidades del usuario en operaciones booleanas.

Por otro lado, Frants y otros (1999) analizaron las críticas efectuadas a este sistema y demostraron que actualmente la mayoría de las opiniones vertidas sobre el sistema de recuperación booleano están dirigidas a su implementación metodológica y no al principio booleano en sí mismo.

3.1.2. Ponderación de términos

Existen muchos refinamientos del modelo booleano para rankear los resultados por algún tipo de relevancia. Los más usados son los modelos de ponderación de términos. Podemos mencionar fundamentalmente dos: por frecuencia de aparición, y por ubicación dentro del documento y símbolos especiales, por ejemplo etiquetas asociadas al término y tipo de fuente. En frecuencia de aparición, a su vez, se pueden distinguir los estrictamente sumatorios ó los que se basan en aplicaciones de la Ley de Zipf[12]. El sumatorio utiliza la cantidad de veces que aparecen las ocurrencias de la búsqueda en cada documento. Por ejemplo, si la consulta contiene dos palabras p_q y p_r y la sumatoria del documento d_a es 7 y el documento d_b es 6, se considerará más relevante a d_a, aunque sólo contenga una ocurrencia del p_q y seis de p_r, y d_b contenga tres ocurrencias de cada término.
Una variación de la Ley de Zipf propuesta por Spack Jones propone dar mayor significación a términos que ocurren en unos pocos documentos[13].
En ponderación por localización y etiquetas especiales el sistema le otorga un peso a cada término de acuerdo a su localización (título, encabezamientos, primer párrafo, metadatos, etc.) y significación tipográfica-semántica dentro del cuerpo del documento (Brin y Page, 1998). Usualmente, un término ubicado en el título o

[12] Existe cierta relación entre la frecuencia de aparición de las palabras y la importancia que éstas tienen para representar el contenido de los documentos.
[13] A mayor frecuencia de una palabra en la colección de la base de datos menor es su representatividad dentro del documento.

encabezado es más importante que el mismo término localizado en el cuerpo del documento. Un término encerrado entre etiquetas o tipeado en fuentes especiales es probable que sea más importante que el mismo término sin alguna etiqueta o fuente especial asociada.

3.1.3. Modelo de espacio vectorial

El modelo de espacio vectorial propuesto por Gerald Salton a principios de los años setenta consiste en otorgar un peso desigual a los términos, o sea, en lugar de valorar con un 1 (coincidencia exacta) o con un 0 (coincidencia nula) la presencia y ausencia de un término, se asignan valores situados entre 1 y 0, según la capacidad discriminatoria del término (Moya Anegón, 1995).

En este acercamiento cada documento es considerado como un vector *n*-dimensional, donde *n* es el número de términos en el documento. Para calcular el valor discriminatorio a cada término se le otorga un peso[14]. Este peso es determinado como la frecuencia inversa del documento *(fid)* por la frecuencia del término *(ft)*.

$$peso_t = fid * ft$$

La *fid* resulta de la división del número de documentos de la colección *(N)* por el número de documentos en los que aparece un término determinado *(fd)*.

$$fid = N / fd = log(N/fd)$$

Esto refleja el hecho que palabras poco comunes en la colección, cuando son especificadas en una consulta, se ponderarán más alto y palabras muy frecuentes se ponderarán con valores bajos. La *ft* es el número de veces que aparece un término determinado. Entonces palabras que ocurren múltiples veces en un documento harán que ese documento sea más relevante que un documento donde el término aparezca una sola vez. También, el término es mejor ponderado según lugar de ubicación dentro del documento.

[14] Idealmente, el peso de un término en un documento debe indicar la importancia del término en la representación del contenido del documento.

Por lo antedicho, también, una consulta *(Q)* puede ser considerada como un vector *z*-dimensional (*z* es el número de términos de la consulta). Los términos son ponderados en base a su *ft* y/o *fid.* Mientras la *ft* es usada en las consultas y los documentos, la *fid* es usualmente usada en las consultas ó en los documentos, pero no en ambos (Yu y Meng,1999).

El vector de la consulta y el vector del documento son comparados por su relevancia usando el coeficiente de similaridad, más específicamente el coeficiente normalizado del coseno, que es el ángulo de dos vectores en el espacio. Cuando el documento es similar a la consulta, el ángulo entre los dos vectores es cero (coseno 0º = 1). Si el documento y la consulta no tienen ningún término en común, el ángulo es de noventa grados (coseno 90º = 0).

La normalización del coeficiente del coseno se refiere al ajuste del producto *fid* ft* dividido por la longitud del documento, que es la raíz cuadrada de la suma de los cuadrados de *fdi*ft* por cada término (Harman,1992).

$$SIM(Q_x D_y) = \frac{\sum_{i=1}^{n} q_{xi} d_{yi}}{\sqrt{\sum_{i=1}^{n} q_{xi}^2 \sum_{i=1}^{n} d_{yi}^2}}$$

donde:
SIM es la medida de similitud entre la consulta x y el documento y
Q_x es el vector de la consulta x
D_y es el vector del documento y
q_{xi} es el peso del término representativo i de la pregunta Q_x
d_{yi} es el peso del término representativo i del documento D_y
n es el número de términos representativos en los vectores

La idea atrás de la normalización es que el peso de cada término se ajuste a la longitud del documento[15]. Así, si una palabra aparece 5 veces en un texto de 430 palabras, el documento deberá ser mejor ponderado que si en un documento de 801 términos la palabra ocurre 5 veces.

[15] Es muy probable que un documento extenso acerca de un tópico tenga una mayor frecuencia de aparición de palabras de la consulta que uno más corto.

3.1.4. Análisis de enlaces

Inspirada en el análisis de citaciones de la literatura científica, el uso de la estructura de enlaces ha emergido recientemente como un nuevo y promisorio acercamiento a la búsqueda de información en la Web (Carriere, Kazman, 1997). Una citación provee una relación entre dos o más documentos, y con frecuencia es la única manera de localizar textos pertinentes relacionados con el tema del documento citante. Un enlace de una página web sirve para un propósito similar, pero hay importantes diferencias entre la citación y un enlace web (Efe y otros, 2000):

- Una citación en la literatura científica es estática y unidireccional. Una vez publicado un trabajo no hay manera de incorporar nuevas referencias. En cambio, en las páginas web se pueden agregar nuevos enlaces a otros documentos creados posteriormente.
- La referencia web generalmente es más subjetiva que en la literatura científica (Weinstock, 1971). Los creadores de documentos web muchas veces no toman en cuenta la relevancia, calidad u objetividad de la información.
- Mientras algunos enlaces en una página web pueden dirigir a documentos relacionados (o no relacionados), otros son creados por razones que no tienen que ver con este "aval". Muchos enlaces existen por propósitos puramente navegacionales (retroceder, ir al inicio) o publicitarios (gane dinero ya, compre autos baratos).

Un hipervínculo en una página web conecta un documento con otro y representa un aval implícito con la página destino. Cuando consideramos dos vínculos, obtenemos un variado número de patrones básicos (Efe y otros, 2000) (fig. 3). Dos páginas apuntándose entre si refuerzan nuestra intuición acerca de su mutua relevancia. Una página referenciando dos páginas distintas (co-citación) sugiere la probabilidad de que estén relacionadas en contenido. Dos documentos vinculando a un tercero producen una elección social (*social choice*), es decir, dos páginas están relacionadas mutuamente sin estar enlazadas entre si. Finalmente aval

transitivo, ocurre cuando la página p_g vincula a p_u, la cual a su vez enlaza a p_v. Transitivamente, p_g es considerada como aval de p_v aunque este sea un aval débil. Y por consiguiente estas estructuras pueden combinarse entre si formando patrones muy complejos.

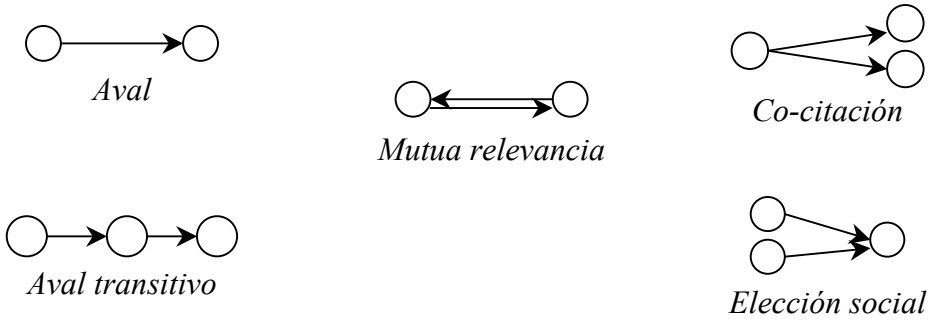
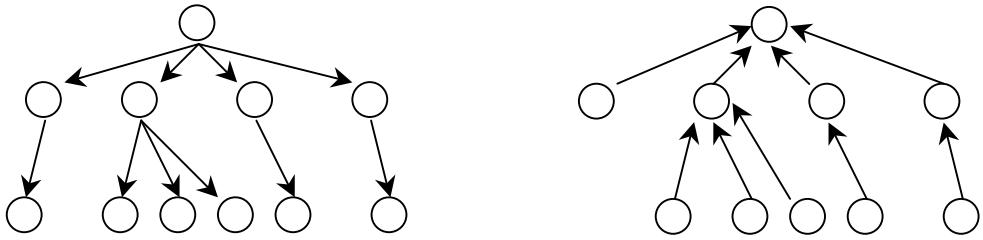


Figura 3 - Patrones básicos formados por dos enlaces

De las múltiples combinaciones posibles dos estructuras arbóreas con grandes grados de entrada y salida han sido de particular interés para la comunidad científica: elección social y co-citación.



La estructura arbórea de entrada nos indica que si muchas páginas diferentes enlazan directamente o transitivamente a un nodo, es probable que este sea un recurso valioso (autoridad) sobre algún tema o de interés compartido por las otras páginas. Esto es análogo a la medición del factor de impacto en la literatura científica por el número de citaciones que recibe una publicación. El interés en la

estructura arbórea de salida está dado por la suposición de que si un nodo vincula a muchas páginas web valiosas en un tópico (*hub*), luego podemos considerarlo como una buena fuente para buscar información relevante.
Podemos decir que autoridades son aquellas páginas prominentes debido a que muchas otras páginas le citan y *hubs* son prominentes por referenciar muchas buenas autoridades. Las autoridades y los *hubs* tienen una mutua relación de aval (Kleinberg, 1998). Por transición, un documento que apunta muchas autoridades "buenas" es un mejor *hub,* mientras que si una página web es referenciada por muchos *hubs* "buenos" es bastante mejor autoridad.

PageRank

Google emplea un método de ordenación por relevancia desarrollado en base a la estructura arbórea de entrada denominado *PageRank* (PR). Los robots de Google recorren continuamente la Red recogiendo nuevas páginas y actualizando las viejas. Estas páginas son comprimidas y almacenadas en un repositorio. La estructura de enlaces de estas páginas es almacenada separadamente de la otra información y es representada como el grafo de la Web. Posteriormente, este grafo es utilizado para calcular el ranking de las páginas (Brin y Page, 1998):

$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

donde se asume que la página A tiene $T1 \ldots Tn$ citaciones, el parámetro d es el factor de *damping* que puede tener un valor entre 0 y 1, y $C(A)$ es el número de enlaces que apuntan a la página A.

De la ecuación planteada se puede decir que una página puede tener un alto PR si hay muchas páginas cuyos enlaces le apuntan y/o si las páginas que la citan tienen un alto PR.

Páginas Similares por co-citación

En este método la entrada del proceso de búsqueda no son los términos de una consulta, sino el URL de una página y la salida es un conjunto de páginas web

relacionadas. Una página relacionada es aquella que contiene el mismo tópico que la página original, pero no son necesariamente idénticas semánticamente.

El algoritmo más implementado es el de co-citación (también lo podemos encontrar como opción - *what's related* - en el navegador Communicator de Netscape). Dean y Henzinger (1999) implementaron esta técnica con buenos resultados. El método usado se basó en encontrar páginas que vinculasen al URL modelo y después determinar las páginas enlazadas por él, obteniendo como respuesta una lista ordenada por autoridad.

3.1.5. Problema de especificidad: búsqueda por ostención y expansión

La recuperación de información sobre la World Wide Web se dificulta por una serie de problemas, uno de ellos es el de especificidad. Si la pregunta es demasiado general, la respuesta es una avalancha de resultados. Se recuperan tantos registros que es pragmáticamente imposible de cerner a través de ellos. Sin embargo, si la pregunta es demasiado específica, la repuesta es nula. Ningún registro es recuperado. Esto es tan inútil como recoger cientos de miles de documentos irrelevantes. Una pregunta muy general, a diferencia de una consulta muy precisa, puede ser redefinida en una más específica. Así quizás reducir la lista a una cantidad manipulable de resultados. No obstante, el proceso de refinamiento puede ser bastante engorroso. Asimismo, pasa a menudo que se refine tanto la pregunta que no devuelva nada. En el caso de preguntas demasiado específicas se prueba otra variación ó se empieza de nuevo.

Con el crecimiento de la colección el problema de especificidad incrementa. Más preguntas pasan al estado de demasiado general. Para oponerse a este problema, se debe usar preguntas más específicas, dando lugar a respuestas vacías.

Como disyuntiva a esta problemática fueron diseñadas dos técnicas de recuperación: búsqueda por ostención y búsqueda por expansión

La búsqueda por ostención o retroalimentación (*query by example*) es una técnica de recuperación implementada recientemente en los motores de búsqueda como alternativa a la redefinición de la expresión de búsqueda.

Los usuarios frecuentemente expresan su necesidad de información con no más de dos palabras (Silverstein y otros, 1998; Jansen, Spink y Saracevic, 2000), lo cual hace difícil para el buscador identificar documentos relevantes. Al mismo tiempo, hay probabilidad de que muchos documentos relevantes no sean recuperados por no contener los términos exactos de la consulta.

Como respuesta a una pregunta realizada en el modo convencional el sistema brinda una lista de aciertos. El usuario, luego de determinar el más representativo puede utilizarlo como documento modelo para buscar más documentos relacionados. El documento muestra provee al algoritmo información acerca de la frecuencia de palabras dentro de la página web seleccionada y otros patrones o frases comunes que pueden ser útiles para mejorar la respuesta.

La búsqueda por expansión (*query expansion*) se basa en la idea de que los términos de la consulta del usuario son indicativos del contenido de los documentos en los cuales está interesado, entonces el sistema a través de adicionar sinónimos o términos relacionados (por co-ocurrencia) "expande" la ecuación de búsqueda representándola en un mejor contexto. Tradicionalmente la expansión se lleva a cabo mediante el uso de tesauros, vocabularios controlados o métodos lingüísticos complejos. No obstante, en el ámbito web se reduce a una mera lista de términos (sin normalización) que el buscador identifica como próximos o semirelacionados con la estrategia de búsqueda del usuario.

3.1.6. Algoritmo de ordenación

La ordenación por relevancia consiste en la ordenación o clasificación de los resultados acorde a una cierta criteria. La criteria puede incluir la frecuencia de los términos, su proximidad, su exactitud, su localización dentro del documento, la longitud del texto, el número de términos equiparados, la popularidad de los enlaces, etc.. La "formula exacta" de como esta criteria es aplicada es el algoritmo de rankeo. Además, las técnicas de ordenación difieren entre los buscadores y metabuscadores.

Según Zhang y Dong (2000) un algoritmo de ordenación racional de documentos web debe ser multidimensional y, al menos, debe incluir las siguientes métricas:

Relevancia: la métrica relevancia significa la distancia entre el contenido del documento web y la consulta del usuario. Ésta es la métrica usada en la mayoría de los motores de búsqueda.

Autoridad: la métrica autoridad significa cuantos documentos web *(b, c ... n)* refieren al documento web a. Supuestamente las páginas son referenciadas por la calidad de su contenido.

Integridad: la métrica integridad significa cuantos documentos web *(b, c ... n)* son apuntados por el documento web a.

Novedad: la métrica novedad significa el grado en el cual un documento web es diferente del resto (por ejemplo, provee información nueva)

Subjetividad: el valor de un documento depende no solo de la consulta, sino también de la nación, edad, cultura, profesión, etc. del usuario.

3.2. El documento

El espacio de la Web, representado por la información accesible vía HTTP, está constituido por una amalgama de documentos de múltiples tipos. Documentos digitales completos codificados en una gran variedad de formatos. Los textos en formatos ASCII (HTML, XHTML, XML, TXT), *Portable Document Format* (PDF), *PostScrip* (PS), Microsoft Word (DOC); las imágenes en formatos *Portable Network Graphic* (PNG*), Joint Photographic experts Group* (JPG), *Graphics Interchange Format* (GIF), BMP; las imágenes animadas Quicktime, MOV, AVI, MPEG; archivos de sonido MP3, Real Player (RAM), WAV son algunos de los tantos utilizados.

Además, las páginas web son compuestas, altamente dinámicas y volátiles (Koehler, 1999)[16], de moderado tamaño[17] y de muy baja estructuración interna. Combinan representaciones gráficas, textuales y sonoras cada vez más interrelacionadas con la evolución de los lenguajes de marcado. El dinamismo

[16] La OCLC determinó una volatilidad, en base a una muestra de direcciones IP tomada en 1998, del 44% en 1999 y del 65% en el año 2000 respectivamente.

[17] En febrero de 1999 el tamaño medio de una página era de 3,9 kb ó 0,98 kb después de reducirla solamente al contenido textual.

antedicho se refiere a los continuos cambios de contenido de muchos de los documentos. La volatilidad, a los cambios de destino de un mismo documento o su desaparición.

Esta mixtura de información textual, audible e icónica en formatos heterogéneos lejos de ser una ventaja se ha convertido en una enorme barrera para la recuperación de información.

La generalidad de los buscadores no indiza la diversidad de formatos de archivos aludidos aunque ofrecen algunas opciones indirectas para localizarlos por su nombre y extensión (por ejemplo, haciendo uso del comando *link en* AltaVista), por el texto alternativo de la etiqueta (comando *anchor*), pero no por su contenido. A tal fin están surgiendo nuevas herramientas de búsqueda. Search PDF[18] creado por la empresa Adobe permite buscar archivos PDF y DOC, CiteSeer ofrece la posibilidad de recuperar trabajos científicos PS y PDF relacionados con las ciencias de la información e ingeniería informática (Lawrence, Bollacker y Giles, 1999), los sistemas comerciales Ereo[19] y LookThatUp[20] permiten buscar imágenes por su contenido (color, textura, forma, tamaño, etc.).

Los motores de búsqueda son la antítesis de las bases de datos tradicionales en lo que respecta al cuerpo documental con el cual trabajan. Manipulan cientos de miles o millones de páginas mal estructuradas, tanto en su formato como en su validación[21] y sin descripción de contenido. Estas deficiencias relegan todo el trabajo de procesamiento e indización al robot, con los consiguientes errores en el tratamiento seudodocumental (ocasionado por malas interpretaciones) del código semántico y las falencias propias del sistema.

La descripción de contenido con metadatos[22] continua siendo un método muy

[18] URL: http://searchpdf.adobe.com
[19] URL: http://www.ereo.com/
[20] URL: http://lookthatup.com/
[21] El formato está asociado a las especificaciones del lenguaje respecto a las reglas sintácticas (no está sujeto a los elementos fijados en un *Document Type Definition*) y la validación sigue una estructura y semántica determinada por un DTD.
[22] Los metadatos o datos representacionales son definidos como datos sobre los datos. Es un conjunto de elementos que poseen una semántica comúnmente aceptada, destinados a ordenar y describir la información contenida en documentos electrónicos (*Document Like Object* - DLO). Se distinguen principalmente dos tipos de metadatos que pueden ser incorporados en un documento web: HTTP-EQUIV y META NAME. El primer tipo indica atributos para navegadores y buscadores

inconsistente, esencialmente por cuatro motivos: 1) los sistemas de recuperación no siempre reconocen o utilizan para rankear sus resultados esta descripción (Sullivan, 2000), 2) empleo inadecuado (intencionado ó impremeditado)[23] 3) su uso no es una práctica habitual, tan solo el 34,2% de los servidores contienen metadatos (etiquetas de descripción y palabras clave) en su página inicial 4) no hay estandarización en su empleo, a comienzos de 1999 fueron encontrados 123 tipos de etiquetas META distintas.

3.3. El usuario

Los estudios de sistemas de recuperación web mayormente han sido enfocados a los aspectos técnicos del sistema, determinando sus ventajas y desventajas; pero han ignorado un factor clave al momento de estudiar a la herramienta de búsqueda: la cualificación del usuario.

Henninger y Belkin (1996) señalan que un sistema de recuperación documental además de ser capaz de proveer una eficiente recuperación, debe también ayudar al usuario cuando este describa un problema que no entiende bien (por ejemplo, ofrecer un buen lenguaje de interrogación y un modelo de diálogo reiterativo).

La popularización de Internet ha producido un crecimiento exponencial en el número de usuarios, que ha dado como resultado más personas sin (o al menos con poco) entrenamiento en la búsqueda de información. Los usuarios emplean las herramientas de búsqueda no por recreación sino con el fin de encontrar la información que desean.

La mayoría de los motores de búsqueda ofrecen al usuario la posibilidad de entrar una rica combinación de palabras clave y/o frases combinadas con operadores booleanos como AND, OR y NOT, y algunas veces otros operadores de proximidad como NEAR u operadores de truncamiento y de distancia. A su vez la adición de paréntesis y filtrado por idioma, fecha, etc. debiera permitirle al usuario

(cookies, actualización cache, fecha de expiración, etc.) y el segundo describe el contenido del documento (autor, fecha publicación, descripción, palabras clave, etc.)

[23] La tarea de la asignación de metadatos al no ser realizada por profesionales de la información, se le ha distorsionado su utilidad con una excesiva sobrecarga de datos (redundancia de información), o repetición de términos (*spamming*) para provocar una mejor posición del documento en la lista de resultados, aunque muchos servicios penalizan esta técnica.

hacer una consulta relativamente precisa e inequívoca. Cabe aquí preguntarnos ¿por qué las opciones de búsqueda simple y avanzada son empleadas en forma incorrecta o no son utilizadas?. Hay al menos tres razones:

- Probablemente la generalidad de los usuarios no posee conocimientos matemáticos o lógicos. Expresiones booleanas como *((batman OR superman) AND marvel)) AND NOT mafalda* son demasiado complicadas para que sean entendidas por los usuarios. Este tipo de comportamiento es confirmado por la investigación hecha por Koeneman y Belkin (1996), en la cual han mostrado que los usuarios sin entrenamiento en la formulación de estrategias de búsqueda tienen dificultades en hacer uso de los operadores disponibles (Koeneman y Belkin, 1996). Hearst (1997) dice que muchos usuarios encuentran a los operadores booleanos confusos, intimidantes o simplemente inútiles.
- Una palabra clave puede tener múltiples significados y un tópico específico puede ser representado por diferentes palabras clave. Los usuarios pueden saber que están buscando, pero carecen del conocimiento y/o preparación para expresar su necesidad de información como enunciado de búsqueda (Henninger y Belkin, 1996). Un problema inherente es que las personas acostumbran usar una cantidad diversa de términos o expresiones para referirse al mismo objeto lingüístico, tal es, que la probabilidad de escoger el mismo término para describir un objeto familiar es menor al 15% (por ejemplo, "accidente" puede ser expresado como "situación desafortunada", "incidente", "percance", "desgracia", "contratiempo", etc.). Esta dificultad presente en la interacción hombre-computadora denominada problema de vocabulario (*vocabulary problem*), ha sido estudiada extensamente por Furnas y otros autores (Furnas , 1987).
- Los seres humanos somos "perezosos", es decir, realizamos el menor esfuerzo posible. Con frecuencia tomamos la salida más fácil. En este caso significa que ingresamos una palabra y ejecutamos inmediatamente la búsqueda. La situación planteada, por supuesto, producirá una avalancha de resultados

mayormente irrelevantes que conlleva a más frustración. Otras actitudes perjudiciales son no reformular la consulta y no ver más allá de la primera página de resultados.

4. Evaluación de herramientas de búsqueda

4.1 Método

La presente evaluación de herramientas de recuperación de información se realiza por medio de un estudio exploratorio preexperimental de tipo transeccional descriptivo.

Universo

Se tomó como universo los motores de búsquedas de mayor cobertura, y los metabuscadores que intercalan y ordenan los resultados por relevancia.

Muestra

La muestra está conformada por 5 (cinco) motores de búsqueda y 3 (tres) metabuscadores de acceso público y gratuito.

Definición de términos

Sistema de recuperación de información: herramienta de recuperación de información web de acceso público y gratuito vía HTTP. Comprende motores de búsqueda, metabuscadores y directorios temáticos.
Motor de búsqueda: herramienta de recuperación de información que indexa automáticamente documentos de la Web y almacena parte de su contenido en una base de datos para su posterior consulta.
Metabuscador: herramienta de recuperación de información constituida únicamente por la interfaz de interrogación y el algoritmo de rankeo. La consulta efectuada al sistema es reenviada simultáneamente a múltiples servicios de búsqueda.

Directorio temático: lista clasificada de recursos web agrupados en categorías temáticas de manera jerárquica y se caracteriza por recopilar, organizar y clasificar manualmente recursos de información completos compuestos por una o más páginas web.

Precisión: Es la relación entre el número de documentos relevantes y el número de documentos recuperados del sistema

Exhaustividad: Es la relación entre el número de documentos relevantes recuperados y el total de documentos relevantes obtenidos por los sistemas.

Cobertura: Es la relación entre el total de documentos obtenidos y el total de documentos recogidos por todos los SRI. Los URLs duplicados no son contados

Ordenamiento por relevancia: Capacidad del motor de búsqueda o metabuscador para ordenar correctamente los resultados de una pesquisa de acuerdo a la criteria que conforma su algoritmo de rankeo.

Variables

- Sistema de recuperación de información web
- Ordenamiento por relevancia
- Capacidad de recuperación de documentos relevantes

4.1.1. Preexperimentos

En el trabajo se efectúan dos preexperimentos. El primero analiza, emulando la conducta del usuario, la ordenación por relevancia (ranking) de los primeros diez y veinte resultados de cinco motores de búsqueda de mayor cobertura (AltaVista[24], Excite[25], Fast[26], Google[27] y Northern Light[28]) en respuesta a ecuaciones de búsqueda no estructuradas. Y el segundo, testea las medidas tradicionales de precisión y exhaustividad, determina la proporción de URLs solapados y similitud

[24] URL: http://www.altavista.com
[25] URL: http://www.excite.com
[26] URL: http://www.alltheweb.com
[27] URL: http://www.google.com
[28] URL: http://www.northernlight.com

entre los cinco buscadores mencionados y en tres metabuscadores (C4[29], Ixquick[30] y MetaCrawler[31]) usando para su interrogación palabras poco frecuentes.

El método utilizado para evaluar la relevancia de los resultados recogidos por los SRI se basó en la similitud consulta-documento. En el primer preexperimento se efectuó una valoración objetiva del contenido de la página web puesto que los sistemas de recuperación no "entienden" el contexto en el cual está interesado el usuario, es decir la presencia o ausencia de los términos de la consulta determinó o no su relevancia. En el segundo preexperimento se hizo una leve valoración subjetiva a modo ver como influye en el rendimiento del sistema. Además de tener en cuenta la presencia o ausencia de las ocurrencias se concedió mayor peso a un contexto adecuado pese a la falencia de los sistemas actuales para reconocer el significado de las palabras.

4.2. Ordenación por relevancia

Los usuarios juzgan los resultados de una forma muy diferente a como se determina la precisión de un sistema de búsqueda. Suelen observar la lista de documentos recuperados y acceder sólo a registros selectos. Un número diverso de factores (objetivos y subjetivos) deciden si recuperarán o no un documento; pero un factor clave es el número de términos equiparados, es decir, los usuarios tienen un sentido intuitivo de como la ordenación por relevancia funciona y el indicador principal de esa satisfacción intuitiva es el número de palabras distintas que contiene un documento respecto a las ocurrencias de la expresión de búsqueda (Koll, 1993).

El usuario, previo a seleccionar un registro, ojea la lista de resultados al efecto de detectar las palabras de la búsqueda. Los registros que presenten los términos en el título o en el sumario serán aparentemente escogidos. También puede verificar la exactitud y proximidad de las ocurrencias en el documento valiéndose de la opción de búsqueda que brindan los navegadores.

[29] URL: http://www.c4.com
[30] URL: http://www.ixquick.com
[31] URL: http://www.metacrawler.com

Para el primer preexperimento fueron elegidos arbitrariamente 5 motores de búsqueda, posicionados entre los de mayor cobertura (Sullivan, oct. 2000): en orden alfabético AltaVista, Excite, Fast, Google y Northern Light. Un conjunto de 10 consultas, 4 en idioma español y 6 en idioma inglés, constituidas por frases[32] de dos y tres palabras no estructuras se sometieron a los buscadores. Los temas de las búsquedas y las consultas han sido inventadas al azar, y sin tener en cuenta los defectos o virtudes de las herramientas a analizar; una sola pregunta fue tomada de un estudio previo.
Las expresiones de búsqueda fueron las siguientes:

1. quantity theory of money
2. death penalty
3. hypertext system
4. matriz insumo producto
5. abuso sexual en niños
6. olympic games sydney
7. effects of nuclear war
8. produccion vitivinicola argentina
9. primeros auxilios
10. citation analysis

A diferencia de gran parte de los estudios precedentes, donde se emplearon consultas estructuras usando operadores booleanos, limitadores +/-, comillas para indicar una frase y a veces algunas estructuras simples con el objeto de obtener el mejor resultado posible; en este test hemos querido representar fielmente el comportamiento del usuario y no usar ningún operador o modificador, dejando hacer la mayor tarea al algoritmo de rankeo[33]. Por ende también las interrogaciones fueron introducidas en los buscadores sin modificar las opciones

[32] En Recuperación de Información una frase es cualquier construcción compuesta no una frase gramatical.
[33] Los usuarios no siempre expresan su demanda de información de manera clara y no aprovechan al máximo las prestaciones de los servicios de búsqueda. Señalamos aquí un proceso inverso: la adaptación del buscador a la consulta del usuario.

por defecto. Optamos por estas alternativas apoyándonos en estudios de usuarios precedentes.

Silverstein y otros (1998) indicaron, por medio del análisis de cerca de un billón de consultas efectuadas en AltaVista, que el 79,6% de los usuarios no usó operadores y/o limitadores en sus búsquedas de un promedio 2,35 palabras[34]; Jansen, Spink y Saracevic (2000) examinaron el comportamiento de 18113 usuarios usando Excite a través de 51473 consultas y mostraron que poco más del 10% de las búsquedas usaban operadores lógicos y que los limitadores +/- y comillas fueron usadas meramente en el 9% de las consultas.

Sander-Beuermann y Schomburg (1998) señalaron en su estudio que el 95% de los usuarios no cambia las opciones de búsqueda por defecto.

Al momento de efectuar las consultas las opciones por defecto eran (Sullivan, oct. 2000; Notess, oct. 2000):

Búsqueda por frase: AltaVista (si la detecta como tal)
Búsqueda por unión: Excite y AltaVista (sino detecta frase)
Búsqueda por intersección: Google, Fast y Northern Light
Truncación automática plurales/singulares: Northern Light
Sensible a mayúsculas y minúsculas: AltaVista

Captura de registros

Cuando se comparan buscadores es importante que la recolección de referencias de las consultas se haga lo más rápido posible para reducir la posibilidad de cambio en el estado de sus bases de datos. Por ello, las búsquedas se efectuaron entre el 2 y 10 de octubre de 2000[35] y, para una misma consulta se realizaron en el mismo día con una diferencia temporal no mayor a 1 minuto. Múltiples ventanas (una por cada motor) fueron abiertas a fin de reducir al máximo el tiempo mencionado.

[34] Otros datos interesantes obtenidos en la investigación muestran que en el 85% de las consultas no se vio más allá de la primera página de resultados y en el 77% de las sesiones se realizó una única búsqueda.
[35] Ver Anexo I.

Para cada búsqueda fueron grabados y descargados los primeros 20 registros (2 páginas). Cabe aclarar que en el caso de Northern Light no se tuvieron en cuenta los registros de su colección especial[36].

La descarga de las páginas se realizó casi automáticamente usando el programa Teleport. En él se introdujeron las listas de aciertos de URLs de los servicios de búsqueda y se obtuvieron las primeras páginas de cada sitio. Las páginas no descargadas por este mecanismo fueron chequeadas manualmente para comprobar con certeza el fallo de la conexión. Esta tarea llevó un período de 2 a 3 días.

Una vez comprobados los 1000 URLs se procedió a procesarlos. El procesamiento se efectuó basándonos en la metodología implementada por Courtois y Berry (1999). A diferencia de ese trabajo se amplió la criteria de 3 a 6 preguntas, se varió algunos condicionantes en las respuestas y se elaboró un análisis más profundo de los datos obtenidos.

Para aplicar la criteria de evaluación fue inspeccionado manualmente el código de marcado de cada documento.

Criteria de evaluación

La criteria elaborada incluyó las siguientes 6 preguntas:

1. Presencia de alguno de los términos

¿Contiene el documento alguna las ocurrencias de la expresión de búsqueda?

2. Presencia de todos los términos

¿Contiene el documento todas las ocurrencias de la expresión de búsqueda?

3. Proximidad

¿Hay al menos una ocurrencia de todos los términos de búsqueda que aparezca en una frase contigua?

4. Localización

[36] La colección especial es integrada por documentos de acceso restringido.

¿Aparecen las ocurrencias de la expresión de búsqueda en el título, etiquetas de encabezamiento H1-H6 o metaetiquetas?

5. Exactitud
¿Aparece la expresión de búsqueda de manera exacta en el documento?

6. Metaetiquetas
¿Están presentes todas las palabras de la consulta en las metaetiquetas palabras clave (*keywords*) y descripción (*description*)?

La respuesta por sí/no de cada pregunta estuvo delimitada por algunos condicionantes. En el caso de las preguntas no. 1, 2, 3, 4 y 6 los plurales[37] de las palabras fueron aceptados sin embargo no otras variaciones de sufijos o variaciones de prefijos.
Para que la pregunta no. 4 recibiera un resultado positivo todos los términos tenían que estar presentes en el título, encabezamientos o metaetiquetas. Por ejemplo, si uno de los términos de una búsqueda de tres ocurrencias estaba presente en el título y los otros dos en las etiquetas de encabezamiento recibía un "sí". Si dos de los términos estaban presentes en el título y el tercero en el cuerpo del documento recibía un "no". Las palabras en las etiquetas META, para los tests de Localización y Metaetiquetas, de los documentos recuperados por Google, Fast y Northern Light no fueron computadas, puesto que estos buscadores no le otorgan ningún peso adicional o no las indizan. En Excite únicamente se tomaron en cuenta aquellas ubicadas en metaetiqueta de descripción (Sullivan, oct. 2000)[38].
Las palabras vacías (sin valor semántico) han sido obviadas para la pregunta no. 5 en Google, Excite y AltaVista, dado que estas palabras no son indexadas. Por ejemplo, si para la búsqueda de "abuso sexual en niños" localizamos un documento con la frase "abuso sexual en los niños" recibía un "sí" en Google, Excite y AltaVista pero un "no" en Fast y Northern Light.

[37] Se aceptó esta variación de sufijo por no alterar la connotación semántica del término.
[38] Excite. (op. cit.) p. 44.

A todas las referencias que no pudieron ser descargadas por fallos de conexión (Error 404, acceso prohibido, Error 603, etc.) se les otorgó un "no" en todos los incisos de la criteria; reflejando así la actitud que tomaría un usuario al estar en contacto con un ítem en estado inactivo. Aunque la referencia sea pertinente la inactividad del enlace induce a su descarte.

Primeramente, para cada búsqueda se identificó la posición de ordenación del ultimo registro con respuesta afirmativa (#a) a la pregunta de la criteria a verificar, tanto para los primeros 10 y 20 resultados. A partir de esa ubicación se determinaron todas las respuestas negativas hasta la posición #1. Y luego se calculó el porcentaje de ítems que dieron una respuesta negativa a la pregunta a testear de la siguiente manera:

$$Prn = (Rrn / Up) \; x \; 100$$

donde:
Prn: Porcentaje de registros con repuesta negativa
Rrn: Número de resultados con respuesta negativa desde #1 hasta #a
Up: Ubicación de #a

Por ejemplo - en los primeros 10 aciertos -, si para la pregunta no. 3 (Proximidad) Fast devuelve el registro #7 como último registro con respuesta positiva a la pregunta y entre la posición #1 y #7 localizamos 5 referencias que no la satisfacen (no presentan dos ocurrencias como una frase contigua), el porcentaje se calcula como 5/7 (71,4%).
Un porcentaje bajo indica un mejor rankeo, es decir, que pocas instancias donde un registro que no cumple con la pregunta de la criteria es mejor ubicado que un ítem que sí la satisface. Entonces una puntuación de 0% señala una ordenación perfecta, todas las referencias que cumplen con la criteria son rankeadas más alto que aquellas que no.
Consiguientemente, para llegar a conocer un poco más o al menos aproximarnos a una apreciación más acertada del funcionamiento del algoritmo de rankeo de los diferentes buscadores, calculamos el porcentaje de URLs (sin tener en cuenta su ubicación) que respondieron negativamente las preguntas de la criteria. Este dato

es imprescindible para determinar a que inciso de la criteria se le otorga mayor peso al momento de presentar y ordenar los resultados[39]. Cuanto menor sea el porcentaje mayor será la cantidad de registros pertinentes a la pregunta evaluada.

4.2.1. Resultados

Una vez evaluados los URLs recogidos en las 50 consultas con la criteria[40] se calculó el valor medio para cada uno de los interrogantes verificados (tabla 1 y 2).

Tabla 1

Valor medio de los ítems con respuesta negativa mejor rankeados respecto a la posición del último acierto. Primeros 10 resultados

Motores de búsqueda	Primeros 10 resultados					
	Presencia de alguno de los términos	Presencia de todos los términos	Proximidad	Localización	Exactitud	Meta - etiquetas
AltaVista	21%	34%	37%	52%	44%	88%
Excite	9%	25%	32%	49%	32%	75%
Fast	11%	11%	24%	22%	46%	n/a
Google	0%	3%	7%	21%	27%	n/a
Northern Light	1%	4%	4%	29%	27%	n/a

Tabla 2

Valor medio de los ítems con respuesta negativa mejor rankeados respecto a la posición del último acierto. Primeros 20 resultados

Motores de búsqueda	Primeros 20 resultados					
	Presencia de alguno de los términos	Presencia de todos los términos	Proximidad	Localización	Exactitud	Meta - etiquetas
AltaVista	20%	35%	42%	56%	46%	89%
Excite	7%	33%	38%	53%	35%	86%
Fast	14%	16%	28%	25%	49%	n/a
Google	2%	7%	10%	36%	28%	n/a
Northern Light	3%	8%	8%	26%	29%	n/a

[39] Es necesario aclarar que los factores intervinientes en el ordenamiento por relevancia evaluados en este test (los considerados por el usuario) no son los únicos existentes, otros tales como la frecuencia de los términos en la longitud del documento y en la base de datos que son parte del algoritmo no son contemplados por ser factores más difíciles que el usuario valore.
[40] Ver Anexo II.

a) Presencia de alguno de los términos

Google produjo un promedio perfecto para los 10 primeros aciertos y el mejor (2%) para los 20 aciertos. Northern Light también realizó una buena tarea ubicándose en el segundo puesto para las 10 y 20 referencias con 1% y 3% respectivamente. Excite fue el único servicio que mejoró ligeramente su puntuación de 9% en los 10 aciertos a 7% en los 20. AltaVista con un pobre 21% y 20% quedó relegado al último puesto en ambas cantidades de URLs. Otra observación que podemos realizar es que el ranking de relevancia es generalmente bueno, considerando el tamaño de las bases de datos y el número de consultas efectuadas al sistema. Mientras es común encontrar documentos que no contienen alguno de los términos de la búsqueda[41] mejor ubicados que documentos que contienen alguna de las ocurrencias su cantidad no es elevada.

b) Presencia de todos los términos

Google (3% y 7%) y Northern Light (4% y 8%) se mantuvieron, en este test, en los primeros puestos para los 10 y 20 registros, su puntuación indica que solamente un ítem que no incluyó todos lo términos de la expresión de búsqueda se ubicó más alto en el ranking que registros que presentan todos los elementos. Fast que también usa el operador lógico AND por defecto dio una puntuación mucho peor (11% y 16%), sugiriendo que esta prestación del operador AND en la búsqueda simple no es tan efectiva como en Google y Northern Light. Excite (25% y 33%) y AltaVista (34% y 35%) dieron los peores promedios que atribuimos al uso de OR como operador por defecto.

c) Proximidad

Northern Light tuvo el mejor rendimiento para este test con 4% y 8% para los 10 y 20 aciertos correspondientemente. Esto indicaría que la Proximidad es un elemento con alto peso en el algoritmo de ordenación de Northern Light. Google rindió bien este test con un 7% para los 10 ítems y 10% para los 20. A pesar que Fast obtuvo un desempeño relativamente bueno en el test de Presencia de Todos

[41] Podemos asignar dicho comportamiento a la presencia de vínculos inactivos, páginas actualizadas en su contenido pero no en la base de datos del motor y fallas propias del sistema.

los Términos tuvo una mala puntuación para Proximidad en los 10 (24%) y 20 (28%) aciertos. Contrariamente con los esperado AltaVista defraudó en el funcionamiento de su búsqueda automática por frase adjudicándose los peores porcentajes (37% y 42%) del test.

d) Localización

La puntuación para esta prueba que varió de 21% a 52% en las 10 primeras referencias y de 26% a 56% en las 20 fue la peor de todas, lo cual insinuaría que la Localización no es un componente de peso en la mayoría de los algoritmos de ordenación. Se puede observar que Google, Northern Light y Fast tuvieron mejor promedio que Excite y AltaVista a pesar de no indizar metaetiquetas, lo que hace pensar que el peso otorgado a la información de los metadatos continua siendo muy bajo. Google mostró la mejor puntuación con un 21% en los 10 ítems y Fast con 25% en los 20.

e) Exactitud

Los valores obtenidos fueron muy altos. Excite fue el único motor en lograr mantener el mismo porcentaje que en Proximidad en los 10 aciertos e incluso mejorar en los 20. El resto de los buscadores empeoraron notablemente su ordenación de resultados respecto a la lograda en Proximidad. Los mejores promedios de esta prueba obtenidos por Google y Northern Light indicarían que dos o tres ítems no satisfactorios se ubicaron más arriba que el último acierto satisfactorio en los primeros 10 resultados y cinco o seis en los 20.

f) Metaetiquetas

En esta verificación pudimos evaluar exclusivamente a Excite y AltaVista. Ambos servicios tuvieron un rendimiento malo. Excite logró un mejor funcionamiento que AltaVista pese a indizar únicamente el campo de metaetiqueta descripción (*description*). Los valores altos obtenidos se deben en parte a la poca utilización de esta descripción de datos y al bajo peso otorgado al rankear los resultados.

Los motores de búsqueda dieron los mejores promedios cuando se testearon los primeros 10 aciertos en comparación con los primeros 20, o sea que el ranking fue

consistentemente más fiable en la primer decena de resultados. En pocos casos las diferencias fueron muy amplias; por ejemplo, la búsqueda "matriz insumo producto" produjo una puntuación de 16,7% para el test de Localización en AltaVista en los 10 primeros documentos y de 87,5% en los 20 aciertos. Sin embargo, para la mayoría de las consultas la puntuación varió entre:

Presencia de Alguno de los Términos: 0% y 5%
Presencia de Todos de los Términos: 0% y 6,3%
Proximidad: 0% y 8%
Exactitud: 0% y 5%

Excite y AltaVista fueron los únicos servicios en mejorar su funcionamiento en el test de Presencia de Alguno de los Términos en los 20 registros. Otras mejoras aunque leves en búsquedas aisladas podemos observar, como por ejemplo, en Fast para la búsqueda "hypertext system" en el test Localización, y las consultas "matriz insumo producto" y "olympic games sydney" en Google para el mismo test.

A partir de los datos analizados se elaboró un ranking con los motores de búsqueda (tabla 3 y 4).

Tabla 3

Ranking de motores de búsqueda para los primeros 10 ítems

Primeros 10 resultados				
Presencia de alguno de los términos	Presencia de todos los términos	Proximidad	Localización	Exactitud
Google	Google	Northern Light	Google	Google / Northern Light
Northern Light	Northern Light	Google	Fast	Excite
Excite	Fast	Fast	Northern Light	AltaVista
Fast	Excite	Excite	Excite	Fast
AltaVista	AltaVista	AltaVista	AltaVista	-

Tabla 4

Ranking de motores de búsqueda para los primeros 20 ítems

Primeros 20 resultados				
Presencia de alguno de los términos	Presencia de todos los términos	Proximidad	Localización	Exactitud
Google	Google	Northern Light	Fast	Google
Northern Light	Northern Light	Google	Northern Light	Northern Light
Excite	Fast	Fast	Google	Excite
Fast	Excite	Excite	Excite	AltaVista
AltaVista	AltaVista	AltaVista	AltaVista	Fast

El ranking no sufrió grandes variaciones entre los 10 y 20 aciertos. Se podría señalar que los algoritmos mantienen mayormente sus pesos de ordenación en las primeras páginas de resultados. Google y Northern Light lideraron casi todos los tests por encima del resto de los buscadores.

URLs precisos recuperados

Luego examinamos la capacidad de los sistemas de recuperación para devolver registros precisos técnicamente[42]. Los promedios de URLs no precisos técnicamente son exhibidos en la tabla 5 y 6. Estos porcentajes indican la proporción de URLs no precisos que cada motor de búsqueda recogió respecto al total de registros recuperados en las diferentes búsquedas.

Al momento de optar por un buscador no sólo es necesario conocer su capacidad para ubicar los registros precisos antes de los no precisos, sino también su habilidad para devolver exclusivamente registros pertinentes a la frase de consulta, por ejemplo, si un SRI recoge tan solo unas pocas referencias satisfactorias y logra ubicarlas al tope de la lista de aciertos consigue una puntuación perfecta en el test de ordenación, por el contrario si otro SRI recupera muchas referencias precisas pero no logra ubicarlas antes de los aciertos no precisos su puntuación decrece bruscamente, lo que ocasiona una valoración

[42] Un registro es preciso técnicamente si responde afirmativamente uno de los incisos de la criteria.

imparcial de esta ultima herramienta sin antes conocer su proporción de URLs precisos con respecto al total obtenido.

Tabla 5

Valor medio de los resultados con respuesta negativa respecto al total de los registros recuperados. Primeros 10 resultados

Motores de búsqueda	Primeros 10 resultados					
	Presencia de alguno de los términos	Presencia de todos los términos	Proximidad	Localización	Exactitud	Meta - etiquetas
AltaVista	22%	41%	46%	65%	53%	94%
Excite	11%	40%	40%	60%	55%	90%
Fast	16%	19%	30%	35%	53%	n/a
Google	0%	4%	9%	29%	26%	n/a
Northern Light	1%	5%	5%	44%	31%	n/a

Tabla 6

Valor medio de los resultados con respuesta negativa respecto al total de los registros recuperados. Primeros 20 resultados

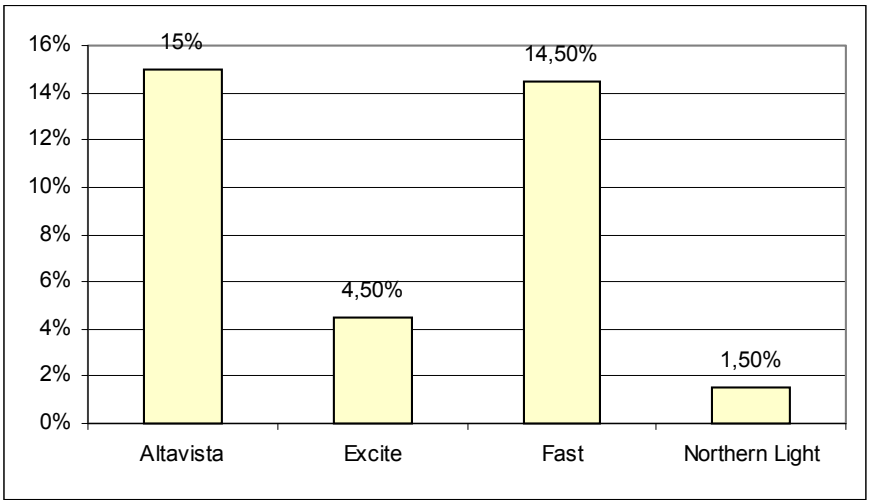
Motores de búsqueda	Primeros 20 resultados					
	Presencia de alguno de los términos	Presencia de todos los términos	Proximidad	Localización	Exactitud	Meta - etiquetas
AltaVista	21%	50%	52%	69%	60%	95%
Excite	11%	48%	51%	68%	63%	92%
Fast	18%	21%	34%	41%	56%	n/a
Google	2%	7%	13%	41%	31%	n/a
Northern Light	5%	9%	9%	50%	40%	n/a

Como se puede apreciar en la tabla 5 y 6 casi todos los SRI mantuvieron la misma prioridad de pesos en la presentación de aciertos e incluso no hubo grandes variaciones con la ordenación de ítems evaluada, lo que reforzaría la idea de una aproximación certera al funcionamiento real del algoritmo en cada motor (en lo referente a los componentes tenidos en cuenta al momento de presentar los resultados) y similitud - en rasgos generales - entre ellos. Primero se ubicó Presencia de Alguno de los Términos, luego Presencia de Todos los Términos

seguida por Proximidad, después Exactitud y finalmente Localización y Metaetiquetas (sólo para Excite y AltaVista).

Cuatro de los cinco buscadores recuperaron referencias que no pudieron ser descargadas, por ejemplo por Error 404 y casos donde el documento había sido movido o eliminado. La excepción a este inconveniente fue Google dado que por medio de su opción Cached se puede tener acceso a documentos que no existen más en la Web (o al menos en el URL indexado), con lo cual superó dicha dificultad logrando una recuperación perfecta. Los porcentajes para vínculos inactivos se muestran en el gráfico 1[43].

Gráfico 1. Porcentaje de vínculos inactivos



AltaVista, junto a Fast, es el que mayor porcentaje de vínculos inactivos presenta. Esto indicaría que la actualización de su base de datos no es están frecuente como sería deseable. La importancia de este indicador está dada en su impacto en el algoritmo de rankeo. Un documento que contenga todos las ocurrencias de la búsqueda puede ser ubicado al inicio de la lista de resultados, no obstante si este documento no está disponible para su consulta es de utilidad nula para el usuario.

[43] Ver Anexo III.

Después de examinar estos datos podemos atribuir (en parte) el rendimiento medio de Fast y bajo de AltaVista (además del uso del operador OR por defecto) a su alta tasa de enlaces inactivos, en cambio a Excite a su denominada búsqueda "conceptual" que permite recuperar tanto frases relacionadas como términos exactos.

4.3. Recuperación de documentos relevantes

Cuando se aplican las medidas de evaluación - precisión y exhaustividad - de los sistemas de recuperación tradicionales a los motores de búsqueda web surgen diversos problemas. Uno de los principales problemas es la cantidad de documentos, tanto el número total de la base de datos como el número de aciertos localizados para una búsqueda determinada. Otra dificultad es el acelerado crecimiento, dinámica e inestabilidad de la base de datos de los buscadores para procesar y manipular el mare mágnum de información (Bar-Ilan, 1999; Rousseau, 1999; Selberg y Etzioni, 2000).

Al no ser factible evaluar la relevancia de todos los documentos tampoco la exhaustividad puede ser computada. Y a su vez, no siempre todos los documentos de la lista de resultados pueden ser normalmente evaluados (puede variar entre cientos y millones)[44]. Para evitar estas dificultades, la lista de aciertos es cortada en cierto punto (en el que el usuario cesa de examinarla) que Blair (1996) denomina grado de persistencia y la precisión es calculada hasta el punto de corte. Los registros ubicados en este umbral son considerados como el número total de documentos relevantes obtenidos.

La exhaustividad no es computada en la mayoría de los estudios y en aquellos donde ha sido calculada es una pobre estimación, usando los documentos relevantes encontrados en la limitada lista de resultados.

En este preexperimento tratamos de superar esta barrera de otra manera. Recurrimos a palabras poco usadas en la literatura como términos de búsqueda, esperando con ello una baja frecuencia de aparición en la base de datos y

consecuentemente una lista reducida de aciertos. Esto hace posible evaluar la relevancia de todos los documentos recogidos y calcular, no la precisión y exhaustividad real del sistema pero sí una aproximación más cercana con datos más estables. Además, permite determinar otras características de su funcionamiento; en este caso haremos hincapié en la proporción de páginas web solapadas, cobertura relativa y similitud entre las herramientas de búsqueda.

A efectos de realizar el segundo estudio a los buscados AltaVista, Excite, Fast, Google y Northern Light se les adicionó tres metabuscadores, C4, Ixquick y MetaCrawler. Su elección, al igual que la efectuada para los motores, fue arbitraria.

MetaCrawler, uno de los servicios pioneros de búsquedas múltiples, cubre 15 SRI (About, AltaVista, DirectHit[45], Excite, FindWhat[46], Google, GoTo[47], Infoseek[48], Kanoodle[49], LookSmart[50], Lycos[51], Internet Keywords[52], Sprinks[53], Thunderstone[54] y Webcrawler[55]), C4 13 (AltaVista, Excite, GoTo, Google, HotBot[56], FindWhat, Infoseek, Snap[57], Yahoo[58], Lycos, WebCrawler, Magellan y Kanoodle) e Ixquick, que se autodenomina el más potente del mundo (aunque está muy lejos de serlo), 14 (AltaVista, Brújula, Chévere[59], Directorio Abierto, EuroSeek[60], Excite, Hispavista[61], Iguana[62], Lycos, México Web[63], Sol[64], Xasa[65], Yahoo y Yupi[66])

---

[44] Ver, por ejemplo, en el Anexo IV la cantidad de registros recuperados y su relación porcentual con la cantidad de ítems evaluados para el primer preexperimento.
[45] URL: http://www.directhit.com
[46] URL: http://findwhat.com
[47] URL: http://www.goto.com
[48] URL: http://www.infoseek.com
[49] URL: http://www.kanoodle.com
[50] URL: http://www.looksmart.com
[51] URL: http://www.lycos.com
[52] URL: http://www.realnames.com
[53] URL: http://www.sprinks.com
[54] URL: http://thunderstone.go2net.com/texis/websearch/
[55] URL: http://www.webcrawler.com
[56] URL: http://www.hotbot.com
[57] URL: http://www.snap.com
[58] URL: http://www.yahoo.com
[59] URL: http://www.terra.com.ve
[60] URL: http://www.euroseek.com
[61] URL: http://www.hipavista.com
[62] URL: http://www.iguana.com.mx
[63] URL: http://mexico.web.com.mx
[64] URL: http://www.sol.es

respectivamente. Como puede observarse MetaCrawler y C4 tiene 9 SRI en común, lo cual presupondría a priori un cierto grado de similitud en su respuesta, no así en su ranking dado que cada metabuscador tiene su propio algoritmo de recuperación; Ixquick integra una variedad de servicios poco conocidos (a nivel de popularidad y cobertura) que podrían devolver documentos que otros motores y/o directorios grandes y conocidos no pudieron localizar.

Realización de las consultas en los sistemas

En el test se formularon 10 preguntas de un único término a cada uno de los 8 SRI seleccionados, por tanto se efectuaron 80 consultas. Los términos utilizados fueron palabras poco frecuentes con una única acepción para evitar cualquier tipo de distorsión semántica en su uso. Y su elección fue hecha al azar a excepción de tres palabras que fueron tomadas de un trabajo anterior.
Los términos remitidos, 3 en lengua inglesa y 7 en española, fueron los siguientes:

1. grigallo
2. reduvio
3. vomitel
4. guapomo
5. huisquil
6. apodyopsis
7. fluctisonant
8. materteral
9. escrocon
10. galopillo

Las consultas se llevaron a cabo entre el 24 y 27 de octubre de 2000[67]. El tiempo empleado para realizar una misma pregunta entre los SRI no superó los 2 minutos.

---

[65] URL: http://www.xasa.com
[66] URL: http://www.yupi.com
[67] Ver Anexo I.

La descarga y verificación (por fallos de conexión) de los URLs se hizo de la misma manera que en el estudio anterior, y demoró 2 días.

Categorización por relevancia

La validación de relevancia de las búsquedas se hizo cuidadosamente desde el documento a texto completo utilizando una escala constituida por cinco categorías[68]:

Enlaces inactivos: son aquellos que dan alguno de los siguientes fallos: Error 404, Error 603, acceso prohibido, el documento ha sido trasladado o no se encuentra.
Enlaces duplicados: presentan el mismo URL básico. Las sedes *mirror* (mismo contenido distinto URL) no son contadas como duplicados. Esta categoría es independiente de las otras cualidades.
Categoría cero: la página es irrelevante por no contener el término.
Categoría uno: el término de búsqueda está presente.
Categoría dos: el término de búsqueda está presente y además está definido o explicado.

4.3.1. Análisis de datos

Un total de 408 URLs fue el resultado de las búsquedas ejecutadas. La tabla 7 muestra el número de aciertos para cada consulta.

[68] Las categorías se elaboraron en base a las propuestas por Leighton y Srivastava (1997) en su trabajo. Sin estas categorías el juicio de relevancia sería binario.

Tabla 7. Cantidad de URLs recuperados

Sistema de búsqueda	Cantidad de URLs recuperados por búsqueda (incluye vínculos inactivos)									
	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
AltaVista	1	1	8	6	15	0	0	2	1	2
C4	6	0	18	9	38	0	7	10	0	0
Excite	0	0	0	2	2	2	0	5	0	0
Fast	5	2	11	5	23	4	12	7	0	2
Google	4	2	10	3	24	4	4	7	0	1
Ixquick	2	0	7	4	10	0	0	0	0	2
MetaCrawler	4	1	7	7	18	2	4	8	1	1
Northern Light	5	2	10	14	15	1	8	5	2	3
Total	27	8	71	50	145	13	35	44	4	11

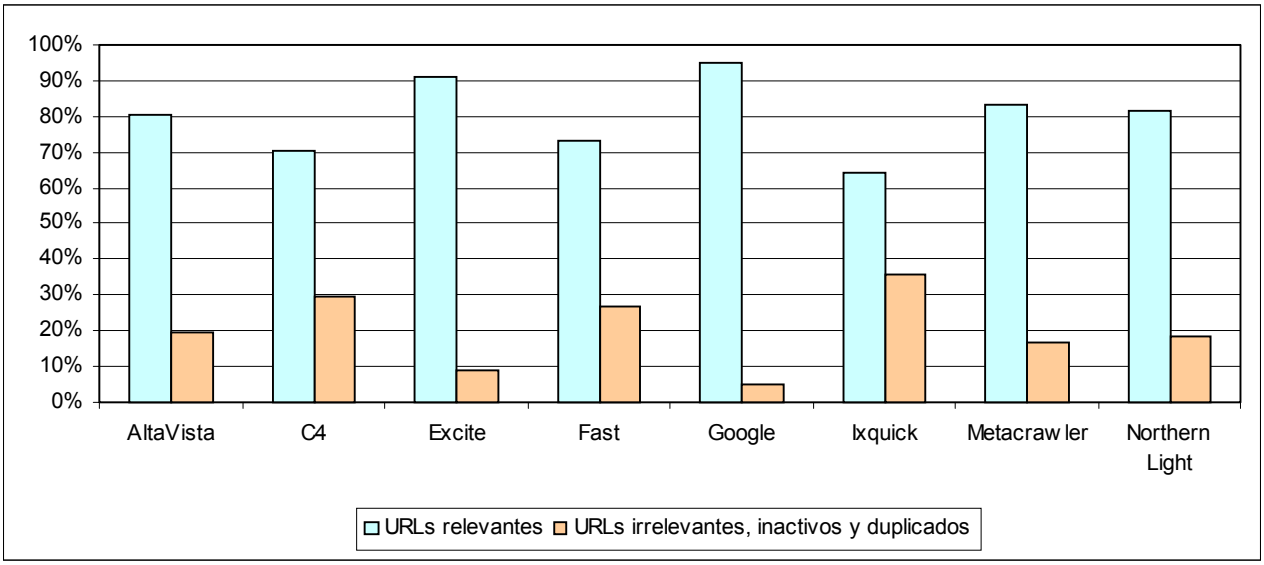
Es evidente, visualizando la tabla 7, que la recuperación de los metabuscadores (al menos los evaluados) está muy lejos de su funcionamiento ideal. ¿Cómo puede ser que C4 y MetaCrawler obtengan menos resultados que la combinación de AltaVista, Excite y Google, e Ixquick de AltaVista y Excite? [69]. Podemos sugerir las siguientes razones. Posiblemente los tiempos de conexión con los diferentes servicios no sean suficientes para recoger todos los resultados (en pos de satisfacer en forma inmediata al usuario). Sobrecarga de información en la red. Fallas del sistema para obtener simultáneamente múltiples ítems en paralelo.

Para cada uno de los resultados obtenidos se determinó su condición: acierto insatisfactorio (enlace inactivo, duplicado o categoría cero) ó acierto satisfactorio (categoría uno o categoría dos). Los vínculos irrelevantes, inactivos o duplicados recibían una puntuación de 0, los relevantes categoría uno percibían un punto y los relevantes categoría dos obtenían dos puntos.

El gráfico 2 exhibe las proporciones de registros satisfactorios e insatisfactorios. La totalidad de los servicios de búsqueda recogió más de 64% de aciertos relevantes. Google (95%) y Excite (91%) obtuvieron el mayor número de registros satisfactorios. Mientras que Ixquick (36%) obtuvo el más alto porcentaje de URLs insatisfactorios seguido por C4 (30%).

[69] Es necesario advertir que esta falencia sólo es visible en este tipo de estudio donde se manipulan reducidas cantidades de aciertos.

Gráfico 2. Porcentajes de documentos satisfactorios e insatisfactorios para las 10 consultas.



La tabla 8 muestra los resultados inactivos y duplicados, y la proporción de URLs que cada servicio de búsqueda recogió en las diez búsquedas ejecutadas.

Tabla 8. Ruido documental

Sistema de Recuperación	Enlaces inactivos		Enlaces duplicados	
	Cantidad	Porcentaje	Cantidad	Porcentaje
AltaVista	7	19,44%	0	0%
C4	10	11,36%	16	18,18%
Excite	1	9,09%	0	0%
Fast	9	12,68%	10	14,08%
Google	0	0%	3	5,08%
Ixquick	7	28%	2	8%
MetaCrawler	5	9,43%	4	7,55%
Northern Light	8	12,31%	4	6,15%

Los buscadores AltaVista (19,44%) y Fast (12,68%), al igual que en el preexperimento precedente, presentan los mayores porcentajes de referencias inactivas[70]. Llama la atención el alto porcentaje de Northern Light (12,31%) en comparación al obtenido previamente, que podríamos atribuir a un desfasaje en la

actualización de la base de datos. C4 (18,18%) y Fast (14,08%) aparecen con los mayores porcentajes de vínculos duplicados. Como señala Olvera Lobo (2000), el usuario, a diferencia del profesional de la información, puede conferir a la repetición de URLs de un mismo sitio una sensación de máxima relevancia en vez de considerarlo como un aspecto negativo del sistema.

En la tabla 9 se presentan las cantidades de URLs (con duplicados) distribuidas en las categorías de relevancia. C4 devolvió el mayor número de referencias relevantes potenciales y óptimas. La categoría cero no obtuvo ningún acierto, o sea, no se recuperaron páginas donde la ocurrencia de búsqueda no estuviera presente. Los gráficos 3-4 muestran la distribución porcentual en las 10 búsquedas.

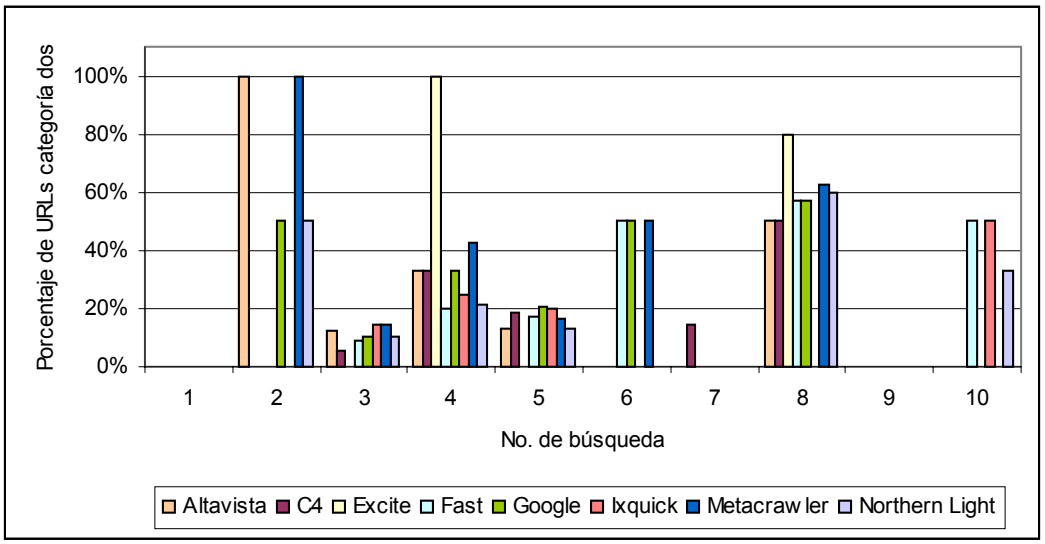
Tabla 9. Categorías de relevancia

Sistema de Recuperación	Escala de Relevancia		
	R0 - No. de URLs	R1 - No. de URLs	R2 - No. de URLs
Altavista	0	21	8
C4	0	61	17
Sistema de Recuperación	Escala de Relevancia		
	R0 - No. de URLs	R1 - No. de URLs	R2 - No. de URLs
Excite	0	7	3
Fast	0	49	13
Google	0	45	14
Ixquick	0	13	5
MetaCrawler	0	34	14
Northern Light	0	46	11

R0: Categoría cero (documento no relevante)
R1: Categoría uno (documento relevante potencial)
R2: Categoría dos (documento relevante óptimo)

[70] No señalamos el alto porcentaje de Ixquick (28%) dado que la recuperación de enlaces muertos no es una falla atribuible al metabuscador.

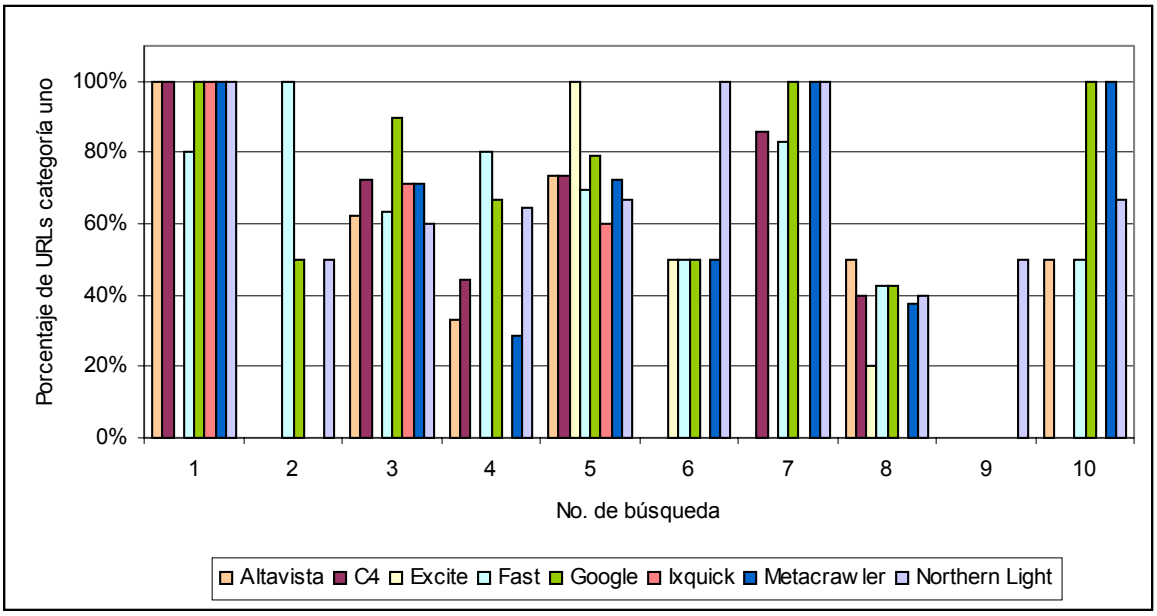
Gráfico 3. Distribución porcentual de URLs relevantes óptimos.



Como se puede ver los resultados óptimos son claramente insatisfactorios en todos los casos, la mediana de documentos es inferior a 16%.

Para la categoría de documentos relevantes potenciales, la más representativa, la media grupal de aciertos es de 42% con una desviación estándar de 20%. Northern Light (70%, desv. típ. 22%) muestra el mayor porcentaje de ítems seguido por Google (68%, desv. típ. 33%), Fast (62%, desv. típ. 28%) y MetaCrawler (56%, desv. típ. 39%). El alto grado de dispersión de los datos con respecto al valor medio es debido a la poca cantidad de registros relevantes computados de acuerdo con la categorización establecida.

Gráfico 4. Distribución porcentual de URLs relevantes potenciales.



En la tabla 10 se presentan algunas estadísticas para los 8 servicios de búsqueda. La cantidad de aciertos encontrados, la cantidad de aciertos no solapados, el número total de registros relevantes (con duplicados) y el número de veces que el sistema de recuperación no halló URLs.

Tabla 10.

Sistema de Recuperación	TDEa	TDEb	TDR	TCC
AltaVista	36	0	29	2
C4	88	10	78	4
Excite	11	2	10	6
Fast	71	16	62	1
Google	59	0	59	1
Ixquick	25	3	18	5
MetaCrawler	53	0	48	0
Northern Light	65	22	57	0

TDEa: Total de documentos encontrados (sin tener en cuenta su categoría). Los URLs localizados por más de un SRI son contados.
TDEb: Total de documentos encontrados (sin tener en cuenta su categoría). Los URLs localizados por más de un SRI no son contados.
TDR: Total de documentos relevantes encontrados (categoría uno y dos con duplicados)
TCC: Total de consultas con cero aciertos

Como era de esperar, un metabuscador (C4), recolectó la mayor cantidad de aciertos, seguido extrañamente por dos motores de búsqueda (Fast y Northern Light) en vez del servicio de búsquedas múltiples Ixquick o MetaCrawler por el número de SRI que abarcan. El bajo número de ítems localizados por Excite e Ixquick se debe a su elevado número de consultas vacías.

En general, las herramientas tuvieron un buen rendimiento en la devolución de registros pertinentes a la consulta. Ixquick fue el único sistema en no sobrepasar el 80% de registros relevantes recuperados.

Northern Light (22) junto a Fast (16) presentan el mayor número de registros únicos (no solapados), debido, en cierta manera a que no son cubiertos por ningún metabuscador analizado.

A modo de exponer el relativamente bajo solapamiento existente entre los motores de búsqueda - confirmando los hallazgos de Lawrence y Giles (1998, 1999) -, se muestra en la tabla 11 el número de URLs solapados entre los motores de búsqueda y en la tabla 12 la relación de solapamiento entre ellos.

La proporción de solapamiento se cuantificó en base a la siguiente fórmula:

$$S(a,b) = (N_{ab} / N_a) \; x \; 100$$

donde:
S : Solapamiento
N_{ab} : Número de aciertos comunes entre el servicio de búsqueda a y b
N_a : Número de aciertos del servicio de búsqueda a.

Los metabuscadores han sido incluidos a fin de visualizar más claramente la recogida incompleta de documentos. En otras palabras, la proporción de solapamiento de un servicio de búsqueda con un metabuscador que lo comprende idealmente tiene que ser del 100%. Aquí podemos ver que la tasa de solapamiento varió de 54% a 85%.

Tabla 11. URLs solapados

Sistema de Recuperación	AltaVista	C4	Excite	Fast	Google	Ixquick	Meta-Crawler	Northern Light
AltaVista	-	30	3	18	20	15	29	18
C4	30	-	7	27	50	20	46	28
Excite	3	7	-	4	5	0	6	4
Fast	18	27	4	-	38	13	30	24
Google	20	50	5	38	-	13	41	23
Ixquick	15	20	0	13	13	-	14	11
MetaCrawler	29	46	6	30	41	14	-	24
Northern Light	18	28	4	24	23	11	24	-

Tabla 12. Relación de solapamiento

SRI - A (fuente)	SRI - B (bajo testeo)	Solapamiento de A en B
AltaVista	C4	83%
	Excite	8%
	Fast	50%
	Google	56%
	Ixquick	42%
	MetaCrawler	80%
	Northern Light	50%
C4	AltaVista	34%
	Excite	64%
	Fast	31%
	Google	57%
	Ixquick	23%
	MetaCrawler	52%
	Northern Light	32%
Excite	AltaVista	27%
	C4	64%
	Fast	36%
	Google	45%
	Ixquick	0%
	MetaCrawler	54%
	Northern Light	36%
Fast	AltaVista	25%
	C4	38%
	Excite	6%
	Google	54%
	Ixquick	18%
	MetaCrawler	42%
	Northern Light	34%
Google	AltaVista	34%
	C4	85%

SRI - A (fuente)	SRI - B (bajo testeo)	Solapamiento de A en B
Google	Excite	8%
	Fast	64%
	Ixquick	22%
	MetaCrawler	69%
	Northern Light	39%
Ixquick	AltaVista	60%
	C4	80%
	Excite	0%
	Fast	52%
	Google	52%
	MetaCrawler	56%
	Northern Light	44%
MetaCrawler	AltaVista	55%
	C4	87%
	Excite	11%
	Fast	56%
	Google	77%
	Ixquick	26%
	Northern Light	45%
Northern Light	AltaVista	28%
	C4	43%
	Excite	6%
	Fast	37%
	Google	35%
	Ixquick	17%
	MetaCrawler	37%

Corroborando a un más la baja tasa de solapamiento antedicha, hallamos que cada URL fue localizado una media de 1,8 veces en los 5 motores y, si le adicionamos los resultados recogidos de los metabuscadores cada URL es encontrado una media de 2,7 veces con una desviación estándar de 1,9. Cerca de la mitad de los URLs fue hallado una sola ocasión y cinco de los sistemas encontraron al menos 2 registros que ninguno de los otros descubrió (tabla 10). Ningún URL fue devuelto por los 8 servicios de búsqueda, y tan sólo 4 URLs (7,5%) fueron ubicados por 7 SRI.

Todo indica que para encontrar la mayor cantidad de registros como sea posible uno debe buscar en diferentes sistemas de recuperación, pero el uso de un metabuscador no es una opción muy confiable como conjunción efectiva de varios motores.

Aplicadas las categorías de relevancia (tabla 9) se les realizaron a los ítems una serie de pruebas. De esta manera se comparó los SRI usando diferentes niveles de relevancia.

La primera prueba mide la precisión y exhaustividad para documentos relevantes asignando 1 a los URLs categoría uno y dos. Este test muestra la capacidad del sistema para entregar referencias que satisfagan la consulta (ver tabla 13).

La segunda prueba mide la precisión y exhaustividad otorgando 1 a los URLs con grado de relevancia dos. Aquí se muestra la capacidad de cada servicio para recuperar enlaces relevantes óptimos a la necesidad de información (ver tabla 14).

La tercera prueba mide la precisión y exhaustividad concediendo 1 a los URLs con relevancia uno. Esto refleja la capacidad de cada SRI para obtener URLs que mínimamente satisfagan la búsqueda (ver tabla 15). Finalmente, la cuarta prueba mide la precisión y exhaustividad para documentos relevantes sin URLs duplicados. Los servicios de búsqueda son penalizados por introducir ruido a la lista de aciertos (ver tabla 16).

Ninguna de las cuatro pruebas computa los enlaces inactivos.

A fin de delinear el rendimiento global de cada SRI en las distintas pruebas calculamos la precisión y exhaustividad como promedio de las 10 consultas[71].

Tabla 13. Documentos relevantes.

SRI	Precisión	Exhaustividad
AltaVista	0,678	0,088
C4	0,538	0,135
Excite	0,350	0,027
Fast	0,823	0,185
Google	0,900	0,157
Ixquick	0,341	0,038
MetaCrawler	0,946	0,147
Northern Light	0,936	0,224

Tabla 14. Documentos relevantes óptimos.

SRI	Precisión	Exhaustividad
AltaVista	0,209	0,064
C4	0,122	0,180

[71] Ver Anexo V.

SRI	Precisión	Exhaustividad
Excite	0,180	0,028
Fast	0,204	0,125
Google	0,221	0,121
Ixquick	0,109	0,062
MetaCrawler	0,286	0,109
Northern Light	0,188	0,111

Tabla 15. Documentos relevantes potenciales.

SRI	Precisión	Exhaustividad
AltaVista	0,469	0,081
C4	0,416	0,135
Excite	0,170	0,022
Fast	0,619	0,206
Google	0,679	0,161
Ixquick	0,231	0,023
MetaCrawler	0,660	0,133
Northern Light	0,748	0,239

Tabla 16. Documentos relevantes sin duplicados.

SRI	Precisión	Exhaustividad
AltaVista	0,678	0,091
C4	0,466	0,128
Excite	0,350	0,029
Fast	0,762	0,198
Google	0,877	0,167
Ixquick	0,341	0,040
MetaCrawler	0,897	0,156
Northern Light	0,877	0,235

Las diferentes pruebas muestran como varía la puntuación dependiendo del grado de relevancia. La prueba 1 obtuvo una regular precisión media global de 0,69 con una desviación típica de 0,25, y una exhaustividad de 0,13 con una desviación típica de 0,07. Aquí, el mejor rendimiento lo consiguió Northern Light y MetaCrawler. En la prueba 2, la más exigente, la precisión media global decreció bruscamente a 0,19 (desv. típ. 0,06), entretanto la exhaustividad se mantuvo en un 0,1 (desv. típ. 0,05). MetaCrawler, Google y Fast fueron los más eficientes en el test. Para la prueba 3 donde se valoran los URLs relevantes potenciales la media global incrementó a 0,5 (desv. típ. 0,22) y la exhaustividad a 0,13 (desv. típ. 0,08).

Los servicios de Northern Light y Google produjeron los resultados más elevados. Anulando las referencias duplicadas en la prueba 4, la precisión media general decae levemente con relación a la prueba 1 a 0,65 (desv. típ. 0,24) y la exhaustividad no varia demasiado. Northern Light y MetaCrawler consiguieron, aquí, el mejor rendimiento.

Un análisis de varianza no mostró una diferencia estadísticamente significativa entre las precisiones a nivel 0,05 en la prueba 2, en tanto que en la prueba 1 (F=5,2), prueba 3 (F=3,94) y prueba 4 (F= 4,8) proporcionaron una diferencia estadísticamente significativa entre los SRI comparado con el valor crítico ($F_{(0,05;7,72)}$=2,14). El análisis de la exhaustividad de igual manera presentó una diferencia significativa en la prueba 1 (F=6,4), prueba 3 (F=7) y prueba 4 (F=7,1) [72]. Señalando que el grado de relevancia tiene un efecto significativo en el rendimiento del sistema.

Los sistemas más precisos han sido MetaCrawler (0,7), Northern Light (0,69) y Google (0,67) y los más exhaustivos Northern Light (0,19), Fast (0,17) y Google (0,15). En todas las pruebas, Excite e Ixquick, son los SRI con peor performance dada a su alta cantidad de búsquedas con cero aciertos.

Cobertura relativa

Para computar la cobertura relativa de cada sistema no se tuvo en cuenta las referencias muertas y duplicadas, y se determinó como promedio de las diez preguntas.

La tabla 17 expone que ninguno de los SRI cubrió más del 21% de las páginas. Nuevamente se puede observar la pobre cobertura de los metabuscadores. Los valores bajos corresponden a sistemas con alta tasa de vínculos inactivos, duplicados o que hallan obtenido un elevado número de consultas con cero aciertos ó la combinación de estas fallas.

[72] Ver Anexo VI.

Tabla 17. Cobertura relativa

SRI	Cobertura
AltaVista	0,100
C4	0,190
Excite	0,034
Fast	0,179
Google	0,148
Ixquick	0,053
MetaCrawler	0,143
Northern Light	0,214

Similitud entre los servicios de búsqueda

Con el objeto de determinar similitudes (o diferencias) entre pares de sistemas de recuperación basándonos en la cantidad de registros encontrados aplicamos el coeficiente de Jaccard. Este índice varía entre 0 (no existe similitud entre las referencias halladas) y 1 (similitud exacta entre ellas).

$$J = N_{ab} / (N_a + N_b - N_{ab})$$

donde:
J: Coeficiente de Jaccard
$N_{ab:}$ URLs presentes en el SRI a y b
$N_{a:}$ URLs encontrados por a
$N_{b:}$ URLs encontrados por b

La tabla 18 muestra las principales tres similitudes entre motores de búsqueda y la tabla 19 las tres disimilitudes. Los metabuscadores fueron obviados por la recogida incompleta de URLs mencionada anteriormente, evitando de esta manera brindar un distorsionado J[73].

Tabla 18. Similitud entre pares de motores de búsqueda

Pares de SRI	Jaccard
Fast - Google	0,413
AltaVista - Google	0,266
Google - Northern Light	0,227

[73] El Anexo VII presenta la lista completa de J.

Tabla 19. Disimilitud entre pares de motores de búsqueda

Pares de SRI	Jaccard
Excite - Fast	0,051
Excite - Northern Light	0,055
Excite - Google	0,076

En casi todos los casos los pares de buscadores presentaron coeficientes de Jaccard bastante bajos, oscilando entre 0,41 y 0,051. Esto indica una gran disimilitud entre los sistemas; lo cual es muy útil al momento de tener que seleccionar herramientas disímiles para obtener una amplia cobertura de la Web.
Se observa, en la tabla 18, que el 41% de los URLs devueltos por Fast y Google fueron encontrados por ambos motores. Altavista y Google tuvieron en común 30% de registros y Google y Northern Light un 23%.
La gran disimilitud que presenta Excite corresponde en parte al reducido número de URLs que recogió.

Asimismo, pudimos comprobar con cierto grado fiabilidad, a causa de las inconsistencias de los sistemas, la premisa expresada al comienzo de este test acerca de una mayor similitud entre los metabuscadores MetaCrawler y C4 que con Ixquick. C4 y MetaCrawler tuvieron un grado de similitud del 48% e Ixquick de 22% con cada uno.

5. Conclusión

La World Wide Web es un espacio anárquico, fuera de control y en crecimiento continuo. Lejos de ser la ansiada biblioteca digital mundial se ha convertido en un caótico repositorio de información heterogénea y desestructurada; en el cual miles de servicios de búsqueda tratan de brindar un medio confiable para su acceso. Dos de las herramientas más empleadas, y analizadas en este estudio son los motores de búsqueda y metabuscadores. Ambos sistemas poseen mecanismos de búsqueda (simples o potentes) que mayormente no se adecuan a las necesidades (o mejor dicho a la conducta) de usuarios inexpertos. A fin de superar

esa barrera y mejorar la relevancia de los resultados se van implementando nuevos métodos de recuperación que aprovechan la información del contexto para incrementar la eficiencia del sistema.

De los dos preexperimentos realizados no podemos obtener sólidas conclusiones aplicables a todos los SRI existentes en respuesta a todo tipo de consulta[74], pero sí señalar ciertas observaciones generales en su funcionamiento.

De la primera evaluación se desprende que los motores seleccionados - en respuesta a consultas constituidas por frases de dos y tres palabras - ordenan por relevancia de una manera más confiable los diez primeros resultados que los primeros veinte. Sin embargo, en pocos casos los veinte primeros aciertos produjeron mejores puntuaciones que la primer decena. Según la criteria elaborada, al momento de rankear las referencias los algoritmos tienen especialmente en cuenta: Presencia de Alguno de los Términos, Presencia de Todos los Términos, Proximidad, luego Exactitud y finalmente Localización y Metaetiquetas. Google y Northern Light tuvieron el más alto rendimiento en casi todos los incisos de la criteria, tanto en los diez como en los veinte primeros URLs, por posicionar en los primeros puestos del ranking los registros satisfactorios y recuperar la mayor cantidad de documentos útiles.

Podemos sugerir que los usuarios hallarán mejor ubicadas las referencias apropiadas que las inapropiadas en las primeras páginas de resultados.

Cuando evaluamos, en el segundo preexperimento, la pertinencia de los resultados comprobamos como varía ampliamente su relevancia de acuerdo a diferentes categorizaciones subjetivas, y afecta a las medidas de precisión y exhaustividad. Al aplicar la categoría de relevancia exigente, los resultados ofrecidos ponen de manifiesto la poca precisión y exhaustividad de los sistemas de recuperación web. El SRI más preciso en esa instancia fue MetaCrawler con un pobre 0,286 y el más exhaustivo C4 con 0,18. En cambio, si tenemos en cuenta la valoración objetiva (categoría 1-2 sin duplicados), los servicios de búsqueda

[74] Para ello se requeriría realizar estudios a gran escala (con continua replicación) donde se efectúen miles de consultas simples y estructuradas a cientos de servicios de búsqueda.

obtienen valores de precisión bastante buenos (exceptuando aquellos sistemas con varias consultas de cero acierto), pero valores bajos de exhaustividad.
En el funcionamiento de los metabuscadores quedó expresa la actitud de los propietarios de privilegiar el factor velocidad por encima de la recuperación global, sino no se comprendería una recogida tan incompleta. Esto conlleva a decir que el uso de este tipo de sistemas es poco confiable como cohesión de múltiples servicios.
Los bajos porcentajes de solapamiento indican poca superposición entre zonas de operación de los robots en la red (reflejada en la colección de las bases de datos), y el coeficiente de Jaccard muestra el reducido grado de similitud entre los SRI, lo cual es un buen indicador para no fiarse de la respuesta de una sóla herramienta de búsqueda. Por último, cabe señalar que la cobertura más amplia no superó el 21% de las páginas.

5.1. Recomendaciones futuras

Teniendo en cuenta las características y comportamiento de los motores de búsqueda y metabuscadores sería interesante que a futuro se puedan originar trabajos en torno a este tema para:

- Verificar la estabilidad de los algoritmos de recuperación en base aun mayor número de consultas y aciertos analizados.
- Comparar la eficiencia de los servicios de búsqueda en respuesta a idénticas preguntas (en palabras usadas) planteadas con diferentes niveles de elaboración (estructuras simples contra estructuras complejas).
- Realizar análisis temporales de comportamiento con un mismo conjunto de búsquedas a fin de determinar el rendimiento, crecimiento y posibles inestabilidades en el funcionamiento
- Contrastar la performance de metabuscadores contra los servicios que agrupa.

6. Bibliografía consultada

1. Aigrain, Philippe; Longueville, Véronique. A model for the evaluation of expansion techniques in information retrieval systems. *Journal of the American Society for Information Science*. 45(4), 1994. p. 225-234.

2. Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier, eds. *Modern information retrieval*. New York : ACM Press, 1999. *citado por* Kobayashi, Mei; Takeda, Koichi. Information retrieval on the Web, 2000. Disponible en: <http://www.trl.ibm.co.jp/projects/s7710/dl/trlrep/rt347.ps>. (Consultado 4 octubre 2000)

3. Bar-llan, Judith. Search engine results over time : a case study on search engine stability. *Cybermetrics*. 2/3, issue 1, paper 1, 1999. Disponible en: <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html>. (Consultado 2 septiembre 2000).

4. Beaulieu, Micheline; Robertson, Stephen; Rasmussen, Edie. Evaluating interactive systems in TREC. *Journal of the American Society for Information Science*. 47(1), 1996. p. 85-94.

5. Bharat, Krishna; Broder, Andrei. Mirror, mirror on the Web : a study of host pairs with replicated content, 1999. *Proceedings of Eighth International World Wide Web Conference*, (1999, may 11-14 : Toronto). Disponible en: <http://decweb.ethz.ch/WWW8/data/2147/html/index.htm>. (Consultado 20 abril 2000).

6. Blair, David C. STAIRS redux : thoghts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*. 47(1), 1996. p. 4-22.

7. Brin, Sergey : Page, Lawrence. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems.* 30, 1998. p. 107-117.

8. Broder, Andrei [y otros]. Graph structure in the web: experiments and models, 2000. *Proceedings of the Ninth International World Wide Web Conference*, (2000, may 15-19 : Amsterdan). Disponible en: <http://www9.org/w9cdrom/160/160.html> (Consultado 7 octubre 2000).

9. Carriere, J; Kazman, R. Webquery : searching and visualizing the web through connectivity, 1997. *Proceedings of the Sixth International World Wide Web Conference*, (1997, april 7-11 : Santa Clara). Disponible en: <http://decweb.ethz.ch/WWW6/Technical/Paper096/Paper96.html>. (Consultado 30 febrero 1999).

10. Chakrabarti, Soumen [y otros]. Mining the link structure of the World Wide Web. *IEEE Computer*. 32(8), 1999.

11. Chen, Hsinchun [y otros]. *A concept space approach to addressing the vocabulary problem in scientific information retrieval : an experiment on the Worm Community System*, 1996. Disponible en: <http://ai.bpa.arizona.edu/papers/wcs96/wcs96.html> (Consultado 2 octubre 2000).

12. Chu, Heting T.; Rosenthal, Marilyn. Search engines for the World Wide Web : a comparative study and evaluation methodology, 1996. *Proceedings of the 59th ASIS Annual Meeting*, (1996, oct. 21-24 : Baltimore). Disponible en: <http://www.asis.org/annual-96/ElectronicProceedings/chu.html>. (Consultado 14 febrero 1999).

13. Courtois, Martin P.; Berry, Michael W. Results ranking in web search engines. *Online*. 23(3), 1999, p. 39-46.

14. De Vocht, Josep. *Experiments for the characterization of hypertext structures*, Eindhoven University of Technology, 1994. Disponible en: <http://wwwis.win.tue.nl/~debra/joep/ >. (Consultado 10 marzo 1999).

15. Dean, J., and Henzinger, M. R. Finding related pages in the World Wide Web, 1999. *Proceedings of Eighth International World Wide Web Conference*, (1999, may 11-14 : Toronto). Disponible en: <http://decweb.ethz.ch/WWW8/data/2148/html/index.htm>. (Consultado 20 abril 2000).

16. Efe, Kemal [y otros]. *The shape of the Web and its implicationes for searching the Web*, 2000. Disponible en: <http://www.cacs.usl.edu/Publications/Raghavan/ERCB00.ps.Z>. (Consultado 20 septiembre 2000).

17. Feldman, Susan. NLP meets the jabberwocky: natural language processing in information retrieval. *Online*. May 1999. Disponible en: <http://www.onlineinc.com/onlinemag/OL1999/feldman5.html>. (Consultado 11 julio 2000).

18. Frants, Valery I. [y otros]. Boolean search : current state and perspectives. *Journal of the American Society for Information Science*. 50(1), 1999. p. 86-95.

19. Gauch, Susan; Wang, Guijun. *Information fusion with ProFusion*, 1996. Disponible en: <http://www.ittc.ukans.edu/~sgauch/papers/WebNet96.ps>. (Consultado 28 junio 2000).

20. Gil Leiva, Isidoro. *La automatización de la indización de documentos*. Gijón : TREA, 1999.

21. Graphic, Visualization, and Usability Center. *GVU's tenth WWW user survey report*, 1998. Disponible en: <http://www.gvu.gatech.edu/user_surveys/survey-1998-10/>. (Consultado 15 abril 1999).

22. Gudivada, Ventat N. [y otros]. Information retrieval on the World Wide Web. *IEEE Internet Computing*. 1(5), 1997. p. 58-68.

23. Harman, D. Relevance feedback and others query modification techniques. Information retrieval : data structures and algorithms, New Jersey : Prentice-Hall, 1992.

24. Harter, Stephen P. Psychological relevance and information science. *Journal of the American Society for Information Science*. 43(9), 1992. p. 602-615.

25. Hawking, David [y otros]. Results and challenges in web search evaluation, 1999. *Proceedings of Eighth International World Wide Web Conference*, (1999, may 11-14 : Toronto). Disponible en: <http://decweb.ethz.ch/WWW8/data/2150/ html/index.htm >. (Consultado 20 abril 2000).

26. Hearst, Martin A. Interfaces for searching the Web. *Scientific American*. (3), 1997. Disponible en: <http://www.sciam.com/0397issue/0397hearst.html>. (Consultado 3 marzo 1999).

27. Heinonen, Oskari; Hätönen, Kimmo; Klemettinen, Mika. *WWW robots and search engines*. April 1996. Disponible en: <http://www.cs.helsinki.fi/ ~oheinone/publications/ WWW_Robots_and_ Search _Engines.ps.gz>. (Consultado 3 marzo 1999).

28. Henninger, Scott; Belkin, Nicholas J. Interface issues and interaction strategies for information retrieval systems, 1996. *Proceedings of CHI '96*, (1996, april 13-18 : Vancouver). Disponible en: <http://www.acm.org/sigchi/chi96/proceedings/ tutorial/Henninger/njd_txt.htm>. (Consultado 23 noviembre 1999).

29. Hölscher, Christoph; Strube, Gerhard. Web search behavior of Internet experts and newbies, 2000. *Proceedings of the Ninth International World Wide Web Conference*, (2000, may 15-19 : Amsterdan). Disponible en: <http://www9.org/ w9cdrom/81/81.html>. (Consul-tado 7 octubre 2000).

30. Hou, M. *Comparison of three Internet search tools : Yahoo, AltaVista, Lycos*, 1998. Disponible en: <http://vered.rose.utoronto.ca/people/ming/report.html>. (Consultado 13 junio 2000)

31. Huberman, Bernando A.; Adamic, Lada A. *Evolutionary dynamics of the World Wide Web*, 1999. Disponible en: <http://www.parc.xerox.com/istl/groups/iea/ www/growth.html>. (Consultado 4 marzo 2000).

32. Jansen, Bernard J.; Spink, Amanda; Saracevic, Tefko. Real life, real users, and real needs : a study and analysis of user queries on the web. *Information Processing and Management*. 36(2), 2000. p. 207-227.

33. Kleinberg, Jon M. Authoritative sources in hyperlinked enviroment. *Proceedings of ACM-SIAM Symposium on discrete algorithms*, 1998. p. 668-677.

34. Kobayashi, Mei; Takeda, Koichi. Information retrieval on the Web, 2000. Disponible en: <http://www.trl.ibm.co.jp/projects/s7710/dl/trlrep/rt347.ps>. (Consultado 4 octubre 2000).

35. Koehler, Wallace. An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*. 50(2), 1999. p. 162-180.

36. Koeneman, Jürgen; Belkin, Nicholas J. A case for interaction : study of interactive information retrieval behavior and effectiveness, 1996. *Proceedings of CHI '96*, (1996, april 13-18 : Vancouver). Disponible en: <http://www.acm.org /sigchi/chi96/proceedings/papers/Koeneman/jkl_txt.htm>. (Consultado 23 noviembre 1999).

37. Koll, Matthew B. Automatic relevance ranking : a searcher's complement to indexing. *Proceedings of the 25th Annual Meeting of the American Society of Indexers*, (1993, may 20-22 : Alexandria), 1993. p. 55-60, *citado por* Courtois, Martin P.; Berry, Michael W. Results ranking in web search engines. *Online*. 23(3), 1999, p. 39-46

38. Koster, Martijn. Robots in the Web: threat or treat?. *ConneXions*. 9(4), april 1995. Disponible en: <http://info.webcrawler.com/mak/projects/robots/ threat-or-treat.html>. (Consultado 14 marzo 1999).

39. Landoni, Monica; Bell, Steven. Information retrieval techniques for evaluating search engines : a critical overview. *Aslib Proceedings*. 52(3), 2000. p. 124-129.

40. Lawrence, Steve; Bollacker, Kurt; Giles, C. Lee. Indexing and retrieval of scientific literature, 1999. *Eighth International Conference on Information and Knowledge Management*. (1999, november 2-6 : Kansas City). p. 139-146.

41. Lawrence, Steve; Giles, C. Lee. Accessibility of information on the web. *Nature*. 400, july 1999. p. 107-109.

42. Lawrence, Steve; Giles, C. Lee. Searching the Web : general and scientific information acces. *IEEE Communications*. 37(1), 1999. p. 116-122.

43. Lawrence, Steve; Giles, C. Lee. Searching the World Wide Web. *Science*. 280(5360), 1998. p. 98-100.

44. Leighton, H. Vernon; Srivastava, Jaideep. First 20 precision among World Wide Web search services (search engines). *Journal of the American Society for Information Science*. 50(10), 1999. p. 870-881.

45. Leighton, H. Vernon; Srivastava, Jaideep. *Precision among World Wide Web search services (search engines) : Altavista, excite, HotBot, Infoseek, Lycos*, 1997. Disponible en: <http://www.winona.msus.edu/library/webind2/webind2.htm >. (Consultado 27 febrero 1999)

46. Leighton, H. Vernon. *Performance of four World Wide Web (WWW) index services : Infoseek, Lycos, WebCrawler and WWWorm*, 1995. Disponible en: <http://www.winona.msus.edu/library/webind.htm>. (Consultado 27 febrero 1999).

47. Liddy, Elizabeth D. Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science*. 24(4), 1988. Disponible en: <http://www.asis.org/Bulletin/Apr-98/liddy.html> (Consul- tado 11 julio 2000).

48. Ljosland, Mildridz. *Evaluation of web search engines and the search for better ranking algorithms*, July 1999. Disponible en: <http://www.dei.unipd.it/~ims/sigir99/papers/4-ljosland.ps>. (Consultado 10 abril 2000).

49. Moore, Alvin; Murray, Brian H. *Sizing the Internet*, 2000. Disponible en: <http://www.cyveillance.com >. (Consultado 13 octubre 2000).

50. Moya Anegón, Félix de. *Los sistemas integrados de gestión bibliotecaria : estructuras de datos y recuperación de información*. Madrid : ANABAD, 1995.

51. Nicholson, Scott. Raising of web search tool reseach through replication and chaos theory. *Journal of the American Society for Information Science*. 51(8), 2000. p. 724-729.

52. Notess, Greg R. *Search engine showdown : the users' guide to web searching*. Disponible en: <http://searchengineshowdown.com/>. (Consultado 1999-2000).

53. OCLC. *Web Characterization Project*, 2000. Disponible en: <http://wcp.oclc.org/>. (Consultado 7 octubre 2000).

54. Olvera Lobo, María D. Métodos y técnicas para la indización y recuperación de los recursos de la World Wide Web. *Boletín de la Asociación Andaluza de Bibliotecarios*. (57), 1999. p. 11-22.

55. Olvera Lobo, María D. Rendimiento de los sistemas de recuperación de información en la Web : evaluación de servicios de búsqueda (search engines). *Revista Española de Documentación Científica*. 23(3), 2000. p. 303-317.

56. Olvera Lobo, María D. Rendimiento de los sistemas de recuperación de información en la World Wide Web : revisión metodológica. *Revista Española de Documentación Científica*. 23(1), 2000. p. 63-77.

57. Rousseau, Ronald. Daily time series of common single word searches in AltaVista y Northern Light. *Cybermetrics*. 2/3, issue 1, paper 2, 1999. Disponible en: <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html>. (Consultado 2 septiembre 2000).

58. Sander-Beuermann, Wolfgang; Schomburg, Mario. Internet information retrieval : the further development of meta-search engine technology, 1998. *Proceedings of Internet Society Conference*, (1998, july). Disponible en: <http://www.isoc.org/inet98/proceedings/lc/lc_2.htm>. (Consultado 10 abril 2000).

59. Schlichting, Carsten; Nillsen, Erik. *Signal detection of WWW search engines*. Disponible en: http://www.lclark.edu/~nilsen/ms/searchengine.HTM (Consultado 30 marzo 2000).

60. Schwartz, Candy. Web search engines. *Journal of the American Society for Information Science*. 49(11), 1999. p. 973-982.

61. Selberg, Erik; Etzioni, Oren. Multi-service search and comparison using the MetaCrawler, 1995. *Proceedings Fourth International World Wide Web Conference*, (1995, octuber). Disponible en: <http://w3j.com/1/selberg.169/ paper/169.html>. (Consultado 27 febrero 1999).

62. Selberg, Erik; Etzioni, Oren. *On the instability of web search engines*, 2000. Disponible en: <http:// www.cs.washington.edu/homes/pjallen/papers/riao2.ps>. (Consultado 18 enero 2001).

63. Silverstein, Craig [y otros]. *Analysis of a very large AltaVista query log*, 1998. Disponible en: <ftp://ftp.digital.com/pub/DEC/SRC/technical-notes/SRC-1998-014.ps.gz>. (Consultado 18 agosto 2000).

64. Spinak, Ernesto. *Diccionario enciclopédico de bibliometría, cienciometría e informetría*. Caracas : Unesco, 1996.

65. Su, Louise T. The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*. 45(3), 1994. p. 207-217.

66. Sullivan, Daniel. *How to use meta HTML tags?*. Disponible en: <http://www.searchenginewatch.com/webmasters/meta.html>. Consultado (17 septiembre 2000).

67. Sullivan, Danny, ed. *Search engine watch : tips about Internet search engines & search engine submission*. Disponible en: <http://www.searchenginewatch.com>. (Consultado 1999-2000).

68. van Rijsbergen, C.J. *Information retrieval*. London : Butterworth, 1979.

69. Weinstock, Melvin. Citation indexes. *Encyclopedia of library and information science*. New York : Marcel Dekker. Vol.5, c1971. p. 16-40.

70. Wiley, Deborah L. Beyond information retrieval : ways to provide content in context. *Database.* August 1998. Disponible en: <http://www.onlineinc.com/database/DB1998/wiley8.html>. (Consultado 11 julio 2000).

71. Yu, Clement; Meng, Weiyi. *Search engine*, [1999]. Disponible en: <http://panda.cs.binghamton.edu/~meng/pub.d/se.ps.gz>. (Consultado 9 septiembre 2000)

72. Zhang, Dell; Dong, Yisheng. An efficient algorithm to rank web resources, 2000. *Proceedings of the Ninth International World Wide Web Conference*, (2000, may 15-19 : Amsterdan). Disponible en: <http://www9.org/w9cdrom/251/251.html> (Consultado 7 octubre 2000).

73. Zorn, Peggy [y otros]. Advanced searching : tricks of the trade. *Online*. 21(3), 1996. Disponible en: <http://www.onlineinc.com/onlinemag/MayOL/zorn5.html> (Consultado 29 agosto 2000).

7. Anexos

7.1. Anexo I: cronograma de las búsquedas

Preexperimento 1.

Fecha	Expresión de búsqueda
02/10/2000	quantity theory of money
04/10/2000	death penalty
05/10/2000	hypertext system
05/10/2000	matriz insumo producto
09/10/2000	abuso sexual en niños
	olympic games sydney
	effects of nuclear war
	produccion vitivinicola argentina
10/10/2000	primeros auxilios
	citation analysis

Preexperimento 2.

Fecha	Expresión de búsqueda
24/10/2000	grigallo
	reduvio
	vomitel
	guapomo
	huisquil
26/10/2000	apodyopsis
	fluctisonant
27/10/2000	materteral
	escrocon
	galopillo

7.2. Anexo II: ordenación y ubicación de URLs

PRMRRN: Porcentaje de Resultados Mejor Rankeados con Respuesta Negativa respecto a la posición del último acierto con respuesta positiva.
PRRN: Porcentaje de Resultados con Respuesta Negativa respecto al total de registros recuperados.

AltaVista

Expresión de búsqueda	Primeros diez resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	20%	20%	44,4%	50%	30%	30%
death penalty	20%	20%	20%	20%	20%	20%
hypertext system	0%	0%	50%	50%	85,7%	90%
matriz insumo producto	11,1%	20%	22,2%	30%	22,2%	30%
abuso sexual en niños	10%	10%	20%	60%	0%	50%
olympic games sydney	50%	50%	70%	70%	70%	70%
effects of nuclear war	30%	30%	33,3%	40%	33,3%	40%
produccion vitivinicola argentina	30%	30%	37,5%	50%	60%	80%
primeros auxilios	30%	30%	30%	30%	30%	30%
citation analysis	10%	10%	10%	10%	20%	20%

Expresión de búsqueda	Primeros diez resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	44,5%	50%	44,5%	50%	50%	90%
death penalty	20%	20%	20%	20%	66,7%	80%
hypertext system	85,7%	90%	85,7%	90%	100%	100%
matriz insumo producto	16,7%	80%	22,2%	30%	100%	100%
abuso sexual en niños	20%	60%	20%	60%	100%	100%
olympic games sydney	80%	80%	66,7%	90%	90%	90%
effects of nuclear war	28,6%	50%	33,3%	40%	100%	100%
produccion vitivinicola argentina	100%	100%	100%	100%	100%	100%
primeros auxilios	80%	80%	30%	30%	100%	100%
citation analysis	40%	40%	20%	20%	77,8%	80%

AltaVista

Expresión de búsqueda	Primeros veinte resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	20%	20%	35%	35%	25%	25%
death penalty	15%	15%	20%	20%	20%	20%
hypertext system	15%	15%	56,3%	65%	87,5%	90%
matriz insumo producto	10%	10%	36,4%	60%	50%	55%
abuso sexual en niños	10%	10%	20%	80%	0%	75%
olympic games sydney	30%	30%	75%	80%	75%	80%
effects of nuclear war	21,1%	25%	27,8%	35%	27,8%	35%
produccion vitivinicola argentina	30%	30%	37,5%	75%	81,3%	85%
primeros auxilios	26,3%	30%	26,3%	30%	26,3%	30%
citation analysis	20%	20%	20%	20%	25%	25%

Expresión de búsqueda	Primeros veinte resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	35%	35%	38,9%	45%	50%	95%
death penalty	20%	20%	20%	20%	66,7%	75%
hypertext system	85,7%	95%	85,7%	95%	100%	100%
matriz insumo producto	85,7%	90%	36,4%	60%	100%	100%
abuso sexual en niños	20%	80%	20%	80%	100%	100%
olympic games sydney	81,3%	85%	66,7%	95%	90%	95%
effects of nuclear war	28,6%	75%	44,4%	50%	100%	100%
produccion vitivinicola argentina	100%	100%	100%	100%	100%	100%
primeros auxilios	57,9%	60%	26,3%	30%	100%	100%
citation analysis	50%	50%	25%	25%	84,2%	85%

Excite

Expresión de búsqueda	Primeros diez resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	0%	0%	0%	0%	0%
death penalty	0%	10%	0%	10%	0%	10%
hypertext system	10%	10%	14,3%	40%	75%	90%
matriz insumo producto	10%	10%	0%	70%	12,5%	30%
abuso sexual en niños	0%	0%	75%	80%	40%	40%
olympic games sydney	10%	10%	16,7%	50%	16,7%	50%
effects of nuclear war	10%	10%	20%	20%	20%	20%
produccion vitivinicola argentina	33,3%	40%	85,7%	90%	85,7%	90%
primeros auxilios	0%	0%	0%	0%	0%	0%
citation analysis	20%	20%	40%	40%	70%	70%

Expresión de búsqueda	Primeros diez resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	0%	0%	0%	100%	100%
death penalty	11,1%	20%	0%	10%	25%	60%
hypertext system	83,3%	90%	75%	90%	100%	100%
matriz insumo producto	100%	100%	0%	70%	100%	100%
abuso sexual en niños	87,5%	90%	0%	90%	87,5%	90%
olympic games sydney	25%	70%	16,7%	50%	10%	90%
effects of nuclear war	33,3%	80%	57,1%	70%	100%	100%
produccion vitivinicola argentina	100%	100%	100%	100%	100%	100%
primeros auxilios	0%	0%	0%	0%	57,1%	70%
citation analysis	50%	50%	70%	70%	75%	90%

Excite

Expresión de búsqueda	Primeros veinte resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	0%	0%	0%	0%	0%
death penalty	5%	5%	5,3%	5%	5,3%	5%
hypertext system	5%	5%	35%	35%	84,6%	90%
matriz insumo producto	12,5%	30%	0%	85%	12,5%	65%
abuso sexual en niños	0%	5%	80%	85%	40%	55%
olympic games sydney	5,3%	10%	64,7%	70%	57,9%	60%
effects of nuclear war	10%	10%	26,3%	30%	26,3%	30%
produccion vitivinicola argentina	22,2%	30%	85,7%	95%	87,5%	95%
primeros auxilios	0%	0%	0%	40%	0%	40%
citation analysis	10%	10%	35%	35%	66,7%	70%

Expresión de búsqueda	Primeros veinte resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	0%	0%	0%	100%	100%
death penalty	15%	15%	5%	5%	55%	55%
hypertext system	86,7%	90%	92,3%	90%	100%	100%
matriz insumo producto	100%	100%	0%	85%	100%	100%
abuso sexual en niños	87,5%	95%	0%	95%	87,5%	95%
olympic games sydney	25%	85%	16,7%	75%	75%	95%
effects of nuclear war	72,7%	85%	66,7%	70%	100%	100%
produccion vitivinicola argentina	100%	100%	100%	100%	100%	100%
primeros auxilios	0%	65%	0%	40%	57,1%	85%
citation analysis	42,1%	45%	66,7%	70%	88,2%	90%

Fast

Expresión de búsqueda	Primeros diez resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	22,2%	30%	22,2%	30%	22,2%	30%
death penalty	0%	0%	0%	0%	0%	0%
hypertext system	10%	10%	10%	10%	100%	100%
matriz insumo producto	10%	10%	0%	30%	11,1%	20%
abuso sexual en niños	20%	20%	30%	30%	20%	20%
olympic games sydney	0%	0%	0%	0%	0%	0%
effects of nuclear war	0%	0%	0%	0%	0%	0%
produccion vitivinicola argentina	50%	80%	50%	80%	50%	80%
primeros auxilios	0%	10%	0%	10%	0%	10%
citation analysis	0%	0%	0%	0%	33,3%	40%

Expresión de búsqueda	Primeros diez resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	22,2%	30%	22,2%	30%	n/a	n/a
death penalty	0%	0%	0%	0%	n/a	n/a
hypertext system	10%	10%	100%	100%	n/a	n/a
matriz insumo producto	16,7%	50%	42,9%	60%	n/a	n/a
abuso sexual en niños	30%	30%	100%	100%	n/a	n/a
olympic games sydney	0%	80%	33,3%	60%	n/a	n/a
effects of nuclear war	30%	30%	30%	30%	n/a	n/a
produccion vitivinicola argentina	100%	100%	100%	100%	n/a	n/a
primeros auxilios	0%	10%	0%	10%	n/a	n/a
citation analysis	10%	10%	33,3%	40%	n/a	n/a

Fast

Expresión de búsqueda	Primeros veinte resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	25%	25%	25%	25%	25%	25%
death penalty	0%	0%	0%	0%	0%	0%
hypertext system	10%	10%	10%	10%	88,9%	90%
matriz insumo producto	10%	10%	25%	25%	35%	35%
abuso sexual en niños	30%	30%	35%	35%	30%	30%
olympic games sydney	5%	5%	5%	5%	10%	10%
effects of nuclear war	0%	0%	0%	0%	5%	5%
produccion vitivinicola argentina	50%	90%	50%	90%	50%	90%
primeros auxilios	10%	10%	10,5%	15%	10,5%	15%
citation analysis	0%	0%	0%	0%	27,8%	35%

Expresión de búsqueda	Primeros veinte resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	30%	30%	25%	25%	n/a	n/a
death penalty	0%	0%	0%	0%	n/a	n/a
hypertext system	15%	15%	88,9%	90%	n/a	n/a
matriz insumo producto	16,7%	75%	64,7%	70%	n/a	n/a
abuso sexual en niños	35%	35%	100%	100%	n/a	n/a
olympic games sydney	0%	90%	33,3%	80%	n/a	n/a
effects of nuclear war	36,8%	40%	42,1%	45%	n/a	n/a
produccion vitivinicola argentina	100%	100%	100%	100%	n/a	n/a
primeros auxilios	10,5%	15%	10,5%	15%	n/a	n/a
citation analysis	10%	10%	27,8%	35%	n/a	n/a

Google

Expresión de búsqueda	Primeros diez resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	0%	0%	0%	0%	0%
death penalty	0%	0%	0%	0%	0%	0%
hypertext system	0%	0%	0%	0%	44,4%	50%
matriz insumo producto	0%	0%	0%	0%	0%	0%
abuso sexual en niños	0%	0%	0%	0%	0%	0%
olympic games sydney	0%	0%	25%	40%	25%	40%
effects of nuclear war	0%	0%	0%	0%	0%	0%
produccion vitivinicola argentina	0%	0%	0%	0%	0%	0%
primeros auxilios	0%	0%	0%	0%	0%	0%
citation analysis	0%	0%	0%	0%	0%	0%

Expresión de búsqueda	Primeros diez resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	10%	11,1%	20%	n/a	n/a
death penalty	0%	0%	0%	0%	n/a	n/a
hypertext system	0%	0%	80%	50%	n/a	n/a
matriz insumo producto	33,3%	60%	0%	0%	n/a	n/a
abuso sexual en niños	20%	20%	22,2%	30%	n/a	n/a
olympic games sydney	16,7%	50%	62,5%	70%	n/a	n/a
effects of nuclear war	37,5%	50%	0%	0%	n/a	n/a
produccion vitivinicola argentina	100%	100%	90%	90%	n/a	n/a
primeros auxilios	0%	0%	0%	0%	n/a	n/a
citation analysis	0%	0%	0%	0%	n/a	n/a

Google

Expresión de búsqueda	Primeros veinte resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	0%	0%	0%	0%	0%
death penalty	0%	0%	0%	0%	0%	0%
hypertext system	0%	0%	0%	0%	35%	35%
matriz insumo producto	0%	0%	0%	0%	10%	10%
abuso sexual en niños	0%	0%	0%	0%	0%	0%
olympic games sydney	5%	5%	45%	45%	40%	40%
effects of nuclear war	5%	5%	10%	10%	5%	5%
produccion vitivinicola argentina	10%	10%	10%	10%	14,3%	40%
primeros auxilios	0%	0%	0%	0%	0%	0%
citation analysis	0%	0%	0%	0%	0%	0%

Expresión de búsqueda	Primeros veinte resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	5%	5%	15%	15%	n/a	n/a
death penalty	0%	0%	0%	0%	n/a	n/a
hypertext system	0%	0%	35%	35%	n/a	n/a
matriz insumo producto	66,7%	75%	5%	5%	n/a	n/a
abuso sexual en niños	30,8%	55%	50%	50%	n/a	n/a
olympic games sydney	76,5%	70%	60%	60%	n/a	n/a
effects of nuclear war	42,9%	60%	12,5%	30%	n/a	n/a
produccion vitivinicola argentina	100%	100%	90%	95%	n/a	n/a
primeros auxilios	10%	10%	0%	0%	n/a	n/a
citation analysis	0%	0%	0%	0%	n/a	n/a

Northern Light

Expresión de búsqueda	Primeros diez resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	0%	0%	0%	0%	0%
death penalty	0%	0%	0%	0%	0%	0%
hypertext system	0%	0%	10%	10%	10%	10%
matriz insumo producto	0%	0%	0%	0%	0%	0%
abuso sexual en niños	0%	0%	10%	10%	0%	0%
olympic games sydney	0%	0%	10%	10%	10%	10%
effects of nuclear war	0%	0%	0%	0%	0%	0%
produccion vitivinicola argentina	10%	10%	11,1%	20%	22,2%	30%
primeros auxilios	0%	0%	0%	0%	0%	0%
citation analysis	0%	0%	0%	0%	0%	0%

Expresión de búsqueda	Primeros diez resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	0%	0%	0%	n/a	n/a
death penalty	0%	0%	0%	0%	n/a	n/a
hypertext system	10%	10%	50%	50%	n/a	n/a
matriz insumo producto	0%	90%	30%	30%	n/a	n/a
abuso sexual en niños	55,6%	60%	62,5%	70%	n/a	n/a
olympic games sydney	33,3%	80%	11,1%	20%	n/a	n/a
effects of nuclear war	71,4%	80%	57,1%	70%	n/a	n/a
produccion vitivinicola argentina	100%	100%	57,1%	70%	n/a	n/a
primeros auxilios	10%	10%	0%	0%	n/a	n/a
citation analysis	10%	10%	0%	0%	n/a	n/a

Northern Light

Expresión de búsqueda	Primeros veinte resultados					
	Presencia de alguno de los términos		Presencia de todos los términos		Proximidad	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	0%	5%	0%	5%	0%	5%
death penalty	0%	0%	0%	0%	0%	0%
hypertext system	0%	0%	5%	5%	5%	5%
matriz insumo producto	5%	5%	5%	5%	5%	5%
abuso sexual en niños	5%	5%	15%	15%	5%	5%
olympic games sydney	5%	5%	20%	20%	20%	20%
effects of nuclear war	0%	0%	5%	5%	0%	0%
produccion vitivinicola argentina	10%	10%	20%	20%	30%	30%
primeros auxilios	5,6%	15%	5,6%	15%	5,6%	15%
citation analysis	0%	0%	0%	0%	5%	5%

Expresión de búsqueda	Primeros veinte resultados					
	Localización		Exactitud		Metaetiquetas	
	PRMRRN	PRRN	PRMRRN	PRRN	PRMRRN	PRRN
quantity theory of money	10,5%	15%	16,7%	25%	n/a	n/a
death penalty	0%	0%	0%	0%	n/a	n/a
hypertext system	5%	5%	35%	35%	n/a	n/a
matriz insumo producto	0%	95%	20%	20%	n/a	n/a
abuso sexual en niños	25%	80%	76,5%	80%	n/a	n/a
olympic games sydney	33,3%	90%	11,1%	60%	n/a	n/a
effects of nuclear war	71,4%	90%	57,1%	85%	n/a	n/a
produccion vitivinicola argentina	100%	100%	61,5%	75%	n/a	n/a
primeros auxilios	11,1%	20%	5,6%	15%	n/a	n/a
citation analysis	5%	5%	5%	5%	n/a	n/a

7.3. Anexo III: URLs inactivos

Porcentaje de registros inactivos recuperados.

Expresión de búsqueda	Primeros diez resultados				
	AltaVista	Excite	Fast	Google	Northern Light
quantity theory of money	20%	0%	30%	0%	0%
death penalty	10%	10%	0%	0%	0%
hypertext system	0%	0%	10%	0%	0%
matriz insumo producto	20%	10%	0%	0%	0%
abuso sexual en niños	10%	0%	0%	0%	0%
olympic games sydney	30%	0%	0%	0%	0%
effects of nuclear war	30%	10%	10%	0%	0%
produccion vitivinicola argentina	20%	20%	80%	0%	0%
primeros auxilios	10%	0%	10%	0%	0%
citation analysis	0%	0%	0%	0%	0%

Expresión de búsqueda	Primeros veinte resultados				
	AltaVista	Excite	Fast	Google	Northern Light
quantity theory of money	20%	0%	25%	0%	5%
death penalty	10%	5%	0%	0%	0%
hypertext system	15%	0%	5%	0%	0%
matriz insumo producto	10%	5%	5%	0%	5%
abuso sexual en niños	5%	0%	0%	0%	0%
olympic games sydney	15%	0%	5%	0%	0%
effects of nuclear war	25%	10%	5%	0%	0%
produccion vitivinicola argentina	10%	20%	90%	0%	5%
primeros auxilios	20%	0%	15%	0%	15%
citation analysis	15%	0%	0%	0%	0%

7.4. Anexo IV: cantidad real de URLs recuperados

Relación porcentual entre la cantidad de registros recuperados (RR) y la cantidad de registros evaluados.

Expresión de búsqueda	AltaVista			Excite		
	Total RR	10 Primeros	20 Primeros	Total RR	P. 10	P. 20
quantity theory of money	4722540	0,0002%	0,0004%	2380000	0,0004%	0,0008%
death penalty	189018	0,0053%	0,0106%	1110000	0,0009%	0,0018%
hypertext system	1072647	0,0009%	0,0019%	4700000	0,0002%	0,0004%
matriz insumo producto	30315	0,0330%	0,0660%	5530	0,1808%	0,3617%
abuso sexual en niños	687207	0,0015%	0,0029%	2540	0,3937%	0,7874%
olympic games sydney	1110615	0,0009%	0,0018%	2430000	0,0004%	0,0008%
effects of nuclear war	1998366	0,0005%	0,0010%	1670000	0,0006%	0,0012%
produccion vitivinicola argentina	1561698	0,0006%	0,0013%	656000	0,0015%	0,0030%
primeros auxilios	11644	0,0859%	0,1718%	1650	0,6061%	1,2121%
citation analysis	2833	0,3530%	0,7060%	1420000	0,0007%	0,0014%

Expresión de búsqueda	Fast			Google		
	Total RR	10 Primeros	20 Primeros	Total RR	P. 10	P. 20
quantity theory of money	43496	0,0230%	0,0460%	82200	0,0122%	0,0243%
death penalty	352639	0,0028%	0,0057%	469000	0,0021%	0,0043%
hypertext system	226749	0,0044%	0,0088%	402000	0,0025%	0,0050%
matriz insumo producto	630	1,5873%	3,1746%	538	1,8587%	3,7175%
abuso sexual en niños	8455	0,1183%	0,2365%	5590	0,1789%	0,3578%
olympic games sydney	408	2,4510%	4,9020%	265	3,7736%	7,5472%
effects of nuclear war	105620	0,0095%	0,0189%	153000	0,0065%	0,0131%
produccion vitivinicola argentina	104	9,6154%	19,2308%	51	19,6078%	39,2157%
primeros auxilios	26373	0,0379%	0,0758%	40100	0,0249%	0,0499%
citation analysis	207578	0,0048%	0,0096%	255000	0,0039%	0,0078%

Expresión de búsqueda	Northern Light		
	Total RR	10 Primeros	20 Primeros
quantity theory of money	72108	0,0139%	0,0277%
death penalty	529447	0,0019%	0,0038%
hypertext system	297388	0,0034%	0,0067%
matriz insumo producto	1214	0,8237%	1,6474%
abuso sexual en niños	7700	0,1299%	0,2597%
olympic games sydney	301	3,3223%	6,6445%
effects of nuclear war	160875	0,0062%	0,0124%
produccion vitivinicola argentina	679	1,4728%	2,9455%
primeros auxilios	29142	0,0343%	0,0686%
citation analysis	434138	0,0023%	0,0046%

7.5. Anexo V: precisión-exhaustividad

Precisión

Documentos relevantes.

Término de búsqueda	AltaVista	C4	Excite	Fast	Google	Ixquick	MetaCrawler	Northern Light
grigallo	1,00	1,00	0	0,80	1,00	1,00	1,00	1,00
reduvio	1,00	0	0	1,00	1,00	0	1,00	1,00
vomitel	0,75	0,78	0	0,73	1,00	0,86	0,86	0,70
guapomo	0,67	0,78	1,00	1,00	1,00	0,25	0,71	0,86
huisquil	0,87	0,92	1,00	0,87	1,00	0,80	0,89	0,80
apodyopsis	0	0	0,50	1,00	1,00	0	1,00	1,00
fluctisonant	0	1,00	0	0,83	1,00	0	1,00	1,00
materteral	1,00	0,90	1,00	1,00	1,00	0	1,00	1,00
escrocon	1,00	0	0	0	0	0	1,00	1,00
galopillo	0,50	0	0	1,00	1,00	0,50	1,00	1,00

Documentos relevantes potenciales.

Término de búsqueda	AltaVista	C4	Excite	Fast	Google	Ixquick	MetaCrawler	Northern Light
grigallo	1,00	1,00	0	0,80	1,00	1,00	1,00	1,00
reduvio	0	0	0	1,00	0,50	0	0	0,50
vomitel	0,63	0,72	0	0,64	0,90	0,71	0,71	0,60
guapomo	0,33	0,44	0	0,80	0,67	0	0,29	0,64
huisquil	0,73	0,74	1,00	0,70	0,79	0,60	0,72	0,67
apodyopsis	0	0	0,50	0,50	0,50	0	0,50	1,00
fluctisonant	0	0,86	0	0,83	1,00	0	1,00	1,00
materteral	0,50	0,40	0,20	0,43	0,43	0	0,38	0,40
escrocon	1,00	0	0	0	0	0	1,00	1,00
galopillo	0,50	0	0	0,50	1,00	0	1,00	0,67

Documentos relevantes óptimos.

Término de búsqueda	AltaVista	C4	Excite	Fast	Google	Ixquick	MetaCrawler	Northern Light
grigallo	0	0	0	0	0	0	0	0
reduvio	1,00	0	0	0	0,50	0	1,00	0,50
vomitel	0,13	0,06	0	0,09	0,10	0,14	0,14	0,10
guapomo	0,33	0,33	1,00	0,20	0,33	0,25	0,43	0,21
huisquil	0,13	0,18	0	0,17	0,21	0,20	0,17	0,13
apodyopsis	0	0	0	0,50	0,50	0	0,50	0
fluctisonant	0	0,14	0	0	0	0	0	0
materteral	0,50	0,50	0,80	0,57	0,57	0	0,63	0,60
escrocon	0	0	0	0	0	0	0	0
galopillo	0	0	0	0,50	0	0,50	0	0,33

Documentos relevantes sin duplicados.

Término de búsqueda	AltaVista	C4	Excite	Fast	Google	Ixquick	MetaCrawler	Northern Light
grigallo	1,00	0,83	0	0,80	1,00	1,00	0,75	0,80
reduvio	1,00	0	0	1,00	1,00	0	1,00	1,00
vomitel	0,75	0,61	0	0,73	1,00	0,86	0,86	0,70
guapomo	0,67	0,78	1,00	1,00	1,00	0,25	0,71	0,86
huisquil	0,87	0,74	1,00	0,74	0,92	0,80	0,78	0,73
apodyopsis	0	0	0,50	1,00	1,00	0	1,00	1,00
fluctisonant	0	1,00	0	0,50	1,00	0	1,00	0,88
materteral	1,00	0,70	1,00	0,86	0,86	0	0,88	0,80
escrocon	1,00	0	0	0	0	0	1,00	1,00
galopillo	0,50	0	0	1,00	1,00	0,50	1,00	1,00

Exhaustividad

Documentos relevantes.

Término de búsqueda	AltaVista	C4	Excite	Fast	Google	Ixquick	MetaCrawler	Northern Light
grigallo	0,04	0,23	0	0,15	0,15	0,08	0,15	0,19
reduvio	0,13	0	0	0,25	0,25	0	0,13	0,25
vomitel	0,11	0,25	0	0,14	0,18	0,11	0,11	0,12
guapomo	0,10	0,18	0,05	0,13	0,08	0,03	0,13	0,31
huisquil	0,10	0,27	0,02	0,15	0,18	0,06	0,12	0,09
apodyopsis	0	0	0,08	0,33	0,33	0	0,17	0,08
fluctisonant	0	0,21	0	0,30	0,12	0	0,12	0,24
materteral	0,05	0,21	0,12	0,16	0,16	0	0,19	0,12
escrocon	0,25	0	0	0	0	0	0,25	0,50
galopillo	0,11	0	0	0,22	0,11	0,11	0,11	0,33

Documentos relevantes potenciales.

Término de búsqueda	AltaVista	C4	Excite	Fast	Google	Ixquick	MetaCrawler	Northern Light
grigallo	0,04	0,23	0	0,15	0,15	0,08	0,15	0,19
reduvio	0	0	0	0,50	0,25	0	0	0,25
vomitel	0,10	0,26	0	0,14	0,18	0,10	0,10	0,12
guapomo	0,09	0,17	0	0,17	0,09	0	0,09	0,39
huisquil	0,10	0,27	0,02	0,15	0,18	0,06	0,12	0,10
apodyopsis	0	0	0,14	0,29	0,29	0	0,14	0,14
fluctisonant	0	0,19	0	0,31	0,13	0	0,13	0,25
materteral	0,06	0,24	0,06	0,18	0,18	0	0,18	0,12
escrocon	0,25	0	0	0	0	0	0,25	0,50
galopillo	0,17	0	0	0,17	0,17	0	0,17	0,33

Documentos relevantes óptimos.

Término de búsqueda	AltaVista	C4	Excite	Fast	Google	Ixquick	MetaCrawler	Northern Light
grigallo	0	0	0	0	0	0	0	0
reduvio	0,25	0	0	0	0,25	0	0,25	0,25
vomitel	0,14	0,14	0	0,14	0,14	0,14	0,14	0,14
guapomo	0,13	0,19	0,13	0,06	0,06	0,06	0,19	0,19
huisquil	0,08	0,28	0	0,16	0,20	0,08	0,12	0,08
apodyopsis	0	0	0	0,40	0,40	0	0,20	0
fluctisonant	0	1,00	0	0	0	0	0	0
materteral	0,04	0,19	0,15	0,15	0,15	0	0,19	0,12
escrocon	0	0	0	0	0	0	0	0
galopillo	0	0	0	0,33	0	0,33	0	0,33

Documentos relevantes sin duplicados.

Término de búsqueda	AltaVista	C4	Excite	Fast	Google	Ixquick	MetaCrawler	Northern Light
grigallo	0,04	0,22	0	0,17	0,17	0,09	0,17	0,22
reduvio	0,13	0	0	0,25	0,25	0	0,13	0,25
vomitel	0,11	0,20	0	0,15	0,19	0,11	0,11	0,13
guapomo	0,10	0,18	0,05	0,13	0,08	0,03	0,13	0,31
huisquil	0,11	0,24	0,02	0,17	0,21	0,07	0,14	0,10
apodyopsis	0	0	0,08	0,33	0,33	0	0,17	0,08
fluctisonant	0	0,25	0	0,36	0,14	0	0,14	0,29
materteral	0,05	0,19	0,14	0,19	0,19	0	0,22	0,14
escrocon	0,25	0	0	0	0	0	0,25	0,50
galopillo	0,11	0	0	0,22	0,11	0,11	0,11	0,33

7.6. Anexo VI: análisis de varianza

Precisión

Documentos relevantes.

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Valor crítico para F
Entre grupos	4,49	7	0,64	5,20	2,14
Dentro de los grupos	8,88	72	0,12		
Total	13,36	79			

Documentos relevantes potenciales.

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Valor crítico para F
Entre grupos	3,22	7	0,46	3,94	2,14
Dentro de los grupos	8,40	72	0,12		
Total	11,62	79			

Documentos relevantes óptimos.

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Valor crítico para F
Entre grupos	0,22	7	0,03	0,43	2,14
Dentro de los grupos	5,28	72	0,07		
Total	5,50	79			

Documentos relevantes sin duplicados.

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Valor crítico para F
Entre grupos	3,97	7	0,57	4,80	2,14
Dentro de los grupos	8,51	72	0,12		
Total	12,48	79			

Exhaustividad

Documentos relevantes.

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Valor crítico para F
Entre grupos	0,34	7	0,05	6,37	2,14
Dentro de los grupos	0,54	72	0,01		
Total	0,88	79			

Documentos relevantes potenciales.

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Valor crítico para F
Entre grupos	0,44	7	0,06	7,09	2,14
Dentro de los grupos	0,64	72	0,01		
Total	1,08	79			

Documentos relevantes óptimos.

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Valor crítico para F
Entre grupos	0,16	7	0,02	1,00	2,14
Dentro de los grupos	1,61	72	0,02		
Total	1,77	79			

Documentos relevantes sin duplicados.

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Valor crítico para F
Entre grupos	0,37	7	0,05	7,08	2,14
Dentro de los grupos	0,54	72	0,01		
Total	0,92	79			

7.7. Anexo VII: coeficiente de Jaccard

Pares de SRI	J
AltaVista - Excite	0,07
AltaVista - C4	0,32
AltaVista - Fast	0,20
AltaVista - Google	0,27
AltaVista - Ixquick	0,33
AltaVista - MetaCrawler	0,48
AltaVista - Northern Light	0,22
C4 - Excite	0,08
C4 - Fast	0,20
C4 - Google	0,52
C4 - Ixquick	0,22
C4 - MetaCrawler	0,48
C4 - Northern Light	0,22
Excite - Fast	0,05
Excite - Google	0,08
Excite - Ixquick	0,00
Excite - MetaCrawler	0,10
Excite - Northern Light	0,06
Fast - Google	0,41
Fast - Ixquick	0,16
Fast - MetaCrawler	0,32
Fast - Northern Light	0,21
Google - Ixquick	0,18
Google - MetaCrawler	0,58
Google - Northern Light	0,23
Ixquick - MetaCrawler	0,22
Ixquick - Northern Light	0,14
MetaCrawler - Northern Light	0,26