ANÁLISIS DE LOS SISTEMAS DE BÚSQUEDA, PLANTEAMIENTO Y APLICACIÓN DE UN MODELO DE EVALUACION EN LA RECUPERACIÓN DE LA INFORMACIÓN EN LA INTERNET

NANCY PATRICIA GUTIERREZ SANCHEZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER

ESCUELA DE INGENIERIAS ELECTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

FACULTAD DE CIENCIAS FISICOMECÁNICAS

BUCARAMANGA

2005

ANÁLISIS DE LOS SISTEMAS DE BÚSQUEDA, PLANTEAMIENTO Y APLICACIÓN DE UN MODELO DE EVALUACION EN LA RECUPERACIÓN DE LA INFORMACIÓN EN LA INTERNET

NANCY PATRICIA GUTIERREZ SANCHEZ

Monografía presentado como requisito para optar el título de
Especialista en Telecomunicaciones

DIRECTOR DE PROYECTO: JORGE HERNANDO RAMÓN

UNIVERSIDAD INDUSTRIAL DE SANTANDER

ESCUELA DE INGENIERIAS ELECTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

FACULTAD DE CIENCIAS FISICOMECÁNICAS

BUCARAMANGA

2005

CONTENIDO

Pág

INTRODUCCION i
1. PRESENTACION DE LA MONOGRAFIA 4
1.1 TITULO 4
1.2 PLANTEAMIENTO DEL PROBLEMA 4
1.3 OBJETIVOS……5
1.3.1. Objetivo General 5
1.3.2. Objetivo Específicos 5
1.4 JUSTIFICACION DEL PLAN PROPUESTO 6
2. NECESIDAD DE EVALUAR SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN 9
2.1 Parámetros de Evaluación para los Sistemas de Recuperación en la Web……..10
3. PLANTEAMIENTO DEL MODELO DE EVALUACIÓN DE LOS SISTEMAS DE RECUPERACIÓN DE LA INFORMACIÓN EN LA WEB……33
3.1 Fases de desarrollo para el estudio de campo……34
3.1.1 Selección de Sistemas de Búsqueda a evaluar…….. 35
3.1.2 Determinación de necesidades informativas a aplicar en los buscadores seleccionados…… ……..…….....35
3.1.3 Descripción del tipo de búsqueda mediante una sintaxis específica……..36
3.1.4 Cronograma de realización de las consultas en los sistemas de búsqueda.……..39
3.1.5 Valoración de la relevancia durante el análisis de los documentos recuperado.....40
3.1.6 Análisis de los resultados: “Criterios de Evaluación”…….....41
4. APLICACIÓN MODELO DE EVALUACION Y ANALISIS DE RESULTADOS……..…52
4.1 Criterio basado en el tamaño promedio del índice……..……52
4.2. Criterio basado en el tiempo promedio de respuesta…..……..……...54
4.3 Criterio basado en el Ruido Documenta……56
4.4 Criterio según el promedio de relevancia, en función de pruebas…….57

4.5 Criterio según el promedio de relevancia en función al tipo de consulta: Especializada y General……..……62
4.6 Criterio basado en la relevancia, en base a la Sintaxis Búsqueda Planteadas…….64
4.7 Criterio basado en el promedio de Exhaustividad-Presición…… ……..67
4.8 Criterio basado en la Similitud de Resultados……..……..69
4.9 Criterio basado en el Análisis de Agrupamiento (Clustering)……..71
APENDICE A. CONCEPTUALIZACIÓN GENERAL DE RECUPERACION DE INFORMACION
A.1 Inconvenientes previsibles en la recuperación de información…… ……76
A.1.1 Recuperación de la Información Vs Recuperación de Datos…… . ……..77
A.2 Enfoque Evolutivo y Clasificatorio de los Sistemas de Recuperación de Información de Documentos textuales……80
A.2.1 Registro Evolutivo de los Sistemas de Recuperación de información……..85
A.2.2 Modelos Conceptuales para la Recuperación de Información……...88
A.2.3 Estructura de Ficheros……96
A.2.4 Operaciones de Consulta, Operaciones sobre Términos y Operaciones sobre Documentos……98
A.3 Escenario fundamental de la Recuperación de la Información en Internet……100
A.3.1 Perspectiva Evolutiva de los Sistemas de Recuperación de la Información en la web……...102
A.3.2 Enfoques Expuestos sobre la Recuperación de Información en la Internet……..107
A.3.3 Limitaciones en la Recuperación de la Información en Internet……..111
A.3.4 Metodología y Estrategias para el Proceso de Recuperación de Información en Internet……..112
APENDICE B. PRINCIPALES HERRAMIENTAS DE BÚSQUEDA EN INTERNET
B.1 Directorios o Índices Temáticos……...126
B.2 Motores de Búsqueda……..……130
B.3 Metabuscadores……..140
B.4 Agentes Inteligentes……..144
B.5 Colecciones de Herramientas de Búsqueda……...148
B.6 Portales……..148

B.7 Cómo acceder a la porción de Internet que no es rastreada por los motores de búsqueda……..14

9APENDICE C. PRESTACIONES DE LOS SISTEMAS DE BÚSQUEDA COMO APOYO TANGIBLE EN LA RECUPERACIÓN DE LA INFORMACIÓN EN INTERNET (CASO DE ESTUDIO).

C.1 Motor de búsqueda " LYCOS "……151

C.2 Motor de búsqueda "GOOGLE "……160

C.3 Motor de búsqueda " YAHOO "……173

C.4 Motor de búsqueda "ALTAVISTA "……184

C.5 Motor de búsqueda " HOTBOT "……191

C.6 Motor de búsqueda " EXCITE "……200

C.7 Motor de búsqueda " MSN SEARCH "……208

CONCLUSIONES……217

PROPUESTAS DE EXPLORACION FUTURAS……222

BIBLIOGRAFIAS……224

LISTA DE FIGURAS

Pág

FIGURA1. Intersección de conjuntos similares (Similitud)……..…..26

FIGURA2. Comportamiento Exhaustividad-Precisión llevado a cabo por Altavista sobre los primeros 10 documentos recuperados, para la primera consulta de búsqueda realizada……47

FIGURA3. Resultados de cobertura promedio de índice, en periodos relativamente distantes……..…52

FIGURA4. Comportamiento manifestado por el índice de cada motor en relación con la pregunta formulada en la dos pruebas llevadas a cabo……53

FIGURA5. Resultados de tiempo promedio de respuesta, en periodos relativamente distantes……..55

FIGURA6. Resultados de ruido documental presente en los motores de búsqueda……56

FIGURA7. Comportamiento promedio de cada motor, según relevancia (0, 1, 2 y 3) sobre los 10, 20 y 30 documentos recuperados y 18 preguntas evaluadas……58

FIGURA8. Análisis de resultados de documentos relevantes Vs documentos permisiblemente relevantes u óptimos (en base 10 primeros documentos recuperados)……..59

FIGURA9. Análisis de resultados de documentos relevantes Vs documentos permisiblemente relevantes u óptimos (en base 20 primeros documentos recuperados)……..60

FIGURA10. Análisis de resultados de documentos relevantes Vs documentos permisiblemente relevantes u óptimos (en base 30 primeros documentos recuperados).....61

FIGURA11. Análisis de resultados promedio según la relevancia por tipo de consulta, Especifica y General (en base a los 10 primeros documentos recuperados).....62

FIGURA12. Análisis de resultados promedio según la relevancia por tipo de consulta, Específica y General (en base a los 20 primeros documentos recuperados)……...63

FIGURA13. Análisis de resultados según la relevancia (2 ó 3), en base a los 10 primeros documentos recuperados y a una pregunta representativa al tipo de sintaxis de consulta…..65

FIGURA14. Análisis de resultados según la relevancia (2 ó 3), en base a los 30 primeros documentos recuperados y a una pregunta representativa al tipo de sintaxis de consulta.....66

FIGURA15. Indicador de Rendimiento, producto del cálculo de valores medios de cada búsqueda, sobre los 10 primeros resultados para las 18 consultas (nivel de relev. 2 ó 3)……68

FIGURA16. Distribución radial de la distancia media de los buscadores respecto a un centro que constituye la mejor colección de documentos……..71

FIGURA17. Agrupamiento de Motores de Búsqueda según su afinidad sobre los diez documentos recuperados en base a una distribución radial de la distancia media existente entre ellos……72

FIGURA18. Proceso de Recuperación de información según Baeza-Yates y Ribeiro-Neto, 1999……76

FIGURA 19. Vista lógica de la entrada de los documentos a los SRI, de Baeza-Yates y Ribeiro-Neto……83

FIGURA20. Vista funcional o lógica asociada con un tipo común de SRI basado en el modelo Booleano……84

FIGURA21. Taxonomía de los modelos de RI [Baeza-Yates y Ribeiro-Neto 1999]……..90

FIGURA22. Representación de la estructura de un fichero inverso……97

FIGURA23. Estructura y funcionamiento de un Directorio temático……....128

FIGURA24. Estructura y funcionamiento de un motor de búsqueda……132

FIGURA25. Criterio de Ponderación……..…139

FIGURA26. Estructura y funcionamiento de un Metabuscador……..141

LISTA DE TABLAS

Pág.

TABLA1. Resultados de Exhaustividad- Precisión, sobre los primeros 10 documentos……..45

TABLA 2. Ejemplo aplicativo de la técnica de agrupamiento “Promedio Aritmético”……..…51

TABLA3. Resultados de similitudes medias obtenidas en este experimento para cada par de motores con 10 documentos analizados……..70

TABLA4. Similitud y Distancia media de cada motor con respecto al resto……..70

TABLA5. Distancias Medias de cada unos de los motores de búsqueda……..72

TABLA6. Agrupamientos y distancias entre motores con las muestras obtenidas en el análisis de los diez primeros documentos recuperados……72

TABLA7. Sinopsis de Criterios Desarrollados…..……74

TABLA8. Diferencias existentes entre recuperación de datos o recuperación de información……..80

TABLA9. Clasif. de los Modelos de Recuperación de Información según Dominich……...89

TABLA10. Clasificación de los modelos de Recuperación de información según Baeza-Yates…….....90

TABLA11. Modelos de Recuperación y atributos que los define ……92

TABLA12. Características del modelo Booleano ……..……93

TABLA13. Características del modelo Vectorial ……..……95

TABLA14. Características del modelo Probabilística ……..……96

TABLA15. Operadores de Consulta…… ……...…118

TABLA16. Ventajas e inconvenientes del uso de Índices Temáticos..……...…130

TABLA17. Ventajas e inconvenientes del uso de los Motores de búsqueda..……136

TABLA18. Ventajas e inconvenientes del uso de los Metabuscadores……144

TABLA19. Características de los Agentes Inteligentes……..147

TABLA20. Ventajas e inconvenientes del uso de Colecciones de herramientas……148

TABLA21. Características de los Portales……149

TABLA22. Prestaciones e información general del Motor de búsqueda de Lycos……157

TABLA23. Estructuras de consulta y operaciones soportadas por Lycos……160

TABLA24. Prestaciones e información general del Motor de búsqueda de Google……169

TABLA25. Estructuras de consulta y operaciones soportadas por Google……172

TABLA26. Prestaciones e información general del Motor de búsqueda de Yahoo……180

TABLA27. Estructuras de consulta y operaciones soportadas por Yahoo……183

TABLA28. Prestaciones e información general del Motor de búsqueda de Altavista……188

TABLA29. Estructuras de consulta y operaciones soportadas por Altavista……191

TABLA30. Prestaciones e información general del Motor de búsqueda de Hotbot……197

TABLA31. Estructuras de consulta y operaciones soportadas por Hotbot……200

TABLA32. Prestaciones e información general del Motor de búsqueda de Excite……205

TABLA33. Estructuras de consulta y operaciones soportadas por Excite……208

TABLA34. Prestaciones e información general del Motor de búsqueda de MSN Search…213

TABLA35. Estructuras de consulta y operaciones soportadas por MSN Search……216

LISTA DE ANEXOS

Pág.

ANEXO1. FORMATO DE CONSULTAS REALIZADAS.....235
ANEXO2. TABLA DE DATOS-CRITERIO DE INDICE……238
ANEXO3. TABLA DE DATOS-CRITERIO DE TIEMPO DE RESPUESTA……239
ANEXO4. TABLA DE DATOS-CRITERIO DE RUIDO DOCUMENTAL……240
ANEXO5. TABLA DE DATOS (1-10, 11-20 Y 21-30 DOCUMENTOS RECUPERADOS)….240
ANEXO6. REPRESENTACION GRAFICA, RENDIMIENTO PROMEDIO EXHAUSTIVIDAD PRECISION……241
ANEXO7. PROCEDIMIENTO DE AGRUPAMIENTO……242
ANEXO8. REPRESENTACION DE LA TOMA DE DATOS Y CÁLCULO DE SIMILITUD (PAR DE MOTORES DE BÚSQUEDA ALTAVISTA-EXCITE)……244

TITULO

ANÁLISIS DE LOS SISTEMAS DE BÚSQUEDA, PLANTEAMIENTO Y APLICACIÓN DE UN MODELO DE EVALUACION EN LA RECUPERACIÓN DE LA INFORMACIÓN EN LA INTERNET*

AUTOR

NANCY PATRICIA GUTIERREZ SANCHEZ**

PALABRAS CLAVES

Sistemas de recuperación de información, motores de búsqueda, indexación, evaluación, prestaciones, World Wide Web.

DESCRIPCION O CONTENIDO

La información es uno de los recursos mas valorados en la sociedad actual. La necesidad de encontrar información pertinente, precisa y en el momento oportuno, en tal volumen de documentos, que además de ser heterogéneos, conduce a un análisis, acerca de cómo los sistemas de recuperación de información desencadenan nuevos métodos que permitan agregar información semántica a los documentos, nuevas formas de indexación y nuevos servicios que son imprescindibles en estos momentos de crucial cambio.

El interés de la monografía radica en la elaboración de una propuesta de desarrollo de un modelo de evaluación que permita analizar la efectividad de la recuperación de la información efectuada por los motores de búsqueda en Internet. El propósito con este presente trabajo, se halla en sacar provecho del manejo y uso que brindan los motores de búsqueda, como sistemas de recuperación de información web, conocer las prestaciones por lo cuales son analizados, exponer una metodología de búsqueda sencilla pero eficaz, y finalmente realizar un estudio puramente experimental, conducente a analizar de forma inductiva, la actuación general de los motores de búsqueda, con la idea de establecer un comportamiento común entre ellos.

---

* Proyecto de Investigación.

** Escuela de Ingenierías Eléctrica, Electrónica y Telecomunicaciones. Especialización en Telecomunicaciones. Director Jorge Hernando Ramón

TITLE

ANALYSIS OF THE SEARCH ENGINES, APPROACH AND APPLICATION OF A MODEL OF EVALUATION IN THE RECOVERY OF THE INFORMATION IN THE INTERNET[*].

AUTHOR

NANCY PATRICIA GUTIERREZ SANCHEZ[**]

KEY WORDS

Systems of recovery of information, search engines, indexation, evaluation, performance, World Wide Web.

DESCRIPTION OR CONTENT

The information is one of the resources more valued for the current society. The need to find pertinent, precise information and in the opportune moment, in such a volume of documents, which beside being heterogeneous, leads to an analysis, brings over of how the systems of recovery of information unleash new methods that allow to add semantic information to the documents, new forms of indexation and new services that are indispensable in these moments of crucial change.

The interest of the monograph lies in the elaboration of an offer of development of a model of evaluation who allows to analyze the efficiency of the recovery of the information carried out by the engines of search in Internet. The intention with this present work, is situated in extracting profit of the managing and use that the engines of search offer, as systems of recovery of information web, know the presentations for which are analyzed, to expose a methodology of simple but effective search, and finally to realize a purely experimental, conducive study to analyzing of inductive form, the performance of the engines of search, with the idea of establishing a common behaviour among them.

---

[*] Investigation Project.

[**] Electrical, Electronic school of Engineerings and Telecommunications. Specialization in Telecommunications. Director Jorge Hernando Ramón

INTRODUCCION

La búsqueda de información cada día resulta más complicada como consecuencia del fuerte crecimiento que está sufriendo la Red, a ingentes volúmenes de información, cuyo crecimiento exponencial surge a raíz del florecimiento de la Internet, y a su caótica distribución estructural que exige "tener acceso a la información relevante", como ser capaz de "descartar lo irrelevante".

El problema radica por un lado, en el considerable número de los motores de búsqueda que aseguran poseer la mayor cantidad de recursos debidamente indexados y accesibles a través de ellos pero cuya capacidad coyuntural para tratar dicho volumen de datos es limitada y por el otro el desconocimiento e incertidumbre por parte del usuario sobre el uso y escogencia de la herramienta adecuada, convierte este inconveniente, en una de las tareas más arduas para encontrar lo que se está buscando, por esa razón muchas veces se termina optando por la navegación al azar.

Es claro, que los sistemas de recuperación de información en la Internet son cada día mas importantes para la sociedad de la información en la que vivimos, por tanto la necesidad de evaluar éstas herramientas con el fin de que de que los usuarios encuentren en ellas las condiciones para valorar su efectividad, y de este modo, adquieran confianza en los mismos. Sin embargo, este tipo de evaluaciones sobre estos sistemas de búsqueda requieren un enfoque multidimensional, que permita abarcar una serie de medidas que represente su efectividad e incluso un constante seguimiento debido a su propia naturaleza determinista y al grado de susceptibilidad a cambios en lo que respecta al contexto web(dinamismo y volatilidad)[1] sobre los cuales se encuentran sujetos, y por ende, les exige estar a la vanguardia para dar cobertura eficiente sobre las necesidades informativas del usuario en su interacción con ellos.

[1] **El dinamismo se refiere a los continuos cambios de contenido de muchos de los documentos de Internet y la volatilidad, a los cambios de destino de un mismo documento.**

La elaboración de esta monografía, es conducida hacia un desarrollo empírico y multidimencional de evaluación, destinada a proporcionar un proceso de análisis que permita verificar la viabilidad por medio del contraste de resultados obtenidos en nuestro estudio con los aportados en otras investigaciones que serán posteriormente evidenciadas.

En el trabajo presentado, reúne la apreciación de siete principales herramientas para la recuperación de la información en Internet (Altavista, Excite, Google, Hotbot, Lycos, MSN Search y Yahoo). Se manejaron 18 búsquedas o consultas en cada una de estos sistemas, a partir de una propuesta de evaluación claramente determinada por una serie de criterios sobre los cuales se operaron parámetros como cobertura del sistema ante una estrategia de búsqueda, tiempo de respuesta, ruido documental, relevancia en la recuperación de información de cada herramienta de búsqueda, precisión y exhaustividad con el fin de reflejar el funcionamiento que un usuario puede esperar obtener de un sistema de búsqueda, similitud documental en los sistemas de recuperación, con el fin de establecer rasgos de cercanía en cuanto a documentos recuperados se refiere, dentro de una colección de documentos examinados y finalmente se utilizó una técnica de clustering, con el fin de identificar el comportamiento en lo relacionado con los agrupamientos de los motores de búsqueda que corroboren el grado de coincidencia en sus contenidos unos contra otros.

Cabe mencionar, que en el apéndice A, B y C, se presenta de carácter meramente conceptual información suficientemente amplia acerca de los sistemas de recuperación de información tanto tradicionales (datos) como sistemas de recuperación web. De igual manera, se ha establecido una metodología básica pero lo suficientemente eficaz para llevar a cabo búsquedas de cualquier índole, se ha dado información amplia sobre los diferentes sistemas de recuperación hoy en día existentes, describiendo su funcionamiento y las ventajas por los cuales se distinguen frente a los otros. Además, se ha profundizado sobre información concerniente, a la evolución que ha trascendido sobre los siete motores de búsqueda que son soporte de nuestra actual propuesta de evaluación en donde se expone de manera resumida el origen al cual tuvo lugar como herramientas de recuperación de información y todo su proceso de cambio al cual los caracteriza hoy en día como unos de los mejores y mas usados motores de búsquedas. Asimismo, y no siendo menos importante, se condensa mediante un cuadro sinóptico (ó tabla de síntesis) las prestaciones que

actualmente cada sistema de búsqueda ofrece y así aprovechar todas las posibilidades de búsquedas en los mismos.

1. PRESENTACION DE LA MONOGRAFIA

1.1 TITULO

Análisis de los Sistemas de Búsqueda, Planteamiento y Aplicación de un Modelo de Evaluación en la Recuperación de la Información en la Internet.

1.2 PLANTEAMIENTO DEL PROBLEMA

Vivimos en un medio de constante transición tecnológica, y es a partir de este panorama que surge una constante necesidad de mantener y establecer nuevos puentes que cobren conciencia de que la información no es un fin en si misma, sino que la información es solo una condición para el conocimiento y que su importancia radica, indiscutiblemente en el efectivo acceso y manejo que hagamos de ella.

Actualmente, las posibilidades de información inmersos en la Internet, ha alterado profundamente el tejido de nuestra sociedad y han cambiado en gran medida nuestros modos de proceder en la llamada “Sociedad de la información y del conocimiento”. Sin embargo, a partir de los logros conseguidos y de otros iniciados, se puede entrever el efecto de incertidumbre frente a la recuperación de la información.

La importancia que posee la recuperación de la información puede inferirse tras su traslado al contexto de la Web. Este nuevo escenario ha vislumbrado una serie de problemas e inquietudes dando lugar, al desarrollo de nuevos sistemas o mecanismos que permitan unos niveles elevados de precisión, posibilitando al usuario la localización de la información que precise en cualquier momento, de la manera mas relevante y eficaz posible, en pocas palabras que permita devolver la máxima señal con el mínimo ruido.

La necesidad de un análisis crítico de la evaluación de los sistemas de recuperación de información subyace específicamente en analizar la viabilidad de los datos, que pueden ser de carácter genérico, semánticamente ambiguo y no necesariamente estructurado. A partir de esto, podemos decir, que el interés central a analizar sobre estos sistemas radica en la relevancia, la cual separa lo que realmente nos interesa de lo que no, siendo

éste un espectro totalmente diferente para cada usuario y para cada instante de tiempo dado.

Hoy por hoy, resolver la ambigüedad semántica no es fácil, la dificultad estriba por un lado, en la manera como los sistemas interpretan en forma sólida lo que pedimos; por otro lado está, en muchos casos el desconocimiento evidente que manifiesta el usuario frente a cómo expresar lo que busca.

Podemos concluir que la planeación efectiva de procesos de evaluación que se pretende ahondar, consiste en dar a conocer el comportamiento que manifiesta los sistemas de recuperación de información, específicamente en la Web; analizar, e investigar en mayor o menor grado el adecuado funcionamiento de gestión de éstos sistemas, encaminados a identificar la eficacia y dinamismo que son capaces de proporcionar de manera coherente y relevante de acuerdo a la necesidad informativa del usuario tras una búsqueda dada.

Por supuesto, en este proyecto, además de canalizar información especifica sobre el funcionamiento de los sistemas de recuperación de información; con base en el proceso de estudio planteado, se fundamenta la viabilidad de los resultados obtenidos y se exponen posibles mejoras a resolver en futuras monografías con respecto al modelo de evaluación aplicado.

1.3 OBJETIVOS

1.3.1. Objetivo General

Analizar en profundidad el funcionamiento los Sistemas de Búsqueda y elaborar una propuesta de desarrollo que permita evaluar la efectividad de la recuperación de la información efectuada por los Motores de búsqueda en la Internet

1.3.2. Objetivos Específicos

- Proporcionar de manera detallada características y prestaciones de los principales sistemas de búsqueda.
- Proponer una metodología general y eficaz para la utilización de los Sistemas de Búsqueda en la Internet.
- Plantear un modelo de evaluación que permita:
    - Investigar en qué medida los sistemas de búsqueda, presentan los mejores resultados de exhaustividad y precisión en relación con el volumen de documentos que ofrecen como respuesta.
    - Evaluar grados de relevancia, concernientes con el alineamiento de los documentos como respuesta a una búsqueda determinada.
    - Determinar el ruido documental presente en estos sistemas de recuperación Web.
    - Establecimiento de la similitud existente entre los índices de los motores en función del contenido y alineamiento de su respuesta.
    - Determinar un método de agrupamiento, conducente a identificar los motores de búsqueda mas afines uno con respecto a otros.

1.4. JUSTIFICACIÓN DEL PLAN PROPUESTO

El paradigma que emana los sistemas de recuperación de información, específicamente los motores de búsqueda emerge en los usuarios una serie de incógnitas en la apreciación de la efectividad de la recuperación de la información que éstos ofrecen en pro de satisfacer respuestas de calidad e idónea tras alguna información en particular. Es precisamente esa incertidumbre lo que lleva a reflexionar si lo que en realidad se retorna es una información sesgada frente a una respuesta ideal.

Para nadie es desconocido que por mas de dos décadas la gran red global “la Internet” y la gran ingente información producto de la creación de la World Wide Web, ha suscitado la creación, implantación y funcionamiento de los Sistemas de Búsqueda con el fin de solventar la compleja manipulación de la información que disponemos.

Sin embargo el constante avance y materialización de infinidad de investigaciones sobre estos sistemas constituyen para la recuperación de la información un acreciente reproducción substancial de niveles de incertidumbre que permitan interpretar la efectividad de su funcionamiento.

Gran parte de nuestro tiempo es invertido frente a nuestro computador buscando información, barrera que puede ser solventada aprendiendo cómo obtenerla y otra gran parte examinando y entreviendo datos completamente irrelevantes para nosotros. Las herramientas más utilizadas en Internet se basan en el empleo de patrones de búsqueda. El acceso a bases de datos se hace mediante sistemas de consulta más potentes y cómodos. La experiencia y habilidad para usarlos correctamente es mucho más importante si deseamos encontrar datos que nos sean útiles.

La crisis sobre la búsqueda de la información se debe a la incapacidad de una sola persona de retener ni siquiera una mínima parte de la información que le es útil. El problema se agrava cuando la información útil es a su vez una parte ínfima de toda la que recibe e inapreciablemente a toda la que tiene a su disposición.

Con tanta información disponible lo que debemos saber es cual nos interesa, donde encontrarla, como obtenerla y la mejor manera de gestionarla. Con la ayuda de los navegadores de la World Wide Web, se facilitan las tareas de localización y obtención, pero saber qué es lo que hay, lo que nos interesa y cómo utilizarlo son cuestiones más complejas.

Para facilitar la tarea, cada día hay más sitios dedicados a la creación y mantenimiento de listas y directorios de recursos organizados según criterios geográficos o temáticos denominados generalmente “Directorios o catálogos”. Pueden ser los puntos de partida ideales cuando buscamos información sobre un tema y queremos saber que hay en la Red o buscamos algún servicio siguiendo criterios geográficos. A partir de las entradas del catálogo podemos ir descendiendo hasta encontrar lo que realmente nos interesa, simplemente siguiendo la estructura lógica que define el servicio.

De todos modos, el empleo de los catálogos deja de ser adecuado cuando lo que buscamos es más específico, ya que encontrarlo simplemente navegando puede resultar bastante difícil. Nos encontramos entonces con la necesidad de emplear herramientas para la búsqueda y selección de la información disponible en la Red. Estamos hablando de lo que hemos denominado "Motores de Búsqueda".

Preguntas como: "¿Cuál seria la forma de evaluar la calidad de la información que ofrecen los Motores de Búsqueda? " Inquirir frente a este planteamiento lleva a analizar a que sólo porque un documento aparece en línea no significa que contiene información válida. De hecho, la finalidad de esta monografía es ahondar sobre que tipos de sistemas de recuperación contamos, profundizar sobre el funcionamiento de algunos motores de búsqueda y finalmente plantear y aplicar un modelo de evaluación, enmarcada en la reunión empírica de datos, dirigida a un estudio inductivo, permitiendo de esta manera, reflejar de modo comparativa las características y comportamientos predominantes (relevancia de operaciones de búsqueda, calidad de respuesta, funcionamiento global del sistema…) de cada uno de los sistemas con un detallado escrutinio.

2. NECESIDAD DE EVALUAR LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

Los Sistemas de Recuperación de Información Web, resultan susceptibles, como cualquier otro sistema, a ser sometidos a evaluación, para que sus usuarios puedan valorar su efectividad. La tradición de la evaluación es tan antigua casi como el desarrollo de los primeros Sistemas de Recuperación de Información, encontrándose estrechamente vinculadas con la investigación y el desarrollo de la recuperación de información. Realmente, "la propia naturaleza de los Sistemas de Recuperación de Información, propicia su necesidad crítica de evaluación, justo como cualquier otro campo de trabajo que aspire a ser clasificado como campo científico"[2].

Si bien es cierto, el incremento de motores de búsqueda en la web ha dado lugar a una serie de características de gran discusión referente a los mecanismos llevados a cabo para la evaluación de estos sistemas, como es la presencia eventual de estudios realizados, debido al poco interés existente de personal especializado en esta línea de trabajo, cuya ausencia de periodicidad a incidido de manera negativa en la calidad de la evaluaciones y al seguimiento y establecimiento de nuevos parámetros que permitan llevar a cabo estos procesos de evaluación.

Varios estudios de tipo divulgativo sobre esta área de investigación, se han manejado de forma aislada y con resultados muy poco confiables, dispares y dispersos[2]. Usualmente se trata de evaluaciones a muy pequeña escala, que abarcan pocos sistemas de búsqueda, esbozan pocas consultas o examinan muy pocos de los resultados logrados. En su mayoría presentan un perfil simplemente descriptivo, y en ellos se llega a conclusiones contradictorias. Los estudios con un enfoque más cuantitativo son mucho más escasos y tampoco suelen indicar detalladamente la metodología empleada en el experimento, lo cual les resta fiabilidad. Cuando sí la indican, el método es, a menudo, deductivamente poco coherente o inductivamente poco riguroso.

[2] Revistas divulgativas, donde el personal encargado de realizar este tipo de estudios, son personas no profesionales o ajenos a las Ciencias de la información.

Para nadie es desconocido, que los numerosos sistemas de recuperación de la información, compiten por atraer a nuevos usuarios y su variedad de prestaciones y características propician su evaluación. Es así que las evaluaciones juegan un papel fundamental en el planteamiento de nuevos retos e incentivos para la investigación sobre como optimizar la recuperación de información y por ende alcanzar una eficiente calidad informativa que personifica el entorno Web. No obstante, el método de estimación seguido, la propia naturaleza de los sistemas de búsqueda y el carácter dinámico de sus bases de datos, en constante cambio y crecimiento, dificultan enormemente su evaluación, por tanto la importancia de hacer seguimiento continuo a este tipo de estudios.

2.1 PARAMETROS GENERALES UTILIZADOS PARA LA EVALUACIÓN DE SISTEMAS DE RECUPERACION WEB

Para comprender el enfoque evaluativo utilizado en los sistemas de recuperación de información vale la pena exponer la siguiente reflexión proporcionada por Baeza-Yates que manifiesta que “un Sistema de Recuperación de Información puede ser evaluado bajo diversos criterios, incluyendo entre los mismos: la eficacia en la ejecución, el efectivo almacenamiento de los datos, la efectividad en la recuperación de la información y la serie de características que ofrece el sistema al usuario”. Estos criterios no deben confundirse, la eficacia en la ejecución es la medida del tiempo para realizar una operación, la eficiencia del almacenamiento es el espacio que se precisa para almacenar los datos y por último está la efectividad de la recuperación “normalmente basada en la relevancia de los documentos recuperados”[1].

Antes de dar a conocer algunos de los parámetros utilizados en anteriores investigaciones, es importante mencionar la existencia de tres grandes grupos a los cuales se han visto sometidos los sistemas de recuperación web, cuyo análisis va enfocado a algún tipo de ensayo o experimentación orientados a proponer una técnica evaluativo de carácter global.

Estudios explícitos

Son en realidad pocos los trabajos realizados en este grupo de investigación, en los que se pone en consideración distintas variables de evaluación, como son:
Prestaciones en la recuperación de información, amigabilidad del Interface, precisión, número de ocurrencias, formato de presentación, documentación, frecuencia de incorporación de nuevos documentos, audiencia, Porción de página indexada, legibilidad de los documentos, tamaño del índice.

Algunos publicaciones realizados sobre este grupo evaluaciones proceden sus inicios a partir de 1995, donde claramente se destacan Chu y Rosenthal[3][6], de la Universidad de Long Island en Nueva York, cuyos estudios se encuentran basados en las características formales del propio motor de búsqueda y en las descripciones técnicas que proporciona el sistema (lógica booleana, truncamiento, búsqueda por campos, por palabra o por frase), en el que concluyentemente señaló a Altavista, como el mejor sistema de búsqueda en cuanto aciertos, cantidad de opciones avanzadas así como la facilidad de su uso. Así mismo, en su trabajo mencionan a Courtois, Baer y Stara[9], quienes destacan la potencialidad del motor Webcrawler en todo lo relacionado con su flexibilidad a la hora de plantear las ecuaciones de búsqueda y su rápida respuesta; resaltando además su interface, al cual consideran muy adecuado para usuarios con escasos conocimientos.

Basándose también en la idea de la flexibilidad del interface, Scoville [41] apuesta por los motores Excite, Infoseek y Lycos por la profundidad en cuanto posibilidades de recuperación de información ofrecían. Análogo es el estudio propuesto por Davis [11], quien toma en consideración el tamaño de índice del motor y las posibilidades de recuperación de información, destacándose por Alta Vista, Hotbot e Infoseek, sobre un total de siete motores evaluados.

En otro estudio de 1995, Winship[55] estudia cuatro motores de búsqueda: World Wide Web Worm, Webcrawler, Lycos y Harvest; junto a dos directorios: Yahoo y Galaxy,

[3] Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology, 1995

centrando su análisis en la porción de página indexada, en el interface del sistema, en las capacidades de búsqueda, en los formatos de presentación de los documentos y en el número de documentos recuperados, destacando en primer lugar a Lycos, ligeramente por encima de Harvest.

En 1996 un estudio sencillo es llevado a cabo por Lebedev[25] que consistió en introducir ocho palabras clave, relacionadas con la Química y con la Física, en varios motores de búsqueda y suma el total de ocurrencias de las ocho búsquedas. Para él, el número de documentos es el indicador principal, pero indicador al fin y al cabo, de la calidad de los distintos motores de búsqueda. Por otro lado, propone un segundo criterio, el de la incorporación de nuevas páginas al motor, este factor lo denomina "dinámica de los motores de búsqueda", concluyendo que, bajo su punto de vista, Alta Vista en el mejor motor de búsqueda, seguido muy de cerca por Hotbot. Llama la atención en este experimento el pésimo resultado de Yahoo, Galaxy y Webcrawler.

Otro estudio desarrollado en 1996, se le atribuye, a Stobart y Kerridge [44] de la Universidad de Sunderland, donde su investigación se basó sobre un conjunto de cuatrocientos usuarios de Internet (profesores e investigadores universitarios en su mayor parte), que completaron un cuestionario. La primera de las preguntas consistía en indicar cuáles eran los motores de búsqueda que más empleaban, obteniéndose como respuesta Alta Vista, Yahoo y Lycos (por ese orden) de forma preferente. En segundo lugar se preguntaba a quienes aseguraban usar más de un motor de búsqueda cuál era al que acudían primero, siendo Alta Vista el preferido como una amplísima diferencia. En tercer lugar se preguntaban las posibles causas de esa fidelidad a un motor de las cuales se destacaron en forma descendente: velocidad 26%, tamaño de índice 21%, hábito 18%, exactitud 14%, amigabilidad 13% y otras causas 8%.

Westera [53] presenta de igual forma aportes en este tipo de estudio donde se centra en la realización de varios análisis de las prestaciones de los distintos motores de búsqueda. Uno de ellos se encuentra dirigido en las capacidades de su interface a la hora de realizar las operaciones de recuperación de información. Una particularidad interesante de este estudio de Westera es que el mismo se actualiza constantemente desde 1996. Divide las

capacidades de búsquedas en dos grupos, básicas y especiales. En el primer grupo destacan Alta Vista y Google; en el segundo Alta Vista y Hotbot.

Entre otros autores, se destaca Greg R. Notess, que al igual que los demás ha seguido una línea de trabajo muy similar, para él la obsolescencia que posee estas exploraciones, implica la necesidad de actualizar los estudios con cierta frecuencia, tal como lo expone en la web Search Engine Showdown. En sus publicaciones se puede observar los diferentes análisis realizados en diferentes periodos de tiempo, y donde el último trabajo expuesto data de diciembre del 2002[36], donde calcula, entre otras estadísticas, los documentos recuperados en distintos motores (nueve exactamente), tras realizarle una serie de veinticinco preguntas. En él se destaca en primer lugar Google como motor de búsqueda que más documentos recuperó y le sigue en un segundo lugar Alltheweb[4].

Otro criterio explícito, lo encontramos en la popularidad o "audiencia" del motor de búsqueda, como es el caso del trabajo de Sullivan[48],[49] quien ha realizado varias publicaciones que incluye también número de consultas realizadas (proporciona una clasificación del número de consultas que reciben cada día los motores de búsqueda, aportando como componente suplementario la utilidad de cada motor para sus usuarios, en donde Google claramente mantiene su posición de predominio, seguido por Inktomi), tráfico redirigido (analiza los registros de accesos de múltiples sitios web mas de cien mil, para verificar de donde proceden las visitas recibidas, donde concluye como los mas destacados a Yahoo y Google) con respecto a varios motores de búsqueda. Sus dos últimos artículos se encuentran titulados como "Nielsen NetRatings Search Engine Ratings" y "comScore Media Metrix Search Engine Ratings"[5], donde clasifica a los distintos motores según un estudio de audiencia sobre los accesos que más de cien mil usuarios de Internet realizan a los distintos motores de búsqueda, proporcionándonos como motor más popular a Google (47.3%) seguido por Yahoo (29%) y MSN Search por un (13.6%).

[4] Search Engine Statistics: Relative Size Showdown, http://www.searchengineshowdown.com/stats/size.shtml
[5] comScore Media Metrix Search Engine Ratings, por Danny Sullivan, Editor February 11, 2005 y Nielsen NetRatings Search Engine Ratings, en abril 22, 2005 < http://searchenginewatch.com/reports/article.php/2156451>

Por otro lado, tenemos a Gordon y Pathak [18] que consideran que la evaluación procedente de la observación de las características más técnicas de los motores de búsqueda no deja de ser una mera percepción testimonial. En su trabajo, ellos recogen también las conclusiones de algunos de estos estudios pero afirman que "aunque los tests testimoniales pueden proporcionarnos datos útiles para la toma de decisiones sobre qué motor de búsqueda emplear en un determinado momento, únicamente nos pueden proporcionar información indirecta sobre cuál es el más efectivo en la recuperación de información", por lo cual proponen proceder a otro tipo de estudios más relacionados con esa efectividad.

Así mismo, C|net[7], empresa especializada en evaluar productos en línea, realizó un estudio comparativo de diecinueve motores de búsqueda considerando el acierto en sus búsquedas, la facilidad de su uso y la cantidad de opciones avanzadas que proporcionaba cada motor. En su estudio, claramente destaca a Google y Yahoo como los mejores de todos.

Chang[5], en cambio, se centró en realizar estudios relacionados con la estimación del tamaño del índice del motor de búsqueda y en el que expone que debido al constante cambio y expansión de la web, provoca que ninguno de los motores de búsqueda pueden indexar la totalidad de los documentos. Entre sus resultados determinó que los motores recopilan entre el 5% y el 30% de la totalidad de documentos de la web, y la unión de los once principales motores de búsqueda no alcanza el 50%.

El tamaño del índice del motor, el número de documentos devueltos, características relacionadas con el modo de almacenar las referencias y el formato de salida de los documentos, constituyen el objeto del trabajo de Peterson [39]. Este autor tomó datos en tres períodos de tiempo distintos (febrero, mayo y noviembre del año 1996), interrogando a ocho motores por medio de dos expresiones de búsqueda, una conformada por un término individual y la otra por una frase literal. El resultado de esta parte del experimento mostraba al motor Hotbot como el de mayor número de documentos.

En realidad, son muchos lo estudios realizados, sin embargo en este apartado se nombran solo algunos de ellos, lo que si es importante mencionar a cerca de estas investigaciones, es la prominente disparidad en sus resultados lo que ha servido para revelar el alto grado de variedad y divergencia de opiniones detectado y a las distintas consideraciones evaluadas por los distintos investigadores, lo que ha dado lugar a considerar el uso de nuevos parámetros más objetivos que permitan extraer conclusiones más comunes.

Estudios Experimentales

Estos estudios son un conjunto de publicaciones que van enfocado a una serie de tests o experimentaciones que van más allá de las simples consideraciones de las características externas de cada motor. En ellas se plantean parámetros como composición de los índices, donde se estudia la cobertura, la frecuencia de actualización y la porción de página indexada; las capacidades de búsqueda donde se analizan las prestaciones que los distintos motores poseen a la hora de recuperar información como son operadores booleanos, búsquedas por frase literal, truncamiento y refinamiento de búsquedas; ejecución de la recuperación de información donde se emplean parámetros como la precisión, la exhaustividad y el tiempo de respuesta; y finalmente se tiene el esfuerzo del usuario, donde se valoran la documentación y el interface del sistema.

Este tipo de estudios, mucho más rigurosos que los incluidos en el apartado anterior, sin embargo tienden a proporcionar una interesante cantidad de datos pero, al mismo tiempo presentan un problema de gran consideración: un alto grado de esfuerzo y dedicación lo que conlleva implícito una escasa producción de los mismos y por tanto, un riesgo alto de obsolescencia al evolucionar con tanta rapidez los motores de búsqueda en la web. Por ende tiene el inconveniente de poder abarcar, por lo general, una gran variedad motores de búsqueda.

Si bien es cierto, la curiosidad comparativa frente a la aparición de numerosos buscadores ha llevado al logro por parte de varios investigadores al estudio de esta línea de trabajo. Sin embargo, de este tipo de disertaciones realizadas se ha caracterizado por dar a conocer una metodología que carece de profundidad, y en la mayoría de los

casos son más de tipo lacónico, es decir breves, cuyo análisis de datos son a pequeña escala y cuya continuidad es poco puesta en práctica.

Algunas investigaciones que engloban en este grupo, podemos citar a Marchionini, Barlow y Hill[31] realizaron un estudio comparativo de WAIS y Dialog, ejemplo de un sistema basado en la lógica booleana, en una base de datos de 200.000 registros bibliográficos y, aunque usaron los clásicos instrumentos de Cranfield para su evaluación, comentaron en sus conclusiones que se constataba la necesidad de nuevas medidas de evaluación orientadas a la naturaleza interactiva de sistemas como WAIS y ya aludieron a la falta de estudios sistemáticos sobre el funcionamiento de la recuperación de los sistemas en red, indicando el reto de analizar sistemas que aplicaran técnicas e interfaces diferentes, así como evaluar esos sistemas desde una perspectiva formativa y comparativa.

El estudio de Ding y Marchionini [13], que evaluaba InfoSeek, Lycos y OpenText, se basó en Cranfield, introduciendo algunas variaciones. Realizaron una detallada comparación de las características de los sistemas y una evaluación de la eficacia de la recuperación. Se utilizaron sólo cinco preguntas, considerando los primeros 20 registros recuperados por consulta en cada motor de búsqueda y asignando sus propios juicios de relevancia, con seis niveles. El funcionamiento se evaluó usando la precisión (una nueva medida propuesta por los autores), definida como la suma de las puntuaciones de relevancia de las referencias obtenidas en cada servicio dividida por la suma de la puntuación para todos los servicios. No utilizaron la exhaustividad. Lycos y OpenText fueron considerados superiores a Infoseek.

Courtois, Baer y Stara [9] evaluaron el funcionamiento de varios buscadores mediante tres preguntas. Para ello procedieron a identificar recursos de información que ellos esperaban que los motores serían capaces de identificar, basándose en la experiencia de los autores, más que en una evaluación cuantitativa de la exhaustividad. Su comparación de CUIW3 Catalog, Harvest, Lycos, OpenText, WebCrawler, W3Worm y Yahoo mostró que sólo Lycos y Opentext Index identificaban toda la lista de recursos esperados.

Leighton y Srivastava [26] plantearon consultas de diferente dificultad, comparando cinco motores de búsqueda: Altavista, Excite, Hotbot, Infoseek y Lycos. Usaron para ello diez preguntas planteadas en el servicio de referencia de una biblioteca universitaria y añadieron cinco preguntas adicionales de otros estudiantes. Aplicaron sus propios juicios de relevancia, penalizó enlaces inactivos y compararon los resultados utilizando varias medidas de eficacia basadas en la relevancia, pero no usaron la exhaustividad porque consideraban prácticamente imposible determinar el número total de páginas relevantes a una pregunta concreta existente en la W3. Además, efectuaron un riguroso análisis estadístico de todos los datos que se presentaban. En este caso. fueron Altavista, Excite e Infoseek los que ofrecieron resultados más relevantes. En 1999 los autores publican nuevamente este estudio sin introducir revisiones ni modificaciones al método que habían propuesto con lo que demuestra su plena vigencia[27].

Por otro lado, Chu y Rosenthal[6] reutilizaron las medidas propuestas por Cleverdon (cobertura, exhaustividad, precisión, tiempo de respuesta, esfuerzo del usuario y formato de presentación de los documentos), adaptadas al entorno de la web, considerando que siguen siendo válidas casi cuatro décadas después de enunciarse. Usaron diez preguntas de referencia reales de diferentes niveles de complejidad, obtenidas a partir de consultas planteadas en bibliotecas universitarias y seleccionadas con el fin de analizar las siguientes características de los servicios de búsqueda: tiempo de respuesta, precisión (calculada para los primeros 10 documentos resultantes de cada consulta), opciones de presentación de los resultados, documentación del sistema e interfaz, capacidades de búsqueda como prestaciones que los distintos SB poseen a la hora de recuperar información y operadores de búsqueda. La eficacia del sistema fue medida a partir del tiempo de respuesta y la precisión usando las valoraciones de relevancia binaria (sí/no) de los registros recuperados. Tampoco consideraron la exhaustividad, por las mismas razones que Leighton y Srivastava [26]. Su estudio, se caracterizó por el escaso número de Sistemas de Búsqueda (Altavista, Lycos y Excite) en el cual, concluyeron que Altavista es la mejor elección para usuarios que necesitan alta precisión, mientras que factores tales como documentación e interfaz de usuario pueden estar basados en preferencias personales.

Gordon y Pathak[18] aportan otras conclusiones dirigidas a exponer que los índices de los motores poseen tamaños muy diferentes (algunos son diez veces más grandes que otros) y (aunque algún motor, en su publicidad afirme lo contrario), ninguno pretende realmente indexar toda la web. Asimismo, los motores también difieren en la actualización periódica de los datos, en la posibilidad de que los usuarios añadan páginas por su cuenta, en el plazo de tiempo para incorporar una nueva página indexada tras tener noticia de su existencia y en el seguimiento de la disponibilidad de los enlaces. De igual manera estudiaron el grado de solapamiento existente entre los documentos recuperados y los documentos recuperados considerados relevantes. Es decir, su estudio se asentó sobre el rendimiento en la recuperación de varios buscadores basándose en las medidas de la exhaustividad y precisión para los veinte primeros documentos recuperados, si bien utiliza una técnica de evaluación que extrapola esos resultados a doscientos documentos para cada SRI analizado. Los autores analizan ocho buscadores (Altavista, Excite, InforSeek, Open Text, Hotbot, Lycos y Magellan) utilizando treinta y tres preguntas, y realizando diversos tests. De forma global Altavista y Opentext obtuvieron los mejores resultados frente a Yahoo como el peor situado.

En esta misma línea encontramos un completo estudio de Ming[35], donde contempla el análisis de diversos parámetros tales como precisión, el tiempo de respuesta del sistema, la interface de usuario, el número de aciertos e introduce un factor que denomina sensibilidad, factor sobre el cual incide la calidad de los enlaces devueltos por el motor de búsqueda. Por desgracia, este estudio sólo afecta a Yahoo, Alta Vista y Lycos, y ofrece como resultados más destacados el hecho de no encontrar diferencias significativas en el número de documentos devuelto por cada sistema; prefiere el interface y el tiempo de respuesta ofrecido por Yahoo (aunque reconoce que es una opinión subjetiva). En lo relacionado con el análisis de la precisión, destaca que el valor medio de los tres motores evaluados en los primeros diez documentos devueltos es ligeramente superior al obtenido cuando sólo se toman en cuenta los cinco primeros documentos devueltos. En relación al parámetro de la sensibilidad, Ming afirma que Yahoo supera a los otros dos, resaltando los pobres valores de Alta Vista.

Desai[12] comparó trece motores de búsqueda mediante una sola pregunta(su propio nombre) y aunque tenía en consideración 24 documentos relevantes, no calculó en su

estudio exhaustividad, ni precisión; sólo se limitó a ofrecer directamente un recuento de las referencias obtenidas.

Tomaiuolo y Packer [62], realizaron uno de lose estudios mas extensos en cuanto al número de de consultas empleadas, donde se consideraron más de 200 temas. Las consultas fueron planteadas a Magullan Point, Lycos, Infoseek y Altavista teniendo en cuenta las primeras referencias resultantes de cada consulta, para los que determinaba su relevancia. Altavista era el que ofrecía los mejores resultados.

Westera [52] también ha estudiado la precisión de varios motores de búsqueda, destacando a Alta Vista, Lycos e Infoseek en su estudio, donde también ha introducido juicios sobre la calidad de los enlaces suministrados y sobre el crecimiento de los índices.

Wishard [54] lleva a cabo un estudio de la precisión de distintos motores a los que pregunta cuestiones relacionadas con la Geología y no encuentra diferencias significativas entre todos ellos.

Ljosland[30], realiza otro interesante y completo estudio para calcular la precisión en veinte motores de búsqueda sobre los que realiza diez búsquedas simples utilizando términos de poco uso (él los denomina "raros", como es el caso de, por ejemplo, la palabra haliography" o la palabra "peleidou"). En sus resultados indica que los motores que más documentos recuperaron fueron Fast, AskJeeves y Northern Light; en cobertura destaca a AskJeeves, Fast, Excite y Northern Light; en precisión Euroseek y WebCrawler obtienen el máximo valor (100%) y Lycos obtiene un promedio del 95%; por último, en lo relacionado con la exhaustividad, el orden es AskJeeves, Fast y Excite. Este mismo autor lleva a cabo posteriormente otro estudio comparando únicamente tres motores: Alta Vista, Google y AlltheWeb, siendo el primero un motor de consolidada posición y reconocido prestigio y los otros dos nuevos proyectos que buscan hacerse un hueco dentro de este amplio conjunto. Ljosland introduce una variante sobre estudios anteriores al considerar la posibilidad de que un documento sea "relevante parcialmente", en lugar del binomio tradicionalmente empleado de "relevante/no relevante". El experimento arroja como resultado que, cuando no se considera la relevancia parcial, se obtiene un valor de 0.4 para Alta Vista, 0.7 para Google y de 0.4 para AllTheWeb, cuando se considera la

relevancia parcial suben un poco todos los valores anteriores: 0.5 para Alta Vista, 0.9 para Google y 0.5 para AllTheWeb. Por tanto, Google se presenta como el mejor motor de los tres analizados frente a la posibilidad de encontrar un documento relevante en primer lugar de la lista de documentos devueltos.

Gwizdka, J. and Chignell[19], ponen en duda la tradicional asignación binaria de los valores de la relevancia (documento relevante-documento no relevante), ya que un documento difícilmente será relevante o no relevante en términos absolutos. En su estudio analizan un considerable número de aspectos: la precisión (entendida en los términos de la función definida por Leighton y Srivastava, el alineamiento de los documentos (estableciendo una función diferencial de la precisión conforme aparezcan alineados), el esfuerzo del usuario, la longitud esperada de búsqueda, el número de enlaces erróneos y el número de enlaces duplicados. Como resultados globales de este experimento, Alta Vista presenta mejores resultados que los otros dos motores analizados, en términos de precisión y de diferencial de precisión. El solapamiento es bajo, fundamentalmente porque los motores "emplean diferentes procedimientos de localización de los documentos y porque sus métodos de recopilación e indización son sustancialmente distintos". El experimento no detecta efectos significativos de la precisión entre los motores y los dominios geográficos analizados, siendo Alta Vista el que presenta mejores índices de cobertura en todos ellos.

Notess[36], dentro del campo de los estudios experimentales, proporciona cálculos de solapamiento, acierto único y enlaces fallidos. El solapamiento detectado es ligeramente superior que en el estudio de Gordon y Pathak, debido al crecimiento de los índices y al alto número de motores analizados. Notess, analiza ocho motores de búsqueda (Alta Vista, Excite, Northern Light, Google, Hotbot, AlltheWeb (Fast), MSN (Inktomi), y Anzwers) y obtiene como resultados, que Altavista, resulta ser ligeramente el motor cuyo porcentaje de enlaces fallidos es el mas elevado, seguido por Excite y MSN cuyos porcentajes son 13,7%, 8,7% y 5,7% respectivamente.

M.ª Dolores Olvera[[37], quien toma en consideración el parámetro de exhaustividad y precisión como elemento de importancia para la evaluación de los sistemas de recuperación de información. Su estudio se enfoca sobre 10 motores de búsqueda, que

en términos generales, establece que los 3 mejores en ranking de rendimiento se destaca Excite, Infoseek, y Hotbot.

Martínez F[34]. destaca su gran aporte, al evaluar 6 motores de búsqueda entre ellos tenemos Alltheweb, Altavista, Excite, Google, Nothern Light y Microsoft Network sobre los cuales analiza enlaces erróneos y duplicados, evalúa juicio de relevancia, aplica criterio de normalización de precisión y exhaustividad, determinación de similitud y agrupamiento. En su estudio destaca, a Google como el motor de búsqueda que mejor comportamiento experimenta en lo relacionado con la efectividad de la recuperación de la información, así como el que presenta el comportamiento mas estable en relación a la modalidad de búsqueda, tamaño de búsqueda analizada lo cual constituye una garantía de rendimiento eficiente para los usuarios y justifica el enorme éxito y popularidad del mimo.

Estudios Divulgativos

Dentro de este tercer grupo, se recoge un amplio conjunto de evaluaciones realizadas por revistas divulgativas especializadas en Informática y que poseen una gran aceptación por parte de los usuarios de Internet. En ellos se desprenden dos factores importantes como son el contacto directo con los usuarios con las ventajas inherentes que se desprenden de esa cercanía y, en segundo lugar, la actualidad de sus estudios, que suelen repetirse cada cierto tiempo y permiten recoger en ellos la presencia de nuevos motores que se vienen desarrollando. El enfoque de estos estudios no vienen avalados por una elaborada experimentación científica previa, aunque no por ello, su valor va a ser menor que cualquiera de los estudios englobados en los dos apartados anteriores. Básicamente su información va orientado a ofrecer una serie de sugerencias sobre el motor más apropiado a elegir dependiendo del tipo de búsqueda que deseemos realizar.

Algunos representantes de este grupo informativo sobre sistemas de búsqueda tenemos a C|net, empresa que viene estableciendo rankings entre los distintos motores. En su estudio de julio de 1999 establecía el siguiente orden de preferencias: Hot Bot, Alta Vista, Excite, Infoseek y Lycos; en cambio, en junio de 2000 modifica sustancialmente sus resultados al establecer el siguiente orden[7]: Google, Yahoo, Microsoft Network, Alta Vista, Lycos y Netscape Netcenter.

La Web Harlingen College del Instituto Técnico de Texas (26), ofrece una serie de sugerencias sobre el motor más apropiado a elegir. En éste se exaltaa Alta Vista como el mejor para labores de investigación inicial gracias a su alto número de documentos indexados.

Sherman [59], también mantiene en websearch.about.com una serie de interesantes documentos relacionados con los motores de búsqueda, y entre ellos, destaca la guía "How to choose the best general-purpose search tool" donde nos indica qué motor es, según sus criterios y estudios previos, el mejor para distintas situaciones de búsqueda.

En sí existen una variedad de publicaciones y recursos web relacionados con los aspectos comerciales, de gestión y financiación de los servicios de búsqueda. Algunas revistas de amplia difusión, centradas en el análisis de diversos aspectos sobre tecnologías de información, han incluido artículos sobre buscadores entre ellas encontramos PC World, PC Magazine, PC Computer, PC Week, entre otras. No obstante hoy en día se ha abierto a nuevos medios de proliferación relacionados con variedad de investigaciones sobre Sistemas de Recuperacón de Información como JASIS e Information Processing and Management, aunque también en Aslib Proceedings, Electronic Library, Computers in Library, y en el terreno español en El Profesional de la Información o la Revista Española de Documentación Científica(CINDOC-CSIC) o bien en páginas web a cargo de bibliotecas de gran reconocimiento, usuarios individuales o las empresas propietarias de los motores de búsqueda.

OTROS PARAMETROS A TENER EN CUENTA

El porcentaje de enlaces inactivos, constituye en uno de los factores más importantes a tener en cuenta a la hora de evaluar a los distintos motores de búsqueda, teniendo presente que esto incide en la percepción negativa que un usuario de un sitio web puede alcanzar si en el mismo proliferan los enlaces negativos.

Otro parámetro es la frecuencia de actualización del índice de cada motor de recuperación, valor que tiende a ser grande a medida que crece el tamaño del mismo. Ese aumento también puede ser un factor a considerar, aunque no tan determinante como los

anteriores, sin embargo esto incurre a que éste adquiera cierto grado de popularidad debido a su alcance de información documental que puede influir de alguna manera en la escogencia como herramienta de búsqueda.

El tiempo de respuesta, es otro de los parámetros de gran importancia a la hora de evaluar un sistema de búsqueda. "La habilidad de un sistema de recuperación de información es precisamente obtener rápida y eficientemente la información requerida por los usuarios. Subsecuentemente cuanto más difícil sea y más tiempo demande su recuperación, menos se preferirá usarlo"[62]. Por tanto, un sistema de búsqueda tenderá a no ser usado en cualquier momento, cuando se torne problemático para el usuario obtener información de éste.

Otros factores considerados en algunos análisis de este tipo son el solapamiento entre los motores - coincidencias de documentos recogidos por los distintos motores y la referencia única - es decir, cuándo un documento aparece recogido exclusivamente por un motor de búsqueda.

Relevancia o Pertinencia. Relevancia es en realidad una medida abstracta que cuantifica cómo un documento recuperado satisface una determinada consulta. Idílicamente un sistema debería recuperar todos los documentos relevantes, pero desafortunadamente esta es una medida subjetiva y difícil de cuantificar. Un documento puede considerarse relevante si el contenido del mismo posee alguna significación o importancia con motivo de la pregunta realizada por el usuario, es decir con su necesidad de información. La relevancia queda asociada con el concepto de la relación existente entre los contenidos de un documento con una temática determinada. Por otro lado el término de Pertinencia hace referencia al punto de vista del usuario final que realiza una operación de recuperación de información, asociada a la relación de utilidad existente entre un documento recuperado y una necesidad de información concreta e individual. Un documento pertinente es aquel que añade nueva información a la previa del usuario y que le resulta útil.

Este parámetro de evaluación ha sido tema de grandes discusiones entre varios investigadores[6]: Según Lancaster, 1993[24] señala que un mismo documento puede parecer relevante o no a la misma persona en momentos diferentes de tiempo. Por otro lado Blair, 1990[2] dice que el concepto de la relevancia, puede estar afectado de una gran dosis de subjetividad, por tanto, dentro del contexto de una búsqueda en un sistema de recuperación de información este concepto puede ser precisado de muchas maneras distintas por todos aquellos que realicen búsquedas.

Para algunos autores, surge entonces el concepto de "relevancia parcial", debido a que, en realidad, la relevancia no puede medirse en términos binarios (sí/no), sino que puede adquirir muchos valores intermedios (muy relevante, relevante, escasamente relevante, mínimamente relevante, etc.), lo que propicia que la relevancia pueda medirse en términos de función continua en lugar de una función binaria (que sólo admite dos estados).

Debido a todos estas aserciones y condiciones reflejados a cerca de la viabilidad de la relevancia para constituirse en un criterio de evaluación de la recuperación de la información. Cooper, 1973[8] plantea una manera diferente de definir a la relevancia, él manifiesta a ésta en términos de la percepción que un usuario posee ante un documento recuperado, es decir: "si el mismo le va a ser útil o no". Así, a partir de este punto de vista, posibilita asumir que un usuario tendrá problemas a la hora de definir qué es relevante y qué no lo es, pero tendrá pocos problemas a la hora de decidir si el documento le parece o no útil.

La importancia del concepto de "utilidad" lleva a Blair, 1990[2] a concluir que este término "simplifica el objetivo de un sistemas de recuperación de información y, aunque su evaluación es subjetiva, es posible medirla de mejor manera que si no se aplica este criterio, en tanto que es una noción primitiva que denota la realización de una actividad"

Por otro lado, Frants, 1997[15] plantea la utilidad en términos de "eficiencia funcional", un sistema de recuperación de información, alcanzará altos niveles de este valor cuando la mayoría de los documentos recuperados satisfagan la demanda de información del usuario, es decir, le resulten útiles.

[6] Fuente: http://irsweb.blogspot.com/2004_10_01_irsweb_archive.html

Por su lado Gerard Salton, 1983[41] considera que el conjunto pertinente de documentos recuperados puede definirse como "el subconjunto de los documentos almacenados en el sistema que es apropiado para la necesidad de información del usuario"

Finalmente se puede concluir que la aplicación de este criterio como acción en la recuperación de la información, debe ser siempre enfocada a que todo documento será considerado relevante ante una necesidad informativa siempre y cuando su contenido aporte información útil. Por tanto el término de Relevancia siempre va ligada al término de Pertinencia.

Exhaustividad y Precisión. El término de Exhaustividad mide la capacidad del sistema para recuperar documentos útiles, mientras que la precisión mide la habilidad de rechazar material no relevante. La precisión se halla dividiendo el número de documentos relevantes recuperados entre el número de documentos recuperados. La exhaustividad se calcula dividiendo el número de documentos relevantes recuperados entre el número de documentos relevantes existentes en la colección.

El par de medidas Precisión-Exhaustividad, han sido utilizadas por muchos investigadores como Lancaster[23], Oppenheim[60], Chu-Rosenthal[6], Gordon[18], Francisco Martinez[34] entre otros. No obstante, es importante destacar que hallar el parámetro de precisión genera pocos problemas, excepto cuando en una determinada búsqueda no se recupera ningún documento. La exhaustividad en cambio presenta muchas complicaciones, incluso cuando se trata de pequeñas colecciones, ya que hallar este parámetro, que es inevitablemente un valor relativo, requiere que cada documento de la colección sea contrastado en relación con cada consulta sobre un tema determinado, esto por tanto, exige el conocimiento del número total de documentos relevantes de la colección con respecto a la pregunta. Ante la imposibilidad de determinar el volumen de documentos que tiene una base de datos sobre un tema en particular, diferentes especialistas han propuesto métodos alternativos que tratan de salvar la dificultad para calcular este valor.

Lancaster [24] utiliza en dos ocasiones un método de muestreo para estimar el índice de exhaustividad de una gran base de datos, es decir, midiendo la relevancia de un conjunto de documentos de la colección.

Salton[40] por su parte, propone el mismo procedimiento y sostiene que la exhaustividad no es un valor exacto, sino una estimación del número total de documentos relevantes de la colección: La valoración de la relevancia se hace en base a un subconjunto de documentos de la colección. Alternativamente, una consulta dada puede .ser procesada por una variedad de diferentes métodos de búsqueda y recuperación, dando por supuesta que todos los documentos relevantes van a .ser recuperados por medio de dichas búsquedas. Los resultados se combinan entonces en una única lista de resultados. La lista de documentos relevantes .se obtiene mediante la valoración de la relevancia de esa lista de resultados.

Van Slype[57] considera dos posibilidades para determinar la exhaustividad: ya sea sistemáticamente, examinando las referencias una a una (lo que tiene el riesgo de durar mucho), o bien interrogando de nuevo con una serie de ecuaciones muy amplias (con pocos o ningún operador Y, con muchos operadores O).

Matriz de Similaridad. La función de similitud sirve para determinar que tanto se parecen dos elementos. Los coeficientes o índices de similitud más usuales consideran dos conjuntos con un cierto nivel de intersección (Figura 1) La similitud entre ambos conjuntos depende siempre del tamaño de esa intersección bien respecto del tamaño total de los dos conjuntos o bien de parte de ellos.

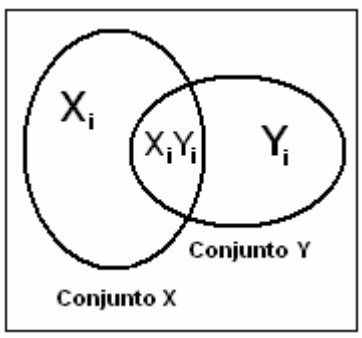


Figura1. Intersección de conjuntos similares

Para un modelo de escalamiento multidimensional, [Salton, 1981], define la fórmula del coseno como el ángulo formado entre dos vectores asociados X y Y como se expresa a continuación:

$$\text{Cosine}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)*(\sum_{i=1}^{n} y_i^2)}}$$

Expresión1. Función del coseno

Con ella se trata de medir el grado de semejanza que existe entre cada vector. Su aplicación es muy extendida en el campo de la recuperación de la información y es considerada la mas adecuada debido a que genera una menor dispersión y una mayor compactación de los resultados[7].

Los valores de similaridad vienen expresados en una escala (0,1) en el que los valores más cercanos al 0 expresan una gradual disimilaridad, mientras que los más cercanos al 1 demuestran un mayor grado de similitud, hasta el punto de llegar a un punto de similaridad total donde S(xi,yi)=1. Por otro lado, valga decir que al tratarse de un producto escalar, o sea, una multiplicación entre dos vectores, el orden no altera el producto.

Francisco J. Martinez[8], explica con más detalle el modelo de espacio vectorial, propuesto por Salton, (función de similitud del coseno existente entre dos vectores de resultados). La dimensión del espacio vectorial es igual al número total de términos utilizado en la caracterización del problema (en este caso, el número total de respuestas del cuestionario). En el modelo se hace la suposición básica de que la distancia relativa entre dos vectores en el espacio n-dimensional, representa la diferencia entre los perfiles que se

7 SALTON, G., MCGILL, M. (1981). Introduction to Modern Information Retrieval. New York: McGraw-Hill.Fuente: http://internetlab.cindoc.csic.es/cv/11/ANALISIS_DE_REFERENCIAS_BASADO_EN_UN_MODELO_DE_ESPACIOS_VECTORIALES_LA_INVESTIGACION_EN_HISTORIA_CONTEMPORANEA_EN_JAEN_DURANTE_1990_1995.PDF

8 Francisco Javier Martinez,Técnicas y Métodos Avanzados de Recuperación de Información. Universidad de Murcia. Fuente: http://www.um.es/~gtiweb/fjmm/tmari/tmari-prac4.pdf

han utilizado para configurar dichos vectores. La función de similitud, función coseno viene dada por la expresión1, anteriormente mencionada.

Donde X es el vector problema de un caso almacenado en la Base de Casos y Y es el vector problema del nuevo caso. El resultado de la función Coseno proporciona valores entre 0 y 1. En el caso de que el valor de Cos (X,Y) fuese 1, indicaría que ambos casos (X,Y) son idénticos y por lo tanto que los dos motores vectores de resultados son similares. Consideremos las componentes del vector X = {1; 1; 1; 1; 0; 0; 0; 0}, y a continuación, las componentes del vector Y = {0; 0; 1;1; 0; 0; 1;1}. El cálculo de la función de similitud del coseno queda reflejado en la siguiente tabla:

X	Y	ESCALAR	X^2	Y^2
1	0	0	1	0
1	0	0	1	0
1	1	1	1	1
1	1	1	1	1
0	0	0	0	0
0	0	0	0	0
0	1	0	0	1
0	1	0	0	1
	Sumas	2	4	4
	Coseno	0,5		

Es importante, comentar que la del coseno no es la única función de similitud. Existen otras, entre las que destacan las de Dice y Jaccard, pero que pueden resultar algo más complejas no sólo de calcular sino más bien de interpretar y que por tanto son menos aplicadas en la Recuperación de Información[9].

Matriz de Distancias. Una vez calculadas las similitudes pasamos a representar las distancias que existen entre cada vector en este espacio n-dimensional. Para ello transformaremos las similitudes en distancias, restándoles 1 a las similitudes. De esta forma tendremos cuanto difieren unos autores de otros[10].

Según una publicación presentada por francisco J. Martínez[11], un concepto muy relacionado con Similitud es el de Distancia, tan relacionado como que dos vectores distantes son aquellos que poseen entre ellos un escaso valor de similitud. De hecho, la

[9] Fuente: http://web.udl.es/dept/dal/sepln/sepln99.ppt#303,27,Medidas de similitud.
[10] Fuente: http://internetlab.cindoc.csic.es/cv/11/ANALISIS_DE_REFERENCIAS_BASADO_EN_UN_MODELO_DE_ESPACIOS_VECTORIALES_LA_INVESTIGACION_EN_HISTORIA_CONTEMPORANEA_EN_JAEN_DURANTE_1990_1995.PDF
[11] Fuente: http://irsweb.blogspot.com/2005/03/el-modelo-del-espacio-vectorial-ii_22.html

distancia para muchos autores es equivalente a la disimilitud, es decir |Distancia| = 1 - |Similitud|.

Proceso Agrupamiento o Clustering. Las técnicas de agrupamiento se basan fundamentalmente en el concepto de similitud (o disimilitud) entre ejemplos y agrupaciones. Muchas veces se utilizan métricas (o distancias) para medir la similitud entre ejemplos. Existen muchas técnicas de agrupamiento, las más usadas son las de agrupamiento jerárquico y las de agrupamiento dinámico que a su vez pueden ser por reunión(aglomeración) o separación(división).

En nuestro estudio nos centraremos en los Algoritmos jerárquicos aglomerativos[12], estos algoritmos producen una sucesión de conglomerados de tal manera que en cada paso el número de conglomerados va disminuyendo, es decir funcionan de manera que los elementos existentes son asignados sucesivamente a grupos. El algoritmo AGNES (Agglomerative Nesting) que está dentro de está categoría, se explica a continuación:

Este algoritmo construye una jerarquía en forma de árbol que contiene implícitamente todos los valores de k, comenzando con N conglomerados y siguiendo con fusiones sucesivas hasta obtener un sólo conglomerado con todos los objetos [10]. Podríamos esquematizar el funcionamiento del algoritmo como sigue:

1) Si n es el número de elementos se comienza con tantas clases como elementos. Las distancias entre clases son las distancias entre los elementos originales.
2) Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos una clase.
3) Sustituir los dos elementos utilizados en (2) para definir la clase por un nuevo elemento que represente la clase construida. Las distancias entre este nuevo elemento y los anteriores se calculan con cualquiera de los criterios que comentamos a continuación.
4)) Volver a (2) y (3) hasta que tengamos todos los elementos agrupados en un una sola clase.
5) Cortar el árbol donde se considere conveniente.

[12] Jaimes Luis Gabriel. Uso de técnicas de clasificación en conglomerados para describir perfiles en grandes bases de datos educativas, Universidad de Puerto Rico Mayagüez Campus, 2004, pp. 32-37.

Para el enfoque aglomerativo, hay diferentes medidas de proximidad entre conglomerados estás se derivan de varias estrategias de fusión. Estas son conocidas como: encadenamiento simple(o vecino más próximo) encadenamiento completo (o vecino más lejano), encadenamiento de media de grupos, encadenamiento de "ward", y encadenamiento del centroide[11].

Ejemplo del funcionamiento del AGNES: Supóngase que se tiene la siguiente matriz de distancias, de un conjunto de cinco elementos con dos variables, y utilizando la métrica euclidiana.

	a	*b*	*c*	*d*	*e*
a	0	2	6	10	9
b	2	0	5	9	8
c	6	5	0	4	5
d	10	9	4	0	3
e	9	8	5	3	0

El primer paso consiste en escoger los dos objetos más cercanos o más similares, es decir, donde la distancia sea más pequeña y fusionarlos, excepto los de la diagonal. Cómo se puede ver, el más pequeño es dos, así que unimos a y b para formar el conglomerado {a, b}. En este primer paso se formaron los conglomerados {a, b}, {c}, {d}, {e}. En el siguiente paso se unen los dos más cercanos, pero ahora para medir las distancias no lo tenemos que hacer de objeto a objeto sino entre conglomerados. Se usará entonces el Encadenamiento de media de grupos, discutido anteriormente.

Las distancias de los nuevos conglomerados serán:

$$d(\{a,b\},\{c\}) = \frac{1}{2}\left[d(a,c) + d(b,c)\right] = 5.5$$

$$d(\{a,b\},\{d\}) = \frac{1}{2}\left[d(a,d) + d(b,d)\right] = 9.5$$

$$d(\{a,b\},\{e\}) = \frac{1}{2}\left[d(a,e) + d(b,e)\right] = 8.5$$

Se puede construir una nueva matriz de distancias (disimilaridades) entre los cuatro conglomerados {a, b}, {c}, {d}, {e}. La matriz de distancia es:

	{a,b}	{c}	{d}	{e}
{a, b}	0	5.5	9.5	8.5
{c}	5.5	0	4	5
{d}	9.5	4	0	3
{e}	8.5	5	3	0

Continuando con el procedimiento se buscan los más similares y puede verse que la distancia
entre {d} y {e} es la entrada más pequeña de la matriz. Los cálculos de distancia para los nuevos conglomerados son:

$$d(\{d,e\},\{c\}) = \frac{1}{2}[d(d,c) + d(e,c)] = 4.5$$

$$d(\{d,e\},\{a,b\}) = \frac{1}{4}[d(d,a) + d(d,b) + d(e,a) + d(e,b)] = 9.0$$

De nuevo se tiene otra matriz de distancias ahora con {a,b},{c}, {d,e}, como se muestra a continuación.

	{a,b}	{c}	{d,e}
{a,b}	0	5.5	9.0
{c}	5.5	0	4.5
{d,e}	9.0	4.5	0

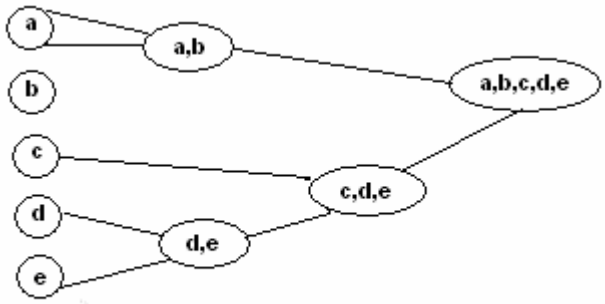
El termino más pequeño ahora es 4.5, luego podemos fusionar {d, e} y {c}, el cálculo de distancias entre los conglomerados resultantes es:

$$d(\{c,d,e\},\{a,b\}) = \frac{1}{6}[d(c,a) + d(c,b) + d(d,a) + d(d,b) + d(e,a) + d(e,b)] = 7.83$$

Lo que conduce a la siguiente matriz de distancias como se muestra a continuación:

	{a,b}	{c,d,e}
{a,b}	0	7.83
{c,d,e}	7.83	0

El paso final consiste en unir los dos últimos conglomerados en uno sólo. La siguiente Figura ilustra el proceso de fusión que sufrieron los elementos.



Proceso de fusión con el algoritmo AGNES

3. PLANTEAMIENTO DEL MODELO DE EVALUACIÓN DE LOS SISTEMAS DE RECUPERACIÓN DE LA INFORMACIÓN EN LA WEB.

La dirección de la presente propuesta de evaluación, se centra básicamente en un modelo de análisis multidimensional, orientado hacia el establecimiento de un marco inductivo, que permita identificar con base en un proceso de selección de criterios de muestra previamente establecida, la investigación de la viabilidad del funcionamiento de los motores de búsqueda; es decir, se procederá a la evaluación de la efectividad de los motores de búsqueda en Internet mediante la aplicación de una metodología determinada de análisis, así, visualizar unos motores frente a otros, con la idea de concebir un planteamiento integral de evaluación en los sistemas de recuperación de información en la web.

Es importante, hacer notar, que debido a la constante evolución de estas herramientas o servicios de búsqueda, perfectamente documentada en investigaciones mencionadas en el apartado anterior, los datos que arrojaron las búsquedas tienen hoy, al igual que en la mayor parte de análisis de este tipo, un valor "histórico". Por ello, los resultados que en este estudio se plantea no busca ayudar a un usuario a elegir el buscador web a utilizar; ya que en realidad esta no es en ningún momento la finalidad de la investigación, la cual está desarrollada básicamente en el establecimiento y/o seguimiento del uso de algunos parámetros usados por anteriores autores, de tal forma explicar en base a este estudio de campo la metodología llevada a cabo y la correspondiente aplicación de este método de evaluación que permita dar un análisis sujeto a los resultados obtenidos.

El método de evaluación se organizó en torno a varias etapas que permitieran establecer paso a paso el proceso llevado a cabo para el mismo; se previeron posibles medidas que posibiliten contrarrestar el ruido comúnmente identificado en un entorno tan dinámico y disperso como es la World Wide Web, posibilitando la consistencia y validez de las especificaciones concluyentes a obtener en la efectividad manifiesta en estos sistemas de estudio y finalmente a partir de las necesidades de información planteadas, se elaboraron las ecuaciones de búsqueda mediante la sintaxis que se consideró adecuada para fines de evaluación y comparación de los sistemas de recuperación en la web.

El esquema de evaluación propuesto fue desarrollado a partir de conocimientos teóricos de anteriores investigaciones y de un marco de trabajo empírico (experimental), emprendido para probar su viabilidad.

El análisis de los buscadores generales más sobresalientes de la World Wide Web se realizó con el objeto de ilustrar el proceso de evaluación y no para determinar cuál es el mejor buscador de Internet, ya que tanto los Sistemas de búsqueda como la propia información, son muy dinámicos en el entorno Web y las conclusiones a las que se pueden llegar son poco perdurables.

3.1 FASES DE DESARROLLO PARA EL ESTUDIO DE CAMPO:

Antecedentes:

Las investigaciones realizadas por diversos autores, han revelado características particulares en cuanto a la metodología de evaluación aplicada. Se ha observado evaluaciones con fines de simple interrogación claramente examinado por Desai[12] y Leighton- Srivastava[26] cuyo uso puede llevar a un sesgo o falta de imparcialidad en los resultados; otros el desarrollo de su estudio fue en colaboración con asesores externos, como proyectos TREC[13] publicados por Harman D. K.; otros enfocaron sus ecuaciones de consulta en base a usuarios reales de información (estudiantes, referencias de biblioteca) como Chu-Rosenthal [6] y Leighton[61]; Otros investigadores como Leighton[27], Winship[55] señalan la manera de interrogar estos sistemas (ya que de esto depende el éxito o fracaso de las pruebas): preguntas en la que se encuentren recursos en la web, establecer preguntas "fáciles", con un alto nivel de respuesta y "difíciles", con resultados más restringidos en relación con la cantidad de documentos que sobre éstos se puedan recuperar; que unas preguntas sean de temas académicos y/o especializados y otras de temas más habituales y que se trate con temas diversos. Por otro lado, otros manejan un número de preguntas bastante reducida como Desai[12] con una sola pregunta, Courtois-Baer-Stark[9] manejó tres preguntas y Ding- Marchionini[13] usó cinco, lo que resulta muy insuficiente para estas pruebas; en contraste esta Tomaiuolo,-Packer[62] cuyo estudio se basó en 200 preguntas. En esta monografía se pretende evaluar un número de dieciocho

[13] TREC(Conferencias de Recuperación de texto)

preguntas aunque sigue siendo un número no tan ideal para este tipo de investigaciones donde se requiere un estudio mucho mas exhaustivo para llegar a análisis mas definitivos y certeros; esta muestra, no obstante resulta ser adecuada y representativa para este estudio, y permitirá observar los alcances en cuanto a funcionamiento se refiere, proporciona cada motor de búsqueda.

3.1.1 SELECCIÓN DE SISTEMAS DE BÚSQUEDA A EVALUAR

Éstos deberán cumplir los siguientes requisitos:

a) Se manejarán buscadores generales de manera que su base de datos incluya información sobre los más variados temas.

b) Se hará uso de motores de búsqueda de carácter internacional, es decir, que no limita en la información a ninguna zona geográfica.

c) La popularidad de los sistemas de búsqueda, se ve reflejada ampliamente en la familiaridad de su utilización entre los usuarios y entre investigadores de Internet(revista electrónicas), para que la muestra fuera lo más representativa posible.

Los buscadores objeto de este estudio, fueron seleccionados con base en su popularidad. Se examinaron un total de 8 artículos publicados en revistas especializadas entre enero de 2000 y julio de 2005, donde se describían, comparaban o evaluaban las características y funcionamiento de diferentes servicios de búsqueda. Una vez realizado el análisis se encontró que los más estudiados eran: Altavista, Excite, Google, Hotbot, MSN Search, Lycos y Yahoo!, por cual son los que en este trabajo se han considerado. Es su popularidad durante el período de muestra la que justifica su inclusión, independientemente del servicio que en ellos se destaque.

3.1.2 DETERMINACIÓN DE NECESIDADES INFORMATIVAS A APLICAR EN LOS BUSCADORES SELECCIONADOS.

Para llevar a cabo el estudio se estableció de manera espontánea temas de cualquier índole, relacionados de una u otra forma en algunos casos con información en idioma español y en otros en idioma inglés. Esta decisión responde básicamente a dos motivaciones:

a) En primer lugar, poner en situación adversa a buscadores generales e internacionales, ya que el idioma español representa, y aún más en la fecha, una muy pequeña parte en Internet, y los sistemas de recuperación deben responder a preguntas relativas a temas no estrictamente relativos al idioma inglés siendo éste el de mayor predominio en la W3.

b) En segundo lugar, la voluntad de averiguar de cierta manera en las búsquedas, el comportamiento de estos motores en lo que respecta a la calidad y homogeneidad en los resultados, dada la magnitud y variedad de documentos incluidos en sus bases de datos.

Cabe mencionar que determinar la necesidad de información en forma explícita, con la que se pretende resolver una búsqueda en Internet, no es una tarea fácil, ya que la cantidad de temas que se pueden plantear pueden ser de carácter heterogéneo, debido a que pueden ser representativos de diversos tipos de búsqueda. Incluso pueden dar lugar a un nivel de respuesta potencialmente alto y otras con resultados más restringidos.

La propuesta de estudio, se determinó para llevar a cabo un total de 18 preguntas con el fin de analizar la eficacia en la recuperación de la información de algunos de los más reconocidos motores de búsqueda y de esta manera propiciar objetividad en el proceso de evaluación.

3.1.3 DESCRIPCIÓN DEL TIPO DE BÚSQUEDA MEDIANTE UNA SINTAXIS ESPECÍFICA.

Cada buscador presenta un motor de búsqueda y unas prestaciones de cierta manera diferentes debido al algoritmo(o robot) que manejen para la recuperación de documentos, éstas fueron presentadas en detalle en el apéndice C del documento, por ello, las ecuaciones de búsqueda entre ellos pueden diferir en algún momento, sin embargo en lo que respecta a manejo de operaciones booleanas, casi todos los motores ofrecen similares prestaciones. Por lo general los más utilizados (AND, OR, NOT) aunque algunos con mayor variedad que otros. No obstante, para contribuir a la homogeneidad de los resultados y posibilitar su comparación, se adoptaron una serie de criterios.

La primera decisión importante fue la de realizar las búsquedas en inglés y en español de manera mas o menos intercalada, ya que por un lado la primera es una lengua de uso mayoritario en Internet, lo que debía aumentar las posibilidades de encontrar información en las búsquedas planteadas y por otro lado el idioma español que de manera contraria, corresponde a una porción pequeña en Internet, precise comprometer al motor de recuperación responder a la magnitud y variedad de documentos que indexa. La idea de tratar ambos idiomas es contar con resultados de búsquedas más significativos.

La naturaleza de las preguntas demanda una sintaxis de búsqueda diferente (booleana, de frase literal, de un término o palabra, etc.,) y se escogió en cada caso la que resultaba intuitivamente más adecuada. Se ha optado por seleccionar la sintaxis y el método de funcionamiento del motor en formatos simples, que permitan recuperar donde aparezcan algunos de los términos que forman parte de la expresión de consulta, mediante el uso preferiblemente de operadores de exactitud (+ y -); por otro lado, se llevará a cabo la búsqueda avanzada, donde la búsqueda simple no permite realizar la búsqueda en los mejores niveles de búsqueda, haciendo uso de operadores booleanas, frase literal y en ocasiones se elegirá alguna opción específica del menú de búsqueda para plantear la consulta, como búsqueda de todas las palabras, búsqueda de frase, entre otras opciones que se requieran. Por diversas razones no se usó el truncamiento: por un lado porque algunos de los motores no ofrecen el uso de este servicio, o por otro lado el número de palabras a buscar era excesivo, y donde algunos motores se ven limitados a un número máximo de caracteres o realizan las llamadas búsquedas inteligentes o búsqueda por conceptos que ya incluyen esta posibilidad.

Las características de las consultas presentan las siguientes características: Con el fin de reflejar cierto grado de heterogeneidad, asimismo previsto en nuestra cotidianidad en el uso de estos sistemas de recuperación de información web, hemos supuesto abarcar diversos temas en las búsquedas:

- Especificación de temas generales y específicos.
- Formulación de términos tanto en idioma ingles como español.
- Se hace uso de lógica booleana, búsquedas por frase literal es decir, que la serie de términos especificados aparezcan juntos y en ese orden en el documento recuperado,

manejo de lenguaje natural, donde se formula la pregunta de una forma normal como si se estuviera estableciendo una comunicación con una persona.

- Se manejan términos únicos, nombres de enfermedad, entre otras especificaciones.

Intuitivamente se consideró la mejor formulación a cada pregunta. En el anexo 1. Se describen los términos de consulta y el tipo o criterio de consulta a manejar.

El número de preguntas a formular son 18, de los cuales se considerará los siguientes criterios:

- LN: Lenguaje Natural → Who is Elkin Patarroyo? (Pregunta:3)
- FL (Frase literal) ó BA (Búsqueda avanzada: se podrá elegir All of these words o todas las palabras o with exact phrase dependiendo del motor de búsqueda) → "Violencia Intrafamiliar" o All of these words: Violencia Intrafamiliar. (Preguntas:6, 8, 11, 13, 16)
- Inclusión/Exclusión (+/-) → +milky +way. (Pregunta:5,10)
- Operadores booleanos (AND ó OR): Se hará uso de operadores booleanos, es decir, se elegirá la opción más a fin al uso de este operador lógico dependiendo de la consulta a ejecutar → +causas AND maremotos OR tsunamis. (Preguntas: 7,15, 12, 7, 4)
- Se omitirá la funcionalidad de filtro por idioma, debido a que no todos los motores aquí en estudio lo permiten, y el hacerlo podría poner en desventaja entre ellos. Por tanto, el idioma que se maneje en cualquiera de las preguntas, se introducirá en el formulario de consulta de manera directa, con el único objetivo de potencializar el poder de documentos devueltos por cada uno de estos sistemas de recuperación web.
- En algunos casos se hará combinación de criterios (frase literal- Operador booleano). (Preguntas: 2, 9, 14, 18)
- Se determinará según la pregunta si el tema en consideración es evaluado como específico o general (ES/GE), característica de importancia durante el proceso de evaluación en el siguiente capitulo. (ES →Preguntas: 1, 3, 6, 9, 11, 12, 14, 15, 16, 17); (GE → Preguntas: 2, 4, 5, 7, 8, 10, 13, 18).

3.1.4 CRONOGRAMA DE LA REALIZACIÓN DE LAS CONSULTAS EN LOS SISTEMAS DE BÚSQUEDA.

Es claro que en el entorno web, las páginas se pueden modificar su contenido y ubicación en la red con facilidad. La evaluación de los buscadores se ve dificultada por el sistema de compilación de información que siguen, las características de sus motores de búsqueda y el carácter dinámico de sus bases datos, en constante mutación y crecimiento. Por tanto, para conseguir un estudio de estas características, que fuera realmente rigurosa, se empleo un intervalo de tiempo mínimo para realizar el análisis de los documentos recuperados.

En primer término, se evaluará la cobertura y el tiempo que dura cada motor en responder a una consulta dada. Se establece el rango de dos horas y media aproximadamente para llevar a cabo esta operación sobre los diferentes motores de búsqueda en estudio. Se realizaran dos muestras, donde se pone como periodo de la prueba dos meses aproximadamente de diferencia, la finalidad del mismo es evaluar, si al introducir la mismas 18 preguntas formuladas con una diferencia de 60 días aproximadamente, poder identificar si la proporción frente a los resultados obtenidos en la primera prueba se consigue en la segunda o existe una posible variación del mismo (es decir, identificar si los resultados del motor que mayor numero de paginas indizó en este primer periodo de tiempo se mantiene en la misma posición frente a los demás en el siguiente). Es importante tener presente que la cobertura se realiza en base al análisis de ecuaciones formuladas y en ningún momento se pretende obtener la cobertura total de la Base de datos que cada motor en realidad indiza debido a que el tipo de metodología llevada a cabo en cada uno de ellos no se especifica en forma detallada.

En segundo término, se establece un marco en concreto, en el que se considera un periodo de tres a cinco días (en periodos de 4h) como máximo, para las búsquedas de una misma pregunta, sobre cada motor. Esto incluye la ejecución de la pregunta y la correspondiente inspección de los documentos recuperados, alineamiento, duplicados, inactivos e irrelevantes; sobre cada documento se dará análisis al documento completo, con el fin de prevenir posibles modificaciones, eliminaciones o cambio de localización de las paginas recuperadas y hacer mas fiable el análisis. En sí, se pretende como

supuesto, la formulación de 18 preguntas en cada uno de los siete motores de búsqueda escogidos y mencionados en apartados anteriores; de igual forma se establece la relevancia de los primeros 30 resultados de cada consulta.

3.1.5 VALORACIÓN DE LA RELEVANCIA DURANTE EL ANÁLISIS DE LOS DOCUMENTOS RECUPERADOS.

Se considerará como documento relevante todo aquél que haga referencia sobre el tema de la pregunta, es decir, que responda a las necesidades de información tal y como habían sido expresadas. Será examinado y juzgado la relevancia de los documentos a texto completo, donde se considerará la siguiente valoración:

- Enlaces duplicados, enlaces inactivos o irrelevantes todos ellos con un valor de 0.

Con el objeto de obtener valores más reales en este proceso de evaluación, y evitar posibles particularidades presentes en el web se establece las siguientes consideraciones:

- Con respecto a duplicados, si el enlace en cuestión tiene una misma respuesta en una URL genérica y una URL específica, se lo considera duplicado, independientemente de sus otras cualidades (inactivas, irrelevantes o válidas). Los espejos (alias ó mirror sites), servidores idénticos que tienen direcciones IP o directorios diferentes, incluso cuando dos archivos son el mismo o versiones ligeramente diferentes, no se consideran como duplicados.
- Se consideran inactivos, Error 404[14]: el servidor ha sido contactado pero no se consigue localizar ese fichero, Error 601- 609: el servidor no responde, se comprueban los enlaces varias veces, por ejemplo, en un periodo de tres días a una semana, mensajes que indican que el acceso a la página está prohibido o que se necesita clave de acceso y finalmente mensajes que anuncian que la página deseada ha sido eliminada o trasladada a otro servidor.

- Enlaces técnicamente relevantes se considerarán con valor de 1. Es decir, el documento contiene los términos de consulta pero no en el contexto adecuado a la necesidad informativa, o bien se hace mención del tema de manera superficial.

[14] Mensajes de error: http://www.charitydays.net/support/faq/12.shtml

- Enlaces potencialmente útiles se les asignará un valor de 2. Hace referencia al documento con información sobre el algún aspecto de la consulta y con enlaces hacia éstos en el mismo sitio web.
- Enlaces probablemente más útiles, recibirán un valor de 3. Se refiere a un documento que ofrece un estudio general y/o más profundo sobre el tema conteniendo enlaces a diversos documentos alojados en distintos sitios en la web que profundicen el tema de búsqueda o además pueden se sitios que suministren información bibliográfica (libros o páginas web) que pueden servir de gran ayuda para ampliación del tema de investigación.

En seguida se da un ejemplo, de la forma como se llevará a cabo cada consulta.

1. Se plantea la pregunta

PREGUNTA: Dyslexia, Información sobre cuál es el origen del problema, o en qué consiste, como se detecta, quiénes la padecen y existencia de tratamientos.

2. Tipo de expresión a evaluar

TIPO	Búsqueda de una sola palabra, Información especifica.

3. Estimación de la relevancia

Relevancia	3	Documento que da información con profundidad sobre la raíz del problema de dyslexia y/o ofrece posibilidades de listado de páginas con contenido similar alojados en otros sitios web, como también información deposibles libros que pueden ayudar a ampliar aún mas nuestra curiosidad sobre el tema.
	2	Documento con información sobre algún aspecto del tema o con enlaces hacia otras paginas del mismo sitio.
	1	Documento que contiene los términos de consulta, pero su información es muy superficial.
	0	Página que no contiene todas las palabras de búsqueda o estan dispersas, pero no refleja para nada el tema de interés.

Tabla23. Ejemplo explicativo sobre los niveles de relevancia manejados en este estudio.

3.1.6 ANÁLISIS DE LOS RESULTADOS: "CRITERIOS DE EVALUACION"

Para cada uno de treinta resultados recuperados, sobre los cuales, cada uno de los motores suministrará, se identificará su condición de inactivo, duplicado o la puntuación de

relevancia, mediante la valoración anteriormente expuesta (0, 1, 2 ó 3). En forma resumida se pretende especificar claramente los criterios a evaluar.

- Criterio basado en el tamaño relativo del índice promedio, en base a las 18 preguntas previamente formuladas (dos muestras serán llevadas a cabo).

- Criterio basado en el tiempo de respuesta promedio que demora cada motor de búsqueda desde que se envía la pregunta hasta la entrega de resultados. Debido a que no todos los sistemas de búsqueda ofrecen esta información en su página de resultados, se resolvió hacer uso de un cronómetro de alta precisión para determinar de forma más homogénea el tiempo de respuesta sobre el cual, actúa cada sistema de búsqueda en base a cada consulta definida. Sin embargo, se deja para estudios posteriores, la investigación del cálculo y/o monitoreo del tiempo que tarda un sistema de búsqueda desde que envía la petición o consulta y accede al Servidor en búsqueda de la misma; y el tiempo transcurrido en producir y entregar los resultados al usuario. Este enfoque da garantía de percibir un tiempo de respuesta más real de cada sistema. Cabe anotar, que en este criterio, se llevará a cabo la formulación de 18 ecuaciones de búsqueda y dos pruebas las cuales serán realizadas en periodos de tiempo distintas.

- Criterio basado en el ruido documental, donde se recogerá información sobre el número de resultados inactivos, duplicados e irrelevantes presentes en cada motor, correspondientes a los 30 primeros resultados.

- Criterio según la relevancia de Prueba.

Para la realización de este criterio de evaluación se trató de valorar los documentos con la mayor objetividad posible. No obstante, expertos conocedores señalan, que la medida de la relevancia es una estimación subjetiva que depende del nivel de conocimientos de cada individuo. El problema señala Lancaster, 1993[15] sobre el hecho de que un mismo documento puede ser considerado relevante, o no relevante, por dos personas distintas en función de los motivos que producen la necesidad de información o del grado de

[15] Lancaster, F.W. and Warner, A.J. Information Retrieval Today. Arington Virginia:Information Resources, 1993. < http://irsweb.blogspot.com/2004/10/relevancia-o-pertinencia.html>

conocimiento que sobre la materia posean ambos. Sin embargo en este estudio aunque se establece un marco de referencia para estimar la relevancia (por niveles o grados), la percepción que cada persona pueda tener frente a un documento no deja de tener una valoración propia por lo que la coincidencia nunca puede ser el 100% exacta.

En esta experiencia, básicamente, se busca representar en dos gráficas. La primera se analiza el promedio de documentos relevantes con duplicados, para las 18 preguntas en tres primeras páginas de resultados (con el fin de ver el comportamiento global en base a esta muestra. Para ello se tuvo en cuenta los diferentes niveles de relevancia (0, 1, 2 y 3): documentos irrelevantes, poco relevantes, documentos relevantes y documentos considerados óptimos respectivamente. En la segunda grafica se busca analizar con mas detalle el promedio de relevancia sobre los 10, 20 y 30 primeros documentos para condensar la información y dar un análisis mas específico (evitar posibles distorsiones) frente a los documentos recuperados en cada sistema de búsqueda.

- Criterio basado en el promedio de resultados relevantes, en base al tipo de pregunta formulada ya sea por tema (específica/general). En este caso únicamente se evalúa la relevancia de los documentos recuperados, que es lo que realmente hace un usuario cuando consulta una base de datos, no de todos los documentos. Se examina, las primeras diez, veinte referencias de cada lista de resultados para determinar los valores de relevancia correspondientes a cada documento recuperado por los diferentes buscadores en respuesta a cada pregunta. La idea en esta prueba es centrar la atención en documentos que brinden información relevante (1, 2, 3) y/o potencialmente relevante u óptima (2 o 3) y sacar el promedio de valores sobre todos lo temas de tipo específico u temas de tipo general.

- Criterio según la relevancia por tipo de sintaxis (booleana/frase literal). En este criterio se evaluará sobre las 18 ecuaciones formuladas, el promedio de cada tipo de sintaxis de búsqueda a excepción, del tipo de consulta por un solo término o palabra y lenguaje natural, las cuales sólo se determinó una ecuación de consulta (por ser el tipo de consulta menos empleada debido a los resultados tan amplios y/o tergiversados que pueden incitar ante una necesidad informativa dada). La idea de este criterio es con fines

meramente representativos que permitan evaluar su comportamiento en relación con el alineamiento por relevancia, en la cual, en este caso de estudio se abarcará información precisa sobre el caso de relevancia (1, 2, 3) y (2 o 3) para los diez y treinta primeros documentos recuperados.

- Criterio basado en el Promedio de Exhaustividad-Precisión.

El principal objetivo de los motores de búsqueda es localizar y recuperar los documentos contenidos en la red de la forma más eficaz posible. Por tanto su valor dependerá de la capacidad para identificar la información relevante, de la versatilidad del método que utilicen y de la facilidad para rechazar documentos extraños. Un motor de búsqueda será más eficaz cuanto más y mejor sea capaz de satisfacer las necesidades informativas del usuario, desde el punto de vista de su concepto de relevancia, entendiendo ésta como la adecuación de las representaciones documentales ofrecidas por el buscador a la representación de las necesidades informativas que el usuario ha hecho explícitas mediante una consulta.

Es importante aclarar, que se analizará la información del nivel de relevancia (2 ó 3), ya mencionados en apartados anteriores, es decir, los considerados permisiblemente relevantes u óptimos en base al promedio de los pares de valores Exhaustividad-Precisión entre las 18 preguntas, para los 10 primeros resultados suministrados por cada buscador. Para ser mas concreta se calcula los valores medios de cada búsqueda.

Para calcular los pares de valores Exhaustividad-Precisión de cada búsqueda realizada vamos a seguir el método propuesto por Salton [40], cuya fuente es explicada por Martínez J[16]. Teniendo en cuenta los resultados de cada motor de búsqueda según la relevancia de los documentos a la pregunta. Salton claramente explica que la Exhaustividad no puede calcularse de forma exacta, sino de una forma aproximada. El método de Salton calcula los pares de valores Exhaustividad-Precisión a partir de una muestra pequeña de la amplia colección de documentos de la base de datos. En este caso se considera como número total de documentos relevantes los obtenidos entre los 10

[16] Francisco Javier Martinez,Técnicas y Métodos Avanzados de Recuperación de Información. Universidad de Murcia. Fuente: http://www.um.es/~gtiweb/fjmm/tmari/tmari-prac3.pdf

primeros. El par de valores exhaustividad-Precisión se calcula para cada posición en la lista de resultados, usando un rango o muestra como un nivel de recuperación.

En este proceso, se pretende es observar el indicador de rendimiento entre Exhaustividad-Precisión de cada motor de búsqueda. Para realizar esta parte de evaluación se ha tenido en cuenta:

a) Indicador de pertinencia o precisión. Tras obtener la respuesta a una búsqueda, su resultado es producto de un cociente que resulta de dividir el:

$$\frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos recuperados}}$$

b) Indicador de exhaustividad o de respuesta.

$$\frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos relevantes existentes en la colección}}$$

Para aclarar, se dará una pequeña explicación de la metodología llevada a cabo. De acuerdo al anexo 5, podemos ver los resultados obtenidos en nuestro estudio, en donde se realiza la clasificación por niveles de relevancia (recordemos que haremos uso de los niveles 2 ó 3). Bien, ahora veamos el procedimiento con el motor de búsqueda Altavista, supongamos que analizamos los 10 primeros resultados, de la primera ecuación de búsqueda planteada, en ella, 8 son los documentos relevantes, siguiendo lo indicado por Salton los valores de exhaustividad y precisión calculados son los siguientes:

N	Relevantes	E	P
1	X	0,125	1
2	X	0,25	1
3	X	0,375	1
4		0,375	0,75
5	X	0,5	0,8
6	X	0,625	0,83
7		0,625	0,71
8	X	0,75	0,75
9	X	0,875	0,77
10	X	1	0,8

Tabla1. Resultados de Exhaustividad- Precisión, sobre los primeros 10 documentos.

Salton entiende que los cálculos Exhaustividad-Precisión, deben realizarse documento a documento recuperado, es decir, no son iguales el par de valores Exhaustividad-Precisión en el primer documento que en el segundo. Cuando realizamos los cálculos en el primer documento (1), se ha recuperado un único documento que es pertinente y, por tanto, la precisión valer uno (un acierto en un intento) y la exhaustividad (resultado de dividir el valor de uno entre el total de documentos relevantes de la muestra, valor que sí conocemos de antemano y es ocho), vale 0.125(1/8).

Así, el documento 1 tiene asignado el par de valores E-P (0.125, 1). A continuación, procedemos a calcular el par de valores Exhaustividad-Precisión del documento 2, también relevante, aquí la precisión será el resultado de dividir el valor de dos documentos relevantes recuperados (1 y 2) entre el total de documentos recuperados hasta el momento (dos también), por lo que adquiere de nuevo el valor de la unidad; la exhaustividad será el resultado de dividir el valor de dos (ambos son relevantes) entre el total de documentos relevantes de la muestra (ocho), obteniéndose un valor de 0.25, por lo que al documento 2 se le asignaría el par de valores E-P (0.25,1). Siguiendo este método se determinan el resto de los pares de valores E-P para los seis restantes documentos recuperados. Este conjunto de diez pares de valores caracterizará, en principio, a la búsqueda Q, se puede representar en un gráfico Exhaustividad-Precisión reflejado en la siguiente ilustración:

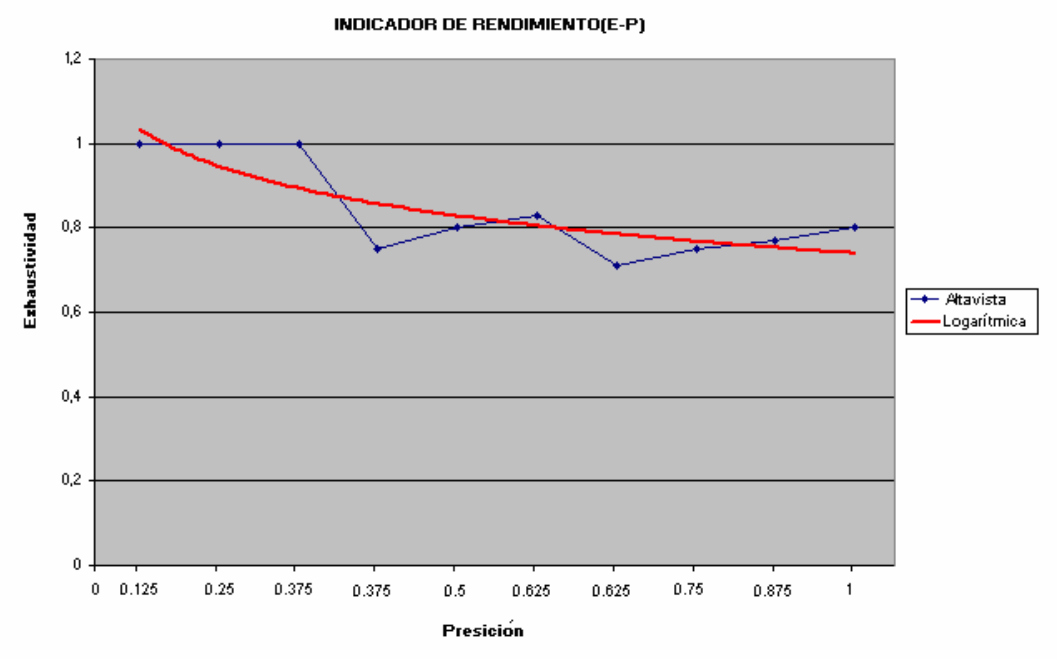


Figura2. Comportamiento Exhaustividad-Precisión llevado a cabo por Altavista sobre los primeros 10 documentos recuperados, para la primera consulta de búsqueda realizada.

- Criterio basado en análisis de similitud de resultados.

Este criterio de similitud también es conocido como solapamiento, que se entiende como el porcentaje de colección común que existe entre los motores, aunque yendo mas allá en la web debemos extender esa definición al porcentaje de documentos comunes presentes en los primeros diez, veinte o treinta documentos de la respuesta de los motores (que es lo que verdaderamente los usuarios de la web van a consultar). En nuestro caso se analizarán los 10 primeros resultados recuperados. Este criterio se fundamenta en que la lista de direcciones que ofrece un motor puede entenderse, dentro de una perspectiva vectorial, como un vector de direcciones que puede compararse con otra lista de direcciones(vector) producida por otro motor en respuesta a la misma pregunta. Esta comparación se puede llevar a cabo por medio de una función de similitud. Existen varias medidas de similitud[17], la cual la más aplicada es conocida como función del coseno.

La determinación de esta semejanza nos proporcionará información sobre su función de alineamiento, ya que si dos motores muestran unos altos grados de similitud, una vez comparados los resultados de varias búsquedas, quiere decir que en sus índices existen

[17] Medidas de similitud < http://web.udl.es/dept/dal/sepln/sepln99.ppt#303,27,Medidas de similitud>

muchos documentos comunes y que además sus algoritmos de alineamiento han respondido de forma similar.

Para esta prueba, se ha considerado el uso del modelo vectorial que permite medir la similitud de dos documentos. Para calcular la medida de similitud existente entre los diferentes vectores de resultado propuestos para cada motor. El análisis será evaluado sobre los documentos cuyo nivel de relevancia es poco relevante (solo en caso de que la existencia del enlace se encuentre en ambos motores de búsqueda en caso contrario será ignorado para nuestros cálculos), permisiblemente relevante u óptimo. Esta similitud será evaluada la similitud sobre los primeros 10 documentos recuperados por cada motor de búsqueda.

Para determinar la similitud existente entre dos vectores (entre dos motores de búsqueda), se hará uso de la función de similitud del coseno propuesta por Salton y Harman, cuya función viene dada por la expresión del ángulo formado por su dos vectores asociados.

$$Cos(P,N) = \sum(pi * ni) \Big/ \sqrt{\sum pi^2 - \sum ni^2}$$

En el modelo se hace la suposición básica de que la distancia relativa entre dos vectores en el espacio n-dimensional, representa la diferencia entre los perfiles que se han utilizado para configurar dichos vectores.

La dimensión del espacio vectorial es igual al número total de términos utilizado en la caracterización del problema (en este caso particular el número de respuestas recuperadas). Donde P es el vector problema de un caso almacenado en la Base de Casos y N es el vector del problema del nuevo caso. El resultado de la función coseno proporciona valores entre 0 y 1. En el cado de que el valor de Cos (P, N) fuese 1, indicaría que ambos casos (P, N) son idénticos y por tanto que los dos vectores de resultados para los motores son similares.

Para el procedimiento del cálculo de similitud, de dos motores de nuestro estudio, y llevando a cabo el modelo anteriormente mencionado, se tendrán en cuenta dos requerimientos:

El primero será, que durante la evaluación de similitud de un motor con respecto al otro y viceversa, se manejarán valores de 1, 0.9, 0.8 que se relacionan con la posición o alineamiento en que los primeros diez documentos recuperados por un motor, se perciben en el otro en la primera, segunda o tercera página respectivamente(es decir esta es simplemente una penalización con el fin de reflejar la posición en que fue encontrado el documento), por otro lado si el documento no es encontrado entre los 30 primeros resultados que son los que comúnmente un usuario examina este se valorará con un 0 las cuales no influirán en nada el resultado final.

Como segundo requerimiento, se llevará a cabo mediante el llenado del siguiente formato para la aplicación del modelo de similitud. Este es básicamente una representación de un modelo de espacio vectorial, cuyos términos están conformados por dos vectores (Altavista y Excite) y cuya caracterización del problema es determinar las similitud existente entre estos sistemas de búsqueda.

		M_A	M_E	$M_A * M_E$	$(M_A)^2$	$(M_E)^2$
P1	http://www.dyslexia.com/	1	1	1	1	1
	http://www.dislexia.net/	1	0	0	1	0
	http://www.interdys.org/	1	0,9	0,9	1	0,81
	http://www.dyslexia-teacher.com/	1	1	1	1	1
	http://www.bda-dyslexia.org.uk/main/home/index.asp	1	0	0	1	0
	http://www.audiblox2000.com/dyslexia_dyslexic/dyslexia.htm	1	0,9	0,9	1	0,81
	http://www.bonnieterrylearning.com/	1	1	1	1	1
	http://www.readingupgrade.com/html/rudyslexia.htm	1	0,9	0,9	1	0,81
	http://www.gow.org/?src=overture	1	0,9	0,9	1	0,81
	http://www.schwablearning.org/articles.asp?r=43&g=1	1	0,8	0,8	1	0,64
	http://infoscouts.com/health/dyslexia.htm	1	1	1	1	1

En este formato como se observa tiene varias columnas: la primera hace referencia al número de expresiones de búsqueda que en nuestro caso se manejarán un total de 18 preguntas; la segunda columna que se encuentra dividida en dos partes, corresponde a los documentos relevantes recuperados en los 10 primeros resultados proveídos por cada

motor. La tercera columna corresponde al vector resultado del motor de búsqueda M_A (Altavista) y la cuarta al vector resultado del segundo motor M_H (Hotbot). Aquí, cada URL engloba el vector resultado, el cual se ve sometido a juicio de relevancia y se le asignará un valor de acuerdo a la ponderación previamente definida. La quinta columna corresponde al producto escalar de los vectores resultados, las dos últimos, corresponde al cuadrado de sus componentes. Una vez que se tenga disponible de todos los datos se aplica finalmente la Función de Similitud del coseno, que se determina con la raíz cuadrada de los de los últimos valores y se divide por ese valor el producto escalar, obteniéndose el valor que representa el porcentaje de coincidencia de los motores búsqueda comparados.

Cabe sugerir, que para efectos de similitud solo interesan los componentes de relevancia mayor de cero, ya que el objetivo de esta prueba es precisamente analizar la similitud de la pertinencia de la respuesta de dos motores, es decir de los resultados relevantes recuperados sujeta a la búsqueda planteada.

Una vez obtenido comparativamente los resultados de las búsquedas tomaremos en consideración los siguientes razonamientos: El análisis de similitud de los resultados obtenidos en este estudio, va dirigido a examinar los documentos relevantes (nivel de relevancia 2 ó 3) de los diez primeros documentos, con el fin de calcular la coincidencia de contenidos en los resultados. Verificar de forma individual las similitudes obtenidas por cada motor de búsqueda en las 18 consultas realizadas, con el fin de detectar algún comportamiento con respecto a este criterio para evaluar si sigue un patrón en concreto o no.

- Criterio basado en análisis de Agrupamiento (Clustering).

Para nuestro caso de estudio, la función de este parámetro es identificar grupos de motores mas afines unos a otros siempre que las distancias medias establecidas entre los componentes de estos grupos más próximos alcancen valores pequeños.

A continuación se aplica el método de agrupamiento por el promedio aritmético a las distancias existentes entre los siete motores de búsqueda, cuando se analizan los diez primeros documentos recuperados. Este método es muy similar al planteado en el capítulo

anterior (El algoritmo AGNES), la diferencia radica simplemente en vez de manejar la matriz completa, se maneja la matriz triangular superior, esta es una forma mas sencilla de utilizar este técnica de agrupación expuesto IIAP (Instituto de Investigaciones de la amazonía Peruana)[18]. Para iniciar con la aplicación de esta técnica de clustering o agrupamiento, se da por dispuesto el cálculo de las similitudes promedio motor a motor, posteriormente se recoge las distancias medias que se han determinado par los siete motores de búsqueda y luego se continua con el proceso normal del método de promedio aritmético (o bien encadenamiento de media de grupos, algoritmo de AGNES).

En la clasificación por el promedio aritmético, la formación de un grupo nuevo se calcula con la distancia promedio desde este grupo a todos los otros con base en las distancias de los enlaces que este grupo tiene con los otros. La próxima unión se hace en la distancia promedio más corta entre los grupos. El promedio es simplemente la suma de las distancias de enlaces dividida por el número de enlaces. La distancia mínima, que es la distancia que resulta de la unión, está escrita en negrita en cada tabla. Debido a las uniones, cada tabla nueva es siempre más pequeña que la anterior. Las distancias nuevas después de la unión, más las formas de calcularlas, están representadas en cada tabla nueva.

	B	C	D	E
A	17,5	47,4	43,0	57,0
B		35,8	27,9	39,2
C			14,4	37,0
D				24,4

	B	C+D	E
A	17,5	42,2=(47,4+43)/2	57,0
B		31,85=(35,8+27,9)/2	39,2
C+D			30,7=(37+24,4)/2

	C+D	E
A+B	38,53= (47,4+43+35,8+27,9)/4	48,1=(57+39,2)/2
C+D		30,7=(37+24,4)/2

	C+D+E
A+B	41,72 =(47,4+43+35,8+ 27,9+57+39,2)/6

Tabla 2. Ejemplo aplicativo de la técnica de agrupamiento “Promedio Aritmético”

4. APLICACIÓN MODELO DE EVALUACION Y ANALISIS DE RESULTADOS.

[18] Guia para etudiar patrones de distribución de especies amazonicas. Entidades ejecutoras: Instituto de Investigaciones de la Amazonía Peruana, Universidad de Turku, Finlandia, Biota BD Oy, Finlandia, 2004, pp. 60-62.

4.1 CRITERIO BASADO EN EL TAMAÑO PROMEDIO DEL INDICE

Se hicieron dos pruebas, las cuales fueron manejadas en base a 18 preguntas formuladas previamente, la primera se realizó el 21 de febrero del 2005 y la segunda el 30 de abril de este mismo año (Ver anexo2). Los resultados obtenidos en cada uno de las fechas se representan a continuación:

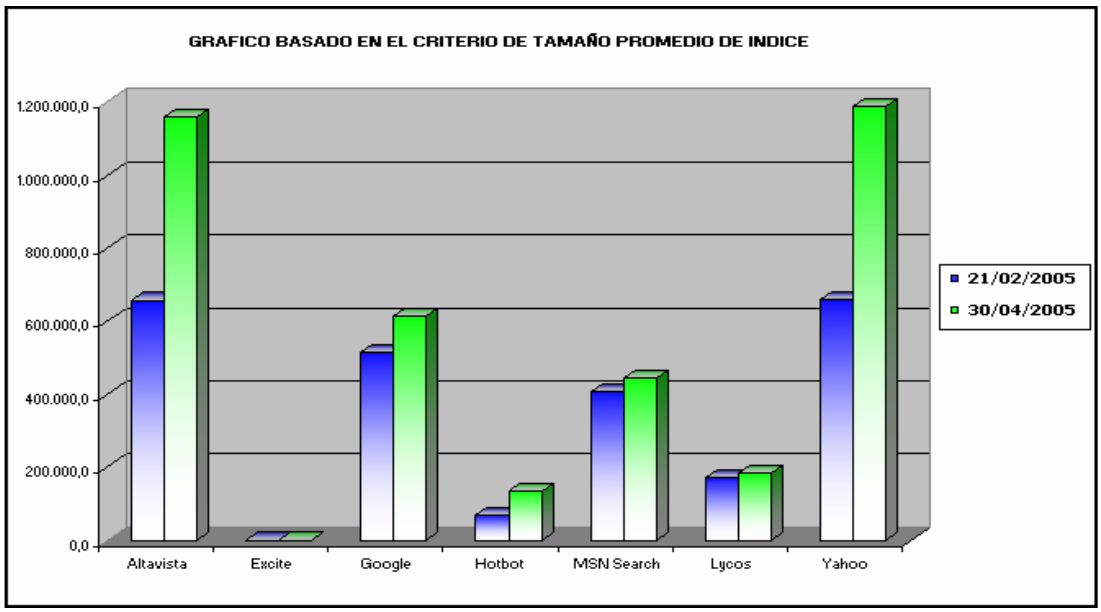
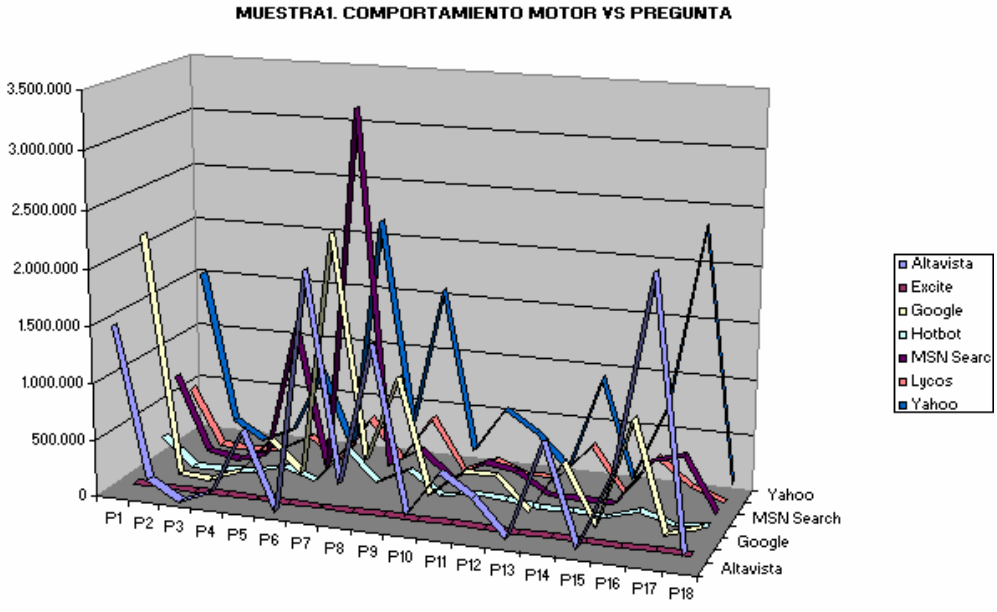


Figura3. Resultados de cobertura promedio de índice, en periodos relativamente distantes.

Seguidamente, se presenta una visualización mas pormenorizada, del comportamiento sobre el cual, cada motor respondió al ser interrogado por medio de una expresión de búsqueda particular:

Prueba 1



Prueba 2

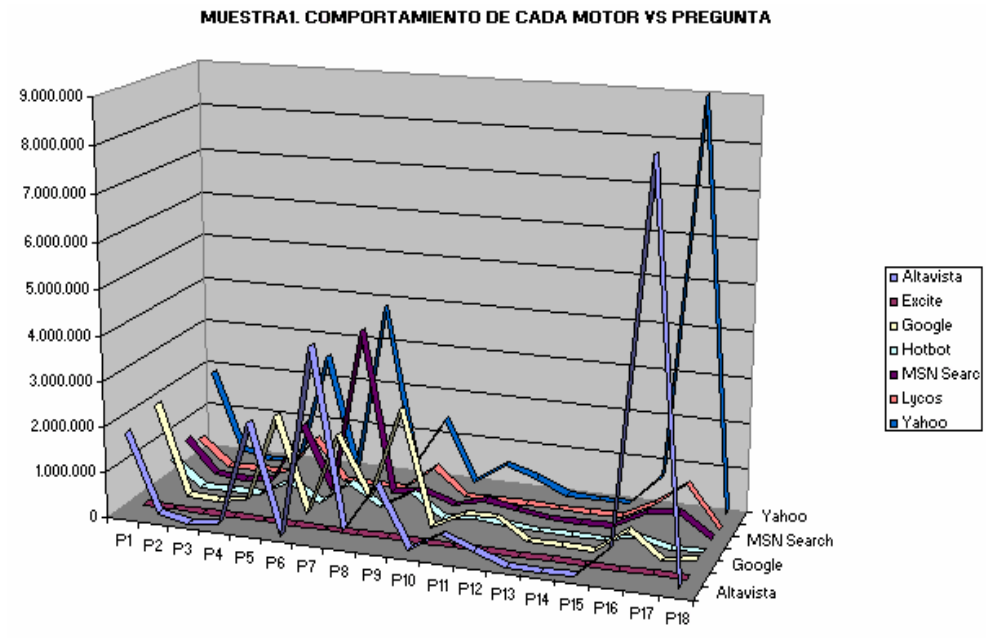


Figura4. Comportamiento manifestado por el índice de cada motor en relación con la pregunta formulada en la dos pruebas llevadas a cabo.

Básicamente el interés de evaluar este criterio era dar una comparación en pequeña escala de resultados reales de búsqueda y de esta manera dar a conocer una mejor

representación del tamaño comparativo de las bases de datos, en cuanto la indización de documentos se refiere. Cerca de mas 2 meses de diferencia fueron tomados estos datos, aunque resulta este análisis algo premeditado para sacar conclusiones radicales al respecto, se puede evaluar este estudio de manera subjetiva con base en juicios expresamente expuestos en esta investigación.

De acuerdo a los resultados obtenidos, se observa la existencia de una cierta correlación entre el tamaño del índice, entre los diferentes motores de búsqueda, aunque no sean claramente proporcionales. La calidad en la recuperación de información en Yahoo! resulta ser superior, seguido por Altavista y Google, no obstante se manifestó durante las dos pruebas realizadas, un incremento sustancial en los documentos recuperados por Yahoo! y Altavista, un muy buen funcionamiento del índice en respuesta promedio a las diferentes expresiones de búsqueda evaluadas.

Cabe señalar, que de acuerdo a información investigada sobre el motor de búsqueda Excite, su capacidad de indexación es relativamente pequeña, y ésta es precisamente apreciable en la grafica presentada, cuyos resultados en comparación al resto de los motores evaluados éste obtiene el menor grado de recuperación documental. Aunque es claro que los resultados comparativos entre estos sistemas de recuperación web no deben basarse en solo su tamaño, es muy interesante y provechoso como primer paso de evaluación, tener alguna idea del tamaño relativo de las bases de datos de cada motor de búsqueda.

4.2 CRITERIO BASADO EN EL TIEMPO PROMEDIO DE RESPUESTA

En este análisis se consideró evaluar el tiempo promedio que demora cada motor en responder una necesidad informativa proporcionada, es decir, el tiempo que tarda desde el envió de la pregunta hasta la entrega de resultados. Los datos presentados de aquí en adelante corresponden a pruebas realizadas durante dos periodos de tiempo cuya diferencia fue aproximadamente más de 60 días.

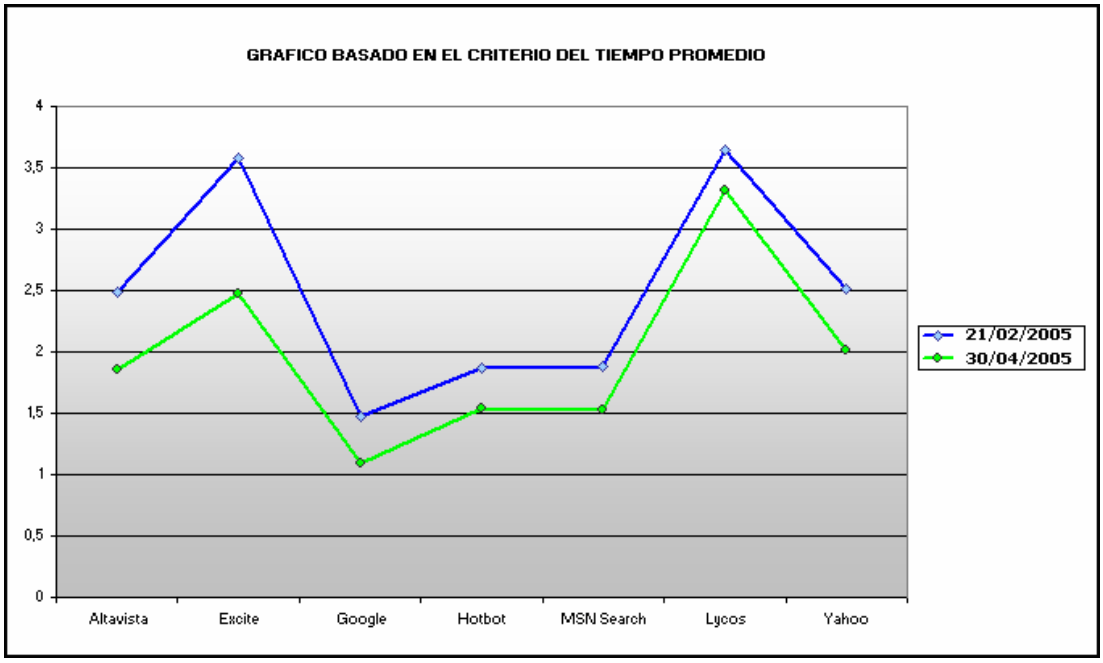


Figura5. Resultados de tiempo promedio de respuesta, en periodos relativamente distantes.

El criterio basado en el tiempo, fue tenido en cuenta para corroborar, si la velocidad de respuesta de cada motor actúa de manera proporcional y/o dependiente al número de documentos recuperados.

Durante la toma de datos se observó que el tiempo de respuesta actúa como una variable independiente al número de documentos recuperados, casi en la mayoría de buscadores (ver anexo3). Sin embargo, aún cuando durante la toma de datos se hicieron en horarios diferentes (una de ellas en horario donde posiblemente se presenta una mayor congestión en la red horas del medio día), en ambos intervalos de tiempo se conservó una relativa asiduidad en la actuación de cada motor. Es decir, en la gráfica se puede observar con claridad que Google, siguiéndole Hotbot y MSN Search fueron los motores de búsqueda cuyo indicador temporal, presentaron la menor saturación de respuesta al servicio de recuperación de documentos, durante su acceso a sus bases de datos. Esto hace pensar que sin importar el número de usuarios que se conecten al motor (índice), existe una buena calidad del software de búsqueda, una adecuada velocidad del hardware de servidor y un buen soporte del servidor para el balanceo de carga.

4.3 CRITERIO BASADO EN EL RUIDO DOCUMENTAL

Como se mencionó la calidad de recuperación de información de un motor no solo se ve revelado en el tamaño del índice sino que también es necesario corroborar la calidad informativa que ofrece según necesidades particulares del usuario. El anexo 4 muestra los resultados para los 30 primeros resultados obtenidos para las 18 consultas formuladas, siendo éste el número de saltos que principalmente el usuario consulta, por ser este considerado la cantidad de documentos de mayor relevancia que cada motor arroja. Por tanto, el análisis que se pretende dar en primera instancia, es visualizar resultados inactivos, duplicados e irrelevantes presentes en cada motor.

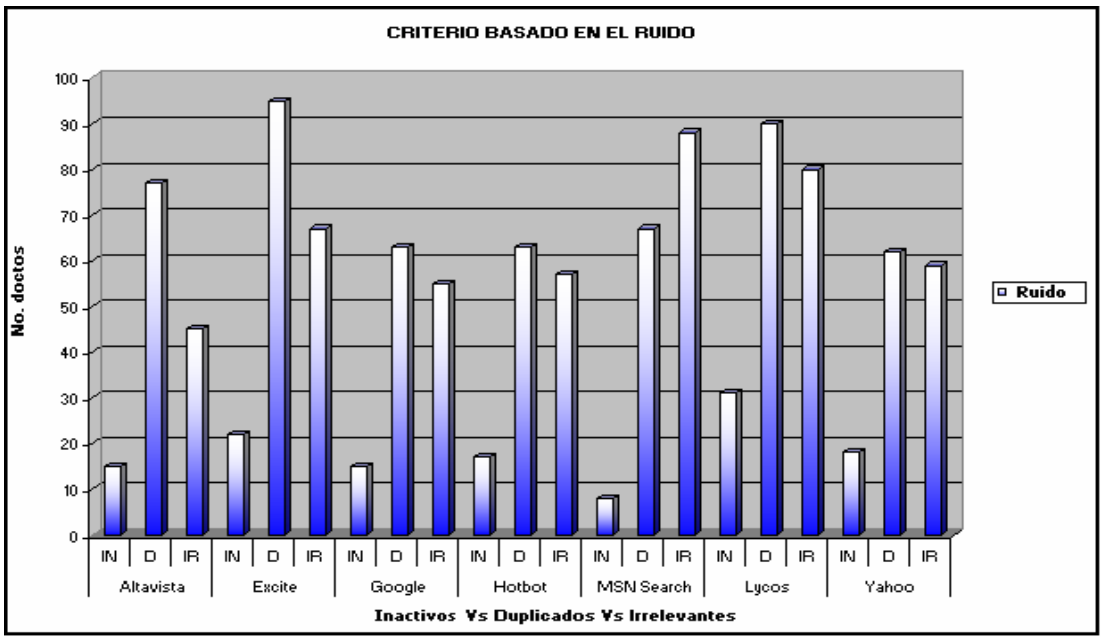


Figura6. Resultados de ruido documental presente en los motores de búsqueda

PORCENTAJES DE RUIDO DOCUMENTAL

	Altavista			Excite			Google			Hotbot			MSN Search			Lycos			Yahoo		
	IN	D	IR	IN	D	IR	IN	D	IR	IN	D	IR	IN	D	IR	IN	D	IR	IN	D	IR
No	15	77	45	22	95	67	15	63	55	17	63	57	8	67	88	31	90	80	18	62	59
%	2,78	14,3	8,33	4,07	17,6	12,4	2,778	11,7	10,2	3,15	11,7	10,6	1,48	12,4	16,3	5,74	16,7	14,8	3,33	11,5	10,9

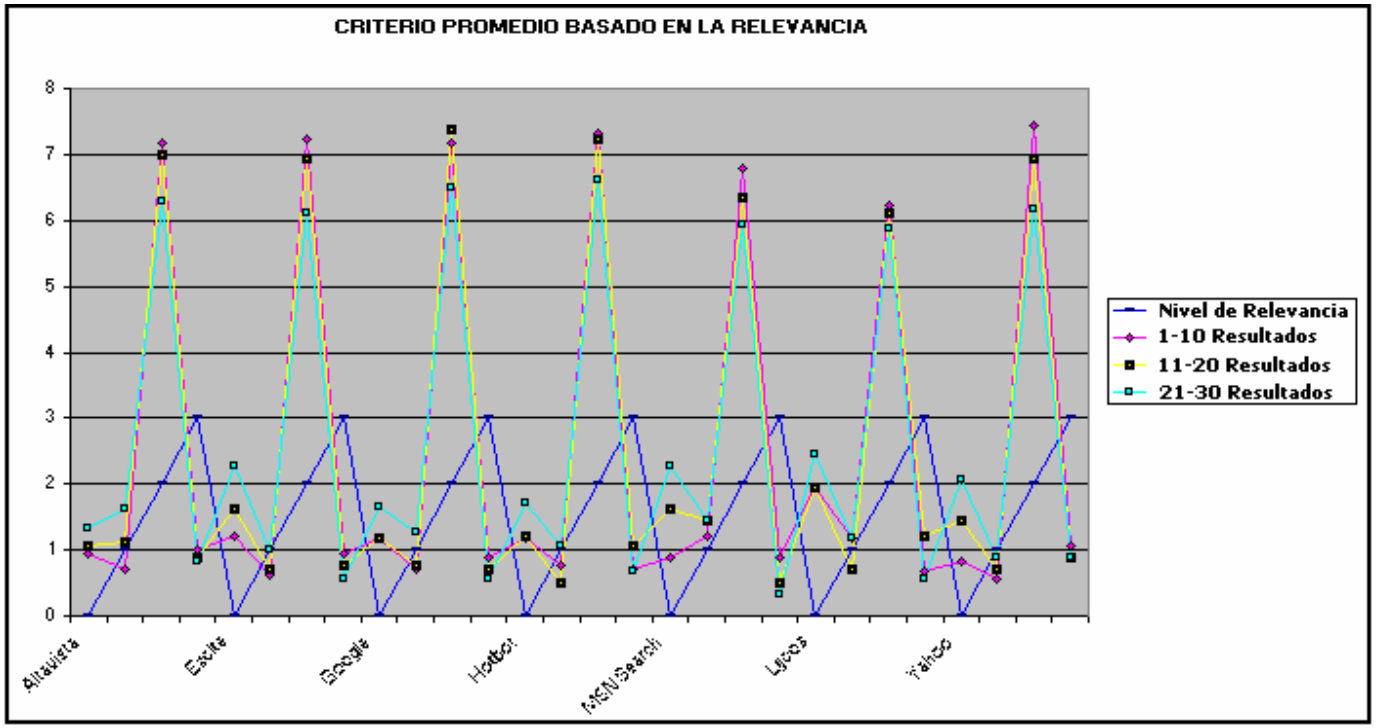
TOTAL DE PAGINAS/PREGUNTA:	30	
NUMERO DE CONSULTAS:	18	
TOTAL DE PAGINAS A EVALUAR:	540	100%

En este estudio, se notó que el ruido documental promedio sobre los primeros 30 documentos recuperados, por cada uno de los 7 motores de búsqueda, objeto del experimento, fue muy pequeña en relación con otros estudios realizados por otros autores. No obstante, es importante recalcar que los estudios realizados anteriormente y hasta la fecha tienen una diferencia tiempo muy acentuado, así como el enfoque de las consultas realizadas en ese entonces, lo que puede diferir considerablemente en los resultados actuales.

Entrando al análisis, en la gráfica se pudo observar que Lycos, es el motor con mas ruido documental ya que sobresale sobre todos los demás con enlaces inactivos de un 5.74%, enlaces irrelevantes (enlaces considerados adecuados pero no útiles para la necesidad informativa) con un 14,8% y en duplicados lo supera Excite con un 17,6%. MSN Search le corresponde el segundo lugar con respecto al ruido tanto en documentos irrelevantes como duplicados con 16.3 y 12,4% respectivamente. Altavista sobresale por un alto grado de duplicados. Todas estas consideraciones observadas en esta prueba, hace pensar que sus bases de datos no se actualiza con tanta frecuencia como sería esperado y también puede llegar asumirse un posible escaso índice de respuesta sobre algunas búsquedas. Esta particularidad vista en nuestro estudio fue viable, sin embargo es importante entender que no puede adjudicarse del todo incuestionable por contarse en esta disertación con una muestra pequeña de búsquedas. No obstante, aún cuando para algunos investigadores la cantidad de duplicados se percibe como un aspecto negativo en los sistemas de búsqueda, en contraste desde el punto de vista del usuario, puede percibirse como una impresión de relevancia frente al documento.

4.4 CRITERIO SEGÚN EL PROMEDIO DE RELEVANCIA, EN FUNCION DE PRUEBAS.

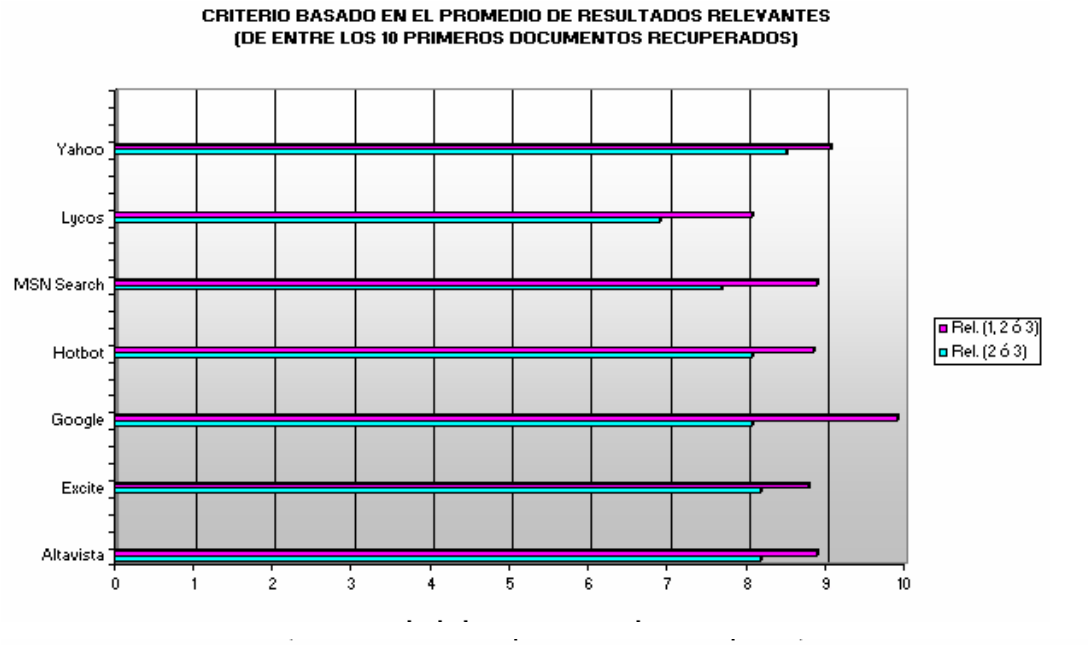
En las graficas vemos que la línea azul representa los diferentes niveles de relevancia y la línea de color representa los resultados recuperados por cada motor según este criterio.



	Altavista				Excite				Google				Hotbot				MSN Search				Lycos				Yahoo			
	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Prom. 1-10 Dctos	0,094	0,072	0,717	0,1	0,122	0,061	0,722	0,094	0,117	0,072	0,717	0,089	0,117	0,078	0,733	0,072	0,089	0,122	0,678	0,089	0,194	0,117	0,622	0,067	0,083	0,056	0,744	0,106
Prom. 11-20 Dctos	0,106	0,111	0,7	0,089	0,161	0,072	0,694	0,078	0,117	0,078	0,739	0,072	0,122	0,05	0,722	0,106	0,161	0,144	0,633	0,05	0,194	0,072	0,611	0,122	0,144	0,072	0,694	0,089
Prom. 21-30 Dctos	0,133	0,161	0,628	0,083	0,228	0,1	0,611	0,056	0,167	0,128	0,65	0,056	0,172	0,106	0,661	0,067	0,228	0,144	0,594	0,033	0,244	0,117	0,589	0,056	0,206	0,089	0,617	0,089

Figura7: Comportamiento promedio de cada motor, según relevancia (0, 1, 2 y 3) sobre los 10, 20 y 30 documentos recuperados y en base a las 18 preguntas objeto de estudio.

En la figura7 se presenta el comportamiento experimentado por cada uno de los motores de búsqueda, en cuanto relevancia promedio se refiere entre la primera, segunda y tercera página de resultados: información considerada irrelevante posición (0), poco relevante (1), permisiblemente relevante (2) y muy relevante u óptimo (3). Puede parecer algo extraño del por qué el nivel de relevancia 2 supera considerablemente al 3 esto es debido a la manera como fue determinada o delimitado este parámetro de evaluación. Antes de proseguir, con el análisis con un mayor detalle en la siguientes gráficas, cabe hacer anotación que en esta prueba adoptamos como nuestra la idea de Lancaster sobre relevancia, es decir "un documento va a resultar relevante cuando su contenido tiene que ver con el objeto de la pregunta y, además, le resulta útil"[23].

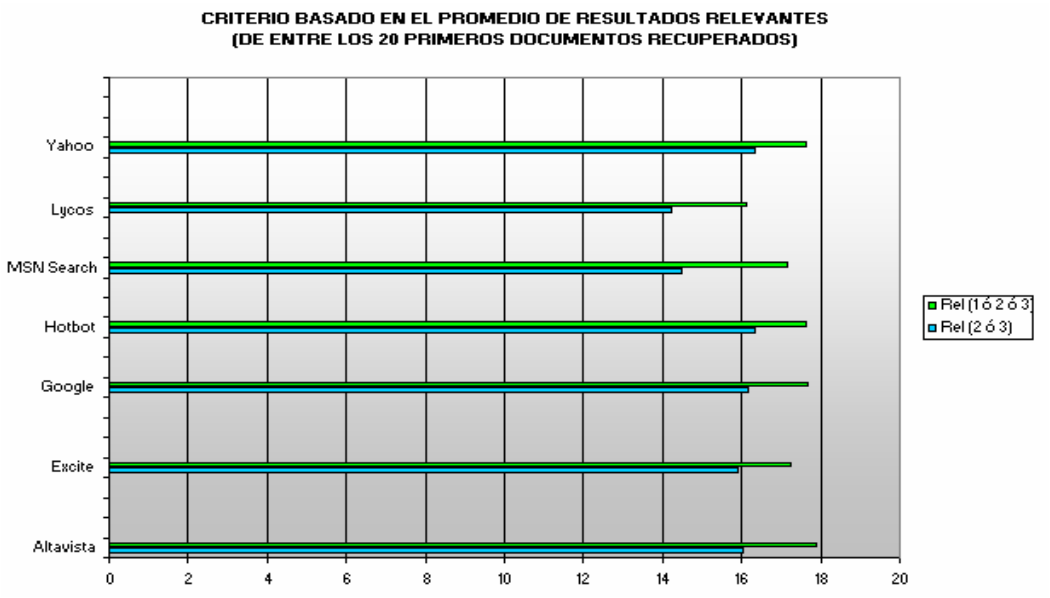


	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
Rel. (2 ó 3)	0,816666667	0,816666667	0,805555556	0,805555556	0,766666667	0,68888889	0,85
Rel. (1, 2 ó 3)	0,888888889	0,877777778	0,988888889	0,883333333	0,888888889	0,80555556	0,905555556

Figura8: Análisis de resultados de documentos relevantes Vs documentos permisiblemente relevantes u óptimos (en base 10 primeros documentos recuperados).

Al analizar los primeros 10 documentos, Google es indudablemente el motor de búsqueda que mejores resultados ha obtenido en cuanto información relevante (nivel de relevancia (1, 2 ó 3) con un promedio del 98% sobre la totalidad de las 18 consultas formuladas y en cuanto a información muy útil u óptima (siendo estos los que representan el mayor grado de significación o importancia con motivo de la pregunta realizada, es

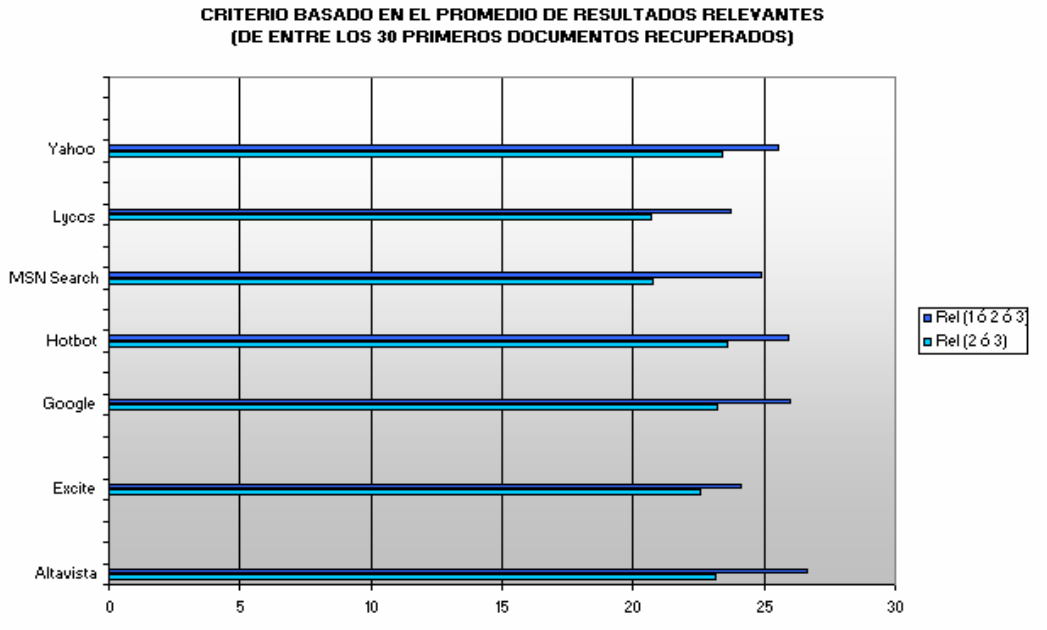
decir, con la necesidad de información), se destaca Yahoo con un promedio de relevancia de 85% seguido por Altavista y Excite con porcentaje promedio similar de 81,6%.



	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
Rel. (2 ó 3)	0,802777778	0,79444444	0,808333333	0,816666667	0,725	0,71111111	0,816666667
Rel. (1 ó 2 ó 3)	0,894444444	0,86111111	0,883333333	0,880555556	0,858333333	0,80555556	0,880555556

Figura9: Análisis de resultados de documentos relevantes Vs documentos permisiblemente relevantes u óptimos (en base 20 primeros documentos recuperados).

Los mejores resultados los ha obtenido Altavista con un 89% y seguido en una proporción mínima de diferencia Yahoo, Google y Hotbot en un rango comprendido entre el 88% (en el nivel de relevancia (1, 2 ó 3) al analizar los 20 documentos. En el nivel de relevancia útil u óptimo (Rel. 2 ó 3) los mejores resultados le son concedidos a Yahoo y Hotbot con un rango de valores de 81% seguido por una ligera diferencia con Altavista y Google con un 80%.

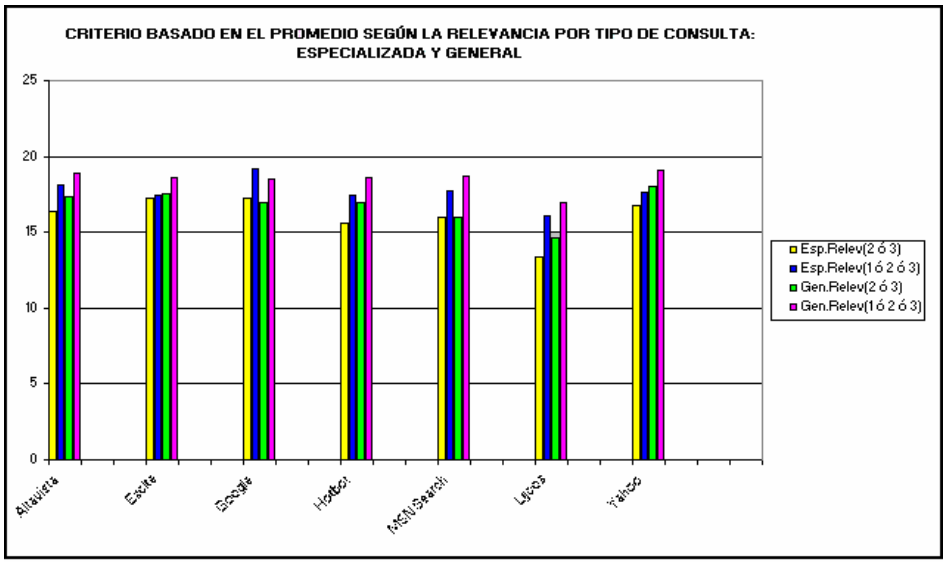


	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
Rel. (2 ó 3)	0,772222222	0,75185185	0,774074074	0,787037037	0,692592593	0,68888889	0,77962963
Rel. (1 ó 2 ó 3)	0,887037037	0,8037037	0,866666667	0,864814815	0,82962963	0,79074074	0,851851852

Figura10: Análisis de resultados de documentos relevantes Vs documentos permisiblemente relevantes u óptimos (en base 30 primeros documentos recuperados).

Para concluir con este criterio, vemos que el rango de valores entre Altavista, Google, Hotbot y Yahoo es ligeramente diferente lo que nos permite reafirmar aun más que estos cuatro motores de búsqueda presentan valores muy cercanos entre sí, en los tres análisis realizados sobre los diez, veinte y treinta documentos recuperados, logrando un mejor alineamiento en la estimación del mayor numero de enlaces técnicamente útiles y/o óptimos. Sin embargo, podemos decir que Excite les sigue casi los pasos en la relevancia en sus resultados (nivel 2 ó 3) considerados como relevancia útil u óptima en los motores de búsqueda previamente mencionados.

4.5 CRITERIO SEGÚN EL PROMEDIO DE RELEVANCIA EN FUNCIÓN AL TIPO DE TEMA DE CONSULTA: ESPECIALIZADA Y GENERAL



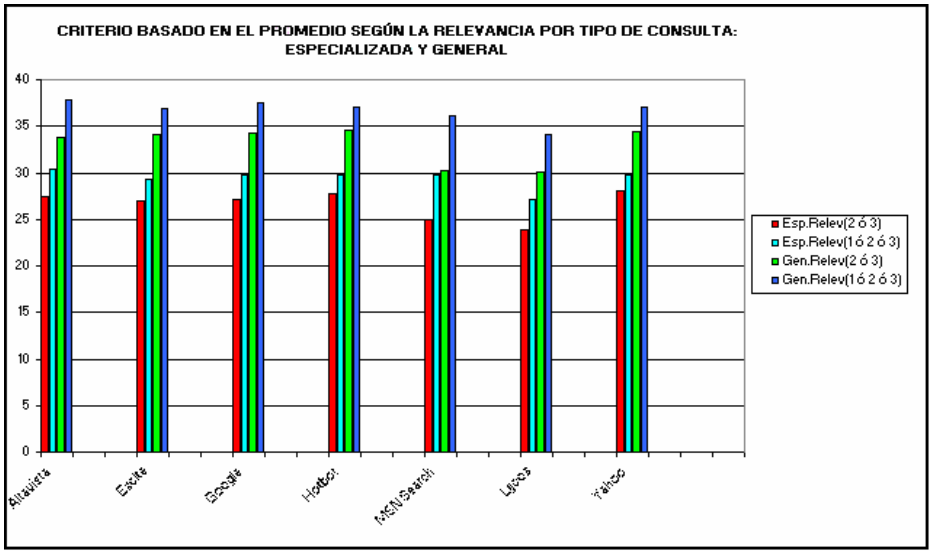
	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
Esp.Relev(2 ó 3)	0,830	0,930	0,910	0,820	0,870	0,660	0,840
Esp.Relev(1 ó 2 ó 3)	0,900	0,890	0,860	0,880	0,880	0,800	0,930
Gen.Relev(2 ó 3)	0,800	0,8125	0,838	0,788	0,775	0,725	0,763
Gen.Relev(1 ó 2 ó 3)	0,875	0,863	0,900	0,888	0,900	0,813	0,875

Figura11: Análisis de resultados promedio según la relevancia por tipo de consulta, Especifica y General (en base a los 10 primeros documentos recuperados).

Los resultados que se presentan en este criterio, corresponden al promedio de información(documentos) considerados relevantes (1, 2 ó 3) e información muy relevantes u óptimo (2 ó 3), de acuerdo a las preguntas, referenciadas por el tipo de tema de consulta, es decir, expresiones formuladas a titulo orientativo como “Especializado” o “General”, en función a los 10 primeros documentos recuperados.

Según se observa en la figura11, los mejores promedios de relevancia óptima, para las consultas de tipo especializadas se destaca Excite claramente con un porcentaje del 93% sobre los demás. Sin embargo, Google y MSN Search le siguen los pasos con un porcentaje promedio de 91% y 87% respectivamente. Por otro lado, en preguntas de este mismo tipo de tema especializado, pero con nivel de relevancia útil (1, 2 ó 3), se distinguen los motores Yahoo y Altavista con un 93% y 90%.

Ahora bien, en lo que respecta al tipo de consultas generales, el que mejor se destaca con un promedio mayor de documentos relevantes (1, 2 ó 3) recuperados se lo concedemos a Google y MSN Search con un porcentaje similar del 90%. Y en resultados de relevancia óptima con respecto al tipo de consulta general, Google y Excite quienes obtienen los mejores promedios con porcentajes de 83% y 81% respectivamente.



	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
Esp.Relev(2 ó 3)	0,815	0,790	0,800	0,855	0,72	0,705	0,875
Esp.Relev(1 ó 2 ó 3)	0,910	0,860	0,875	0,905	0,855	0,795	0,885
Gen.Relev(2 ó 3)	0,7875	0,800	0,819	0,769	0,731	0,719	0,74375
Gen.Relev(1 ó 2 ó 3)	0,875	0,863	0,89375	0,8625	0,8625	0,81875	0,875

Figura12: Análisis de resultados promedio según la relevancia por tipo de consulta, Específica y General (en base a los 20 primeros documentos recuperados).

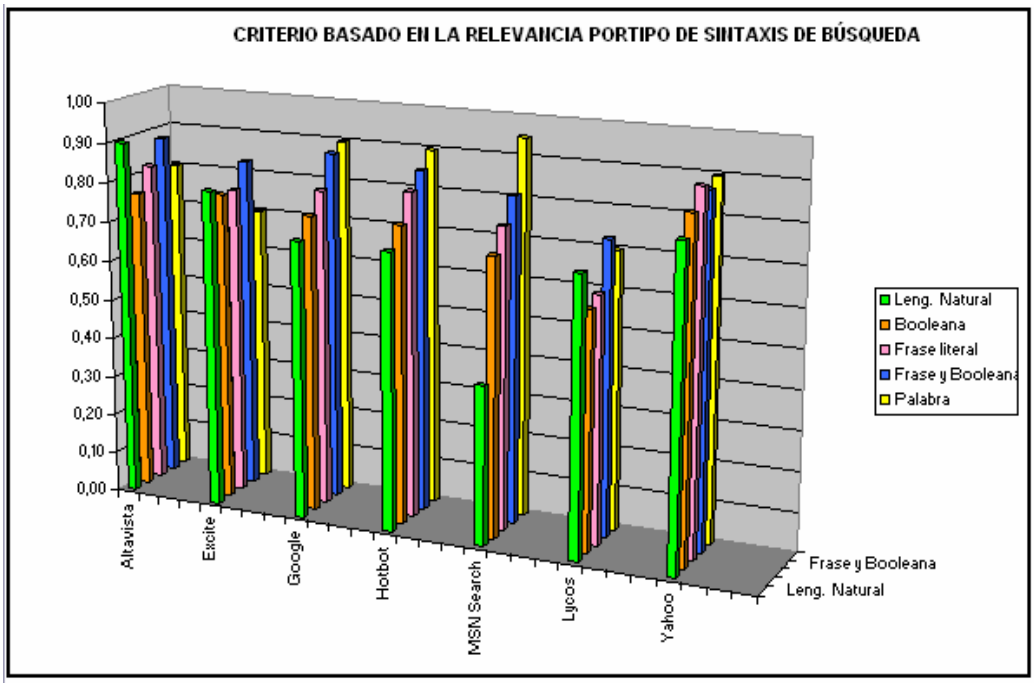
En la figura12, se puede observar, los datos resultantes sobre los 20 primeros resultados analizados. Se destacan con los mejores resultados en consultas de tipo especializado: en información relevante Altavista y Hotbot con 97% y 90% respectivamente; en información considerada técnicamente relevante u óptimo, Yahoo y Hotbot con porcentajes de 87% y 86%. En preguntas de tipo general, se obtienen los mejores promedios en información relevante Google con 89% y Altavista y Yahoo con 87.9% y en información considerada útil u óptimo Google con 82% y Excite con 80%.

Otro aspecto destacable presente en los buscadores, es el hecho de alcanzar mejores resultados en consultas especializadas en comparación con las preguntas generales lo

que hace pensar, posiblemente que las preguntas de este tipo suelen estar mejor definidos en la ecuación de búsqueda, lo que favorece una mejor ordenación por relevancia y por ende mejores resultados tanto para búsquedas precisas como exhaustivas

4.6 CRITERIO BASADO EN LA RELEVANCIA, TOMANDO COMO PARTIDA LA SINTAXIS DE LAS DIFERENTES TIPOS DE EXPRESIONES DE BÚSQUEDA PLANTEADAS.

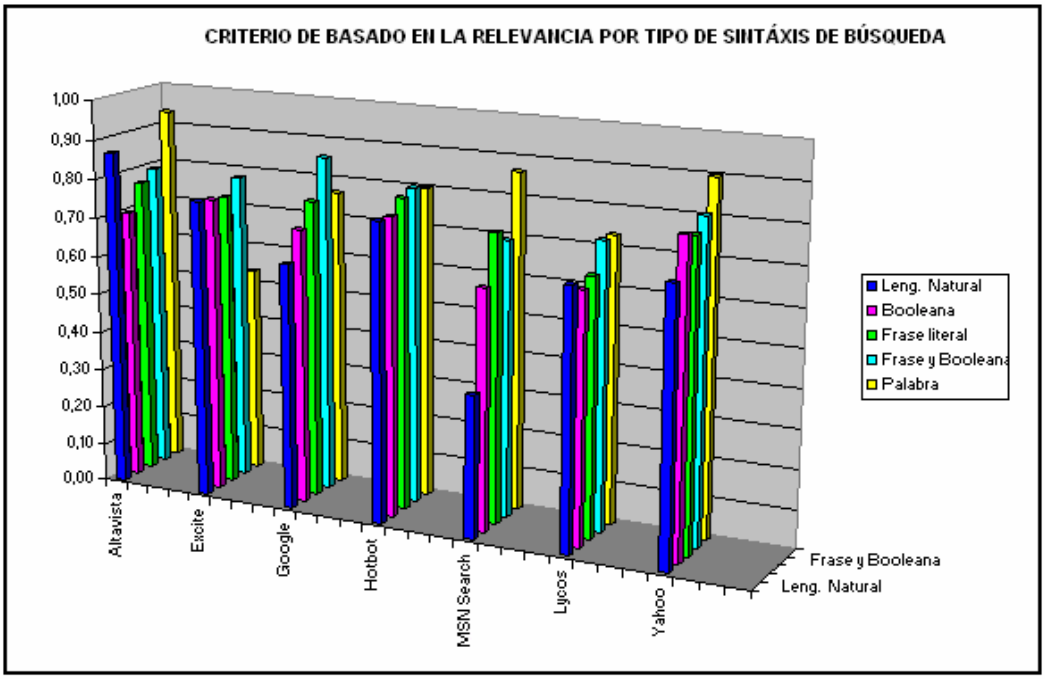
En realidad, el interés de este criterio a evaluar, es sin duda percibir el comportamiento promedio en los resultados, que cada sistema de búsqueda expone frente a un tipo de consulta definida.



	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
Palabra	0,80	0,70	0,90	0,90	0,95	0,70	0,90
Leng. Natural	0,90	0,80	0,70	0,70	0,40	0,70	0,80
Booleana	0,76	0,78	0,75	0,75	0,70	0,60	0,85
Frase literal	0,82	0,78	0,80	0,82	0,76	0,62	0,90
Frase y Booleana	0,88	0,84	0,88	0,86	0,82	0,74	0,88

Figura13: Análisis de resultados según la relevancia (2 ó 3), en base a los 10 primeros documentos recuperados y a una pregunta representativa al tipo de sintaxis de consulta.

Según se observa, los mejores resultados en esta muestra representativa al tipo de consulta, sobre los 10 primeros resultados, se le atribuye a Yahoo, recuperando un gran número de documentos relevantes en los diferentes tipos de sintaxis de consulta, percibiéndose una recuperación de dos tipos consultas de un 90%. A pesar de esto, Hotbot, Google y Altavista, le siguen muy cercanamente a Yahoo, con resultados son de igual manera importantes. En la tabla de resultados se puede observar el tipo de consulta de lenguaje Natural aplicada a los motores búsqueda; información que fue desconocida durante la investigación realizada (prestaciones de los motores de búsqueda, suministradas en posteriores capítulos), pero como objeto de curiosidad en este estudio, se quiso llevar a cabo para su evaluación. La sintaxis con menos relevancia en los primeros resultados precisamente fue ésta, presentándose valores de un 70% en Google, Hotbot y Lycos y un 40% le corresponde a MSN Search siendo este el mas bajo en relación con los obtenidos en las otras consultas. Además se notó en este caso que la consulta Literal superó en valores en relación con la sintaxis Booleana, siendo este resultado coherente ya que exige mayor precisión en los resultados; además se observó que la combinación de ambas sintaxis dió como resultado un alcance de documentos relevantes más significativos. Excite siempre mantiene una relación medianamente alta (intermedia) frente a estos resultados, Lycos por su lado, resulta tener una ponderación inferior frente a sus contendores.



	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
Palabra	0,93	0,53	0,77	0,80	0,87	0,73	0,90
Leng. Natural	0,87	0,77	0,63	0,77	0,37	0,67	0,70
Booleana	0,70	0,76	0,71	0,77	0,62	0,64	0,80
Frase literal	0,77	0,75	0,77	0,80	0,74	0,66	0,79
Frase y Booleana	0,79	0,79	0,87	0,81	0,71	0,73	0,82

Figura14: Análisis de resultados según la relevancia (2 ó 3), en base a los 30 primeros documentos recuperados y a una pregunta representativa al tipo de sintaxis de consulta.

En cuanto a los resultados obtenidos sobre los treinta documentos recuperados vemos que Altavista y Yahoo obtienen los mejores resultados en cuanto a las diferentes tipo de consultas evaluadas. Vemos que la consulta de lenguaje Natural sigue afectando en cierto porcentaje a Google (63% en relación con los demás buscadores) pero sobre todo a MSN Search quien obtiene un porcentaje del 37%. En relación con la sintaxis de consulta booleana y frase literal, se distingue con un 80% Yahoo en la búsqueda booleana y Hotbot con un 80% en frase literal. Sin embargo, algo que si se ha destacado en ambas consultas que para resultados más precisos, la frase literal predomina con mejores resultados en los diez y treinta documentos recuperados, y para búsquedas exhaustivas los operadores booleanas superan en mayor grado. Otro aspecto a destacar en ésta análisis, es como la combinación de ambas sintaxis (booleana y frase literal) se obtienen mejores resultados de búsquedas que en consultas manejadas individualmente.

4.7 CRITERIO BASADO EN EL PROMEDIO DE EXHAUSTIVIDAD-PRESICION.

La idea de esta prueba es reflejar el funcionamiento que un usuario puede esperar obtener de un sistema de búsqueda. Para medir la exhaustividad y la precisión se tuvo en cuenta el número de documentos recuperados y el número de documentos relevantes. Se ha utilizado la regresión logarítmica como forma de representación de los valores de Exhaustividad-Precisión por ser la función que ofrece mejores resultados de ajuste (en el anexo6, se puede observar la grafica original y los resultados promedios obtenidos). Los cálculos de precisión y exhaustividad se realizan según el método propuesto por Salton y McGill (39) ya explicado en el capítulo anterior, para cada uno de los sistemas de búsqueda que ordenan los resultados según la relevancia de los documentos a la pregunta.

En la Figura15, se puede observar, un indicador de rendimiento de cada motor de búsqueda; en última instancia, no es ni la exhaustividad ni la precisión, sino el producto entre ambas, ya que muestra la relación existente entre lo que el usuario espera encontrar (la pregunta) y lo que realmente encuentra (el resultado). En ella, se comprueba visiblemente una relación inversa existente entre la exhaustividad y la precisión. Sin embargo podemos decir en esta prueba que los resultados aquí obtenidos ponen una vez más de manifiesto que los buscadores en su mayoría no son sistemas perdurables en cuanto a recuperación de información se refiere, es decir son menos precisos pero si más exhaustivos.

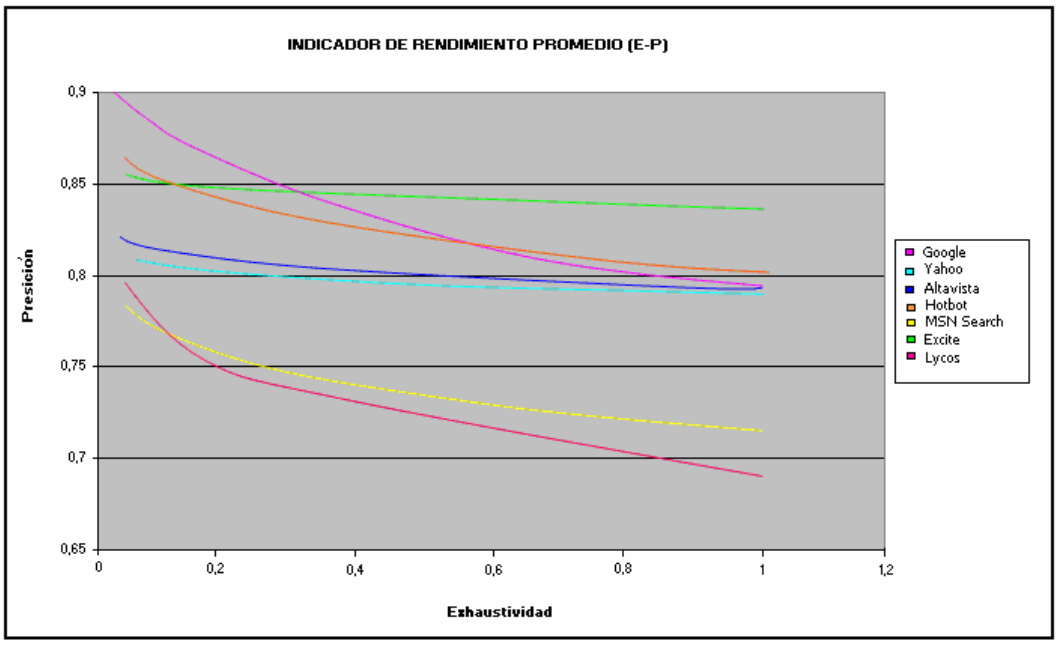


Figura15. Indicador de Rendimiento, producto del cálculo de valores medios de cada búsqueda, sobre los 10 primeros resultados para las 18 consultas (nivel de relev. 2 ó 3).

Un aspecto importante que se aprecia en los datos obtenidos (Ver anexo6) y sobre todo en la representación gráfica, es que sobre los primeros diez documentos recuperados, existe un porcentaje muy bueno en cuanto obtener búsquedas precisas y relevantes. Debido a que el mínimo porcentaje de precisión presenta valores inferiores 0.7 pero superiores a 0.5 característica que se le atribuye a Lycos; es precisamente este último porcentaje el que es considerado adecuado, y por el que siempre se espera conseguir sobre los motores de búsqueda para suponer como viables su efectividad en la recuperación de información ante una muestra examinada por el usuario.

No obstante ante esta afirmación expuesta, es importante hacer un análisis "comparativo" con el fin de determinar cual resulta ser el motor de búsqueda que ofrece los mejores resultados en esta prueba. Se puede percibir, que Lycos a pesar de que es el motor de búsqueda que ofrece las prestaciones de búsqueda más inferiores en balance con el resto de motores de búsqueda, se observa una inclinación muy acentuada, con la tendencia a aumentar la precisión cuando la exhaustividad disminuye, lo que hace reflexionar, la propensión de suministrar información útil ante una proporción de información existente.

Por otro lado se observa en esta prueba que Google es el sistema de búsqueda que mejor desempeño obtiene en cuanto a información pertinente recupera, aproximándose al 100% de suministrar documentos útiles; luego le sigue Hotbot y Excite los que los hace considerar como motores con grandes probabilidades de recuperar documentos de posible interés. Sin embargo se puede observa que Excite a pesar de tener un buen nivel de pertinencia de 0.85, presenta un comportamiento no muy acentuado cuando la exhaustividad disminuye, lo que hace pensar definitivamente que sus búsquedas tienden a ser mas exhaustivas.

Altavista y Yahoo, pueden percibirse como sistemas ideales para búsquedas tanto precisas como exhaustivas. MSN Search al igual que Lycos, también es considerado como el motor que no ofrece los mejores niveles de Precisión en relación a sus contendientes, sin embargo se puede resaltar la presencia de valores satisfactorios siempre que la exhaustividad es baja, en caso contrario se observa una curva mas acentuada para búsquedas mas exhaustivas, lo que significa que esta propensa a recuperar documentos con mas aciertos falsos, lo que hace pensar su ordenamiento por relevancia al igual que Lycos puede verse afectada de cierta manera.

Finalmente se puede concluir, que de acuerdo a los resultados presentes en este experimento se considera que todos los motores de búsqueda tienden a mantener un grado de precisión ideal, de acuerdo a estudios anteriores entre ellos Salton 1983, donde considera un motor con resultados pobres cuando su precisión tiende a ser mucho inferior a 0.5, característica que todos los buscadores superan ampliamente.

4.8 CRITERIO BASADO EN LA SIMILITUD DE RESULTADOS

En la Tabla3, se ha conseguido los valores medios de las similitudes obtenidas para cada par de motores con una muestra de diez documentos recuperados y cuyos documentos cumplen los niveles de relevancia (2 ó 3), en caso de existir un documento considerado poco relevante (nivel 1) en ambos motores, éste será tenido en cuenta como parte del cálculo de similitud, en caso contrario, será excluido.

PROMEDIO DE LAS SIMILITUDES OBTENIDAS MOTOR A MOTOR

M_A-M_E	M_A-M_G	M_A-M_H	M_A-M_M	M_A-M_L	M_A-M_Y	M_E-M_G	M_E-M_H	M_E-M_M	M_E-M_L
0,789	0,654	0,706	0,689	0,572	0,922	0,607	0,612	0,504	0,682

M_E-M_Y	M_G-M_H	M_G-M_L	M_G-M_M	M_G-M_Y	M_H-M_M	M_H-M_L	M_H-M_Y	M_M-M_L	M_M-M_Y	M_L-M_Y
0,794	0,891	0,595	0,567	0,644	0,583	0,56	0,619	0,559	0,545	0,553

Tabla3. Resultados de similitudes medias obtenidas en este experimento para cada par de motores con 10 documentos analizados.
(M_A: Altavista, M_E: Excite, M_G: Google, M_H: Hotbot, M_L: Lycos, M_Y: Yahoo)

Como podemos ver, los valores más similares y por tanto más próximos en un espacio vectorial, son lo motores Yahoo-Altavista y Google-Hotbot, con una similitud de 92,2% y 89,1% respectivamente, lo que resulta lógico debido a que comparten un porcentaje elevado de documentos y emplean o comparten sus bases de datos.

Vemos que Excite y MSN Search presentan el valor mínimo de toda la distribución, es decir se encuentran alejados a la distancia mas grande que se ha detectado en esta prueba.

En la tabla4, se obtuvo la distancia media de cada motor con respecto a todos los demás, par su cálculo se ha obtenido la media obtenidas por un motor con respectos a los seis, con valores correspondientes a los diez primeros documentos recuperados.

	SIMILITUD	DISTANCIA
Altavista	0,722	0,278
Excite	0,664	0,34
Google	0,686	0,313
Hotbot	0,685	0,315
MSN Search	0,574	0,426
Lycos	0,587	0,413
Yahoo	0,679	0,32

Tabla4. Similitud y Distancia media de cada motor con respecto al resto.

La distancia media resulta de restar la unidad menos la similitud (1-Similitud). Se agrupan las distancias obtenidas, en un gráfico radial, estableciendo en torno a un hipotético centro de gravedad, que podría considerarse como el núcleo de la colección de documentos de la Web, se obtendría la siguiente grafica:

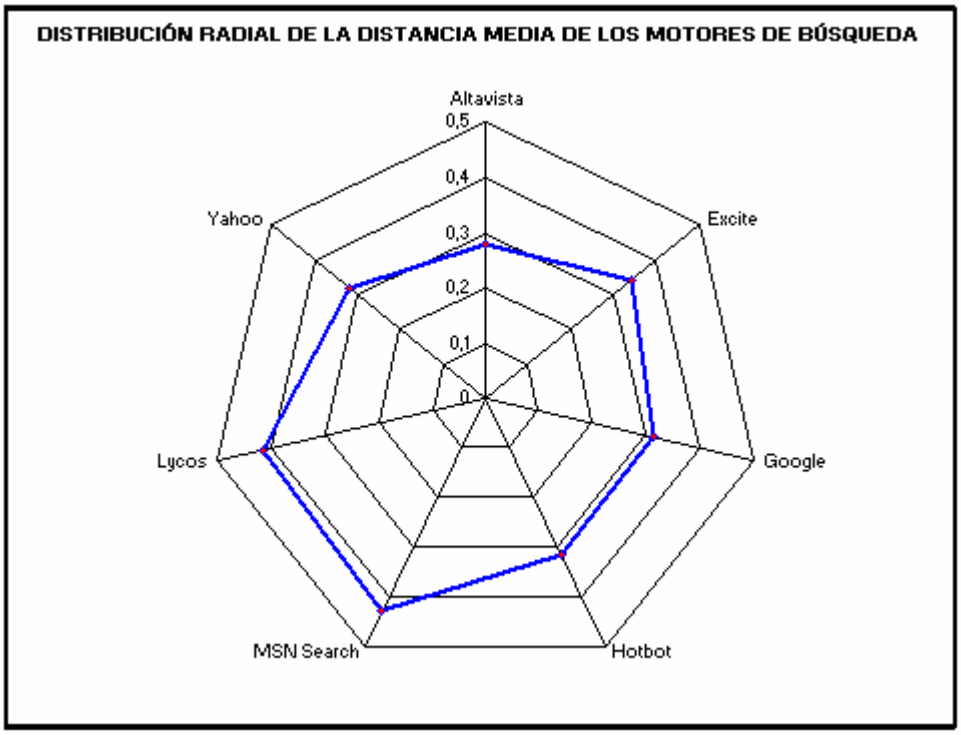


Figura16. Distribución radial de la distancia media de los buscadores respecto a un centro que constituye la mejor colección de documentos.

El interés de esta prueba es observar cuál es el motor de búsqueda cuya distancia media se acerca al centro, que constituye la mejor colección de documentos posible a recuperar en la web sobre los diez primeros documentos recuperados. Vemos que las distancias

con respecto a esa colección ideal no son realmente tan grandes. Por tanto en la figura16, podemos observar que el motor más cercano al centro es Altavista, seguido de muy cerca Yahoo, Google y Hotbot. De igual manera los resultados más distantes los obtiene Lycos y MSN Search.

4.9 CRITERIO BASADO EN EL ANÁLISIS DE AGRUPAMIENTO (CLUSTERING)

En seguida se aplica el método de agrupamiento del "Promedio Aritmético", a las distancias existentes entre los motores de búsqueda en estudio, cuando se analizan los diez primeros documentos recuperados.

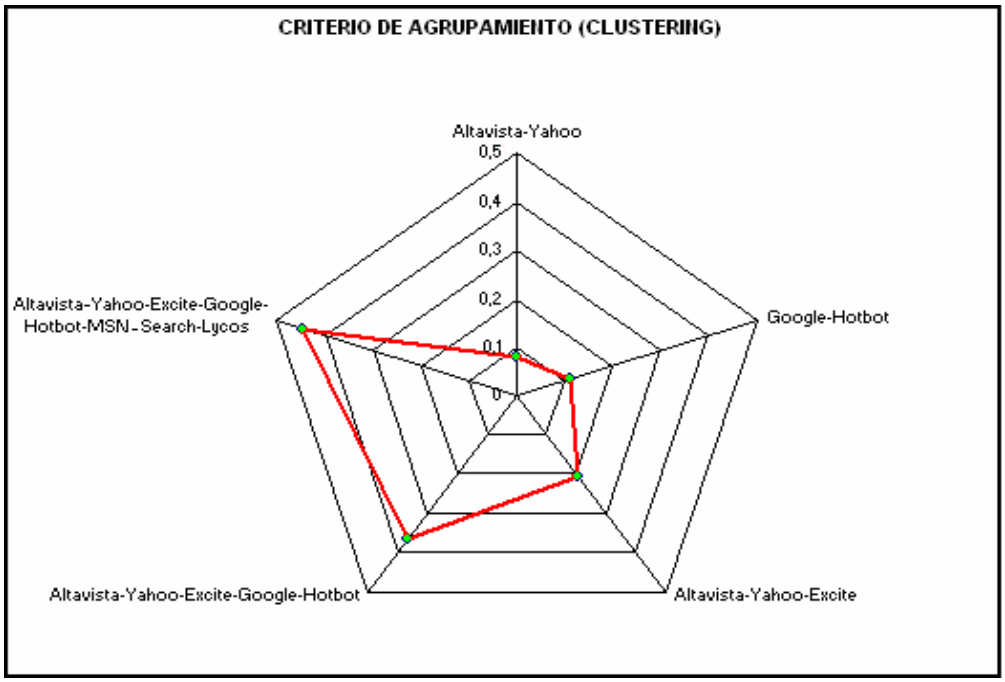


Figura17. Agrupamiento de Motores de Búsqueda según su afinidad sobre los diez documentos recuperados en base a una distribución radial de la distancia media existente entre ellos.

DISTANCIAS PROMEDIO OBTENIDAS MOTOR A MOTOR

	B	C	D	E	F	G
A	0,21	0,346	0,294	0,311	0,428	0,078
B		0,393	0,388	0,496	0,318	0,206
C			0,109	0,405	0,433	0,356
D				0,417	0,44	0,381
E					0,441	0,455
F						0,447

Tabla5. Distancias Medias de cada unos de los motores de búsqueda: A: Altavista, B: Excite, C: Google, D: Hotbot, E: MSN Search, F: Lycos, G=Yahoo

El algoritmo de agrupamiento aplicado sobre las distancias medias de la tabla anterior proporciona los siguientes agrupamientos:

Agrup.	Motores	Distancia
1	Altavista-Yahoo	0,078
2	Google-Hotbot	0,109
3	Altavista-Yahoo-Excite	0,206
4	Altavista-Yahoo-Excite-Google-Hotbot	0,368
5	Altavista-Yahoo-Excite-Google-Hotbot-MSN Search-Lycos	0,447

Tabla6. Agrupamientos y distancias entre motores con las muestras obtenidas en el análisis de los diez primeros documentos recuperados.

Como podemos observar fueron 5 agrupamientos logrados, y las distancias del último agrupamiento fue inferior del 0.50 lo que significa que el 45% (debido a la distancia de 0.447) de los índices de los motores los hacen absolutamente diferentes. Otro aspecto a resaltar es que los valores de distancia a los que se llevaron a cabo los agrupamientos son pequeños, lo que ratifica la suposición planteada sobre la relativa similitud existente entre los vectores resultados. Se observa que con respecto a una hipotética respuesta ideal, vemos que MSN Search y Lycos como los últimos sistemas de búsqueda que se añaden a los agrupamientos, confirmándose una vez más, que éstos son los que poseen las mayores distancias frente a los demás sistemas comparados.

Finalmente podemos mencionar que el comportamiento relacionado con los agrupamientos de los motores de búsqueda de este estudio, muestran distancias que oscilan de 0.078 y 10,9 siendo estos los valores mas pequeños, reiterando aun más sobre los criterios realizados anteriormente, el alto grado de coincidencia de contenidos percibido entre Altavista-Yahoo, Google-Hotbot respectivamente lo que confirma que comparten gran parte de su base de datos debido al nivel de semejanza que en los resultados refleja estos sistemas. En la grafica 17 se ve claramente la distribución radial de las distancias medias a la que se ve representada cada uno de los agrupamientos frente a un núcleo o colección de datos ideal. (Ver anexo7 donde se muestra todo el procedimiento llevado a cabo para los 5 agrupamientos obtenidos).

TABLA7. SINOPSIS DE CRITERIOS DESASARROLLADOS

SIST. DE BUSQUEDA / CRITERIOS	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo	Mejor Desempeño
CRITERIO BASADO EN EL TAMAÑO PROMEDIO DEL INDICE (Dctos)	1'160,688	71,3	614,167	133,144	445,858	184,371	1'188,756	Yahoo y Altavista
CRITERIO BASADO EN EL TIEMPO PROMEDIO DE RESPUESTA (Seg).	1,85	2,48	1,09	1,14	1,53	3,32	2,01	Google
CRITERIO BASADO EN EL RUIDO DOCUMENTAL(conteo sobre los 30 Dctos recuperados en 18 consultas).	Inactivos:15 Duplicados:77 Irrelevantes:45	Inactivos:22 Duplicados:95 Irrelevantes:67	Inactivos:15 Duplicados:63 Irrelevantes:55	Inactivos:17 Duplicados:63 Irrelevantes:57	Inactivos:8 Duplicados:67 Irrelevantes:88	Inactivos:31 Duplicados:90 Irrelevantes:80	Inactivos:18 Duplicados:62 Irrelevantes:59	Altavista, Google, Hotbot y Yahoo
CRITERIO SEGÚN EL PROMEDIO DE RELEVANCIA, EN FUNCION DE PRUEBAS (% promedio sobre los 30 Dctos recuperados, nivel 2 ó 3).	77,7	76,0	77,0	78,3	70,7	68,0	79,3	Yahoo seguido por Hotbot, Altavista y Google.
CRITERIO SEGÚN EL PROMEDIO DE RELEVANCIA EN FUNCIÓN AL TIPO DE CONSULTA: ESPECIALIZADA Y GENERAL(% nivel de relev. 2ó3 y los 10 primeros Dctos recuperados)	Especifica: 83 General: 80	Especifica: 93 General: 81	Especifica: 91 General: 83	Especifica: 82 General: 79	Especifica: 87 General: 78	Especifica: 66 General: 72	Especifica: 84 General: 76	Específica: Google, Excite seguido por MSN y General: Google, Excite, Altavista
CRITERIO SEGÚN EL PROMEDIO DE RELEVANCIA EN FUNCIÓN AL TIPO DE EXPRESIONES DE BÚSQUEDA PLANTEADAS (% sobre los 10 primeros resultados, nivel de relev 2 ó 3).	Palabra: 80 Lenguaje Nat: 90 Booleana:76 Frase literal:82 Frase-Bool.:88	Palabra: 70 Lenguaje Nat: 80 Booleana: 78 Frase literal: 78 Frase-Bool.: 84	Palabra: 90 Lenguaj. Nat: 70 Booleana: 75 Frase literal: 80 Frase-Bool.: 88	Palabra: 90 Lenguaj. Nat: 70 Booleana: 75 Frase literal: 82 Frase-Bool.: 86	Palabra: 95 Lenguaje Nat: 40 Booleana: 70 Frase literal: 76 Frase-Bool.: 82	Palabra: 70 Lenguaj. Nat: 70 Booleana: 60 Frase literal: 62 Frase-Bool.: 74	Palabra: 90 Lenguaje Nat: 80 Booleana: 85 Frase literal: 90 Frase-Bool.: 88	MSN Search en búsqueda de una sola palabra, en LN. sobresale Altavista y finalmente Yahoo sobresale en búsquedas Booleanas , Frase literal y combinacion de ambas sintaxis.
CRITERIO BASADO EN EL PROMEDIO DE EXHAUSTIVIDAD-PRESICION((% nivel de relev. 2ó3 y los 10 primeros Dctos recuperados)	VER ANALISIS EN LA GRAFICA 14.							Google principalmente.
CRITERIO BASADO EN LA SIMILITUD DE RESULTADOS DE CADA MOTOR CON RESPECTO A LOS DEMAS (% sobre los 10 primeros documentos nivel de relv. 2 ó 3)	72	66	69	69	57	59	68	Altavista, seguido por Google, Hotbot y Yahoo
CRITERIO BASADO EN LA DISTANCIA MEDIA MOTOR A MOTOR (nivel de relv. 2 ó 3)	28	34	31	31	43	41	32	Frente a una colección de datos ideal sobresale; Altavista, Google Hotbot y Yahoo.
CRITERIO BASADO EN EL ANÁLISIS DE AGRUPAMIENTO (% sobre los 10 primeros documentos nivel de relv. 2 ó 3)	RESULTADO OBTENIDO 5 AGRUPAMIENTOS							Motores mas afines: Yahoo-Altavista y Google-Hotbot

OBSERVACIONES

En el primero y segundo criterio se exponen los datos obtenidos en la última prueba analizada, por ser esta la mas actualizada. Sin embargo como se vio en el análisis de las dos muestras, no hubo cambio sustancial en la cobertura de cada índice entre una muestra y otra.

En el tercer criterio se observa el mejor desempeño de los sistemas de búsqueda en su proceder ante el indicador total de ruido sobre los 540 documentos recuperados.

Los restantes criterios fueron analizados en base a los 10 primeros documentos recuperados.

APÉNDICE A

CONCEPTUALIZACIÓN GENERAL DE RECUPERACION DE INFORMACION

La recuperación de información es una actividad que el ser humano realiza, consciente e inconscientemente, casi continuamente, y en el marco de cualquier otra actividad. La necesidad de resolver una duda, documentar una información o estudio, son expresiones clásicas de los procesos de recuperación de información. Con el desarrollo de los sistemas digitales de procesamiento de datos y de tratamiento de información, las técnicas de recuperación de información han ido desarrollando un conjunto de teorías y aplicación práctica que subyacen en la actualidad a cualquier actividad en entornos informáticos, y que resultan ser la base del descubrimiento, búsqueda y recuperación de información en Internet.

La importancia de determinar qué se entiende por recuperación de Información (IR), puede inferirse a partir de su propia expresión, que consiste en la representación, almacenamiento, organización y acceso a documentos de la información. La representación y la organización de los documentos de la información deberían proveer al usuario el acceso fácil a la información que precise en un momento dado. Sin embargo, el manifiesto de la "necesidad de la información del usuario", desde el contexto World Wide Web, no es problema simple de manejar, debido a que el usuario claramente debe traducir su necesidad informativa en una pregunta que pueda ser procesada por el motor de búsqueda (o sistema de recuperación de información web, SIR). En su forma más común, esta traducción se fundamenta en una serie de palabras claves (o términos de índice) que resume la descripción de la necesidad de información del usuario. Por tanto, dada la pregunta de usuario, el objetivo determinante del sistema de recuperación de información es precisamente recuperar la información que podría ser útil o relevante al usuario. El énfasis está en la recuperación de la información en comparación con la recuperación de datos.

Actualmente dentro de la recuperación de la información podemos incluir varios temas como son el modelado, la jerarquización, la clasificación de documentos, el filtrado, la

manera en la que se almacena y se accede a la información y los lenguajes que permiten hacer dichas funciones[19]. En la figura1, se ilustra el proceso de RI, en el cual por medio de la interfaz de usuario, se inician las operaciones de texto, las cuales pueden ser utilizadas para el procesamiento de una consulta o la indexación de un documento. Una vez realizado el proceso de una consulta se puede iniciar con el proceso de búsqueda en índices, para así concluir con la jerarquización.

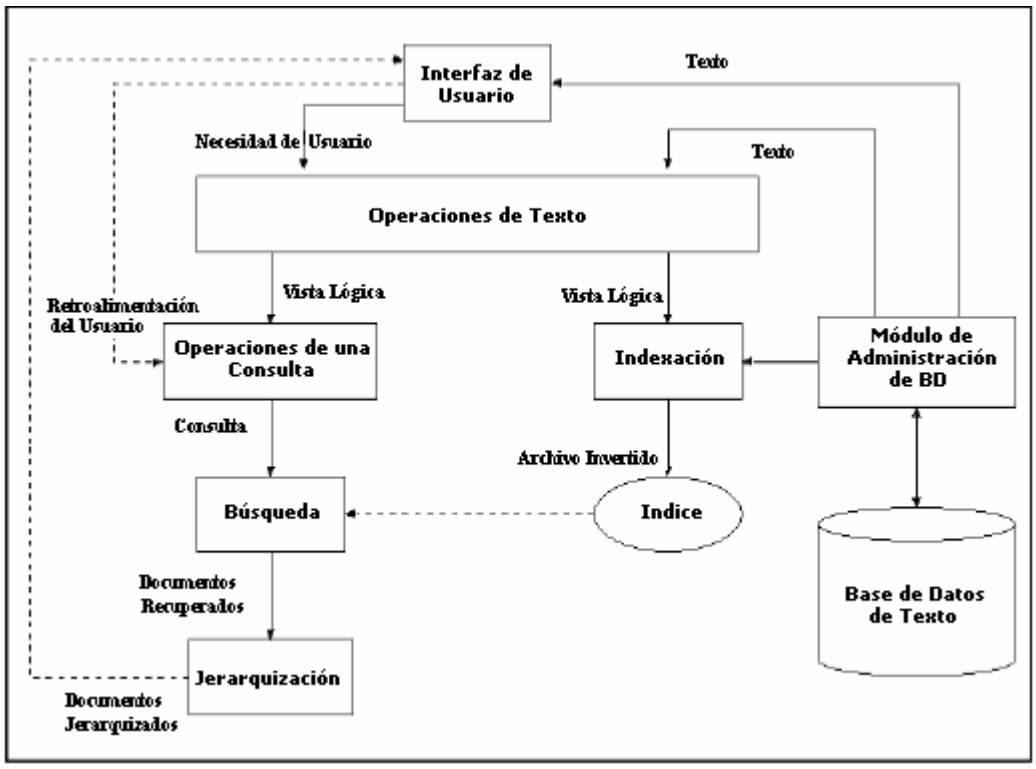


Figura18. Proceso de Recuperación de información según Baeza-Yates y Ribeiro-Neto, 1999

A.1 INCONVENIENTES PREVISIBLES EN LA RECUPERACIÓN DE INFORMACIÓN

El esquema de la recuperación de la información parece sencillo, a simple vista: se representa, se busca y se mira. Sin embargo ante la paradoja de la incertidumbre que rodea este tema de investigación, cabe destacar dos problemas prácticos que inciden en él, con el fin de ayudar al usuario a dar una visión y comprensión del grado de complejidad que le asiste.

[19] Baeza-Yates y Ribeiro-Neto, 1999

A.1.1 RECUPERACION DE LA INFORMACIÓN Vs RECUPERACION DE DATOS.

La recuperación de datos, en el contexto de un sistema de recuperación de información (SIR), consiste principalmente en la determinación de qué documentos de un grupo, contienen las palabras claves a la pregunta expuesta por el usuario con la mayor frecuencia posible, aunque no satisfaga la necesidad de información del usuario. De hecho, un usuario de un sistema de recuperación de información le interesa más el recuperar la información sobre un tema específico, que con el recuperar los datos que satisfacen a una pregunta dada.

Un lenguaje de recuperación de datos tiene como objetivo el recuperar todos los objetos que satisfacen las condiciones claramente definidas tales como una expresión regular (especificación completa) o una expresión algebra relacional (Sistema Gestor de base de datos). Así, que para un sistema de recuperación de datos, un solo objeto erróneo entre mil recuperados significa el fracaso total. Para un sistema de recuperación de información, sin embargo, los objetos que pudieran ser inexactos y con pequeños errores pueden probablemente pasar inadvertidos. La razón principal de esta diferencia es que los sistemas de recuperación de información, manejan texto de lenguaje natural que no siempre esta bien estructurado y podría ser semánticamente ambiguo. De otra parte, un sistema de recuperación de datos (como una base de datos relacional) se ocupa de datos que tienen una estructura y semántica bien definida.

La Recuperación de datos, mientras que proporciona una solución al usuario de un sistema de la base de datos, no soluciona el problema de la recuperación información sobre un tema o asunto. Para ser eficaz en su intento de satisfacer la necesidad de información del usuario, el sistema de recuperación de información debe “interpretar” de alguna manera el contenido de los artículos de la información (documentos) en un grupo y alinearlos según el grado de importancia a la pregunta de usuario. Esta interpretación del contenido del documento implica extraer del texto del documento, información sintáctica y semántica y usar esta información para señalar la necesidad de información del usuario. La dificultad no es solamente saber como extraer esta información, sino también saber

utilizarla para decidir su relevancia. Así, que la noción de "relevancia" está la finalidad de la recuperación de la información. De hecho, el objetivo fundamental de un sistema recuperación de información (SIR) es recuperar todos los documentos que son relevantes a una pregunta de usuario mientras que recupera como sea posible pocos documentos no-relevantes.

Con la aclaración previamente mencionada, la recuperación de información de datos frente a la recuperación de la información son grupos de definiciones con diferencias bien establecidas. Incluso podemos mencionar algunos autores que han señalado criterios frente a este cometido:

Meadow piensa que la recuperación de la información "se trata de una disciplina que involucra la localización de una determinada información dentro de un almacén de información o base de datos"[20]. Este autor, implícitamente asocia el concepto de recuperación de información con el concepto de selectividad, ya que la información ha de extraerse siguiendo algún tipo de criterio discriminatorio (selectivo por tanto).

Pérez-Carballo y Strzalkowski redundan en este tema, indicando que "una típica tarea de la recuperación de información es traer documentos relevantes desde una gran archivo en respuesta a una pregunta formulada por un usuario y ordenar estos documentos de acuerdo con su relevancia"[21].

Grossman y Frieder van más allá cuando indican que "la recuperación de información es encontrar documentos relevantes, no encontrar simples correspondencias a unos patrones de bits"[22].

Feather y Storges ven a la recuperación de información como "el conjunto de actividades necesarias para hacer disponible la información a una comunidad de usuarios" [23].

[20] Meadow, C.T. text information retrieval system. San Diego: Academic Press 1993.
[21] Perez Carballo, J. and Strzalcowski, T. 'Natural language information retrieval: progress report' Information processing and Management 36, 2000. p. 155-178.
[22] Grossman, D.A. and Frieder, O. Information retrieval: algorithms and heuristics. Boston: Kluwer Academia Publishers, 1998.
[23] Internacional Enciclopedia of information & Libray Science. London: Rotledge, 1997

Tramullas Sáez impregna su definición al afirmar que "el planteamiento de la recuperación de información hace referencia a los mecanismos más adecuados para extraer, de un conjunto de documentos, aquellos que fuesen pertinentes a una necesidad informativa dada" [24].

Blair establece una serie de diferencias entre los términos recuperación de datos (data retrieval) y recuperación de información (information retrieval)[25]:

- En recuperación de datos se emplean preguntas altamente formalizadas, cuya respuesta es directamente la información deseada. En cambio, en recuperación de información, las preguntas resultan difíciles de trasladar a un lenguaje normalizado (aunque existen lenguajes para la recuperación de información, son de naturaleza mucho menos formal que los empleados en los sistemas de bases de datos relacionales, por ejemplo) y la respuesta será un conjunto de documentos que pueden contener, sólo probablemente, lo deseado, con un evidente factor de indeterminación.
- Según la relación entre el requerimiento al sistema y la satisfacción de usuario, la recuperación de datos es determinística y en recuperación de información es probabilística, a causa del nivel de incertidumbre presente en la respuesta.
- Según el éxito de la búsqueda. En recuperación de datos el criterio a emplear es la exactitud de lo encontrado, mientras que en recuperación de información, el único criterio de valor es la satisfacción del usuario, basada en un criterio personal de utilidad.
- Según la rapidez de respuesta: En recuperación de datos depende del soporte físico y de la perfección del algoritmo de búsqueda y de los índices. En recuperación de información depende de las decisiones y acciones del usuario durante el proceso de interrogación.

Para diferenciar claramente ambos enfoques, piense en los resultados de formular una consulta contra un sistema de gestión relacional de base de datos. Una consulta en éstos exige un dato en un campo, y la respuesta no deja lugar a duda. Es un sistema cerrado,

[24] Tramillas Saez. J. Introducción a la Documática. Zaragoza: Kronos, 1997.
[25] Blair, D.C. Language and representation in information retrieval. Amsterdam: Elsevie Science Publisher, 1990.

determinado, de respuesta exacta. Si no se obtiene respuesta, es porque no la hay, o porque el usuario a introducido los datos mal. En cambio, en un sistema de recuperación de información textual, no basta con obtener respuestas, hay que valorar si éstas son adecuadas o no, al usuario, y volver a formular la búsqueda en caso de que sea necesario.

	Recuperación de Datos (DR)	*Recuperación de inform.* (IR)
Correspondencia	Exacto	Parcial, la mejor
Inferencia	Deducción (algebraica)	Inducción
Modelo	Determinista	Probabilistica
Clasificación	Monocategórico	Policategórico
Lenguaje de interrogación	Artificial (Fuertemente Estruct.)	Natural (Estructurado o natural)
Especificación de la pregunta	Precisa	Imprecisa
Artículos deseados	Correspondencia	Relevante
Error en la respuesta	Sensible	Insensible

Tabla 8. Diferencias existentes entre recuperación de datos o recuperación de información. [26]

A.2 ENFOQUE EVOLUTIVO Y CLASIFICATORIO DE LOS SISTEMAS DE RECUPERACION DE INFORMACION DE DOCUMENTOS TEXTUALES.

Existen varios modelos de recuperación de información que permiten realizar dicha tarea, y que tienen una gran diferencia con los algoritmos de búsqueda ya que éstos proporcionan un valor que dice que tan relevante puede ser un documento para el usuario.

Dentro de la recuperación de información es muy importante entender dos conceptos básicos como son la tarea de usuario y la vista lógica de los documentos. Ambas son

[26] En la RD se maneja una clasificación monocatgórica, es decir, son clases defincias por objetos que poseen atributos necesarias y suficientes para pertenecer a una clase. En IR se maneja la clasificación policategórica. En tal clasificación cada tipo dentro de una clase poseerá solo una proporción de todas los atributos poseídos por todos los miembros de esa clase. Por lo tanto no hay atributo necesario, ni suficiente para miembros de una clase.
<http://www.dcs.gla.ac.uk/~iain/keith/data/pages/2.htm> [Consulta: 29 de enero del 2005]. Fuente: C.J. van Rijsbergen. Information Retrieval. Department of Computing Science, University of Glasgow, 1999.

importantes dentro del contexto de la recuperación de información ya que permiten tener una vista amplia de los documentos y así como el objetivo que tiene el usuario.

Tarea del Usuario

La tarea del usuario se refiere más precisamente a la actividad que el usuario quiere realizar para obtener la información que requiere. Se relaciona ampliamente con la traducción de las necesidades del usuario en un lenguaje que pueda ser comprendido por el sistema permitiéndole recomendar o recuperar los documentos que contengan dicha información[27].

Para Baeza y Rebeiro la tarea del usuario puede ser de dos tipos: recuperación y navegación, ambas importantes para la decisión de los modelos a utilizar. Cuando se habla de recuperación, se hace referencia más precisamente a la búsqueda de documentos dentro de una colección partiendo de una petición o simplemente de un perfil dado. Sin embargo, la navegación se refiere mas bien al recorrido analítico de los documentos mientras llegamos a lo que en realidad necesitamos, incluso es posible que no tengamos inicialmente un objetivo determinado. Dicha tarea es fácil realizarla mediante los sistemas de hipertexto ya que éstos permiten ligar directamente dichos documentos presentando la información en formato digital.

Es posible realizar la recuperación de dos maneras diferentes, las cuales permiten la utilización de los mismos modelos de recuperación de información para lograr un buen resultado. Grossman y Frieder 1998 [64], mencionan dos modos operacionales como son la recuperación ad hoc y el filtrado de documentos.

En el proceso de recuperación de Información ad hoc la colección de documentos permanece estática mientras el sistema recibe diferentes peticiones de búsqueda, y así se buscan aquellos que parecen ser relevantes dentro de dichas colecciones y al final se realiza un jerarquización de acuerdo al grado de relevancia.

[27] Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval, 1999 <http://www.sims.berkeley.edu/~hearst/irbook/1/node3.html#fig:user-tasks>

Por el contrario, en el proceso de filtrado de documentos dentro de la recuperación de información, las peticiones permanecen estáticas y es la colección de documentos la que está en constantes cambios, es decir ellos llegan al sistema sin necesidad de hacer una petición en específico. Entro de este último proceso, es necesario tener almacenadas las características y gusto s del usuario, para que se comparen con los documentos y así se tome la decisión de que le será mas interesante.

Vista Lógica de los documentos

Como se ha mencionado es muy importante, la manera en la que se almacena información, así como el tipo gramatical que fuesen los términos dentro de los índices que describen a las colecciones.

La vista lógica se refiere a la manera en la que se presenta un documento dentro de sus índices. La forma más sencilla de representar un documento es por medio del conjunto de palabras del texto completo, sin embargo éste puede llegar a ser muy grande y por ello es conveniente reducirlo a una lista con las palabras clave del texto. De ahí se obtiene la primera forma de representación que se llama texto completo. Las vistas lógicas, pueden variar de acuerdo a los diferentes tipos de operaciones que se apliquen al texto, entre las cuales se destacan[28]: la eliminación de palabras que no dan una información de importancia sobre el contenido del documento; La tematización morfológica, que consiste en la reducción de las palabras a su raíz; agrupar las palabras representativas por sinónimos; la extracción manual de términos representantes; representar frases que puedan tener la misma representación dentro del índice, por ejemplo frases como: “recuperación de información”, “recuperando información” o “recuperar información relevante” pueda ser representada como “recupera+información”.

[28] Baeza-Yates, R. and Frakes, W.B. Information retrieval: data structures & algoriths Englewood Cliffs, New Yersey: Prentice Hall, 1992.

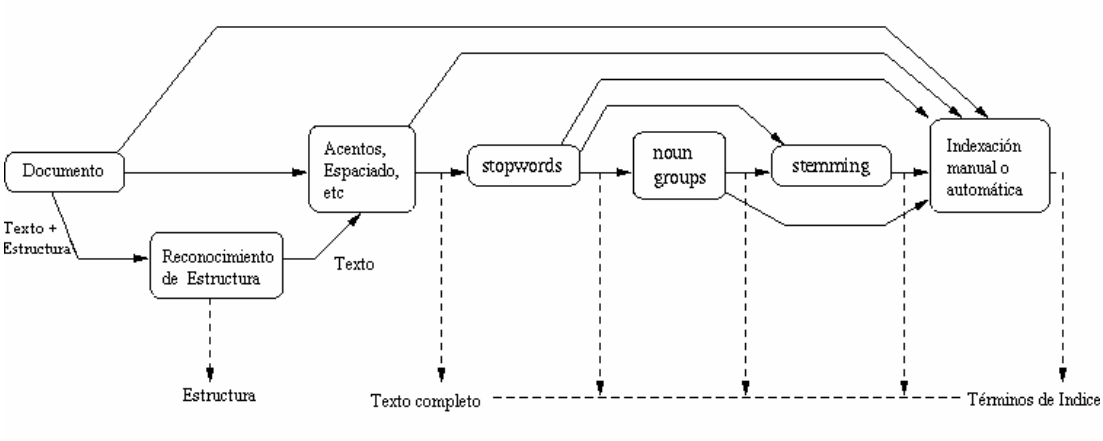


Figura 19: Vista lógica de la entrada de los documentos a los SRI, de Baeza-Yates y Ribeiro-Neto[1]

En la figura19, se observa el proceso de eliminación o extracción de palabras, mediante la Indización (manual o automática), con esta operación, se pretende captar y representar el contenido de los documentos, el cual persigue eliminar la presencia de términos ambiguos en los índices de las bases de datos, contribuyendo a la eficacia de su operatoria y a mejorar su consistencia.

A continuación se presenta otra vista lógica o funcional de un sistema de recuperación de Información. A partir de la entrada del documento textual en el SRI se antepone un preprocesamiento del mismo, cuyas fases están determinadas por:

1. A cada documento que entra se le asigna un Identificador
2. Se identifican o analizan todas y cada una de las palabras contenidas en el documento.
3. Se excluyen las palabras vacías, cuyo objetivo es reducir por filtrado el número de términos a tratar, suponiendo que éstos no añaden información significativa a la representación.
4. Identificación de raíces, proceso llamado stemming, se trata de identificar y aislar las raíces determinantes de las palabras (eliminación de prefijos y sufijos).
5. Se establece un peso de ponderación para cada raíz.
6. Finalmente las raíces debidamente ponderadas se introducen en la base de datos

Cuando el usuario lleva a cabo una operación de recuperación de información, acaecerán los siguientes procesos:

1. El usuario en función de sus necesidades y conveniencias lleva a cabo una serie de juicios de relevancia para confeccionar su ecuación de búsqueda, ayudándose de las prestaciones que le proporciona el Interfaz de Búsqueda.
2. La ecuación de búsqueda, una vez introducida, se descompone en sus partes fundamentales.
3. Los términos clave empleados en la ecuación de búsqueda son "cortados" para extraer de ellos sus raíces y de esta forma proceder a su localización en la base de datos.
4. Una vez localizados los distintos subconjuntos de documentos asociados a los términos clave, se llevan a cabo las operaciones booleanas pertinentes, que han sido introducidas por el usuario en la ecuación de búsqueda.
5. Posteriormente los documentos pueden alinearse para su presentación según un ranking determinado.

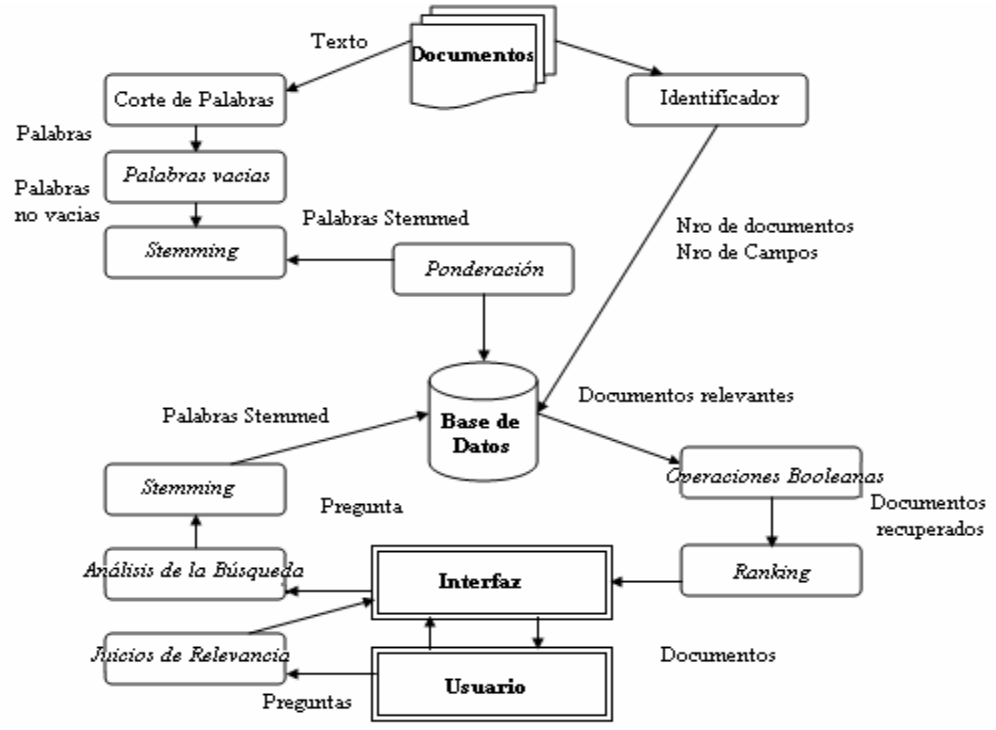


Figura20. Vista funcional o lógica asociada con un tipo común de SRI basado en el modelo Booleano

A.2.1 REGISTRO EVOLUTIVO DE LOS SISTEMAS DE RECUPERACION DE INFORMACION

El devenir asociado con la recuperación de la información, es expuesta de manera resumida por Baeza-Yates en tres fases fundamentales:

Sus principios preceden de antiguos métodos de colección de información cuyo ejemplo típico era el contenido de un libro. Debido al eventual crecimiento del volumen de la información, se hizo necesaria la construcción de estructuras de datos especializados que asegurarán un acceso más rápido a la información almacenada. A partir de esto, surge la aparición de la estructura de datos más antigua y popular denominada "Índice" siendo este el corazón actual de los sistemas de recuperación de datos modernos.

Seguidamente fueron las bibliotecas las primeras instituciones en adoptar sistemas de RI cuyos desarrollos fueron realizados inicialmente en instituciones académicas para posteriormente ser introducidos al mercado comercial. Sus desarrollo tecnológico data desde sistemas basados en una automatización de tecnologías tales como catálogos de tarjeta y búsqueda básica, incluso mas adelante se dio el valor añadido de búsqueda por palabras claves, búsqueda por titulo e implantación de preguntas mas complejas, hasta, el actual despliegue enfocado en interfaces gráficas mejoradas, características hipertexto y arquitecturas de sistema abierto con procesamiento automatizado.

Finalmente Ricardo Baeza concluye con la mención de tres cambios dramáticos y fundamentales que han ocurrido debido a los avances en la tecnología de la informática moderna y el auge del Web. Primero, el acceso a las fuentes de información se hizo mas barato, dando una cobertura de audiencia más amplia a la antes disponible. Segundo, los avances en la comunicación digital permitió el acceso a las redes, implicando que información disponible en localizaciones distantes, estuviera al alcance y de manera rápida. Tercero la facilidad y libertad que posee cualquier usuario, de hacer público cualquier información que considere útil.

Fundamentalmente, la panorámica reflejada han permitido al usuario usar el Web y bibliotecas modernas digitales como un medio sumamente interactivo y de gran conveniencia sobre el servicio que prestan. Sin embargo, tal interactividad que ha llevado a un cambio fundamental, es el actual paradigma en la comunicación.

Hoy por hoy, la recuperación de información se asocia a menudo con motores de búsqueda en Internet. Sin embargo, deriva de una disciplina académica cuyas raíces datan de los 50. Durante su primera década las actividades de investigación normalmente tenían lugar en los departamentos de Ciencias de la computación y las aproximaciones más simples se basaban en las estadísticas de ocurrencia, que tenían una efectividad sorprendente en la recuperación de documentos relevantes. No obstante, un pequeño número de grupos de investigación en Recuperación de Información consiguieron resultados importantes en tres aspectos:

1. Teoría. Se desarrollaron modelos de recuperación probabilística que implicaban una óptima eficacia de recuperación (ver publicaciones de Cooper, Roberston y otros). Más tarde, la recuperación se extendió a otros medios, no solo texto.

2. Sistemas. Recientemente se han intentado varios algoritmos y estructuras de datos, e integrados sistemas de recuperación impracticables (p.e., SMART, Topic, y Sistema Inquiry), así como sistemas de recuperación multimedia.

3. Evaluación. Se han construido colecciones de pruebas consistentes en documentos, consultas y --más importante aún-- de aserciones de relevancia que determinan qué documentos son relevantes para qué consultas. Estas colecciones de pruebas facilitan la comparación de distintos métodos de recuperación en lo concerniente a relevancia y precisión (p.e. colecciones Cranfield, SMART y TREC).

Con el crecimiento de Internet, estos sistemas de Recuperación de Información constituyeron bloques de construcción listos para ser usados. La gran cantidad de datos así, como el espacio abierto de Internet llevado a nuevos y excitantes conceptos, como

enlaces basados en ranking, recuperación XML, integración de fuentes de datos heterogéneas, etc.

Para mencionar con mayor detalle esta evolución, en los últimos 30 años, los SRI han evolucionado de tal manera que podemos hablar de tres generaciones[29]:

Primera generación (1970 - 1985)

- SRI dirigidos a usuarios expertos
- Bases de datos referenciales
- Bases de datos de ciencia y tecnología
- Acceso a las bases de datos a través de servicios en línea
- Consultas en modo comando

Segunda generación (1985 - 1995)

- SRI dirigidos a usuarios expertos y a usuarios finales
- Aumentan las bases de datos a texto completo (ASCII)
- Bases de datos de todas las áreas de conocimiento
- Acceso a bases de datos a través de servicios en línea y sistemas en CDROM
- Consultas en modo comando y a través de menús

Tercera generación (Desde 1995)

- SRI dirigidos a usuarios finales
- Bases de datos e información electrónica multimedia
- Acceso a la información electrónica a través de redes y CD ROM
- Desarrollo de SRI en Internet

[29] Evolución de los SRI. Tipología y formatos. <Evolución de los SRI. Tipología y formatos>

A.2.2 MODELOS CONCEPTUALES PARA LA RECUPERACION DE INFORMACION

Dentro de la recuperación de información el problema central es la determinación de cuales documentos son relevantes y cuáles no. Para lograr esa decisión es importante poder dar un rango de valores que indiquen esa posibilidad y de este modo ordenarlos según un grado de importancia para el usuario.

Al hablar de modelos de recuperación de información, Sparck, K. y Willett, señalan que nos referimos a la aplicación de teorías para generar procesos que nos den esa jerarquización de los documentos y que también sean eficientes y permitan lograr un modelado de varios aspectos de un sistema de recuperación de información [65].

Según [VIL 1997], El diseño de un SRI efectivo, bajo un modelo, ha de incluir técnicas que definan "como se obtienen las representaciones de los documentos y de la consulta, la estrategia para evaluar la relevancia de un documento respecto a una consulta, los métodos para establecer la importancia (orden) de los documentos de salida y los mecanismos que permiten una realimentación por parte del usuario para mejorar la consulta"

Baeza-Yates y Ribeiro-Neto [1], señalan que en un modelo de recuperación no sólo se incluye la forma de encontrar una similitud entre un documento y la petición hecha por el usuario. También es importante considerar aspectos como es la manera en la que el usuario representa dicha necesidad de información y se comunica con el sistema; la interacción humano-computadora permite una mejor comprensión entre la computadora y el usuario la cual asegura el mejor tipo de procesamiento que es necesario hacer a una consulta; los ambientes cognitivos y sociales en los cuales toman lugar la comunicación e interacción con el sistema.

Existen varios niveles en los que los procedimientos de recuperación de información pueden ser modelados. Sin embargo hay que tener presente que dentro de todos ellos existen diferentes modelos que permiten lograr una recuperación mas eficiente o precisa

que otros. Quien mejor describe los modelos es Dominich quien clasifica los modelos de recuperación de una manera clara y concisa, mediante el cual señala 5 agrupaciones [DOM, 2000]:

MODELO	DESCRIPCION
Modelos Clásicos	Incluye los tres modelos más comunmente usados: booleano, espacio vectorial y probabilístico
Modelos Alternativos	Están basado en la lógica Fuzzy. Consiste en introducir una serie de palabras clave, generalmente ordenadas por preferencia. El sistema ordenará los documentos según alguna función de similitud que generalmente dará importancia a la diversidad de las palabras, y a su frecuencia de aparición.
Modelos Lógicos	Modelos basados en lógica formal. La recuperación de información se entiende como un proceso inferencial a través del cual se puede estimar la probabilidad de que una necesidad de información de un usuario, espresada como una o mas consultas, sea satisfecha ofreciendo u documento como "prueba"[VIL,1997].
Modelos basados en la Interactividad	Inclyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la relevancia de los doucumentos recuperados[SAL, 1989].
Modelos basados en la Inteligencia Artificial	Bases de conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural.

Tabla 9. Clasificación de los Modelos de Recuperación de Información según Dominich.[30]

Los modelos de recuperación de información han ido evolucionando partiendo de teorías iniciales y han sido extendidos para lograr una mejor eficiencia y precisión. Un modelo de recuperación de información es evaluado de acuerdo a qué tan bien identifica los documentos relevantes y elimina de la respuesta aquellos que no lo son, de manera que la proporción de documentos que son recuperados frente al numero de documentos que se cree que son relevantes se llama eficiencia y por otro lado, el factor de pertinencia o precisión se basa en la porción de documentos recuperados que son realmente relevantes [Swanson 1998].

[Baeza-Yates y Ribeiro-Neto 1999] distinguen principalmente dos divisiones importantes dentro de dichos modelos para realizar la tarea de recuperación ad hoc y filtrado de documento, que son: los modelos clásicos y los modelos estructurados. A los modelos clásicos pertenecen tres modelos importantes que son la base de muchos otros: modelo booleano, vectorial, probabilística. Y entre los modelos estructurados encontramos las listas coincidentes y el modelo de nodos próximos.

[30] Dominich, S. 'A unified mathematical definition of classical infomation retrieval'. Journal of the American Society for information Science, 2000. p. 614-624.

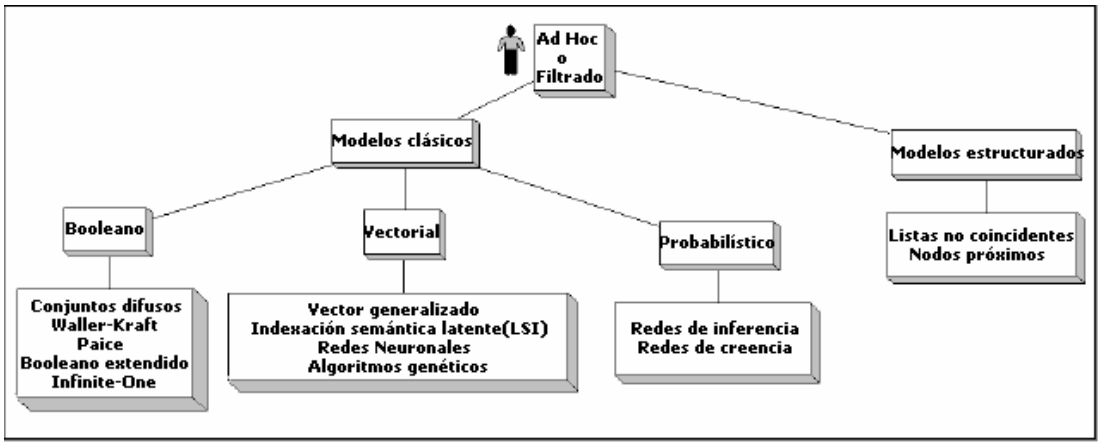


Figura21. Taxonomía de los modelos de RI [adoptado por Baeza-Yates y Ribeiro-Neto 1999]

Estos modelos son la base para el desarrollo de otros nuevos que son la base para el desarrollo de otros nuevos que han permitido mejoras en su eficiencia y precisión. [Lee 1994] enuncia, analiza y compara las diferentes extensiones que ha tenido el modelo booleano como son: el modelo basado en la teoría de conjuntos difusos, el modelo Waller-Kraft, el modelo Paice, el booleano extendido y el Infinite-One. Como extensiones del modelo vectorial encontramos los modelos de vector generalizado, indexación semántica latente (LSI), redes neuronales y algoritmos genéticos. Y por último, en la generalización del modelo probabilística tenemos las redes de inferencia y las redes de creencia.

Tanto los modelos clásicos como sus extensiones permiten realizar búsquedas en documento cuyas vista lógicas van de términos indexados a texto completo; sin embargo los modelos estructurados permiten la recuperación únicamente en documentos estructurados. En la figura anterior se muestra gráficamente la descripción anterior.

[Baeza –Yates] también hace mención sobre modelos basados en la navegación entre páginas web son de tres tipos: estructura plana, estructura guiada e hipertexto.

El primero es una simple lectura de un documento aislado del contexto, el segundo incorpora la posibilidad de facilitar la exploración organizando los documentos en una estructura tipo directorio con jerarquía de clases y subclases y el tercero se basa en la

idea de un sistema de información que de la posibilidad de adquirir información de forma no estrictamente secuencial sino a través de nodos y enlaces [BAE, 1999].

Baeza-Yates, también proporciona una clasificación adicional de estos modelos de recuperación de información, realizada en función de la modalidad de consulta y de la vista lógica de los documentos:

		VISTA LOGICA DE LOS DOCUMENTOS		
MODALIDAD		**Términos de Indice**	**Texto Completo**	**Texto Completo + Estructura**
	Recuperación	Clásicos	Clásicos	Estructurados
		Conjuntos teóricos	Conjuntos teóricos	
		Algebraicos	Algebraicos	
		Probabilísticos	Probabilísticos	
	Navegación			
		Estructura Plana	Estructura Plana	Estructura Plana
			Hipertexto	Hipertexto

Tabla10. Clasificación de los modelos de Recuperación de información según Baeza-Yates.[31]

Mediante la consulta, Watstein y Kesselman opina que un sistema de recuperación de información recibe los datos necesarios para iniciar su función. Las consultas pueden ser expresadas de diferentes formas, por ejemplo, unas incluyen sólo palabras clave, otras consultas aceptan palabras clave y conectivos lógicos y otras se expresan en lenguaje natural [68].

Para poder recuperar información en entornos digitales, es necesario que los sistemas de recuperación de información (SRI) implementen una especial estructura de datos, algoritmos y técnicas de recuperación de información. A continuación se estudia, fundamentos de fondo acerca de las diferencias y similitudes de estos sistemas, expuesto Prieto-Díaz[69].

[31] Baeza-Yates R. and Ribeiro-Neto B. Modern Information retrieval. New Cork: ACM Press: Harlow [etc.]: Addison-Wesley. 1999. 513p.

MODELO CONCEPTUAL	Booleano	Booleano extendido	Probabilístico	Búsqueda de cadenas	Espacio vectorial
ESTRUCTURA DE FICHEROS	Fichero Plano	Fichero Inverso	Patrones de bits	Árbol PAT	Grafos
OPERACIONES DE CONSULTA	Reutilización	Parsing	Booleanas	Clustering	
OPERACIONES SOBRE TERMINOS	Stemming	Ponderación por pesos	Lista de palabras vacías	Truncamiento	Indización
OPERACIONES SOBRE DOCUMENTOS	Visualización docs.	Rango	Ordenación	Enmascaramiento	Asignación Ids

Tabla 11: Modelos de Recuperación y atributos que los define.

Como se observa, la tabla recoge los atributos constantes que define cada uno de los sistemas de recuperación de información a partir de los cuales se encuentran clasificados. Cada atributo refleja un instante determinado en la toma de decisiones efectuada para llevar a cabo el desarrollo de la arquitectura de un SRI. El diseñador del sistema debe elegir, para cada atributo, un valor determinado de entre las alternativas dispuestas.

La recuperación de información clásica está dividida en tres modelos básicos: booleano, vectorial y probabilística, sobre los cuales se han definido diferentes variantes[32]. Estos modelos clásicos se basan en un modelo matemático formal para la recuperación, en donde los documentos están formados por conjuntos de términos que pueden ser individualmente ponderados y manipulados. Las consultas son ejecutadas comparando la representación de la consulta frente a la representación del documento en el espacio, pudiendo recuperar documentos que no contengan necesariamente alguno de los términos de búsqueda.

La mayoría de los sistemas de información son de dos tipos, booleanos y de búsqueda de información por patrones de texto. Las interrogaciones a los sistemas de búsquedas por patrones de texto se llevan a cabo por medio de cadenas de caracteres o por expresiones regulares. Los sistemas de patrones de textos son más utilizados comúnmente en pequeñas colecciones de datos y cuando hay que gestionar grandes volúmenes de documentos destacan mayoritariamente los sistemas booleanos.

El modelo booleano consiste en la utilización de la teoría de conjuntos y el álgebra Booleana. En este modelo cada término indexado es ponderado como presente o no

[32] R. Baeza-Yates, B. Ribeiro-Neto. "Modern Information Retrieval". Addison Wesley, 1999.

presente y todas las consultas se basan en expresiones booleanas[33]. Es el más simple de todos los modelos, si bien únicamente clasifica a los documentos como relevantes o no relevantes (sin una ordenación de los resultados).

El método de representación de los documentos es un conjunto de términos de indexación o palabras claves (palabras con un valor semántico importante), las cuales pueden se extraídas del contenido del documento, de una parte de ésta o de sus meta datos.

Para que un documento sea considerado relevante, debe contener los términos de la consulta[34], por ello se requiere que el usuario tenga cierto conocimiento del tema a buscar, de tal manera que las palabras de la consulta represente de la mejor forma posible y de manera concisa, su solicitud de información; de lo contrario, los resultados no serán satisfactorios.

Características	Desventajas
Basado en la teoría de Conjuntos y en Álgebra de Boole.	Recuperación muy restrictiva basada en un criterio de selección binario (presente/ausente).
Devuelve los documentos que contienen al menos un término de la búsqueda .	Cardinalidad del conjunto recuperado, Muy pequeña o muy grande.
Permite el uso de los operadores lógicos *and* *or* y *not*.	

Tabla12. Características del modelo Booleano

Sobre este modelo conceptual se han desarrollado algunas extensiones que se recogen bajo la denominación de modelo Booleano extendido, la cual agrega conectivos lógicos que enlazan a las palabras claves.

El modelo vectorial es el más utilizado en los sistemas de IR modernos. En este caso, los documentos, términos y consultas se representan mediante vectores en un espacio

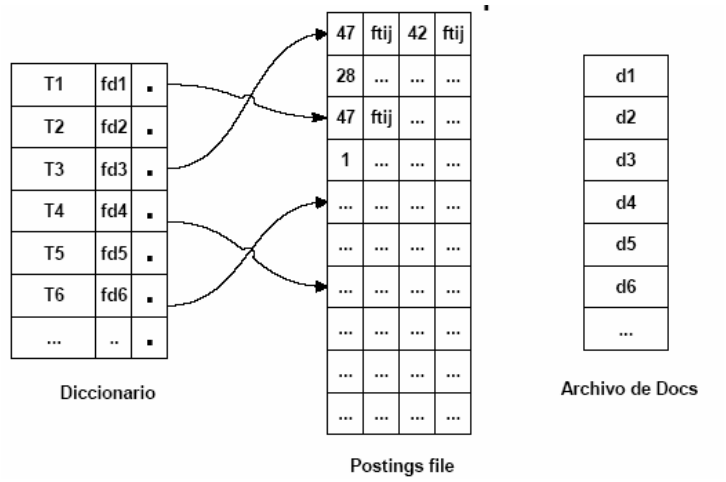
[33] G. Salton, M.J. McGill, "Introduction to Modern Information Retrieval", 1983

[34] Meadow, C.T. Text Information Retrieval Systems. Academic Press, 1992.

multidimensional[35]. Las consultas se resuelven mediante la comparación del vector de consulta frente a los vectores de los documentos, obteniendo una lista ponderada de resultados.

A diferencia de la recuperación booleana, en el modelo los espacios vectoriales se consideran el carácter semántica de los documentos, mediante la asignación de pesos a los términos, que indica su presencia o importancia en el documento o en la colección[36].

Aunque el modelo en sí no señala cómo asignar pesos a los términos, las técnicas comunes son la frecuencia del término, entendida como el número de veces que aparece el término en el documento, la frecuencia inversa del término, calculada como la unidad entre el número de documentos que contienen el término, o el producto de la frecuencia del termino por su frecuencia inversa. Este último caso se conoce como la regla TF X IDF (Term frequency-inverse document frequency)[Writen, 1999][70]. La similitud entre un documento y una consulta se calculan en base la función coseno del ángulo entre sus dos vectores.



[35] G. Salton, A. Wong, C. S. Yang. "A Vector Space Model for Automatic Indexing", 1975.

[36] Balabanovic. Learning to surf: multiagent systems for adaptative web page recommendation. Deparment of Computer Science, University of Stanford, 1998.

Características	Desventajas
Considera un **espacio vectorial** n-dimensional con una dimensión por cada término de la búsqueda.	Hipótesis de independencia entre términos
Cada documento se representa como un **vector** del espacio	
Un término presente en un doc representa una componente > 0, se admiten coincidencias parciales (pesos no binarios). Ausente en un documento representa una componente = 0.	
El peso de la componente: se incrementa con la frecuencia del término y se decrementa con el número de documentos en que aparece.	
Los pesos proporcionan un *índice de similaridad*	
La clasificación final es decreciente por similaridad	

Tabla13. Características del modelo Vectorial:

En el modelo probabilístico en cada consulta se mide para cada documento la probabilidad de que sea relevante para dicha consulta, obteniéndose un primer conjunto de documentos potencialmente relevantes[37]. A continuación, el usuario interactuará con el sistema para indicar aquellos documentos que considera relevantes. El sistema usa esta información para refinar los resultados de la búsqueda, repitiéndose este proceso hasta una adecuada aproximación al conjunto de resultados óptimos.

La función de semejanza es la probabilidad de que un documento sea relevante Sem (p,di) = P(R|di). Para esto, se toma como relevante aquellos documentos en los que su probabilidad de ser relevante es mayor que la de no serlo.

[37] S.E. Robertson, K. Sparck Jones. "Relevante Weighting of Search Terms". Journal of the American Society for Information Science, 1976.

Características	Desventajas
Trata de recuperar el conjunto *ideal* (se asume que existe como subconjunto del total) de documentos relevantes: R	Necesidad de separar inicialmente documentos relevantes y no relevantes
Supone la interacción del usuario con el sistema (refinamientos)	Pesos binarios. No se considera la frecuencia de los términos en los documentos.
Principio Probabilísticos: Dada una *q* y un documento de la colección, el modelo trata de estimar la probabilidad de que el usuario encontrará relevante dicho documento.	Hipótesis de independencia entre términos.
Se obtiene una clasificación ordenada por orden decreciente de probabilidad de relevancia.	

Tabla14. Características del modelo Probabilística

A.2.3 ESTRUCTURA DE FICHEROS

Una decisión fundamental a tomar durante el diseño de los SRI es qué tipo de estructura de ficheros se va a usar para la base de datos subyacente. En la tabla4 hemos visto que el conjunto de estructuras de ficheros es diverso: ficheros planos, ficheros inversos, ficheros de patrones de bits, Árboles PAT y grafos.

Con el uso de ficheros planos, uno o más documentos son almacenados en un fichero (generalmente en formato de texto ASCII), las búsquedas sobre estos ficheros planos se llevan a cabo generalmente por medio de la localización de patrones de texto.

Un fichero inverso es un tipo de fichero índice donde la estructura de cada ítem (o entrada) del fichero es, generalmente: palabra clave, identificador del documento, identificador de campo. Una palabra clave es un término índice que describe al documento, el identificador de documento es único para cada documento y un identificador de campo es un término que nos indica dentro de qué campo del documento aparece la palabra clave. Algunos sistemas incluyen también información acerca de la localización en el documento del párrafo y frase de los términos utilizados para proceder a interrogar la base de datos. La búsqueda se realiza, corrientemente, por medio de la localización de los términos solicitados en el fichero inverso.

(a) Texto de ejemplo.
Cada linea representa un documento.

Document	Text
1	Pease porridge hot, pease porridge cold,
2	Pease porridge in the pot,
3	Nine days old.
4	Some like it hot, some like it cold,
5	Some like it in the pot,
6	Nine days old.

(b) Archivo invertido del texto de ejemplo dado en (a).

Number	Term	Text
1	cold	1,4
2	days	3,6
3	hot	1,4
4	in	2,5
5	it	4,5
6	like	4,5
7	nine	3,6
8	old	3,6
9	pease	1,2
10	porridge	1,2
11	pot	2,5
12	some	4,5
13	the	2,5

Figura22. Representación de la estructura de un fichero inverso

Los ficheros de patrones de bits contienen hileras de dígitos binarios, patrones de bits que representan a los documentos. Existen varias formas de construir estos patrones de bits, un método común consiste en la división de los documentos en bloques lógicos, conteniendo cada uno de ellos un número fijo de distinto significante (una palabra de una lista de términos no vacíos). Cada palabra del bloque es desglosada para ofrecer una hilera de bits (patrón de bits con algunos de los bits "puesto a 1"). Los patrones de bits de cada palabra en un bloque son agrupados para crear un bloque de patrones. Los bloques de signaturas se concatenan posteriormente para producir el patrón de bits del documento. La búsqueda se lleva a cabo por medio de la comparación que se establecerían entre los patrones de bits de las interrogaciones con los patrones de bits de los documentos de la base de datos.

Los árboles PAT, están construidos sobre todas las sistrings de un texto (subcadenas). Si una colección de documentos es concebida como una secuencia numerada de arrays o cadenas de caracteres, una sistring se entiende como una subcadena de caracteres que se define desde un punto determinado del array y se extiende hasta una distancia arbitraria hacia la derecha. Un árbol PAT es un, por tanto, un árbol digital donde los bits individuales de las claves son usados para decidir derivaciones.

Los grafos (o "redes"), son colecciones ordenadas de nodos conectados por arcos; se usan para representar documentos de diversas formas y maneras. Un ejemplo es el grafo denominado red semántica, que representa las relaciones semánticas que se establecen en el texto, y que se pierden (a menudo), en otros sistemas de indización. Aunque constituyen un campo interesante para el estudio, resultan bastante difíciles de llevar a la práctica y requieren excesivo esfuerzo manual para el proceso de la representación de las colecciones de documentos.

A.2.4 OPERACIONES DE CONSULTA, OPERACIONES SOBRE TÉRMINOS Y OPERACIONES SOBRE DOCUMENTOS.

Operaciones de consulta.

Las consultas a los SRI se expresan por medio de sentencias formales de las necesidades de información de los usuarios del sistema. Determinan de forma clara al SRI y permiten diferenciar a unos de otros. Por ejemplo, una de las operaciones de consulta más común es la operación denominada parsing, que consiste en la división de la consulta en sus elementos constituyentes. Las búsquedas booleanas deben ser divididas en sus correspondientes términos de indización o palabras clave y los operadores asociados a ellas para formular la expresión formal de la consulta. El conjunto de los documentos asociados con cada término de consulta es recuperado, y estos conjuntos son, entonces, combinados de acuerdo a los operadores booleanos.

La operación denominada reutilización (en inglés feedback), consiste en la reutilización de una búsqueda anteriormente efectuada. La información sobre el resultado de estas búsquedas es usada para formar parte de las consultas actuales; así, los términos de documentos relevantes encontrados en una consulta previa pueden añadirse a la consulta actual, y los términos correspondientes a documentos no relevantes pueden ser obviados con el factor añadido de no tener que repetir las operaciones anteriores.

Operaciones sobre los términos.

Las operaciones que se pueden llevar a cabo sobre los términos en un SRI conforman el conjunto: {stemming, truncamiento, ponderación por pesos, palabras vacías y tesauros}.

- Con el concepto de *stemming* nos referimos a un proceso de "corte" de las palabras, reduciéndolas normalmente a su forma de raíz más común.

- El *truncamiento* es otro proceso de "corte de palabras" pero realizado de forma manual por el usuario en los procesos de recuperación de información, tal como puede ser la localización de todos los documentos que comiencen por "informa". Otra definición a esto, es decir que el truncamiento es una “mezcla” manual de términos usando caracteres especiales en la palabra, así que el término truncado formará múltiples palabras; en este caso nos referimos a las operaciones de localización de términos con una raíz común.

- Otra forma de asociación de términos relacionados es por medio de la utilización de un *tesauro*, el cual, nos va a ofrecer una lista de términos, sus términos sinónimos y las relaciones semánticas mantenidas entre los términos del mismo.

- *La lista de palabras vacías* es una relación de términos considerados como valores no indizables, usados para eliminar potenciales términos de indización. Los términos de una lista vacía están carentes de todo significado a la hora de recuperar información, como ejemplo podemos tomar el determinante "la", que no posee ninguna funcionalidad a la hora de recuperar documentos, ya que en todos los documentos de la base de datos aparecerá este término de forma casi segura y no nos resalta nada del contenido del documento almacenado. Así, cada término potencial de indización es comprobado previamente, verificándose su presencia en la lista de palabras vacías y es descartado si se encuentra en ella.

En cuanto a la *ponderación* de términos, a éstos se les puede asignar un valor numérico basado en su distribución estadística, o sea, en la frecuencia con la que los términos aparecen en documentos, colecciones de documentos, o en subconjuntos de colecciones de documentos, tales como documentos considerados relevantes en una búsqueda (pregunta).

Operaciones sobre los documentos.

Los documentos son los objetos primarios en un SRI y hay muchas operaciones para ellos. En algunos SRI, a los documentos añadidos a una base de datos se les debe asignar un identificador único, deben dividirse (en partes gramaticales) en sus campos constituyentes, y estos campos deben ser introducidos dentro de identificadores de campos y conjuntos de términos. Una vez en la base de datos, uno a veces quiere *desenmascarar* ciertos campos para buscarlos y mostrarlos, por ejemplo, un investigador puede desear buscar sólo los campos de título y resumen de un documento para una búsqueda dada, o puede desear consultar sólo el título y el autor de los documentos recuperados[38].

- Otra operación común es la de *ordenar* los documentos recuperados por algún campo determinado; por ejemplo el campo autor. La operación de mostrar incluye tanto a la salida impresa de los documentos como a su visualización en la pantalla del ordenador.

- A partir de la información procedente de la *distribución de frecuencias* de los términos, es posible asignar una probabilidad de relevancia a cada documento dentro de un conjunto recuperado, permitiendo que los documentos recuperados sean organizados en orden a esta probable relevancia.

- La información de la distribución de frecuencias de los términos puede ser usada para agrupar documentos similares en un espacio documental, por medio de las técnicas de *clustering.* Otra operación importante a realizar con los documentos es proceder a su visualización. El diseño del interfaz de usuario de un SRI resulta de carácter vital, como en otro tipo de sistemas de información, para conseguir un uso efectivo del mismo.

A.3 ESCENARIO FUNDAMENTAL DE LA RECUPERACIÓN DE LA INFORMACION EN INTERNET.

[38] En algunos sistemas gestores de bases de datos documentales, a este tipo de operación se le denomina búsqueda por referencia cualificada

Internet y, en especial la W3, no se creó en un principio para atender la publicación y recuperación organizada de información. Su amplio desarrollo y crecimiento posterior dificultaban la localización de los documentos pertinentes y suscitó la necesidad de contar con herramientas de búsqueda que facilitaran esa tarea.

Según Chen, H. "Cuando se busca información en la red se puede presentar dos situaciones diferentes: o bien se pretende explorar el espacio de información para familiarizarse con él e identificar algo de interés o bien se pretende buscar y recuperar información relevante de forma mas concreta" [39].

Conceptualmente, como hemos mencionado, la Recuperación de Información (RI) es una operación en la que se interpreta una necesidad de información de un usuario y se seleccionan los documentos más relevantes capaces de solucionarla, es decir, consiste en buscar documentos que exhiban un mayor parecido a la pregunta formulada. En el contexto de la W3, se puede definir el objetivo de la recuperación como la identificación de una o más referencias de páginas web que resulten relevantes para satisfacer una necesidad de información.

El Word Wide Web es una estructura hipertextual e hipermedia de información, cuyo componente fundamental es el texto. Los fundamentos sobre los que construir una técnica de recuperación de información en Internet son el conocimiento de las características propias de los documentos existentes en Internet, y de la teoría de la recuperación de información. Las herramientas de búsqueda aplican sobre el texto los principios que se han aplicado sobre recuperación textual: creación de ficheros inversos, indización automática, compactación…etc., y, consecuentemente, los usuarios disponen de igual forma de las mismas prestaciones para la recuperación, operadores boléanos, de posición, vectorización.

Al tratarse de un entorno abierto y cambiante, las herramientas de búsqueda ofrecen listados de resultados, que dirigen al usuario hacia el documento original. Los cambios que se producen, por la propia dinámica del web, hacen que en ocasiones esa redirección

[39] CHEN, H. Internet Borwsing and Searching: User Evaluations of Category Map and Concept Space Techniques, Journal of the American Society for the Information Science, nº 49, 1998, pp. 582-603.

no ofrezca los resultados esperados, y que en numerosas ocasiones haya que completar la búsqueda mediante procesos de exploración basados en la navegación. Por tanto, el usuario siempre debe pensar que no basta en la recuperación de la información en Internet, con seguir los resultados obtenidos de un motor de búsqueda; esos resultados hay explorarlos, analizarlos, valorarlos, y seleccionarlos como adecuados, o desecharlos como no pertinentes. Las herramientas de recuperación de información en la web son un medio más, una fase intermedia, no un fin.

A.3.1 PERSPECTIVA EVOLUTIVA DE LOS SISTEMAS DE RECUPERACIÓN DE LA INFORMACIÓN EN LA WEB

Los actuales motores y agentes de recuperación de información para Internet no son los primeros ni los únicos instrumentos desarrollados a tal fin. Tan temprano como comenzó el crecimiento geométrico de la información en Internet, comenzaron a desarrollarse instrumentos para facilitar la localización y acceso a los recursos de información[40].

Conceptualmente, como se había mencionado, la recuperación de información es una operación que interpreta una necesidad de información de un usuario y se seleccionan los documentos más relevantes capaces de solucionarla es decir, consiste en buscar documentos que exhiban un mayor parecido a la pregunta formulada. En el contexto de la WWW o W3, se puede definir el objetivo de la recuperación, como la identificación de una o mas referencias de páginas web que resulten relevantes para satisfacer una necesidad de información.

Entre las aplicaciones desarrolladas cabe destacar Archie (como la primera maquina de búsqueda de información en Internet creada en 1990 por un estudiante de la Universidad de McGill de Montreal, consistía en una base de datos de servidores FTP y un motor de búsqueda sencillo que buscaba los ficheros de los servidores FTP que coincidiesen con la búsqueda que realizaba el usuario), luego fue creado Whois, Netfind (para localización de personas), en 1993 fue desarrollado en la Universidad de Nevada Verónica (herramienta

[40] GILSTER, P. Finding It on the Internet. The Internet's Navigator guide to search tools and techniques, New York, Jonh Wiley, 1996.

pensada para realizar una búsqueda similar a Archie pero para servidores Gopher, y fue en ese mismo año, la aplicación más popular en Internet).

Los servidores Gopher eran los más utilizados en ese entonces y se caracterizaban porque organizaban la información en árboles jerárquicos sobre cuyas ramas iban descendiendo en función del tema elegido. De esta manera al conectarnos a un servidor de este tipo, éste mostraba el árbol principal y al seleccionar una rama se conectaba a otro (o al mismo) servidor donde se encontraba el subárbol correspondiente. De esta forma se iría descendiendo hasta acceder al recurso deseado.

Posteriormente surgió en esos primeros años el servicio WAIS que posiblemente sea el menos conocido, pero no por ello menos importante. El poderoso WAIS (Wide Area Information Server, aplicación basada en Z39.50, este servicio indexaba los contenidos de los documentos. Cada servidor WAIS tenía su propio conjunto de índices donde se podían hacer búsquedas). Un hito lo marcó la aplicación Hytelnet, creada por P. Scout, que ofrecía un entorno único de acceso y consulta automatizados de bibliotecas de todo el mundo, basado en Telnet y que posteriormente ha pasado a convertirse en un servicio web.

A pesar de todos estos intentos, la primera máquina de búsqueda pública no llegaría hasta el nacimiento de Excite, en un principio llamada Architext Software. Fue desarrollada en Febrero de 1993 y tenía como objetivo principal, utilizar el análisis estadístico de las relaciones de palabras, para desarrollar búsquedas más eficientes a través de la gran cantidad de información que hay en Internet. Estaba compuesto por una base de datos y se accedía a la información por medio de categorías temáticas.

En la primavera de este mismo año se desarrolló World Wide Web Wanderer, que fue el primer robot de la red y tenía como objetivo contabilizar el número de usuarios de la red para medir el crecimiento de ésta. El éxito del Wanderer y el prominente desarrollo del web inspiraron a numerosos programadores que comenzaron a desarrollar herramientas basadas en la tecnología wanderer y que con el tiempo se han denominado “Spiders” o “Web robots”.

Periodo de los buscadores

Los primeros sistemas de recuperación en la web se reportan en 1989 en el Laboratorio Europeo de Partículas Físicas CERN, Tim Berners Lee creó un sistema de hipertexto basado en la red Internet, cuyo primer prototipo fué puesto en marcha a finales de 1991.

Este fue un avance fundamental para las herramientas de búsqueda, ya que esos links o vínculos serían los que permitirían a los motores de búsqueda relacionar información dispersa.

El sistema de hipertexto permitía crear un vínculo entre dos o más documentos a través de una palabra o una frase. Es lo que comúnmente se conoce como “links” que se activan al pasar el cursor sobre ellos, indicando al lector que hay más información vinculada que puede estar disponible en el mismo texto o en una URL distinta.

En 1993 aparece el primer "navegador gráfico", Mosaic, con lo que se produce un gran avance en el uso de Internet, ya que la navegación se facilita considerablemente. Recordemos que en dicho año ya existen más de dos millones de ordenadores conectados a Internet.

Dado el volumen de información existente en esos momentos surge la necesidad de poseer herramientas que localicen, cataloguen y ordenen la información disponible para posibilitar el acceso a la misma, utilizando para ello las nuevas herramientas disponibles. De esta necesidad surge el nacimiento del primer motor de búsqueda, tal y como lo conocemos hoy en día.

Los buscadores o motores de búsqueda utilizan programas automáticos, genéricamente llamados robot y conocidos por los programadores como "Crawlers», "Bots", "Spiders", 'Warms" etc.

El nombre de Robot, se debe a una obra de Karel Capek's, donde se utiliza el término "robota". Estos programas rastreaban automáticamente Internet analizando páginas y almacenándolas en bases de datos documentales, posteriormente incorporaron una herramienta para hacer búsquedas sobre las bases de datos que previamente habían creado.

Los motores de búsqueda poseían una indudable ventaja. Al ser automatizados disponían de una mayor cantidad de documentos, eran mucho más exhaustivos y estaban permanentemente actualizados. Es a partir de 1990 que aparece el primer motor de búsqueda, el Web Wanderer

En cuanto a los directorios que catalogaban la información de una forma más estructurada, clara o accesible que los buscadores. Permitían "navegar" sobre temas concretos hasta encontrar el documento que se necesitaba, pero al ser una catalogación manual y dado el rápido crecimiento del número de páginas era imposible mantenerlos actualizados, de manera que su información quedaba rápidamente obsoleta. En 1993 se crea el primer directorio, Yahoo!.

Los primeros spiders desarrollados carecían de la suficiente inteligencia para saber que es lo que debían indexar. Para solventar esta deficiencia apareció Galaxy en Enero de 1994 que incorporaba búsquedas a través de Gopher, telnet y Web. En ese mismo año en la Universidad de Standford se creó Yahoo y en la Universidad de Washington Webcrawler que fue la primera máquina que buscaba a texto completo en Internet.

Ya en 1995 aparece Altavista que incluye nuevos aspectos en las máquinas de búsqueda, como por ejemplo la velocidad o el uso de operadores booleanos. Por estas fechas se advierte que cuando se realiza una búsqueda en diferentes buscadores los resultados no son los mismos, es por ello que aparecen los metabuscadores. El primero en aparecer sería Metacrawler, que fue desarrollado por estudiantes de la universidad de Washington en 1995.

Con la aparición de multitud de buscadores y debido a la gran cantidad de información que se encuentra en muchas materias, empezaron a aparecer buscadores que restringían su ámbito de búsqueda a determinados temas, el primero de ellos fue Argos cuya área de información se limitaba a “Historia Medieval y Antigua” y que se desarrollaría en 1996.

A finales de 1997 aparece Google, como un proyecto de investigación de la Universidad de Standford. Su novedad principal es que basa sus búsquedas en un ranking de páginas

ordenadas por su relevancia. El criterio que sigue para ordenar las páginas es el número de enlaces que otras páginas tienen a la página en concreto.

En los últimos años los buscadores están evolucionando con el objetivo de no ser simples máquinas de búsqueda, para ello se están perfeccionando la búsqueda e indexación de documentos, haciendo más sencillas las técnicas de búsqueda y los interfaces y mejorando las técnicas de presentación de los resultados de búsqueda, así como las técnicas de ordenación de relevancia. Algunas de las últimas incorporaciones han sido WiseNut, Teoma, DirectHit, OpenDirectory, en el campo de los motores de búsqueda y Vivísimo en el de los metabuscadores.

Además, actualmente los buscadores no son las únicas herramientas de recuperación de la información disponibles en la red, puesto que se han desarrollado multitud de "agentes inteligentes" que son programas que realizan tareas dirigidas a solventar problemas propios de sus usuarios de forma interactiva, autónoma e independiente, los cuales también tienen su rincón en este estudio.

Es innegable que la Internet, como todos sabemos, se ha convertido en una referencia básica de búsqueda de información y documentación, para todo tipo de perfiles y ámbitos. Desde la investigación más teórica hasta el ocio más popular, Internet y, en concreto, la WWW, es un repositorio de datos como nunca ha conocido la humanidad. Sin embargo, esto no implica que esa información sea fácil de buscar. Prácticamente la totalidad de los recursos existentes en la web han sido generados pensando en la visualización humana, y sin ninguna arquitectura que permita directamente una categorización de esa información.

Cabe señalar el comentario que plantea Mike Yamamoto, CNET.com, Abril 2003: "Hoy en día, la gente espera ser capaz de encontrar la información en un instante y de calidad, una suposición que hubiera sido impensable antes de que la Web se convirtiera en un medio masivo". No obstante, Aunque en la actualidad hay muchos proyectos e iniciativas a este respecto, cierto es que ahora mismo existe una cantidad ingente de información que no puede ser encontrada. Para solucionar este problema, la investigación es una constante primordial queda lugar a la aplicación de las herramientas disponibles que intentan atajar

o, al menos, mitigar esta cuestión y abre de igual manera las puertas al desarrollo de otras nuevas o el mejoramiento de la mismas; siendo estos estudios fuentes vitales en esta era de la información.

A.3.2 ENFOQUES EXPUESTOS SOBRE LA RECUPERACION DE INFORMACION EN LA INTERNET

Principios de Imprecisión y de Fertilidad

La teoría de la recuperación de información en Internet parte de los logros y limitaciones de la teoría de recuperación de información desarrollado desde la década de 1960. Como señaló Sebastiá Salat, la etapa expansiva en la que se encuentra la recuperación de información, de la mano del desarrollo acelerado de Internet, hace necesario replantear las prestaciones de las actuales herramientas, ya que son necesarios cambios que adecuen a los principios de impotencia/impresión y fertilidad en la recuperación de información. Dos son los aspectos que esta autora hace hincapié, y va dirigido al interés por parte del usuario que busca información en Internet:[41]

- Principios de impotencia/impresión: hace referencia a que la necesidad de infamación de un usuario no puede traducirse de manera exacta a una estrategia de búsqueda, por la imposibilidad de abarcar todo el conocimiento humano. Esto supone, que la recuperación de información siempre se sitúa en el ámbito de la hipótesis.

A lo anterior hay que añadir que el proceso de indización automático no asegura que se reproduzca la esencia del documento original representado. Ambas cuestiones hacen que no pueda garantizarse la pertinencia de todos los documentos recuperados tras un proceso de búsqueda.

- Principio de Fertilidad: El uso de los sistemas de información electrónica va a permitir la interconexión, cada vez a mayor escala, entre diversas disciplinas.

La ausencia de Intermediarios

[41] SEBASTIÁ SALAT, M. Reflexiones en torno al software de gestión y el acceso a la información: Aspectos fundamentales de la teoría de la recuperación de la información en la Internet, 1999. pp. 261-267.

La actividad del usuario en la búsqueda y recuperación de información en Internet es independiente y autónoma. No suele existir la presencia de un intermediario entre él y los recursos de información que consulta, a diferencia, por ejemplo, de la existencia de un bibliotecario o del documentalista, en su entorno tradicional. Esto es debido a que el usuario tiene acceso directo a la información y a los documentos, por lo que se convierte en responsable de la misma, sin necesidad de intermediarios.

Es la propia dinámica de la web, la que ha eliminado la presencia de intermediarios especializados, convirtiéndose el usuario al mismo tiempo en creador, lector, localizador y evaluador de documentos. Puede deducirse de manera diáfana que la falta de formación del usuario en los campos de la recuperación de información, de la documentación y de la organización de conocimiento, así como un escaso conocimiento de la estructura, funcionamiento y desarrollo de Internet y de la web, convierten a esta teórica ventaja en un gran inconveniente, ya que el conocimiento en estos campos es un factor de éxito muy importante[42].

Sin embargo, cada vez en mayor medida los especialistas en información (bibliotecarios y documentalistas especialmente) se están convirtiendo en intermediarios especializados para la recuperación de información en Internet. Incluso hoy en día existen algunas empresas especializadas en recuperar información de alto valor, propio y añadido, disponible en Internet. Son numerosos las bibliotecas y centros de información/documentación que ofrecen, entre sus servicios, acceso a Internet y consejo especializado en búsqueda de información, apoyando al usuario en las limitaciones a las que éste, por diversos motivos, tenga que hacer frente.

Los enfoques de creación de recursos de información

La necesidad de desarrollar herramientas que faciliten la localización y acceso a la información en Internet obligó a adoptar, en un primer momento, dos aproximaciones

[42] BRUCE, H. User satisfaction with information seeking on the Internet, Journal of the American Society for information Science, nº 49,6, 2000.

clásicas para la recuperación de la información, similares a las ya existentes en los entornos de documentación automatizada[43]:

1. La creación de listados, índices y catálogos ordenados por áreas o materias de forma que el usuario dispusiese de un conjunto de fuentes seleccionadas en las que empezar a buscar. Un ejemplo claro a este respecto, es Yahoo!, catálogo que comenzó a compilarse y organizarse de forma casi manual, pero el aumento de documentos y páginas web en Internet ha obligado a introducir en su cadena de producción herramientas automáticas de compilación y clasificación. Además han ido añadiendo motores internos a sus prestaciones de forma que permiten consultar mediante ecuaciones sus bases de datos internas.

2. La creación automática de bases de datos basadas en índices o ficheros inversos, mediante unas aplicaciones que rastrean o exploran todo el ámbito Web, llamados robots, spiders o wanderers. Estos robots rastrean la web a la búsqueda de documentos, obtienen una copia, la indizan según los métodos mencionados anteriormente, y usan los enlaces presentes en los mismos para localizar nuevos documentos. Estos sistemas, aunque automático, ofrecen limitaciones, en cuanto a cobertura, nivel de indización del documento y otras cuestiones, como por ejemplo la actualización.

Enfoques según la herramienta de recuperación

La especialización de las herramientas, y los requerimientos de los usuarios, cada vez más exigentes, hacen posibles diferencias o aproximaciones, según la herramienta de recuperación que se utilice en cada caso[44]. Sin embargo, casi todas ellas siguen utilizando, como punto de partida para el proceso de recuperación, los índices y bases de datos indicados en el apartado anterior:

- Directa: El usuario utiliza su navegador para conectarse al servidor web que actúa como interfaz del motor de búsqueda correspondiente a la base de datos que desea consultar. El servidor le envía una página web que actúa como interfaz de interrogación, a través de

[43] BENITO AMAT, C. Recuperación en Internet: cuatro modelos complementarios y una agenda para su integración, Boletín Rediris, nº 48, 1998, pp.36.58.
TRAMULLAS SAZ, J. Recuperación de información en el Word Wide Web: Planteamientos herramientas y perspectivas, Barcelona, 1999, pp. 137-145.
[44] TRAMULLAS J. y OLVERA M. Recuperación de la Información en Internet, Ed. Rama, 2001, pp. 36-37.

la cual formula la consulta, establece los parámetros correspondientes, y la remite al servidor. El servidor la recibe, procesa y la envía como respuesta un nueva página web, generada de forma dinámica, que contiene las diez o veinte respuestas más pertinentes según su criterio, a la cuestión formulada por el usuario. Este actúa directamente sobre el motor de búsqueda y su base de datos.

- Por Intermediario: El usuario utiliza su navegador para conectarse a un servidor web que le ofrece una interfaz de interrogación propia. Esta interfaz le permite interrogar una base de datos correspondiente a un motor de búsqueda situado en un servidor web diferente al que ofrece la interfaz. De esta forma, el servidor que envía la interfaz actúa como intermediario entre el motor de búsqueda de destino y el usuario. Dependiendo del servicio de que se trate, el intermediario recibe la respuesta del motor, varía las características de presentación, y la remite al usuario. El verdadero interés de estos intermediarios se da cuando actúan como interfaz a múltiples motores de búsqueda, ya que de esta forma, el usuario puede remitir su consulta a todos ellos mediante una acción única. Estas interfaces múltiples reciben el nombre de metabuscadores.

- Por agente: El usuario instala en su maquina una aplicación que permite formular las ecuaciones de búsqueda y remitirlas directamente a uno o varios motores de búsqueda. La aplicación lanza conexiones simultáneas al conjunto de motores que se trate, recibe las respuestas, y las entrega al usuario en una presentación única, que puede ofrecer diferentes formas. Dependiendo de las prestaciones del agente, las respuestas pueden ser filtradas, aplicando criterios propios de eliminación de duplicados, reordenación de resultados, etc. Los más avanzados comprueban la existencia real de las páginas web en la dirección de referencia y son capaces de obtener y colocar en la máquina del usuario una copia del documento original.

- Por robot personal: Se trata de aplicaciones que se instalan en el ordenador del usuario, y que son capaces de acceder a un servidor web, construir un mapa de índices de sus contenidos, y utilizar los mismos para acceder a la información que sea interesante para el usuario, obteniendo copias de las paginas o documentos web contenidos en el mismo. El mapa, índice o base de datos creados se almacenan en el ordenador del usuario y pueden ser actualizados regularmente. Cuando el usuario requiere una

información contenida en la misma, el robot lanza al navegador en modo local en busca de la misma. Estos robots, aún escasos, incorporan las prestaciones que ofrecen los agentes.

A.3.3 LIMITACIONES EN LA RECUPERACION DE LA INFORMACION EN INTERNET.

Al igual que en los sistemas clásico de recuperación de información, los sistemas del mismo tipo diseñados para Internet también sufren de las limitaciones y otras derivadas de la estructura hipertextual o de la dinámica de actualización de documentos, o bien de factores externos a las técnicas de recuperación de información y al propio usuario:

- Existe una disfunción entre los procesos de indización automática, la representación del contenido de un documento que se crea como consecuencia del mismo, y el contenido informativo real del mismo.

- La cobertura de los motores no es exhaustiva, por tanto los contenidos de los diferentes motores también se solapa en parte, de tal forma, que al realizar la misma búsqueda en varios, para aumentar la cobertura de la web sobre la cual tiene lugar la consulta, con el riesgo de aumentar las respuesta repetidas.

- la actualización de las bases de datos no es automática. Las variaciones que puede sufrir una página web no son automáticamente reflejadas en los motores. Los robots de indización visitan a intervalos cada vez mayores las páginas indexadas en sus bases de datos, dando prioridad a aquellas que son mas solicitadas en las búsquedas. Esta priorización castiga las paginas web objetos de menos consultas, independientemente de su importancia informativo/documental.

- Como consecuencia a lo anterior, los motores no reflejan adecuadamente la variabilidad espacial y temporal de las páginas web. Esto quiere decir que el usuario puede encontrarse con los típicos errores 404, lo cual no quiere decir que la pagina y su información no exista; puede ser simplemente que éstas hayan cambiado de localización o de estructura hipertextual. Las bases de datos de los motores ofrecen un índice notable de inconsistencia respecto al universo que reflejan.

- La estructura hipertextual de la web no se refleja en la representación de las paginas en los índices y bases de datos. Por ahora, se representan páginas individuales como

objetos diferenciados, sin atender a su posible pertinencia a estructuras mayores, a su contexto informativo.

- No todos los operadores ofrecen los mismos operadores, ni las mismas estructuras y reglas para formular las ecuaciones. No existe un estándar en este campo, por lo que es necesario conocer las particularidades de cada motor, en el caso de tener que desarrollar búsquedas con ecuaciones complejas.

- Las respuestas que ofrecen las herramientas a las ecuaciones formuladas no presuponen fiabilidad ni rigor. Sólo responden a la ecuación planteada. La consideración de si los resultados obtenidos merecen fiabilidad y confianza queda a discreción del usuario. Si este es conocedor del tema, puede discernir estas cuestiones, en caso contrario, debe considerar que la existencia de una página con una información dada en Internet no implica que sea rigurosa ni documentada.

- Los resultados ofrecidos por estas herramientas y aplicaciones no pueden ofrecerse ni tratarse «en bruto». Es necesario e ineludible complementar la respuesta con el acceso al documento, su revisión y su selección o rechazo, en función de criterios de interés establecidos claramente. Esta actividad supone la utilización de la exploración como complementaria a la búsqueda.

A.3.4 METODOLOGIA Y ESTRATEGIAS PARA EL PROCESO DE RECUPERACION DE INFORMACION EN INTERNET

A continuación se plantea una metodología o proceso de localización, recuperación de información en Internet, en el cual se plasma ideas y orientaciones de gran utilidad. Sin embargo no existe una metodología de búsqueda ideal, su continuo uso es el mejor aprendizaje, por tanto, la recuperación de información no es un proceso o actividad exacta, puede haber varias soluciones para el mismo problema.

Básicamente, el hacer uso de herramientas de recuperación de información fundamentadas principalmente en su contenido y posteriormente en su formato, debe estar contextualizado en la conceptualización del “Problema/objetivo” y de sus necesidades de información, el énfasis de una exitosa búsqueda depende en gran medida

del enfoque conceptual que se defina(definir la necesidad, escogencia de la mejor fuente de información, seleccionar la palabras claves o los encabezamientos de los temas que mas apropiadamente defina la información solicitada). Y finalmente, saber, que cada vez que el usuario haga uso de una nueva herramienta de búsqueda, aprenda su uso, el contenido que ofrece, buscando por si mismo cuales son sus principales características, logrando de esta forma información pertinenente y actualizada.

Cabe hacer mención lo que los especialistas de la Biblioteca de la Universidad de California en Berkeley recomiendan: "El usuario debe desarrollar su proceso de búsqueda, siempre con visión periférica" es decir, aprender sobre el tema conforme se busca, variar las estrategias conforme se sabe más, y no abandonar ningún método de búsqueda a menos que se encuentre lo que se busca o se aprenda algo nuevo que conlleve a resultados mas óptimos a los conocidos.

A.3.4.1 PROCESO DE BÚSQUEDA DE INFORMACION EN INTERNET

No existe una metodología de búsqueda ideal. Su continuo uso es el mejor aprendizaje, sin embargo se propone una metodología básica y bastante eficaz, desarrollada a partir de experiencias puramente técnico[45]. Ésta se divide en cuatro pasos, y son los siguientes:

1. Planteamiento del tema de Búsqueda:

En primer lugar, se debe planificar claramente el tema de interés. Hay objetivos que pueden se adecuados, pero que en realidad necesitan ser redefinidos o refinados para obtener los mejores resultados (cada búsqueda es diferente y exige estrategias diferentes). Es importante evitar pensar de manera directa sobre el tema de búsqueda, es ideal plantear posibles situaciones que puedan darse y pensar en varias tácticas alternativas para acercarse al problema. Si se prepara esas tácticas (por ejemplo, el uso de sinónimos o palabras alternativas que son similares en significado), las respuestas que obtenga durante una consulta, sobre temas relacionados con lo que se esta buscando, puede servir de punto de partida para buscar por exploración.

[45] Localización de información en motores de búsqueda en Internet: análisis de la efectividad. Basada en la fuente: <http://www.mcyt.es/asp/publicaciones/revista/numero346/173-182.pdf>

En segundo lugar, se debe establecer cuál es el nivel de conocimientos que se tiene sobre el tema, ya que de éste dependerá el poder abordar el problema de la fiabilidad con mayores garantías. En cualquier situación esta fase debe dar como resultado una formulación clara e inequívoca del objetivo de su búsqueda. De acuerdo con la necesidad informativa es necesario cuestionarse: ¿Qué se sabe sobre éste tema?, lo cual implica una búsqueda retrospectiva o ¿Qué hay de nuevo sobre el tema?, que implica una búsqueda de información actualizada.

2. Identificación de los tipos de información

El web tiene diferentes tipos de información, tanto por el tipo de fichero que los contiene, como por el objetivo y finalidad de las páginas web y de los creadores de las mismas. Por lo tanto, se debe establecer la posible utilidad de cada uno de estos tipos de documentos, y no desdeñar ninguno a priori, ya que por exploración puede encontrar información complementaria que le sea de utilidad.

Es importante saber que “No todo es Internet y en Internet no todo es el World Wide Web”

- Dentro de las fuentes Internet, no buscar sólo en el World Wide Web. Por Es importante utilizar fuentes Internet y no Internet (revistas electrónicas, libros, etc.). Pero incluso en esto puede encontrar ayuda en Internet, ya que algunos buscadores, como Guíame (http://www.guiame.net) incorporan enlaces a fuentes de información que no están en Internet.
- ejemplo utilizar los grupos de debate (más conocidos como "las news" o USENET) o listas de distribución que contienen mensajes de particulares. Conviene leer los objetivos de la lista o grupo de discusión para ver si aceptan este tipo de consultas.
- Como veremos, hay partes del WWW que no están accesibles a los buscadores, el llamado "Web Invisible".

3. Selección de los recursos de información y de las herramientas de consulta.

La selección de los recursos de información, es decir índices, directorios y motores de búsqueda a utilizar, es de gran importancia. Cada vez en mayor número están apareciendo directorios especializados en los más diversos temas; no obstante, el

problema de éstos es que en numerosas ocasiones, ofrecen coberturas muy parciales, aunque los índices que ofrecen tienen un alto nivel de fiabilidad.

Si el área temática a la que pertenece el objeto de la búsqueda está claramente identificada, en primer lugar puede acudirse a una o más herramientas de búsqueda por índice temático (directorios), del tipo de Yahoo!, o de las especializadas en ese tema, si es que existen. Ese tipo de herramientas permitirá obtener información genérica sobre el objeto, es decir, nos ofrecerá una orientación de los posibles resultados.

Posteriormente puede utilizarse alguna herramienta de búsqueda por contenido o motores de búsqueda, del tipo de AltaVista, y más concretamente de su opción de búsqueda avanzada, para obtener información más específica y actualizada sobre el objeto en cuestión.

Como complemento a las dos herramientas anteriores, estrategias se pueden utilizar uno o más Metabuscadores que realicen la búsqueda simultáneamente con distintas herramientas (motores de búsqueda, directorios y buscadores especializados), teniendo en cuenta que estos buscadores en paralelo no suelen tener acceso a toda la potencialidad que esas herramientas ofrecen individualmente, lo que puede dar lugar a búsquedas menos precisas que si se hubiese utilizado en cada herramienta por separado. Como última herramienta se podría usar un agente personal, que agilizará el proceso de consulta de múltiple fuentes de información(Servidores FTP, Bases de datos, etc). Es importante conocer, que el uso de una interfaz web es más lento, pero ofrece todo el potencial de los lenguajes de consulta. Usar agentes es más rápido, pero limita las prestaciones de los lenguajes de interrogación.

4. Transformación entre lenguajes

Al no existir un control terminológico centralizado en Internet, lo que en realidad es una ventaja, el usuario utiliza en las ecuaciones términos que ha identificado en la primera fase. Las palabras claves seleccionadas por el usuario se utilizan directamente en las ecuaciones. En principio, se aconseja evitar todo lo posible la búsqueda simple de palabras o términos muy generales, ya que es casi seguro que de un posible fracaso, cuando use índices, es necesario elegir frases específicas que contengan el término, la

cual deba ser pertinente al tema de búsqueda. Se aconseja el uso de operadores boléanos, y los operadores de frase. Hay que tener presente que los motores de búsqueda trabajan con la presencia/ausencia de texto en determinadas posiciones del documento o pagina web, lo cual quiere decir que la aparición de éstos no significa que los documentos sean pertinentes a lo que esta buscando. De igual forma, es importante saber que los motores no atienden a idiomas si no se les indica: si se quiere recuperar documentos en varias lenguas, lo mejor es hacer uso de términos sinónimos a cada idioma.

5. Formulación de la ecuación.

Unos de los momentos claves corresponden a la formulación de las ecuaciones. Cuando hablamos de formulación de búsqueda o ecuaciones, hacemos referencia al uso de un conjunto de operaciones que tiene por objeto localizar, seleccionar y obtener los documentos que den respuesta a las preguntas formuladas por el usuario en función de sus necesidades de información.

En esta fase se introduce en la interfaz o herramienta que haya seleccionado para desarrollar búsqueda, la expresión que reúne los términos elegidos y los operadores que establecen las relaciones existentes entre ellos. En la tabla8, se presenta una breve explicación de las posibilidades del lenguaje de interrogación que usan los motores de búsqueda, sin embargo esta varia de acuerdo al motor de búsqueda que utilice, así que es necesario conocer a fondo las prestaciones que éstos ofrecen.

Una vez domine un poco el funcionamiento del motor, es preferible utilizar las interfaces avanzadas ya que estas ofrecen un mejor potencial y parámetros que permiten perfeccionar las ecuaciones de búsqueda y a obtener resultados mas óptimos. Si ha adoptado por el uso de un agente, es necesario tener en cuenta, que éste traducirá al lenguaje de cada motor, la expresión que el usuario introduzca, pero es precisamente esta generalización la que hace perder la oportunidad de usar operadores más restrictivos. Algunas reglas básicas a aplicar:

- Si busca un nombre propio o una frase completa, use las opciones y operadores de “frase exacta”, entrecomillando la expresión.

- Si busca palabras muy comunes en muchos contextos, utilice operadores boléanos, especialmente AND (para incluir todos), y los operadores NOT ó AND NOT para excluir palabras que le amplíen demasiado los resultados.

- Si ha optado por usar sinónimos, lo mejor es usar una expresión booleana que relaciones todos los términos sinónimos usando el operador OR.

- Si busca términos de raíz similar, pero diferentes sufijos (por ejemplo singular o plural), use los símbolos de truncamiento.

- Si quiere una primera aproximación exitosa en muchas ocasiones, use el operador posicional de titulo, de forma que recupera los documentos en los cuales la expresión o palabras deseadas aparezcan en el título de la página web.

TABLA15. OPERADORES DE CONSULTA: SIGNOS INCLUSION / EXCLUSION	
-	Excluye términos de los resultados de búsqueda. Utilice el comando menos (-) delante de cualquier palabra o frase para omitir tal término. En algunos buscadores se utilizan en sustitución de los operadores AND NOT. Ejemplo: Buscar: revistas -ordenadores
+	La acción opuesta a prohibir términos de los resultados de la búsqueda es requerir que aparezcan ciertos términos en los documentos que el motor de búsqueda encuentre. Utilice el comando más (+) para identificar las palabras o frases indispensables. En algunos buscadores se utilizan en sustitución de los operadores AND.
OPERADODRES BOOLEANOS: Indican la relación lógica existente entre los términos buscados a nivel de documento o registro	
AND	Indica que se recuperarán los documentos que contengan todas las palabras indicadas en la solicitud de búsqueda. Por esto, se considera muy útil para limitar una búsqueda y reducir el número de registros recuperados, mientras más términos se utilicen más específicos serán los resultados.
OR	Ordena a la base de datos que devuelva todos los documentos que contengan, al menos, una de las palabras claves solicitadas. En este caso, el resultado puede ser: registros, con una, dos o todas las palabras incluidas en el planteamiento de la búsqueda. Es común utilizarlo cuando se puede buscar un término por sus sinónimos. En algunos buscadores se sustituye por el símbolo "|".
NOT o AND NOT	Se utiliza entre dos términos claves y se traduce por "no". Excluye de la búsqueda aquellos documentos que contengan la palabra clave a la que se refiere el operador. Es muy útil para eliminar los problemas causados por la polisemia; generalmente, se utiliza después de haber realizado una primera búsqueda, donde se obtengan resultados irrelevantes con las palabras solicitadas, que aparecen en un contexto diferente al que se busca. En algunos buscadores se sustituye por el símbolo "-" o "!".
OPERADORES POSICIONALES : Definen, cuál es la posición de las palabras claves dentro del documento y las interrelaciona entre ellas, según criterios de proximidad u orden.	
ADJ	Significa "adyacente". Utilice el operador ADJ cuando desee encontrar documentos en los que aparezcan los términos juntos, sea en el orden

	que sea. Los aficionados a los deportes pueden buscar coches ADJ carreras sabiendo que tal búsqueda localizará tanto carreras de coches como coches de carreras.
NEAR	En español significa "cerca". Encuentra documentos que contienen ambas palabras claves, pero que no estén separadas por más de 10 palabras o 100 caracteres (aunque este número puede variar según el buscador). En algunos buscadores se puede sustituir por "~" o por "[]". Los corchetes se usan generalmente para agrupar las expresiones booleanas complejas. Por ejemplo, (cacahuete AND mantequilla) AND (gelatina OR mermelada) encontraremos documentos con las palabras "mantequilla de cacahuete y gelatina" o "mantequilla de cacahuete y mermelada" o ambas.
FOLLOWED BY	En español significa "seguido de". Sus resultados son muy parecidos a los que produce Near, pero marca claramente cuál ha de ser el orden de las palabras claves. No es usado por muchos buscadores.
OPEREDORES DE EXACTITUD: localicen determinados términos tal y como se han introducido en el formulario de búsqueda.	
Comillas (“ “)	La forma más extendida es poniendo entre comillas los términos que se quieren encontrar.
Frase literal	Lo que hace es tratar a las palabras clave como si fueran una frase, deben aparecer en los documentos como han sido introducidas.
OPERADORES DE TRUNCADO : Estos operadores tratan a las palabras claves como cadenas de caracteres, no como palabras completas, con lo cual devolverá aquellos documentos que contengan a la palabra clave, pero también aquellas en la que la palabra clave sea raíz o sufijo. .	
*, #, ?, %	El sistema devuelve aquellos documentos que contengan a la palabra clave, pero también aquellas en la que la palabra clave sea raíz o sufijo. Se indican con símbolos como: *, #, ?, aunque hay buscadores que emplean diferentes símbolos en dependencia de la cantidad de caracteres que estos representen, por ejemplo Northern Light utiliza el * para representar varios caracteres, mientras que usa el % para indicar un solo carácter. Aunque la mayoría de los buscadores permiten solo el truncamiento a la derecha, los símbolos se colocarán al inicio, en el medio o al final de la palabra clave en dependencia de las facilidades permisibles. A esta facilidad de hacer búsqueda por términos truncados, algunos autores la llaman "uso de comodín o wildcards". Por ejemplo, Past* encontrará documentos con "pastel", "pastelero" y "pastelería".

BUSQUEDA POR CAMPOS: La búsqueda por campos es una herramienta tradicional en la búsqueda en bases de datos convencionales. Se considera, en Internet, una de las técnicas más efectivas para restringir los resultados de la búsqueda y aumentar la relevancia.	
anchor:text	Encuentra páginas que contienen la palabra o frase especificada en el texto de un hipervínculo. anchor:empleo +programación encontrará páginas con empleo en un vínculo y con la palabra programación en el contenido de la página. No hay que poner ningún espacio antes ni después de los dos puntos. Debemos repetir la palabra clave para buscar más de una palabra o frase; por ejemplo, anchor:empleo OR anchor:carrera encontraremos páginas con anclas (anchors) que contienen la palabra "empleo" o la palabra "carrera".
applet:class	Encontraremos páginas que contienen un applet de Java especificado. Utilice applet:morph para encontrar páginas que utilicen applets llamados "morph".
object:class	Encuentra páginas que contienen un objeto especificado creado por otro programa (ej. un objeto Flash). Utilice object:dinero para encontrar páginas que utilicen objetos llamados dinero.
domain:domainname	Encuentra páginas dentro del dominio especificado. Se usa domain:uk para encontrar páginas del Reino Unido, o domain:com para encontrar páginas de sitios comerciales.
host:hostname	Generalmente cuando los sitios son muy grandes los buscadores no los rastrean completamente sino que se limitan a buscar en las bases de datos propias de éstos. Esta técnica se utiliza cuando se necesita encontrar información en un sitio muy grande que no tiene un motor de búsqueda interno. Con esta técnica, puede especificarse al motor que busque en todas las páginas de determinado sitio, las palabras claves de interés. Un ejemplo podría ser: host:www.fda.gov +"clinical guides". En este caso el motor de búsqueda rastreará el sitio de la FDA completo en busca de la frase "clinical guides" o la búsqueda host:www.shopping.com encontrará páginas que se hallen en el ordenador Shopping.com, y host:dilbert.unitedmedia.com encontrará páginas en el ordenador llamado "dilbert" dentro de unitedmedia.com.
image:filename	Encuentra páginas con imágenes que tienen un nombre de archivo específico. Se usa image:playas para encontrar páginas con imágenes

	llamadas "playas".
like:URLtext	Encuentra páginas similares o relacionadas con una URL especificada. Por ejemplo, like:www.abebooks.com encuentra sitios web que venden libros de viejo, similares al sitio www.abebooks. like:sfpl.lib.ca.us/ encuentra sitios de bibliotecas públicas o universitarias. like:http://www.indiaxs.com/ encuentra sitios sobre cultura en el subcontinente indio.
link:URLtext	Encuentra páginas con un vínculo a una página con el texto de URL especificado. Se usa link:www.myway.com para encontrar todas las páginas con vínculos a myway.com.
text:text	Encuentra páginas que contienen el texto especificado en cualquier parte de la página excepto las etiquetas de imagen, los vínculos, o las URL. La búsqueda text:graduación encontrará todas las páginas que contengan el término "graduación".
title:text	Indica a la base de datos que debe buscar solo en el campo título, como se muestra en el siguiente ejemplo: title:"Panamerican Health Organization", nótese que no se deben dejar espacios entre (:) y la palabra clave. En este caso, se devolverán todos los sitios que incluyan esta frase en el título.
url:text	Encuentra páginas con una palabra o frase específicas en la URL. Se usa url:jardín para encontrar todas las páginas de todos los servidores que tengan la palabra jardín en cualquier parte del nombre del host, la ruta, o el nombre del archivo.

Algunos buscadores incluyen la posibilidad de realizar una búsqueda expresada en lenguaje natural. Ello permite al usuario utilizar un lenguaje no estructurado (como el inglés) para describir qué está buscando, siendo el motor de búsqueda el responsable de traducir esa búsqueda a un formato estructurado. Ejemplos de servicios que permiten este tipo de búsquedas son Infoseek y AltaVista.

Sea cual sea la forma de expresar la pregunta por parte del usuario, ésta será analizada por el buscador y se traducirá a una representación interna que permita compararla con los términos recogidos en la base de datos y seleccionar así las direcciones URL que sean más relevantes.

Un aspecto a tener en cuenta, es que sea cual sea, la sintaxis de búsqueda a utilizar debe seguir la siguiente regla de oro: 'Utilizar como palabras clave términos que aparezcan con mucha frecuencia en los documentos buscados pero raramente dentro de otros documentos de la colección'.

La búsqueda fuzzy, es otra de los mecanismos de búsqueda simple que ofrecen algunos motores de búsqueda (por ejemplo, Altavista), y consiste en introducir una serie de palabras clave, generalmente ordenadas por preferencia. El sistema ordenará los documentos según alguna función de similitud que generalmente dará importancia a la diversidad de las palabras, y a su frecuencia de aparición. Este tipo de búsqueda suele ser la más usual.

Entre otros servicios para delimitación de la necesidad informativa que ofrecen las herramientas de búsqueda en Internet tenemos:

- Delimitación de la búsqueda bien por tipo de fuente, por campos, por zonas geográficas o idiomas, por fechas, etc. Eso puede ayudar mucho a reducir la cantidad de documentos recuperados y a eliminar ruido.

- En algunos servicios de búsqueda se puede especificar el tipo de fuente que se desea buscar.

- Existen servicios de búsqueda que se han creado específicamente para servir a una comunidad geográfica o lingüística, así como versiones específicas de algunos de los grandes motores de búsqueda internacionales adaptados a un país o a un idioma concreto.

- La delimitación por fechas se basa en la utilización de operadores de comparación. El usuario le indica al sistema si los documentos recuperados deben contener una fecha anterior, igual o posterior a la indicada.

6. Análisis de listado de respuestas. Replanteamiento.

En esta etapa el usuario se enfrenta a algunas decenas de direcciones (URLS), elige una que le parece interesante, la revisa, escoge otra, navega un rato, vuelve atrás, hace una nueva consulta, etc. Tiene múltiples alternativas. Es muy raro que a la primera consulta el buscador le muestre páginas de su interés; lo normal son más de dos consultas por sesión antes de darse por vencido, encontrar lo que se buscaba o cambiar de buscador o de método.

El listado de respuestas que se recibe de cualquier herramienta de búsqueda siempre ofrece unas características muy comunes. Los motores envían resultados agrupados generalmente de diez en diez. Los motores ofrecen un listado general, aunque se puede fijar un límite de respuestas a recibir por parte de cada motor consultado. En ambos casos, los listados de respuestas incluyen:

- El titulo que identifica a la página web en cuestión. Situada dentro de las etiquetas <TITLE> </TITLE>, teóricamente el título de una pagina web es una primera aproximación a su contenido informativo. La pulsación sobre el enlace titulo le lleva al documento original.

- El URL en el que puede localizarse el documento original. Es importante no olvidar que los motores no almacena copias de las paginas web que indizan; solo se consulta sus índices. Cuando quiere ver el documento original, la pulsación de este anclaje lo lleva al mismo.

- Un breve resumen, creado usando las etiquetas <META> </META>, que contienen las primeras frase de la pagina web, o las cabeceras interiores del mismo u otros criterios dependiendo de cada motor.

- La mayoría de motores, se acompañan de enlaces del tipo «More like this…» (mas como éste o paginas similares). Si se trata de un documento especialmente útil para el usuario, la pulsación de este enlace le permitirá obtener un nuevo listado con otros de contenido muy similar.

Una vez que ha preseleccionado las respuestas pertinentes, explorado los documentos originales, se valora la pertinencia o calidad informativa. Si tras analizar las quince o veinte primeras respuesta no ha tenido algún resultado satisfactorio, es necesario cambiar la táctica de búsqueda. El cambio puede referirse a las ecuaciones utilizadas, al motor o herramienta seleccionada o ambas consideraciones.

Una posibilidad de refinamiento es volver a efectuar una búsqueda restringida al conjunto de páginas devueltas en la búsqueda anterior, con lo cual se puede afinar la búsqueda un poco más.

- Si el número e respuestas obtenido es muy elevado y los primeros resultados son poco pertinentes, muy generales, se requiere formular una nueva ecuación con mas condiciones y limitaciones. Esto permitirá reducir el número de respuestas. En el caso contrario, con nulo o escaso número de resultados, puede suceder que: Si la ecuación no es restrictiva, entonces no hay documentos o los documentos no contienen esos términos; o bien, que la ecuación sea demasiado restrictiva, con demasiadas condiciones, es necesario incluir sinónimos o términos alternativos que amplíen el ámbito de búsqueda. Si se recuperaron muchos documentos irrelevantes, excluya términos o frases que causen falsas recuperaciones.

- Si tras probar la modificación de las ecuaciones, en ambos sentidos, sigue sin obtener resultados, entonces se debe pensar en cambiar la herramienta o motor de búsqueda. Como no todos los motores de búsqueda tienen la misma cobertura, ni usan los mismos algoritmos de recuperación y ordenación de resultados, nada impide en obtener resultados positivos en lugares alternativos.

Consejos o trucos para una exitosa búsqueda

Los especialistas en recuperación de información de Abaut.com redactan una interesante columna que ofrece información actualizada sobre técnicas y trucos y desarrollo de herramientas de Internet. De entre todas las notas y guías, se resume a continuación las más interesantes para el usuario final.

- Complemente el uso de índices temáticos con de los motores de búsqueda, cuando los resultados obtenidos sean escasos, o viceversa.

- Usar minúsculas no usar acentos, al menos al principio aumenta las posibilidades de encontrar varios temas, dado que la mayor parte de los buscadores son sensibles a mayúscula o minúsculas.

- Se aconseja buscar pistas y seguir los enlaces encontrados y guardar por si se necesita hacer una revisión posteriormente.

- Estudiar las ayudas de los diferentes motores de búsqueda.

- Usar la regla de los «tres golpes»: si no se encuentre en los tres intentos, cambie la consulta.

- Use lenguaje natural mejor que palabras sueltas.

- Use operadores boléanos de forma selectiva, no como norma.

- Consulte motores especializados

- Manténgase informado de los nuevos desarrollos.

- Tenga precaución con las palabras que tienen varios significados.

- Evite usar palabras vulgares y comunes.

- Cuidado con las frases; no todos los motores entiende igual la proximidad de las palabras.

- No buscar en lugares equivocados: no se puede buscar todo en Internet.

- No use palabras vacías, Stopwords, en las ecuaciones.

APENDICE B

PRINCIPALES HERRAMIENTAS DE BÚSQUEDA EN INTERNET

En Internet resulta difícil encontrar una información pertinente y fiable. Para encontrarla, hay que partir de unos conocimientos previos sobre las herramientas de búsqueda que hay a nuestra disposición, saber qué tipo de información queremos y utilizar la herramienta idónea para el tipo de información buscada.

La multiplicidad de términos con que se alude a los mecanismos de rastreo, indización, recuperación y organización de documentos en la web puede causar confusión al usuario común. Lo cierto es que cada herramienta de búsqueda funciona y tiene un propósito y alcance diferentes, pero cada vez más las diferentes herramientas se combinan dando lugar a híbridos, que pueden dificultar la comprensión del funcionamiento interno de estos mecanismos. Una dificultad adicional es el número creciente de mecanismos disponibles, lo que hace aún más necesario clasificarlos y diferenciarlos.

Hoy en día se pueden distinguir principalmente cuatro tipos de buscadores: índices temáticos, motores de búsqueda, metabuscadores y agentes inteligentes. La diferencia básica entre ellos radica en la forma de conseguir la información para generar la base de datos, sobre la que posteriormente el usuario realizará la búsqueda.

B.1 DIRECTORIOS O INDICES TEMÁTICOS

Según Chen, una posibilidad para mejorar la eficacia de exploración de este gran espacio de información que constituye la W3, es dividirlo en diferentes categorías temáticas[46]. Los índices o directorios, son claros herederos de herramientas de consulta del World Wide Web, constituyen una opción a la búsqueda basada en palabras clave.

[46] CHEN,H. Internet Browsing and SEarching: User evaluations of category Map and Concept Space Tecnhiques, Journal of the American Society for the information Science, nº49,7, 1998, pp 582-603.

Los índices temáticos o directorios son taxonomías jerárquicas que intentan clasificar los distintos temas o áreas del conocimiento; o bien, listados de recursos organizados, incluidos en una base de datos, cuya distribución se basa, atendiendo a algún criterio de clasificación en categorías temáticas. Las categorías temáticas se organizan jerárquicamente en un árbol de materias que permite ver los recursos descendiendo desde los temas más generales a los más específicos.

Las categorías presentan un listado de enlaces a las páginas web referenciadas en el buscador. Cada enlace incluye una breve anotación sobre su contenido. Los recursos Web que contiene el índice son seleccionados y clasificados por indexadores humanos o por los propios autores de las páginas Web referenciadas. Una manera de hacer esto, consiste en rellenar un formulario que los propios buscadores facilitan para que demos de alta nuestro documento. Como es lógico, cuanto más completa sea la toma de datos, las posibilidades de búsqueda serán también mayores.

La mayoría de los índices permiten el acceso a los recursos referenciados a través de dos sistemas: navegación a través de la estructura de las categorías temáticas, y búsquedas por palabras clave sobre el conjunto de referencias del buscador, es lo que se denomina buscadores híbridos.

Cabe anotar, que en la actualidad, la división entre motores de búsqueda e índices temáticos o directorios prácticamente no existe, debido a que ambas herramientas han incorporado servicios que reflejaba antes claramente su distinción.

Algunas características de la esta herramienta: El descubrimiento de los recursos lo realizan las personas; La representación del contenido del documento es clasificado por temas o categorías, una vez recopilada, de forma manual en sus índices; La representación de la consulta es Implícita(mediante navegación por las categorías); Presentación de los resultados se realiza mediante páginas creadas previamente a la consulta; los resultados son poco exhaustivos, pero muy precisos; No realizan las búsquedas en Internet « en vivo », almacenan los datos de los sitios y ofrecen enlace a

éstas; Son muy convenientes para buscar información general, institucional porque devuelve resultados a su páginas principales. Entre otros ejemplos de directorios podemos citar: Galaxy, The Open Directory: (http://dmoz.org) realizado por miles de voluntarios, Looksmart (http://www.looksmart.com), Yahoo (http://www.yahoo.com) y (http://www.yahoo.es).

Los servicios basados en directorios han ido incorporando cada vez mas prestaciones convirtiéndose en una puerta de acceso a todas las posibilidades que ofrece la red Internet. Esta evolución ha dado lugar a lo que hoy en día se denomina “Portales”, el cual es un conjunto de servicios que pretenden satisfacer las necesidades del navegante de Internet, aunque, obviamente es bastante difícil ajustarse a las demandas de millones de usuarios potenciales. Por esto los portales de carácter general son mas adecuados para los usuarios principiantes mientras que los experimentados prefieren los portales temáticos, especializados en un determinado campo de interés.

B.1.1 ESTRUCTURA Y FUNCIONAMIENTO

Los directorios están compuestos de dos partes: La base de datos que es construida con las páginas de los sitios registrados y una estructura jerárquica que facilita la consulta a la base de datos.

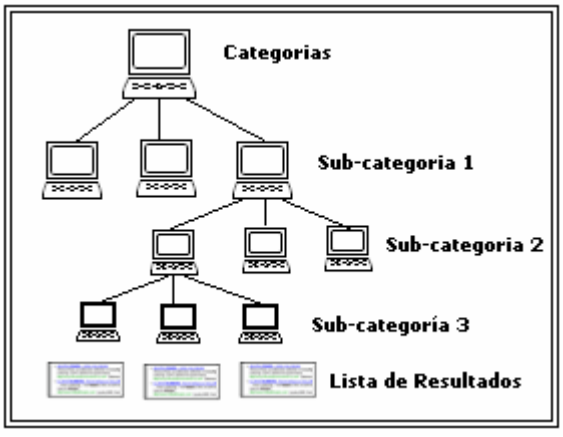


Figura23. Estructura y funcionamiento de un Directorio temático.

Como se ha señalado, los índices temáticos contemplan dos opciones de consulta: la búsqueda por categorías y las búsquedas por palabras claves mediante el motor interno.

La recogida y selección de recursos suele ser manual, aplicando determinados criterios de pertinencia, calidad formal y de contenido para evaluar si un recurso merece ser incluido o no en el directorio.

Al anterior apartado podemos decir, que el modo de indización[47] se basa en operador humano, es quien analiza el documento y asigna los descriptores o encabezamientos de materias que considera convenientes. La asignación de términos de indización consiste en la elección y atribución de términos que aparezcan o no en el texto, para representar los documentos, de acuerdo a un lenguaje documental predeterminado.

Por otro lado, la selección se realiza mediante la navegación (browsing), la cual Cove[48] señala como el arte de no saber lo que uno quiere hasta que lo encuentra. Básicamente lo que hace el usuario es recorrer la estructura ramificada para buscar la información que necesita. De esta manera, el usuario puede descender por los niveles de especificidad hasta encontrar la información adecuada a sus intereses sin necesidad de formular explícitamente su consulta, aunque también se suele ofrecer un formulario de búsqueda para obtener un listado de las categorías pertinentes. Algunos directorios tienen la posibilidad de incluir referencias cruzadas, es decir, desde un punto determinado de la ramificación del árbol de categorías se puede saltar a otra categoría similar ubicada en otro nivel de la jerarquía del directorio.

Si se hace uso de la búsqueda por palabras claves mediante el motor interno, se permite encontrar las páginas web escribiendo los términos de búsqueda en el formulario de consulta del directorio. Seguidamente, se da un resumen acerca de las ventajas e inconvenientes, que caracterizan al los Índices temáticos o los muy conocidos Directorios:

[47] La indización o indexación, es la operación destinada a representar los resultados del análisis de contenido de un documento o de una parte del mismo, mediante elementos(denominándose genéricamente 'términos de indización') de un lenguaje documental o natural, para facilitar la recuperación. El objetivo final de una indización es una recuperación óptima de la información, entendiendo por óptima el encontrar documentos que hablen de una materia determinada (exhaustividad) y sólo aquellos que traten de dicha materia (precisión).

[48] Cove, J.F. and Walsh, B.C."Online text retrieval via browsing". *Information Processing & Management*, 24, 31-37. 1998

TABLA16. VENTAJAS E INCONVENIENTES DEL USO DE LOS INDICES TEMÁTICOS O DIRECTORIOS

Ventajas	Inconvenientes
Facilitan la búsqueda, especialmente para usuarios inexpertos: sólo hay que elegir la categoría principal que más se acerque a la consulta e ir descendiendo por aquellas ramas o nodos (subcategorías) que se aproximen más al interés de búsqueda hasta llegar a las hojas donde encontraremos los enlaces a los recursos de Internet correspondientes. Este filtrado permite ajustar la precisión.	Cubren una proporción relativamente pequeña de los recursos existentes en la Web, por lo que si se quiere información sobre algo muy concreto quizás no se encuentre con un directorio y se deba recurrir a un buscador.
Permiten visualizar de forma global al contenido. Muchos directorios indican en cada uno de los nodos cuántas referencias y subcategorías contienen cada una de las bifurcaciones posibles desde el nodo actual. Eso ayuda al usuario a tener una visión general del volumen y contenido del índice, aspecto que es muy difícil, si no imposible, de determinar en un buscador.	Existe una carencia de criterios homogéneos para la selección y clasificación, y una ausencia también de principios de descripción homogéneos. Es decir, en ocasiones, puede resultar confuso determinar bajo que categoría de temas se habrán de incluido los recursos a los que se esta interesado.
Los términos hallados están dentro del contexto de la categoría en la que efectuemos la búsqueda, lo cual disminuye considerablemente el ruido.	Muchos recursos dejan de ser útiles si no se utilizan mecanismos automáticos para seguir los cambios en sus contenidos, direcciones, aparición o desaparición.
Los recursos disponibles han pasado por un proceso de selección de calidad, generalmente efectuado por documentalistas. Por tanto, las Bases de datos suelen ser mas pequeñas, menos actualizadas, pero mas elaboradas gracias a la precencia del factor humano.	El método navegacional de consulta es lento para encontrar lo deseado, pues exige varios pasos previos.
Las descripciones que muestran acerca de los sitio web suelen estar elaboradas intelectualmente, por lo que son realmente descriptivas de su contenido.	
Resultan muy útiles cuando no se tiene muy perfilada la necesidad de información o bien cuando se buscan recursos de tipo general.	
Presentan también un motor de búsqueda interno para localizar directamente recursos incluidos en sus bases de datos sin que tenga que explorar el directorio temático obligatoriamente, es decir, tambien se pueden ejecutar ecuaciones de búsqueda y plantear consultas mediante palabras clave.	

Hay un gran número de directorios tanto de tipo general como específicos, pero es creciente la dificultad de su existencia como directorios independientes. Solo aquellos directorios que son soportados por grupos de interés como: directorios telefónicos, asociaciones profesionales, grupos de industriales y líneas de productos son los que son más actualizados y visitados. Son los que tienen futuro.

B.2 MOTORES DE BÚSQUEDA

Los motores de búsqueda consisten en bases de datos muy voluminosas generadas como resultado de la indexación de partes significativas de los documentos que han sido analizados previamente en Internet, éstos suelen recoger documentos en formato HTML y otros tipos de recursos.

La tarea para rastrear la estructura hipertextual web y localizar, actualizar y ampliar los recursos que incluirán en su base de datos, es realizado por un programa denominado crawler (Spiders o Robots), que recorren la red de forma automática explorando los servidores a nivel mundial, o en el ámbito de especialización del buscador (geográfico, idiomático o temático). Por ende cada robot rastrea a su manera en la Web, de ahí que la información almacenada en cada base de datos sea diferente.

La recuperación se realiza gracias a un sistema de gestión de bases de datos que permite distintos de tipos de consulta y a la ordenación de los resultados por relevancia en función a la estrategia de consulta. Los motores de búsqueda resultan ser más exhaustivos que los índices en cuanto al volumen de las páginas referenciadas, pero son mucho menos precisos que los índices temáticos o directorios al no ser su contenido objetos de indexación humana.

La participación del usuario en el uso de esta herramienta, se centra en expresar su necesidad de información mediante un formulario. Este puede consistir en una simple caja donde teclear palabras clave (búsqueda simple) hasta un formulario con multitud de opciones para expresar con mayor detalle aquello que se desea buscar (búsqueda avanzada).

Las búsquedas avanzadas suelen ofrecer la posibilidad de utilizar operadores (boléanos, de adyacencia, de existencia, de exactitud) y a veces se puede delimitar la búsqueda por (fechas, por tipo de fuente, por área geográfica o dominio, por idioma etc.). Algunos buscadores incluyen la posibilidad de realizar una búsqueda expresada en lenguaje natural (por ejemplo, altavista,); ello permite al usuario utilizar un lenguaje no estructurado para describir que está buscando, siendo el motor de búsqueda el responsable de traducir esa búsqueda a un formato estructurado. Sea cual sea la forma de expresar la pregunta por parte de usuario, está será analizada por el buscador y se traducirá a una representación interna que permita compararla con los términos recogidos en el fichero interno y seleccionar así las URL que sean más relevantes.

Algunas de las características más notables en los motores de búsqueda es que el descubrimiento de recursos lo hacen principalmente en forma automática mediante robots, la representación del contenido del documento es mediante indización automática, la representación de la consulta es explícita (mediante palabras claves, conceptos, operadores, delimitadores, etc…), presentación de resultados se realiza mediante páginas creadas de forma dinámica para cada consulta.

Existen docenas de motores de búsqueda, cada uno tiene sus características peculiares, por lo que los resultados que proporcionan para una clave de búsqueda idéntica pueden ser muy distintos. El problema, si se quieren manejar varias de estas herramientas, es que el modo de utilizarlas difiere, a veces sustancialmente, de una a otra.

B.2.1 ESTRUCTURA Y FUNCIONAMIENTO

Un motor de búsqueda está compuesto por cuatro componentes: el robot o spider, el motor de indexación, los índices y el motor de búsqueda[49].

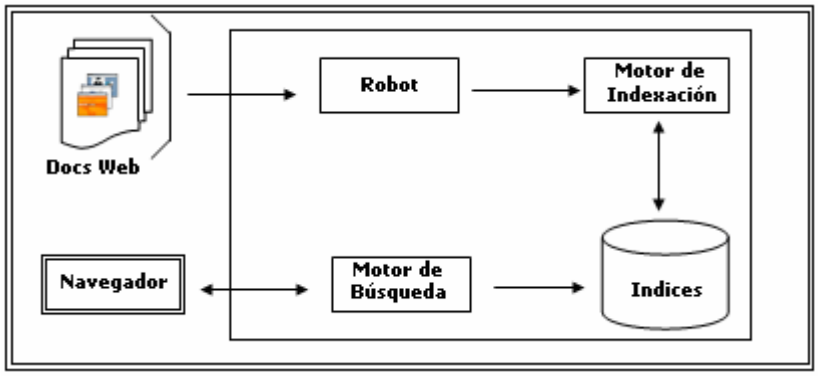


Figura24. Estructura y funcionamiento de un motor de búsqueda.

El robot, también llamado spider, o web crawler, es un programa que se encarga de recorrer web obteniendo la información relevante del buscador de cada una de las páginas que visita. Como primer paso, todo robot pasa de una lista de URLs conocida. La misma puede se producida por el conjunto de usuarios que ha dado de alta en el buscador.

[49] Fernández Leal, Fco. Javier. Diseño e implementación de una Arquitectura multiplataforma para el estudio de Motores de búsqueda en Internet. < http://members.fortunecity.es/javocho/paginas/buscadores.htm>

Luego se elije una URL de la lista y se obtiene el correspondiente documento de la web. La información a recuperar varía de un sistema a otro, algunos almacenan todo el documento, mientras que otros se limita al título, y a las primeras n líneas o palabras. Los enlaces presentes son agregados a lista de URL pendientes, tras lo cual se continúa con la siguiente URL de la lista la forma en que esto enlaces son agregados determinan en gran parte el comportamiento de la búsqueda destacándose las políticas de "Primero en profundidad[50]" y "Primero en anchura[51]" (analiza cuantos enlaces existen hacia otra pagina).

Una vez que se posee la información de la pagina esta debe ser analizada y condensada, para permitir su organización como su posterior presentación a las usuarios del sistema. La parte del sistema encargada de realizar esta tarea es el motor de indexación.

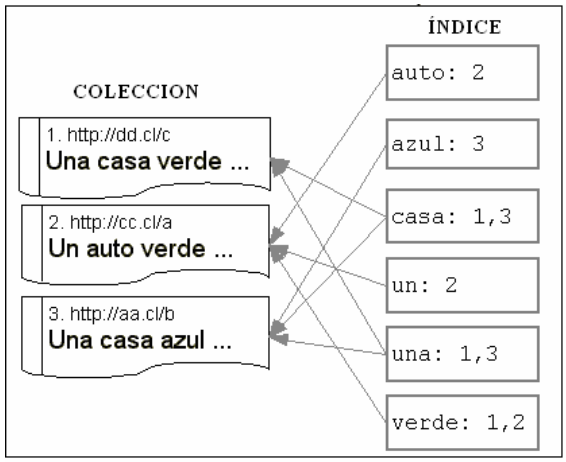
La indexación puede ser por palabras claves, el caso más común, o por conceptos. La primera hace referencia a la indexación por índices inversos de raíces y palabras clave, direcciones, ubicación y frecuencia de apariciones. Este enfoque, esencialmente morfológico y estadístico, basa la recuperación de la información en la similitud formal de las palabras, y las estadísticas de su presencia en documentos y colecciones de documentos. Algunos buscadores obtienen las palabras clave de determinados campos, las metaetiquetas HTML, pero la mayoría indiza el texto completo de las paginas, incluyendo o no las palabras vacías (los artículos, preposiciones, etc.). El segundo método de indexación hace referencia a procedimientos para construir bases de datos basadas en conceptos, algunas de ellas muy complejas y basadas en sofisticadas teorías lingüísticas y de inteligencia artificial. En otros casos, se basa en una aproximación numérica calculando la frecuencia de aparición de ciertas palabras significativas. A partir de análisis estadísticos el buscador determina que conceptos aparecen juntos o

[50] La búsqueda primero en profundidad, es la técnica que consiste en explorar un camino que recorre yendo hacia los grados de profundidad creciente. En este caso de fracaso (una rama conduce a un callejón sin salida), retrocede hasta el nivel nodo-padre y si fuera posible, se buscad otro de sus sucesores y vuelve a marchar en profundidad. La estructura de datos que maneja (LIFO, Last-in First-out), de manera, que se trata con prioridad el nodo mas recientemente memorizado.

[51] La Búsqueda primero en anchura, consiste en desarrolla sucesivamente los nodos de un mismo nivel de profundidad(hay que detenerse cuando se encuentra estado-meta) y en caso que la solución no se haya encontrado, recorrer el proceso en el nivel de profundidad siguiente. La estructura de datos que maneja es (FIFO, First-in First out), de tal forma que el nodo mas antiguo memorizado en el nivel de profundidad es el que se desarrolla con prioridad.

relacionados en textos que se centran en un tema en concreto. Mediante este sistema se pueden recuperar recursos que tratan de un tema dado, incluso aunque las palabras incluidas en el documento no coincidan formalmente con las de la pregunta.

Para permitir las búsquedas en la información recolectada son mantenidas estructuras de datos llamados "Índices o ficheros inversos", mediante los cuales se asocian una palabra a una lista de documentos relacionados con ella. En esto índices cada entrada corresponde a cada una de las palabras distintas que figuran en la base de datos seguidas de una lista de identificadores de cada uno de los documentos que son descritos por dicha palabra y la información estadística respecto a la frecuencia o importancia de esta palabra en ese documento.



El motor de búsqueda es el encargado de procesar las consultas recibidas por los usuarios, para lo cual recorre los índices inversos buscando los términos relacionados con la consulta y obteniendo los identificadores de documentos. El criterio utilizado para ordenar los resultados (ponderación) varia según el motor de búsqueda[52] pero en general se basa en la posición donde aparecen los términos, dando mayor importancia los títulos, las palabras resaltadas y las primera líneas; la frecuencia de ocurrencia de una palabra o frase de consulta, si los términos de algún documento tiene todas las palabras o frases solicitas y la cercanía de los distintos términos dentro de un documento, entre otros.

Exclusión de Robots

[52] Estos documentos serán ordenados usando distintos criterios y heurísticas, con el objeto de indicar al usuario cuál es el documento más relevante.

Como se ha mencionado, los robots funcionan de forma automática, esto ha llevado en muchas ocasiones a producir sobrecarga en la Red. La razón era que si estaban mal diseñados, podían generar sin parar tráfico de páginas web, con el fin de indexarlas. La solución que se pensó fue otorgar en los servidores web la posibilidad de no permitir el acceso a determinados Robots o no permitir la inspección de determinadas rutas de dicho servidor.

El método empleado para que un servidor Web evite la inspección del mismo, por parte de los Robots, se estructura en dos ámbitos: por una parte se facilita al Administrador del Web un mecanismo de exclusión de Robots y por otro se proporciona al propietario de cada página HTML un mecanismo adicional de control del acceso a la misma por parte de los Robots.

El primer mecanismo se denomina “Protocolo de exclusión de Robots”, y permite al Administrador decidir que partes del Web no deben ser indexadas. El medio para conseguirlo es un archivo de texto denominado ROBOTS.TXT, que contiene las instrucciones sobre las páginas visitables y las que no permiten el acceso a los Robots.

El segundo mecanismo se logra mediante la inserción de unas etiquetas HTML denominadas META Tags en las que se indica al Robot si debe o no inspeccionar o indexar cada página HTML individual. Hay que tener en cuenta que estos métodos de protección dependen del buen comportamiento de los Robots, ya que resulta trivial para un Robot pasarlo por alto.

A continuación se presenta las ventajas e inconvenientes que presentan los motores de búsqueda:

TABLA17. VENTAJAS E INCONVENIENTES DEL USO DE LOS MOTORES DE BUSQUEDA	
Ventajas	**Inconvenientes**
Son muy exhaustivos. Dado que el proceso de recogida de recursos y de indexación es automático, eso permite incluir en el buscador información acerca de una cantidad enorme de páginas web, del orden de millones.	Su utilización es bastante más compleja que la de los directorios, requiere un mayor esfuerzo por parte del usuario.
Se utilizan también mecanismos automáticos para seguir los cambios en sus contenidos, direcciones, aparición o desaparición. Algunos buscadores incluso guardan una copia en caché de los documentos tal como estaban en el momento en que fueron explorados.	Cada buscador tiene su propia sintaxis para expresar la consulta que es preciso conocer y diferenciar.
Se pueden utilizar sinónimos y traducciones de los términos significativos para tener una mayor amplitud del campo de búsqueda para lo cual se utilizarán diccionarios de sinónimos y diccionarios de inglés.	Para obtener resultados precisos se requiere formular la consulta cuidadosamente, eligiendo adecuadamente los términos y los operadores, y delimitando adecuadamente la búsqueda.
Se utilizan para buscar información mas escasa, especialidad, actualizada o incluida en páginas personales.	Los recursos indexados por los robots no han pasado generalmente por ningún proceso de elección de calidad por los que entre los resultados puede haber mucha "basura".

CÓMO ES LA SELECCIÓN DE DOCUMENTOS RELEVANTES

Se pueden definir dos tipos fundamentales de relevancia: la relevancia formal y la semántica. La primera se refiere, a lo bien que responden los resultados de una búsqueda a la ecuación planteada, mientras que la segunda, es una medida abstracta de lo bien que satisface un documento la necesidad de información de un usuario. La relevancia semántica es una cualidad relativa y gradual, porque es relativa a los usuarios y a sus necesidades de información, y los documentos la poseen en una forma gradual. Esta función puede adoptar cualquier valor entre 0 (ausencia total de relevancia) y el 1 (completamente relevante). Es una noción subjetiva y difícil de cuantificar.

La información que obtienen los usuarios suele ser muy pobre, tanto formalmente (sus ecuaciones de búsqueda son pobres, o el lugar en el que buscan no es el adecuado), como semánticamente (confunden la información que desean obtener con lo que realmente preguntan). Por tanto, se ha de interrogar adecuadamente al servicio de búsqueda pero además quien pregunta ha de saber realmente cuál es el problema que quiere resolver buscando información. Sin embargo, la verdadera relevancia, es decir obtener información relevante para un problema es mucho más compleja. No consiste simplemente en definir una ecuación de búsqueda correctamente, o en ir a hurgar en una buena base de datos.

Según Stefano Mizzaro[53], la relevancia es una relación entre dos entidades cualesquiera escogidas cada una de ellas entre los componentes de dos grupos distintos. En el primer grupo están:

- El problema que un humano tiene que resolver.
- La necesidad de información, es decir, cómo el humano en cuestión representa en su mente el problema al que se enfrenta.
- La petición de información, o sea, cómo el humano expresa esa necesidad de información a alguien, normalmente en lenguaje natural.
- La interrogación, que consiste en la ecuación de búsqueda a plantear.

En el segundo grupo están:

- El documento, entendido como el soporte físico donde está la información.
- El subrogado, es decir, la representación de ese documento.
- La información, o sea, lo que el usuario capta al leer un documento. Así, se pueden distinguir diversos tipos de relevancia. Por ejemplo, se puede hablar de la relevancia de un documento con respecto a una petición (el documento obtenido responde a la necesidad expresada por el peticionario) o se puede hablar de de un subrogado que es búsqueda en una base de datos satisfacen correctamente la ecuación de búsqueda planteada). Pero ninguna de las dos relevancias garantiza que el documento en cuestión responda al problema real del usuario, que, por otra parte, puede que incluso no conozca bien.

Así que, por una parte, existe el problema de instruir a la gente en cómo reconocer bien los problemas, cómo plantearlos en forma de necesidad de información, y cómo plantear ecuaciones de búsqueda que lleven a documentos 'relevantes' para el problema, aparte de existir la necesidad de conocer bien cuáles son las fuentes de información y cómo se usan.

Por otra parte, en cuanto a los servicios de búsqueda, tenemos los documentos representados en base al análisis documental efectuado (lo que Mizzaro denomina el subrogado) y el sistema ha de devolver aquéllos documentos que mejor respondan a la

[53] Adelaida Delgado. Mecanismos de Recuperación de Información en la WWW. 1998. <http://servidorti.uib.es/adelaida/tice/modul6/memfin.pdf>

ecuación de búsqueda planteada por el usuario (suponiendo que esa ecuación refleja realmente su necesidad de información).

B.2.2 CRITERIOS DE PONDERACION

Los criterios de ponderación varían para cada motor de búsqueda si bien generalmente se basan en métodos estadísticos de frecuencia y en la posición donde aparecen los términos.

Algunos de los criterios incluyen:

- Si las palabras o frases se encuentran en las primeras pocas líneas de un documento, (por ejemplo, en el título de una página web).

- Si la frecuencia de ocurrencia de una palabra o frase de consulta supera un determinado umbral. Son muy pocas las palabras en una consulta que se ponderan más que las palabras comunes (la rareza es determinada por el número de ocurrencias de la palabra en el índice).

- Si todas las palabras o frases especificadas aparecen en un documento. Un documento que contiene las tres palabras especificadas en una búsqueda de tres palabras se clasificará más alto que un documento que contiene únicamente dos o una de las palabras.

- Si las palabras o frases de una consulta múltiple se encuentran cerca una de otra en el documento.

Una aproximación de criterio ponderación es comparar las palabras de la consulta con las palabras que hay en los documentos encontrados. Esto es un poco más complejo que simplemente contar cuantas veces aparece cada palabra de la consulta en cada documento, puesto que hay una variable extra a considerar.

Si la página encontrada contiene una palabra por la que se preguntó al buscador, que no aparece en casi ningún otro documento, entonces eso es una buena evidencia de que la página que estamos mirando es importante. Esto ha sido usado durante años con bastante éxito.

Una opción alternativa es la propuesta en [Kle, 1998]. En síntesis se utilizan los links entre las páginas como evidencia. Una forma simple de comenzar a entenderlo es la siguiente hipótesis: una página con buen contenido, seguramente es referenciada desde muchos buenos índices.

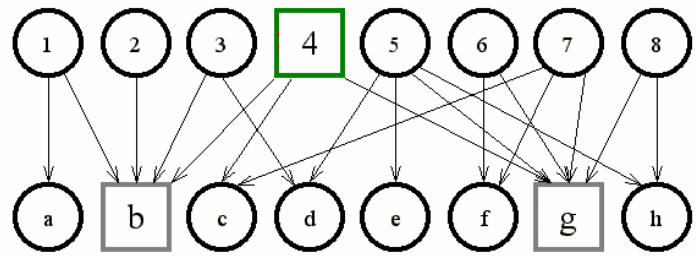


Figura25. Criterio de Ponderación

Por ejemplo, en la figura23, tenemos numeradas algunos potenciales índices, y con letras las potenciales páginas de contenido. Vemos que 'b' y 'g' claramente atraen la mayoría de los enlaces desde otras páginas, por lo cual las marcamos como la "mejores páginas de contenido".

Así mismo, la página '4', (ver fiugra23) se destaca como "mejor página índice", precisamente porque apunta a las dos "mejores páginas de contenido". No es la página con más enlaces necesariamente, de hecho '5' tiene más enlaces, pero no son los más apropiados. Al mismo tiempo, el puntaje como página de contenido de 'c' aumenta, porque aparece en una buena fuente de información. Este procedimiento se repite varias veces, sumando puntos como índice y como contenido.

B.2.3 CRITERIOS PARA LA SELECCIÓN DE UN MOTOR.

Existe una gran cantidad de motores de búsqueda en Internet, cada uno ofrece diferencias en cuanto a volumen de páginas, elementos de cada página que son indexados, interfaz, lenguaje de consulta, algoritmo de cálculo de la relevancia, etc. Estas diferencias provocan que los resultados de aplicar una misma consulta a varios buscadores en ocasiones no coincidan. A la hora de valorar la calidad de un buscador se debe tener en cuenta:

1. La exhaustividad: número de documentos de Internet referenciados que almacena el motor de búsqueda en su base de datos, para las consultas.
2. La calidad y flexibilidad del lenguaje de consulta: indica que tanto se pueden mejorar los resultados de una consulta en base a los operadores con los que cuenta el motor. En este punto también influye mucho la interfase de usuario.
3. La pertinencia de sus resultados (ruido y silencio): el número de resultados arrojados en una consulta no debe ser tan pequeño como para no proporcionar suficiente información, ni tan grande como para no poder definir cuales son los resultados relevantes.
4. Los servicios de valor añadido que incorporan: tales como correo electrónico, compras en Internet, noticias, disco virtual, mensajero electrónico, etc.
5. La periodicidad de actualización de la base de datos: la frecuencia con la que el robot regresa a los sitios que tiene indexados para verificar si alguno de ellos ha actualizado sus páginas, si el sitio ya no existe, o para registrar los sitios nuevos.
6. La velocidad en la recuperación: la velocidad de respuesta a una consulta, es decir, el tiempo que toma el motor de búsqueda en consultar su índice y aplicar el algoritmo para regresar los resultados.
7. Las dificultades de conexión: la facilidad con la cual se puede acceder al sitio del motor de búsqueda.
8. Muchos estudios afirman que el criterio más seguido por los internautas es la costumbre.

B.3 METABUSCADORES

En la actualidad hablar de buscador, obliga necesariamente a hablar de los metabuscadores, innumerables trabajos, abordan variada información sobre el mismo, y es que a pesar, de los beneficios indiscutibles proporcionados por los motores de búsqueda y de los índices temáticos o directorios, su crecimiento condujo a la creación de nuevas herramientas.

Los Metabuscadores, son sistemas de búsqueda desarrollados para mitigar el problema de tener que acceder a varios motores de búsqueda o directorios con el fin de recuperar

una información más completa sobre un tema, ya que son estos mismos sistemas los que se encargan de efectuarlos por el usuario. En realidad el metabuscador tiene como función realizar sus búsquedas de manera simultánea sobre diversos buscadores, de tal forma que éste actúa como almacenamiento intermedio de la información en una base de datos, de tal forma que colecciona las respuestas recibidas y las unifica.

Según Tyner[54], se conoce como metabuscadores, aquellos que permiten interrogar varias bases de datos simultáneamente desde una única interfaz; aunque ellos no ofrecen el mismo nivel de control sobre la lógica y la interfaz de búsqueda que los motores y directorios, la mayoría son bastante rápidos.

B.3.1 FUNCIONAMIENTO

Estos sistemas no tienen sus propias bases de datos, por tanto, no almacenan páginas web, no agregan direcciones, ni clasifican y describen sitios web, en lugar de eso contienen registros de motores de búsqueda e información sobre ellos(es decir, no se sirven de robots, sino que van a buscar directamente a los índices de cada buscador). Básicamente su funcionamiento se basa en enviar la petición del usuario a todos los motores de búsqueda y directorios registrados y obtienen los resultados que les devuelven. Algunos más sofisticados detectan las URL duplicadas provenientes de varios motores de búsqueda y eliminan la redundancia, es decir solo presentan una aparición el mismo recurso al usuario.

Figura26. Estructura y funcionamiento de un Metabuscador

[54] Tyner R. Sink or swim: Internet search tools & techniques, 2001. <http://www.ouc.bc.ca/libr/connect96/search.htm>

Por muy grande y exhaustiva que pudiera llegar a ser la base de datos de un motor de búsqueda o de un directorio, nunca va a cubrir un porcentaje muy elevado del total de la web, por ello, para asegurar una búsqueda sobre una materia suficientemente exhaustiva, se requiere hacer uso de varios de ellos.

Estos sistemas se diferencian unos de otros en la manera en que llevan a cabo el alineamiento de los resultados en el conjunto unificado, y cómo de bien traducen estos sistemas la pregunta formulada por el usuario a los lenguajes específicos de interrogación que maneja cada uno, ya que el lenguaje común a todos será más o menos reducido.

Otra diferencia sustancial existente entre estos sistemas es la presentación de los resultados, con base en esa característica algunos autores hablan de multibuscadores y los metabuscadores. Los multibuscadores ejecutan la consulta contra varios motores de forma simultánea y presentan los resultados en forma concatenada, sin más organización que la derivada de la velocidad de respuesta de cada motor, es decir, para cada motor interrogado se presenta una lista de los resultados obtenidos. Los metabuscadores funcionan de manera similar a los multibuscadores pero, a diferencia de éstos, permiten obtener los resultados de forma integrada, eliminando las referencias duplicadas, agrupan los resultados, indicando qué buscador o buscadores lo han proporcionado y generan nuevos valores de pertinencia para ordenarlos.

En el primer caso el servicio es más rápido ya que a medida que cada uno de los buscadores va devolviendo sus resultados éstos son presentados inmediatamente al usuario; en el segundo caso, hay que esperar que todos los buscadores hayan devuelto sus resultados y que el multibuscador confeccione su propia lista, generalmente ordenada por algún criterio de relevancia.

El principal inconveniente tiene que ver con el hecho ya mencionado de que, como cada motor tiene sus particularidades, éstas no se explotan de forma adecuada en las consultas, necesariamente genéricas, que les pasan los metabuscadores (el denominado "problema del mínimo común denominador"). Además, por regla general sólo tienen en cuenta las primeras referencias que aporta cada motor. Por ello, parecen más indicados para una primera aproximación a un tema específico formulable de manera sencilla.

Uno de los mayores inconvenientes de estos sistemas es que el resultado no tiene porqué ser necesariamente todo el conjunto de páginas sobre la materia preguntada que se encuentran almacenadas en las fuentes del metabuscador, ya que el número de documentos recuperados de cada una de estas fuentes se encuentra generalmente limitado.

Puede sorprender la existencia de esta limitación, pero es importante no olvidar uno de los elementos que, clásicamente, han incidido mucho en los métodos de evaluación: el tiempo de respuesta del sistema. Si un metabuscador devolviera todas las referencias de todos los motores y directorios que le sirven de fuente que tienen relación con la materia objeto de nuestra búsqueda, el tiempo de respuesta del sistema alcanzaría valores que seguramente alejarían a los usuarios del metabuscador por ser excesivamente lento. Es por ello que resulta necesario establecer un número límite de documentos recuperados por cada motor, con el fin de que el tiempo de respuesta (que ya es, por lógica, mayor que cuando se consulta un solo motor o directorio) aumente excesivamente.

Liu[55] expone algunos elementos que son válidos anotar. Él plantea, que es totalmente imposible que estas herramientas puedan unificar todas las ventajas de cada uno de los motores, y que, por consiguiente, las búsquedas booleanas pueden generar resultados diferentes en diversos buscadores, las búsquedas por freses puede que no se ejecuten en algunos de ellos, y otros elementos como el uso de limitadores pueden sacrificarse. Apunta que los metabuscadores no devuelven, desde cada buscador todas las páginas que corresponden con la solicitud sino que toman un rango de 10 a 100 registros por cada uno, por lo que permiten redireccionar la búsqueda, una vez que se escoja el motor, que mas resultados relevantes arrojó, se continúa con la siguiente búsqueda.

Algunos autores aseguran que la mayoría de los grandes motores de búsqueda utilizan lenguajes de interrogación similares. Por esto, los resultados en los metabuscadores pueden verse favorecidos si se conocen, los lenguajes utilizados por los buscadores que estos procesan. Afortunadamente algunos metabuscadores ofrecen al usuario, la opción de escoger dentro de un grupo determinado de buscadores disponibles, cuales incluir en

[55] Liu J. Guide to meta-search engines, 1999. < http://www.indiana.edu/~librcsd/search/meta.html>

su metabúsqueda. Si se escogen aquellos conocidos por su confiabilidad, potencia y rapidez, los resultados de la búsqueda serán los más óptimos.

TABLA18. VENTAJAS E INCONVENIENTES DEL USO DE LOS METABUSCADORES

Ventajas	Inconvenientes
Permiten ejecutar una búsqueda mas extensiva a través de un amplio número de herramientas de búsqueda.	Cuando envían su búsqueda a varios motores de búsqueda, sus resultados tienen cierta dependencia de que estos estén disponible en el momento de la búsquedao se descarguen en el periodo de tiempo permisible.
El usuario solo requiere acceder a una única pagina web para ejecutar la búsqueda y no tiene necesidad de recordar los nombres o direcciondes de los motores de búsqueda.	Son difíciles de usar para búsquedas muy precisas, porque tienen menor control de la búsqueda al interrogar varias bases de datos con interfaces diferentes.
El usuario ha de aprender a utilizar una única interfaz para realizar la búsqueda .	Se recomienda su uso para la búsqueda de información difil de encontrar y de la que poeemos datos específicos.
La formulación de la consulta se realiza una sola vez y el metabuscador remitirá o redireccionará ésta a cada una de las herramientas de búsqueda.	Una limitación se ve reflejada al no presentar opciones de búsqueda avanzada, con lo cual se sacrifica la precisión en los resultados de las búsquedas
Se tiene la posibilidad, de obtener resultados integrados, a partir de varios motores de búsqueda	
Permite indicar el numero máximo de resultados que se desean obtener de cada motor de búsqueda; Salvar la ecuación de búsqueda para poderla ejecutar la proxima vez que se use el servicio; Presenta resultados por orden de relevancia, generalmente basandose en el criterio de "best match".	

B.4 AGENTES INTILIGENTES

Un agente es una entidad que percibe y actúa sobre un entorno. Un tipo de agente particular son los denominados agentes inteligentes.

Los "agentes inteligentes", se tratan de programas basados en técnicas de inteligencia artificial para ser mas precisos, se basan en el uso de tecnología push (servicio de difusión selectiva de información automatizado proporcionado a través de Internet) que intentan ofrecer una información personalizada en función de las necesidades del usuario mediante funciones de búsqueda, discriminación y selección.

El hecho de que el concepto de agente inteligente posea tantas aplicaciones, dificulta el encontrar una definición exacta del mismo. Entre las múltiples definiciones realizadas se pueden destacar:

• Hipola y Vargas-Quesada señala: "Una entidad software que, basándose en su propio conocimiento, realiza un conjunto de operaciones destinadas a satisfacer las necesidades de un usuario o de otro programa, bien por iniciativa propia o porque alguno de éstos lo requiere".

• Maes lo define como "Programas de ordenador capaces de efectuar una tarea o actividad sin la manipulación directa de un usuario humano."

• Stenmark expone: "Un agente inteligente es un software que asiste al cliente y actúa en su nombre"

A partir de estas definiciones, se puede considerar un agente inteligente como, una pieza de software que ejecuta una tarea dada utilizando información recolectada del entorno, para actuar de manera apropiada hasta completar la tarea de manera exitosa.

Los agentes inteligentes tienen la característica de "aprender" de diversas formas:

• Observando e imitando el comportamiento del usuario.

• Recibiendo retroalimentación positiva o negativa por parte del usuario.

• Recibiendo instrucciones explícitas del usuario.

• Pidiendo consejo a otros agentes.

Los agentes inteligentes pueden realizar una serie de tareas sin que los humanos u otros agentes les tengan que decir qué deben de hacer en cada paso que dan. Se diferencian de los buscadores en que éstos son bases de datos estáticas (aunque se actualizan con cierta frecuencia) y responden directamente a las peticiones de los usuarios. Esto no quiere decir que si un buscador cumpliera estas dos premisas ya sería un agente, ya que un agente contiene otras características a mayores. Hoy en día, Se pueden distinguir los siguientes tipos de agentes[56].

a) En cuanto a su ámbito de acción:

• Agentes de escritorio (agentes de sistema operativo, agentes de aplicaciones, etc.)

[56] HÍPOLA, P., VARGAS-QUESADA, B. "Agentes inteligentes: definición y tipología. Los agentes de información". *El profesional e la información*, 1999, vol. 8, n. 4, p.13. Fuente: <http://personales.upv.es/ccarrasc/doc/2002-2003/AgySistemasdeInformac/SRPTrabajo2.htm>

• Agentes Internet (agentes de búsqueda, filtrado, recuperación de información, agentes de notificación, agentes móviles, etc.)

• Agentes Intranet (agentes de customización colaborativa, agentes de bases de datos, agentes de automatización de procesos, etc.)

b) En cuanto a su función:

• Interface Agents (Agentes de Interfaz): La finalidad es proporcionar información a los usuarios, en los sistemas informáticos que dada su complejidad poseen una alta carga de información. Se caracteriza por su capacidad de hacer comprensible la s interfaces.

• System Agents (Agentes de Sistemas): Son agentes que permiten realizar inventario de hardware, interpretar eventos de red, manipular dispositivos de
respaldo y almacenaje, detecta virus, etc., todo esto sobre sistemas complejos como es el caso de los sistemas distribuidos.

• Advisory Agents (Agentes Consejeros): Este tipo de agente proporciona consejos a los usuarios en el caso de utilización de herramientas, o en sistemas de diagnóstico o ayuda.

• Filtering Agents (Agentes de Filtración): Agentes que se usan para reducir la saturación de información mediante el borrado de los datos no deseados (por ejemplo datos que no satisfacen completamente el perfil del usuario). Muchos clientes de e-mail proporcionan estas prestaciones.

• Retrieval Agents (Agentes de Recuperación de Información): Estos agentes se especializan en la búsqueda y recuperación de información, ejecutándose sobre grandes bases de datos, bases de conocimiento o bases de documentos.

• Navigation Agents (Agentes de Navegación): Este tipo de agentes se utilizan para navegar sobre sistemas conectados en red, alunas de sus funciones principales son el recordar sitios y direcciones de interés de manera automática.

• Monitoring Agents (Agentes de Monitoreo): Proporcionan información de manera eficaz y oportuna a los usuarios, cuando ocurre un evento.

• Recomender Agents (Agentes de Recomendación): Estos agentes utilizan bases de datos con información acerca de preferencias de un grupo de usuarios acerca de un ítem o tópico en particular (información, productos, etc.), basando su recomendación en analogías o match con otros usuarios de perfil similar a un usuario dado.

Una de las principales aplicaciones de los agentes esta en la Recuperación y manejo de la información: En esta área los agentes ayudan a los usuarios no solo a buscar automáticamente información disponible en la red a intervalos definidos sin requerir la presencia del usuario, sino también a categorizar, diseminar selectivamente y compartirla bajo criterios colaborativos. Para esto el agente cuenta con un robot, que es quién ejecuta las tareas repetitivas y el agente le da las instrucciones.

Para cada búsqueda estos agentes consultan muchos buscadores de manera simultánea y combinan sus resultados eliminando los duplicados y los enlaces muertos y conservando los documentos más relevantes. Los resultados se pueden ordenar y enviar por correo electrónico, también permiten guardar las estrategias de búsqueda para usarlas con posterioridad. El más conocido es Copernic.

La mayoría de los buscadores utilizan técnicas derivadas de las tecnologías de los agentes inteligentes, sin embargo los motores denominados inteligentes son aquellos que se basan en las tecnologías del lenguaje natural, las técnicas de aprendizaje y las redes neuronales. Los motores inteligentes también se emplean en capturar y almacenar preferencias de los usuarios. El objetivo es saber con precisión sus requerimientos y mejorar el plan de búsqueda.

En la tabla se presenta las principales características que debe tener todo programa, para ser considerado “agente inteligente”.

TABLA19. CARACTERISITICAS DE LOS AGENTES INTELIGENTES:
Han de poseer un nivel de inteligencia suficiente para aprender (se estudiam implementaciones con redes redes bayesianas y redes neuronales). El n ivel de inteligencia del agente determina el método de aprendizaje. Hay varios niveles de inteligencia que van desde la aceptación y ejecución de tareas hasta el aprendizaje y la adaptación al entorno, el establecimiento de relaciones y la prediccion de los de las necesidades de los usuarios.
Han de tener autonomía, la cual dependera del grado de interactividad que se precise entre el usuario y el servidor.
Han de tener movilidad, para poder navegar por las redes y acceder a arquitecturas y plataformas diferentes.
Han de ser capaces de poder reaccionar. Las percepciones captadas de su entorno por parte del agente producen una acción específica (por ejemplo, ante una palabra errónea o mal escrita, determinar qué es a través del contexto.
Han de ser modulares, ello permitirá reutilizar el agente y aplicar estrategia "divide y vencerás" , cuando se enfrente a problemas complejos.
Han de comunicarse con otro agentes para poder trabajar en entornos distribuidos de recuperación de información.
Han de ser fiables, los usuarios solo aceptarán a los agentes si éstos son de confianza.

B.5 COLECCIONES DE HERRAMIENTAS DE BÚSQUEDA

Consisten en una recopilación de formularios de diferentes herramientas de búsqueda web. Pueden cubrir servicios de búsqueda especializada o general y también pueden aparecer éstos clasificados por categorías. Su funcionamiento consiste en redireccionar el servicio de búsqueda correspondiente pasándole la consulta a realizar. Ejemplo: http://www.20search.com.

TABLA20. VENTAJAS E INCONVENIENTES DEL USO DE COLECCIONES DE HERRAMIENTAS

Ventajas	Inconvenientes
El usuario tiene a mano una gran cantidad de servicios de búsqueda sin tener que recordar sus direcciones.	La escasa ayuda que ofrecen estas colecciones en cuanto a la sintaxis apropiada para usar en cada uno de los formularios.
El usuario no tiene que cargar la pagina de presentacion de cada una de las herramientas de búsqueda. Esto supone un ahorro de tiempo teniendo en cuenta que los servicios de búsqueda presentan una interfaz bastante cargada de publicidad.	

B.6 PORTALES

Una última posibilidad de búsqueda en Internet es la que ofrecen los portales. Un portal, como su nombre lo indica es una puerta de entrada a Internet, un lugar en el que usuario ve concentrado una serie de servicios y productos que ofrece, de forma que le permite encontrar en él todo lo que necesita sin tener que salir de dicho website. Su objetivo es atraer a los usuarios y que estos estén cuanto más tiempo posible en el recurso, puesto que el tiempo de conexión y la publicidad son sus principales fuentes de ingresos.

(Pérez de Leza, 2000)[57], define un Portal como: “la página Web que agrega contenidos y funcionalidades, organizados de tal manera que facilitan la navegación y proporcionan al usuario un punto de entrada en la Red con un amplio abanico de opciones”.

Hay dos tipos de portales: horizontales y verticales. Los primeros son aquellos que tratan de englobar de todo y para todos los públicos, son cada vez más escasos debido a la

[57] PÉREZ DE LEZA, J. El valor añadido de un portal. Ecomm, 13. 2000. Fuente: <http://eprints.rclis.org/archive/00002785/01/a12comvirtuales.pdf >

limitada capacidad de absorción de usuarios en este sector; mientras que los verticales focalizan recursos hacia la especialización de contenidos a nivel temático o geográfico, brindando mejores y nuevas posibilidades a las demandas informativas de los usuarios.

Sin lugar a dudas, un portal representa pues, una evolución con respecto al concepto de Buscador, ya que, además de búsquedas de información, el portal ofrece servicios de valor añadido, muchos más contenidos y herramientas como correo electrónico, Chat, agenda, mensajes a teléfonos móviles, espacio web gratuito, software gratuito, grupos de discusión, comercio electrónico, etc. El buscador, en definitiva, es una más de las opciones que ofrece el portal, ya que el usuario, además de acudir a buscar, accede a contenidos y puede hacer cosas a través de los servicios de valor añadido mencionados.

TABLA21. CARACTERÍSTICAS DE LOS PORTALES
Los portales suelen diferenciar entre información y servicios.
Las secciones de información suelen ser relativamente dinámicas adaptándose a las nuevas tendencias en cuanto a las preferencias de los usuarios.
No utilizan estructuras complejas con gran cantidad de niveles.
La información se enlaza entre sí. Las páginas incluyen numerosas enlaces con otros contenidos relacionados.
Los portales permiten ofrecer contenidos de calidad y con una actualización permanente.
Los portales incluyen las guías o buscadores de recursos, permitiendo orientar y formar al usuario en la navegación y en la búsqueda de información que requieran, con independencia del ámbito del portal.
Son puntos de entrada en Internet que ofrecen buscadores, selecciones de recursos, correo eléctrónico y otras muchas posiblidades de las que ofrece la red.
En definitiva la reforma de este herramienta esta centrada en la existencia de portales horizontales como guías generales de Internet y la existencia de portales verticales para cubrir las necesidades de información sobre un tema concreto, así como para el uso de servicios específicos.

B.7 CÓMO ACCEDER A LA PORCIÓN DE INTERNET QUE NO ES RASTREADA POR LOS MOTORES DE BÚSQUEDA

En los últimos años se está prestando mucha atención al llamado Web "invisible" o "profundo". Con esta expresión nos referimos al contenido que no es mostrado por las búsquedas que se llevan a cabo en los motores de búsqueda. Normalmente se trata de lo que está almacenado en bases de datos que son accesibles a través del Web pero que no está disponible a través de los motores de búsqueda por su naturaleza dinámica. También son páginas que los motores de búsqueda deciden no indizar, no por razones técnicas sino por que no les interesa hacerlo por una u otra razón Es decir que el contenido es solo

"invisible" para los motores de búsqueda. La razón es que los robots no pueden o no quieren entrar en las bases de datos y extraer su contenido, tal y como hacen con las páginas habituales que son estáticas.

Con anterioridad estas bases de datos no eran muchas pero hoy en día son bastantes y gran parte de la información que se encuentra en el web está almacenada en ellas. El único modo de acceder a esta información accesible a través de Internet es buscar en las propias bases de datos. Su contenido es muy variado desde lo trivial a lo académico pasando por toda la información que resulta actual y es muy dinámica y cambiante. Pero hay que tenerlo muy en cuenta. Por ello han surgido recursos que permiten acceder a estas bases de datos y una vez en ellas realizar la búsqueda que sea adecuada para nuestra necesidad informativa.

Algunos de estos recursos son los siguientes:

- Direct Search, una amplia compilación de enlaces a entornos de búsqueda de variados recursos de investigación en el Web reunida por Gary Price de la George Washington University.
- A Collection of Search Engines, una colección bastante grande y valiosa de bases de datos.
- Search.Com, docenas de bases de datos sobre diferentes materias compiladas por CNET.
- Internets, gran colección de motores de búsqueda específicos.

APENDICE C

PRESTACIONES DE LOS SISTEMAS DE BÚSQUEDA COMO APOYO TANGIBLE EN LA RECUPERACIÓN DE LA INFORMACIÓN EN INTERNET (CASO DE ESTUDIO).

Como ya hemos señalado existen centenares de motores de búsqueda. Elegir uno u otro depende de lo útil que resulte para el usuario. Por ello hay una feroz competencia entre los diferentes motores de búsqueda que les hace ofrecer mejoras técnicas de modo que consigan ser los preferidos por los Internautas o usuarios de Internet. Obviamente un

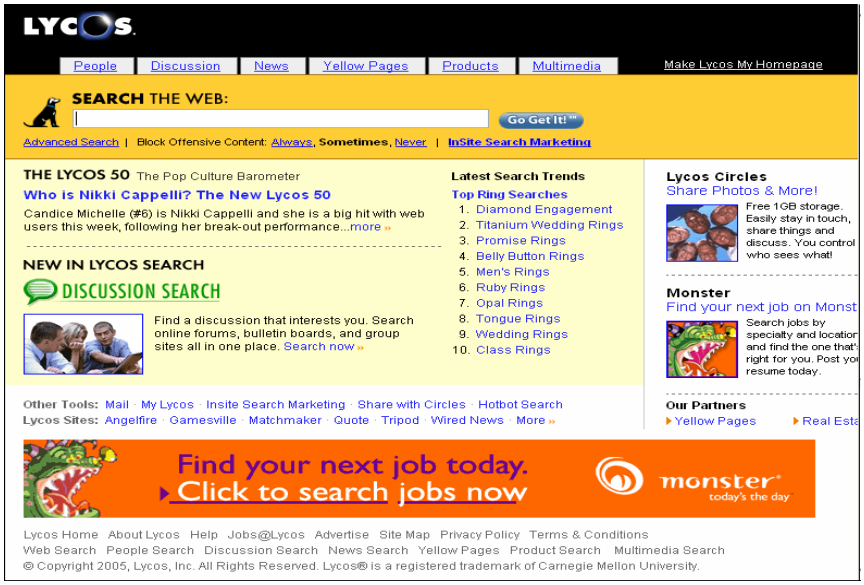
motor de búsqueda es tanto más útil, cuando mejor contesta la necesidad informativa del usuario.

En este apartado, daremos una visión más detallada acerca del entorno de búsqueda que ofrece cada motor que será objeto de estudio para la propuesta de evaluación que se indicará mas adelante. Además, se dará mención acerca de su origen y evolución, funcionamiento que lleva a cabo en la indexación y demás prestaciones en cuanto la utilización de operadores de búsqueda, que permitan establecer con claridad la relación entre los términos de búsqueda a emplear durante su uso.

Cabe mencionar que gran parte de la información fue obtenida por páginas con una gran cantidad de información de última hora acerca de buscadores, metabuscadores, directorios, sus uniones, desapariciones, premios… Además contiene gráficos y estadísticas sobre distintos aspectos de los buscadores. Entre ellos podemos mencionar:
Search Engine Watch (http://www.searchenginewatch.com/).
Search Engine Showdown (http://www.searchengineshowdown.com)
y Pandia Search Central (http://www.pandia.com/)
De igual manera, para el estudio de los buscadores también se consultaron las páginas de cada uno de ellos, así como sus páginas de ayuda.

C.1 MOTOR DE BÚSQUEDA " LYCOS "

El nombre de Lycos proviene de la Araña Lobo, una agresiva araña carnívora que usa su tela para cazar, en honor de las 'arañas' informáticas, programas capaces de recorrer la web automáticamente, recopilando información sobre sus contenidos y volcándola en un depósito donde poder buscarla. Lycos es un Buscador internacional, que entre otros idiomas, tiene un sitio en español. Además posee un Directorio ordenado por categorías. Por ser un robot, Lycos acepta todo tipo de páginas en su base de datos.



Éste motor de búsqueda fue creado como una manera de ayudar a los usuarios a recuperar el control sobre la red. Su objetivo es ofrecer una herramienta simple con una interfaz intuitiva para todo tipo de usuarios, desde expertos hasta novatos.

C.1.1 ORIGEN Y DESCRIPCION

Lycos (http://www.lycos.com, http://www.lycos.es), es uno de los servicios de búsqueda más antiguos de la WWW o W3, surgió en la Carnegie Mellon University de la mano del Dr. Michael Mauldin del Center for Machina Translation, institución dedicada a la investigación y desarrollo de programas de traducción automática. Este centro contaba con una gran base tecnológica de recursos sobre procesamiento del lenguaje natural, algoritmos, técnicas y gran experiencia en este campo.

Como la W3 se ofrecía como una gran base de datos de notable interés para sus proyectos, decidieron aplicar los métodos que utilizaban para el procesamiento del lenguaje natural a los problemas de recuperación y organización de la información. Lycos se dió a conocer en el verano de 1994, cuando la araña diseñada por Jonh Leavitt y Eric Nyberg, originalmente denominada Longlegs, se vinculó al programa de indización desarrollado por Michael Mauldin.

Según su creador, la palabra Lycos proviene de la familia de arácnidos *Lycosidae,* arañas terrestres relativamente grandes que atrapan a su presa persiguiéndolas, en lugar de esperar a que caigan en su telaraña. Estas arañas se caracterizan por su velocidad y por ser especialmente activas por la noche. En opinión de sus diseñadores, Lycos responde a esta descripción.

Pronto se convirtió en uno de los buscadores preferidos por los usuarios, ya que ofrecía la posibilidad de consultar un catálogo relativamente grande mediante palabras del contenido del documento. Cuando Netscape Navigator fue lanzado ampliamente al mercado a finales de 1994, el personal de Netscape Communications Corporation incluyó una página que ofrecía acceso a varias herramientas de búsqueda en Internet. Realizaron una rápida y poco refinada prueba y decidieron que Lycos era el que ofrecía mejores resultados, por lo que optaron por presentarlo el primero de la lista de buscadores. La amplia utilización de este navegador provocó una fama creciente para Lycos, que incluso llegó a sufrir problemas de sobrecarga, pero sirvió para impulsarlo definitivamente como uno de los primeros grandes en la historia de los localizadores de páginas web.

A mediados de 1995, Lycos adquirió Point Communicatios, una compañía reconocida por su colección de críticas acerca de lugares de Internet. Ahora, Point, es llamado 'Top 5%', sigue en funcionamiento como parte del servicio Lycos.

El mayor accionista de Lycos es CMG information Services, proveedor de servicios de mercadeo directo. Cabe destacar, que Lycos fue la primera empresa en Internet en basar su publicidad en CPM, la cual es actualmente un estándar en la industria de Internet.

En Febrero de 1996 se despliega el sitio basado en listado de temas (A2Z directory).

En junio de 1997, Lycos Pro introduce un nuevo algoritmo de búsqueda.

En Abril de 1998, Lycos adquirió la corporación WiseWire la cual es destacada por su software de creación de directorios. Ahora, WiseWire respalda los Directorios Web de Lycos, las cuales son creadas automática y colaborativamente por los usuarios. De igual manera ese mismo año incluye en su base de datos al directorio ODP (Open Directory Project, el directorio de sitios del World Wide Web más grande, organizado en categorías

y totalmente construido de forma manual, por usuarios de Internet. Lycos se hace el motor de búsqueda exclusivo de Planet Oasis (www.lycos.com/software/oasis.html) - un servicio de navegación que dirige a usuarios por el Internet.

En junio de 1999, se une con IntelliSeek para proporcionar acceso a las 10,000 bases de datos de InvisibleWeb.com, que son la parte de lo que es conocido como la web 'invisible'. En septiembre de ese mismo año, anuncia Lycos 50tm, que indica que buscadores son las mas solicitados en la web, además, pone los 50 términos de búsqueda más populares en una lista, surgiendo tendencias y nuevos temas. Seguidamente, lanza 'Lycos Zone', un sitio Web educativo para niños entre 3 a 12 años e incluye el libre acceso a Lycos en cuanto a tecnología de filtración, SearchGuard.

En el 2000 Lycos comenzó a suplir su índice con uno proporcionado por FAST. En el otoño de 2001, Lycos abandonó su propia araña y comenzó a proporcionar sus resultados de la porción exclusivamente de FAST, sin embargo, a pesar de todo esto cambios, éste mantiene su nombre original.

Desde el 1 de abril de 2004[58], Lycos dejó de entregar resultados propios del uso de la base de datos FAST y ahora usa la base de datos Inktomi de Yahoo! .

El spider de Lycos no esta del todo eliminado. Lycos lo conserva para construir colecciones de contenido especial (por ejemplo artículos de noticias), sin embargo, aunque resulta algo confuso entender, sobre qué resultados generales de búsqueda es desarrollada su base de datos, podemos decir, que Lycos proporciona bases de datos diferentes en la sección de Resultados del Web. Ésta puede contener links de bases de datos distintas (resultados contenido de la propia BDs de lycos, resultados de LookSmart: despliga listas comerciales, resultados proporcionados por FAST en pequeña proporción y actualmente el proveedor de los principales resultados es suministrado por Inktomi de Yahoo!).

Su estructura es bastante similar a la del motor Yahoo!: una ventana de búsqueda a través de palabras clave, un índice por temas y una sección de recomendados. Tiene algunas diferencias en la sintaxis empleada para especificar la búsqueda y en la manera de llegar

[58] http://www.searchengineshowdown.com/newsarchive/000769.shtml.

al índice temático, pues este no aparece en la pantalla inicial, como en Yahoo, sino que es necesario hacer clic sobre un botón de enlace (site Map o Directorio) para llegar a él.

Actualmente Lycos es un buscador producido por Lycos Inc. El Centro de Operaciones, que se encuentra ubicado en Waltham, Massachussets, éste centro es quien proporciona guías para encontrar la información sobre el World Wide Web del Internet. La compañía es una red de comunidad y navegación global de Internet dedicada para ayudar a usuarios en línea a localizar, recuperar y manejar la información proporcionando instrumentos de información fáciles de usar. Los ofrecimientos de producto de la compañía incluyen búsqueda de sonidos, compañías en línea, peoplefind, roadmaps, noticias, stockfind, chat, correo electrónico, cityguides, páginas amarillas y guía personal.

Lycos provee a los espectadores de Internet, un destino universal para información, comunicación y servicios de compra en la Web. Los Sitios Web de la compañía se han hecho un medio publicitario extensamente aceptado para las compañías del mundo más prominentes, incluso marcas como Coco-Cola, Disney, General Motors, Hilton, IBM y Visado. La compañía adquirió Metrosplash, el Internet Music Distribution Inc, Quote.com, Gamesville.com y el interés restante al software Valent Corporation Inc en 2000 fiscal. Los ingresos publicitarios explicaron el 66 % de ingresos de 2000 fiscales y comercio electrónico y otros ingresos, el 34 %.

Lycos, se encuentra entre los 10 primeros lugares junto con otros motores de búsqueda, en cuanto posicionamiento frente a los demás analizados, según datos oficiales presentados en mayo del 2004[59].

Según últimas noticias expuestas en este año 2005, Lycos hará uso la tecnología de Ask Jeeves en su motor de búsqueda con la finalidad de facilitar una solución para el motor de búsqueda y dar mayor flexibilidad y capacidad de diferenciar los productos que ofrecen.

C.1.2 FUNCIONAMIENTO Y PRESTACIONES:

Lycos plantea dos formas de búsqueda:

[59] Ref: http://www.submitexpress.com/optimize.html

 La Busqueda Básica:

Para comenzar una búsqueda de Lycos poderosamente simple:

Primero que todo se Localiza la barra de herramientas de Búsqueda en la correspondiente dirección http: // www.lycos.com.

En el formulario presentado (caja de entrada de texto después "Search the WEb", se escribe la palabra o la frase de búsqueda y da click el botón "Ir" (Go Get it!). Después de que usted hace clic el botón "Ir", el software de Lycos descarga toda la información relacionada con su palabra de búsqueda o palabras y arregla que aquellos links aparezcan sobre una Página de Resultados de la Búsqueda. La Página de Resultados de Búsqueda puede ser en sí mismo, un instrumento poderoso para organizar la Web y mas adelante refinar su búsqueda.

 Búsqueda avanzada:

Lycos la Búsqueda Avanzada le ayuda a construir búsquedas más poderosas con instrumentos llamados a operadores Booleanos que influirá en la forma de evaluar la petición. Una vez que se ha escrito los términos de consulta, Lycos le ofrece otras opciones para refinar la búsqueda en caso de que la primera consulta no le haya brindado resultados relevantes.

Todas estas prestaciones de filtrado y demás características de este motor de búsqueda quedan consignadas en la Tabla22. Es importante recordar que para aprovechar las características de búsqueda avanzada, debe hacer click al enlace con el mismo nombre “Advanced Search” y elegir automáticamente en los cuadros de texto, los parámetros de búsqueda a los que se desea usar para filtrar información.

El robot de Lycos indexa especialmente el contenido de la etiqueta TITLE y la muestra como el título del sitio web. Y toma el contenido de la primera descripción que encuentra en el cuerpo del documento para mostrarlo como una descripción de este sitio. El ordenamiento dentro de su Directorio es alfabético y por categorías.

	TABLA22. PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BUSQUEDA (www.lycos.com)
	Información general
Tipo de Servicio (Motor/Directorio):	Motor de búsqueda en la web(proporcionada actualmente por Inktomi) y Directorio Temático interno(lo provee Open directory).
Acceso(libre,comercial):	Libre
Nro de URLs:	2,1 billones de paginas web reportados.
Origen:	Fue anunciado, el 12 de Agosto de 1994, por Michael L. Mauldin, en la Universidad Carnegie Mellon.
Frecuencia de actualizacion de la base de datos(mensual, semanal, quincenal,diaria):	4-12 semanas
Nº de paginas que recoge por dia	Indexa 6 a 10 millones.
Tiempo que dura en dar las paginas web registradas:	2-3 semanas.
Versión analizada:	Ingles. Pero existe la version en castellano.
Portal:	Si. Contenido de portal extenso.
Obedece el protocolo de exclusión(Sirven para indicarle a los robots que la página que los contiene no debe ser indexada y/o que los enlaces de tal página no deben ser seguidos).	robot.txt
	Recolección
Robot:	Spider
Método(Humano/Automático):	Automático
Primero en profundidad:	Si aplica
Primero en Anchura:	No aplica
Tipo de cobertura (www, gopher,WAIS, ftp,telnet,IRC, UseNet News, productos multimedia, etc.,):	Por defecto busca dentro del WWW, y de servidores Gopher y FTP.
Cobertura geográfica:	cobertura internacional con nodos locales en: Alemania, Bélgica, España, Estados Unidos, Francia, Holanda, Italia, Reino Unido, Suecia y Suiza.
Objeto de Cobertura (general o contenido especializado):	General.
Frecuencia de actualizacion para visitar los mismo sitios/documentos:	El equipo de Lycos indexa -las 24 horas del día- información cuidadosamente seleccionada de la Web y la incorpora a su directorio.
Otras bases de datos:	Anuncios (links Patrocinados): Overture y AdBuyer propietario de Lycos; Resultados de Web: Contenido de Red de Lycos, 10 Anuncios LookSmart, Luego base de datos de FAST (AlltheWeb); Noticias: Bases de datos Lycos and FAST News; Directorio Open Directory (sólo accesible desde el link, no de resultados de búsqueda regulares); Imágenes: Disponible bajo "Multimedia". Este es una base de datos de FAST; De audio y de Vídeo: Disponible bajo "Multimedia". Este es una base de datos de FAST.

LYCOS Search	**PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.lycos.com)**
Indexación	
Que indexa (titulos, cabeceras, información de cabecera, Links(URLs),etiquetas HTML, resumenes, texto completo, etc.,):	Este buscador crea un índice con el título, cabecera del documento, de las 100 palabras más utilizadas en el documento y las primeras 20 líneas.
Que no indexa:	Spam, URLs con caracteres especiales(&, ?, =, %, $). No indexa palabras vacías(Stopwords, marcos, mapas de imagen, etiquetas de meta (excepto robot etiqueta de meta), comentarios.
Sistema de Recuperación	
Herramienta de Búsqueda:	Motor de búsqueda.
Tipo de Recuperación (Booleana, Vectorial, Probabilística, otro):	Probabilística.
Estructuras de consulta y operaciones soportadas:	Consignada en la Tabla 23.
Criterio de Búsqueda:	Número de veces que la palabra aparece en el documento, en qué campos aparece (título, cabecera o texto), Número de veces que este documento está referenciado en otros.
Sistema búsca *por default/por campos seleccionables(*URLs, titulos,Resumen, Texto completo, otros.,)	Por defecto en el texto completo.En caso del directorio, la búsqueda es por niveles
Mejora en la Busqueda (reducción de resultados, reformulación, vocabulario controlado, mediante ejemplo,otros.,):	Reformulación o refinamiento(Search within these results).
Incluye textos alternativos(inf. similar):	No aplica
Ordenamiento:	los Sitios son clasificados por orden de relevancia en los resultados. No hay ninguna opcion para clasificar por orden alfabético o por fecha.
Despliegue de resultados: (Título, descripción,URL, tamaño,fecha de alta, Nº total de saltos, correspondencia de términos,orden de relevancia,valor de relevancia,Nº de saltos por pagina, formato variable, detecta novedades, permite traducir,otros.,):	Permite ver título con el enlace, la URL, las primeras palabras a modo de resumen, correspondencia de términos, Nro total de saltos, la salida se compone por defecto de 10 resultados por pagina, el usuario puede elegir los criterios de relevancia en el orden de presentación de los resultados. No tiene disponible información de la fecha, el tamaño del archivo.

PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.lycos.com)

Interfaz de Usuario	
Descripción de la Interfaz	Menus estructurados, desplegables
Claridad de la interfaz y de la pagina de búsqueda:	completamente claro.
Ofrece ayuda en cuanto a la documentacion de búsqueda:	Las ayudas detallan la forma de realizar las búsquedas, los operadores a utilizar y los criterios de relevancia a aplicar en la presentación de resultados.
Busqueda Simple/Avanzada	Ambas(formato:formulario)
Establece preferencias	Si permite establecer preferencias y guardar la configuración.
Idiomas Interfaz:	47 idiomas.
Servicios:	noticias, chat, e-mail gratuito y reporte meteorológico, así como búsqueda de software seguro y diversas opciones enfocadas a negocios y entretenimiento como guías vacacionales, empleos y noticias de primera plana. Ofrece opciones de búsqueda sumamente especializadas y permite buscar ilustraciones y sonidos.
URL Pagina de Ayuda:	http://help.lycos.com/search/search_1_help.asp
URL Pagina deTraduccion:	http://translation.lycos.com
Ranking:	Trabaja con ranking basado en popularidad.

TABLA23. ESTRUCTURAS DE CONSULTA Y OPERACIONES SOPORTADAS

Operador default:	AND
Operador de Existencia: (exclusión/inclusion)	-/+
Operadores Booleanos:	AND(Y), OR(O), NOT(NO)
Operadores de Proximidad:	No aplica
Búsqueda por campos:	domain, link, title, url.
Operadores de exactitud:	Frase literal o uso comillas " ".
Lenguaje Natural:	Si aplica.
Truncamiento:	No. Ningún comodín es admitido.
Agrupacion de términos y operadores:	No aplica
Sencible mayúsculas/minúsculas:	No aplica
Reconoce sinónimos:	Si aplica
Caso sensitivo a acentos, puntuación:	Si aplica.
Filtro por idioma:	Si Aplica.
Filtro por formato:	No aplica.
Filtro URL/Site:	Si Aplica. Multiples dominios/sitios pueden ser especificados separados mediante comas.
Filtro por Region:	Limita los resultados a un específico continente o país.
Filtro por numero de resultados de páginas:	Si Aplica. El numero de resultados que despliega 10, sin embargo no se dispone de la habilidad de ampliar esta lista de resultados.
Filtro por fecha:	Si Aplica. Limita los resultados de paginas publicadas dentro de un periodo de tiempo especificado(un rango de tiempo o especificando a partir de una fecha determinada).
Permite bloqueo de contenido ofensivo:	Si Aplica. (tiene opciones de siempre, algunas veces o nunca)

C.2 MOTOR DE BÚSQUEDA " GOOGLE "

GOOGLE es un Buscador con un robot potentísimo que indexa páginas en todo Internet, sin diferenciaciones, actualmente trabaja en diferentes idiomas: Alemán, Chino Coreano, Danés, Español, Finlandés, Francés, Holandés, Inglés Italiano, Japonés, Noruego, Portugués, Sueco y abarca toda clase de temas de nuestro interés.

Google fue fundado en septiembre de 1998 por Larry Page y Sergey Brin, dos estudiantes de doctorado de Stanford.

El nombre "Google" surgió como un juego de palabras que proviene de la palabra "googol", término acuñado por Milton Sirotta para referirse a un número representado por un 1 seguido de 100 ceros.

Un googol es un número muy largo. Por tanto, éste motor de búsqueda fue creado como una manera de reflejar la misión de organizar la inmensidad infinita que aparenta la World Wide y de esta forma, transformar la manera en que el mundo busca y archiva información.

C.2.1 ORIGEN Y DESCRIPCION

Google (http://www.google.com, http://www.google.com.co), empezó en una primavera de 1995 cuando dos jóvenes universitarios, Sergey Brin, de 23 años, con experiencia en diseño web y titulado en Ingeniería Electrónica, y Larry Page, de 24, experto en tratamiento de datos y licenciado en Informática y Ciencias Matemáticas, que acabarían siendo cofundadores y actuales presidente y CEO de Google, se conocieron en un evento organizado por la Universidad de Standford para los aspirantes a su prestigioso Doctorado en Informática. Allí discutieron arduamente sobre cada tema hablando, sus sólidas

opiniones y divergentes puntos de vista encontraron un objetivo común, conseguir información relevante a partir de una importante cantidad de datos.

Es en otoño de ese mismo año, cuando estos dos estudiantes empiezan a desarrollar un algoritmo para la búsqueda de datos, que utilizarían para el proyecto de "Biblioteca Digital" de la Universidad de Standford (Digital Library Project). En ese momento, comenzó lo que más tarde sería llamado por Larry Page como PageRank(software dedicado a posicionar las páginas web entre los resultados), en 1997.

En enero de 1996, Larry y Sergey empezaron a desarrollar un motor de búsqueda llamado BackRub, nombrado así por su capacidad única de analizar los "enlaces entrantes" (enlaces que provienen de otras páginas) de una página web. Este buscador corría sobre varias máquinas Sun e Intel y su base de datos principal era guardada por un Sun Ultra II con 28Gb de disco y estaba implementado en Java y Python.

Larry empezó a trabajar en la forma de conseguir un entorno para los servidores que funcionara con PCs de gama baja y que no necesitará de potentes máquinas para funcionar. Un año después, en 1997, para ser exactos, la tecnología utilizada por BackRub para analizar los links empezaba a ser conocida en todo el campus, obteniendo una gran reputación. Era la base sobre la que se construiría Google.

Durante los primeros meses de 1998, Larry y Sergey continuaron trabajando para perfeccionar la tecnología de búsqueda que habían desarrollado. Utilizaron sus dormitorios como centro de datos y oficinas. Con esta infraestructura iniciaron la búsqueda de inversores que les ayudaran a financiar su novedosa tecnología, superior a todas las existentes hasta la fecha. A pesar de la fiebre de las “puntocom” en ese momento, Larry y Sergey tenían poco interés en montar una empresa propia cuyo negocio fuera el motor de búsqueda que habían desarrollado.

Entre estos posibles inversores, se encontraba David Filo, amigo de ambos y uno de los fundadores de Yahoo!. Filo les animó a que ellos mismos desarrollaran el proyecto, creando una empresa basada en el buscador cuando estuviera completamente desarrollado. Aunque el potencial que tenía era enorme, se encontraron con la negativa de

muchos portales, que consideraban el hecho de tener un buen buscador como algo secundario en sus objetivos.

Así pues, tomaron la decisión de poner en marcha el proyecto y buscar capital para abandonar las habitaciones y acabar de pagar todo el material que habían comprado para los servidores. Hicieron un plan de empresa y fueron en busca de inversores. Su primera visita fue al amigo de un miembro de la facultad.

Andy Bechtolsheim, uno de los fundadores de Sun Microsystems, enseguida vio que Google tenía un potencial enorme. Sólo pudieron mostrarle una pequeña demo pero fue suficiente para que inmediatamente les diera un cheque por valor de 100.000 $, a nombre de Google Inc. Pero surgió un pequeño problema: no existía aún una empresa llamada Google Inc., por lo tanto no podían cobrar ni ingresar el cheque. Un par de semanas más tarde decidieron buscar nuevos inversores entre familiares, amigos y conocidos para poner en marcha la compañía.

El 7 de septiembre de 1998, Google Inc. ya disponía de oficinas propias en Menlo Park, California, todo un lujo comparado con la situación en la que habían estado hasta entonces. Google.com, todavía se encontraba en fase beta, tenía unas 10,000 búsquedas cada día. La prensa empezaba a hablar del nuevo buscador y de su excelente funcionamiento.

En 1999 consiguieron 25 millones de dólares de dos importantes inversores: Sequoia Capital y Kleiner Perkins Caufield & Buyers. Las modestas oficinas ya eran pequeñas para todos los directivos y trabajadores de Google, así que se trasladaron a Googleplex, la actual sede central de Google en Mountain View, California.

Nuevos e importantes clientes iban llegando, como por ejemplo AOL/Netscape que escogió a Google como su servicio de buscador, haciendo que superase los 3 millones de búsquedas al día. Lo que empezó siendo un proyecto universitario ya era una gran empresa con un crecimiento impresionante. El 21 de septiembre de 1999 desapareció definitivamente de Google.com la etiqueta que lo identificaba como una versión beta.

En el 2000 Google inicia su consolidación como nuevo motor de búsqueda de Yahoo!, reemplazando su tradicional motor de búsqueda Inktomi. Es a partir de junio de este mismo año que Google se gana a pulso el título de sitio más útil de Internet, convirtiéndose en el índice más exhaustivo de la Red.

A principios del 2000, el buscador Google incluye entre las opciones de búsqueda, los números de teléfonos y las direcciones de los ciudadanos estadounidenses, utilizando el llamado servicio Google PhoneBook donde se introduce los datos de la persona como nombre y apellido, código postal o la ciudad y estado donde vive y la página de resultados destaca, en la parte superior, los números y direcciones encontrados, juntamente con mapas de localización geográfica. A diferencia de otras webs que ofrecen este tipo de servicios de directorios de datos personales, Google ofrece la posibilidad de borrarse del banco de datos de forma muy sencilla.

En agosto de este mismo año, NTT DoCoMo incorporaró en su servicio de telefonía móvil 'i-mode' el potente buscador Google.

En el 2002, Google obtiene la patente del metodo de búsqueda. La Oficina de Patentes de Estados Unidos otorga a Google la patente de su método para determinar la relevancia de recursos en Internet mediante su motor de búsqueda. La patente fue tramitada en enero de 2001, y revela la metodología empleada para acceder a documentos web y conducir a los usuarios a los resultados más adecuados a sus parámetros de búsqueda.

Seguidamente, compra la primera empresa de weblogs mundial (Según wikipedia: "los weblogs son sitios web donde se recopilan cronológicamente mensajes de uno o varios autores, sobre una temática o a modo de diario personal"). De ahí, abre sus negocios a las páginas personales (o blogs). Seguidamente lanza una sección dedicada a la búsqueda de productos en venta de diversos supermercados y tiendas on-line.

En junio Llega AdSense, un nuevo sistema de publicidad, basado en insertar publicidad relacionada con la relevancia de palabras clave de cada página web. De esta manera, los operadores de pequeños sitios web puedrian insertar este tipo de anuncios, y cobrar cada

vez que alguien hace 'click' sobre ellos. Este servicio se ofrece a otros Website que cobran una comisión de Google del tráfico dirigido.

Posteriormente aparece en pruebas Google News Alerts, un nuevo servicio que nos permitirá estar siempre al tanto de lo que sucede sobre un tema especifico. A través del correo electrónico y con una gran flexibilidad en la elección de la frecuencia (cada día, o cada vez que surja la noticia), llegará a nuestro buzón de correo electrónico las correspondientes alertas sobre la palabra seleccionada. También en este mes abre sus puertas Google India. Seguidamente, en agosto nace Googlenews, el nuevo servicio rastrea noticias de miles de sitios y periódicos mostrándolos automáticamente en el portal, todo ello sin ningún tipo de intervención humana.

A finales de este año, se pone en funcionamiento el dominio "google.es", aunque por el momento redirige el tráfico hacía la página en inglés. De igual manera, aparece la nueva versión v2.0.102 de la barra de Google. La instalación es ahora más completa, y posibilitando elegir si se quiere que Google sea el buscador por defecto de MS IExplorer.

En el 2003 Google llega a España. A través de 'google.es' se pueden realizar búsquedas solamente de páginas web ubicadas en España. Y también se puede acceder a las versiones de Google en tres idiomas oficiales en este país: catalán, gallego y euskera.

En ese mismo año, Google lanza un servicio de búsqueda de noticias en castellano: 'news.google.es', que se basa en tecnologías de clusterización y análisis para agrupar en diversas categorías ('Internacional', 'España', 'Economía', 'Tecnología', 'Deportes', 'Espectáculos' y 'Salud') titulares procedentes de distintas publicaciones en castellano presentes en la Red.

En agosto del 2004, El poder de Google en el sector de internet se atribuye no sólo a que cuenta con uno de los buscadores más potentes del mercado, sino a su capacidad de innovación y ofrecer servicios de valor añadido. La compañía ha sacado ventaja a sus competidores con el lanzamiento de Google Desktop Search, una herramienta de búsqueda de información dentro del ordenador; el Gmail Google, correo electrónico de alta capacidad que se ofrece gratis a los usuarios y, en Estados Unidos, un acuerdo con las principales bibliotecas para poner en la red un gran catálogo de libros.

En este año 2005, el motor de búsqueda en Internet Google anunció el lanzamiento en versión experimental en inglés, de una nueva herramienta que permite obtener en línea las imágenes de video o la programación televisiva relacionada con una palabra clave. Además hace el lanzamiento para presentar una nueva versión, la 3.0.120.7 beta, de Google Toolbar, la cual incorpora entre sus principales novedades una utilidad para corregir faltas de ortografía en formularios web. Por otro lado también funcionalidades de traducción, facilitando la traducción de textos del inglés entre otras características.[60]

C.2.2 FUNCIONAMIENTO Y PRESTACIONES

Google s un buscador con robot y su funcionamiento difiere bastante de los otros motores de búsqueda. Su sistema de selección de resultados otorga una relevancia mayúscula a la popularidad de un sitio.

Su base de datos está desarrollada de forma que, ante todo, le sea asignado a cada sitio un índice de popularidad. Sólo después, ante una consulta determinada, utiliza la comparación textual y lista las coincidencias conservando el ordenamiento que cada una de ellas tiene en la escala del buscador.

Google, el algoritmo que utiliza para mostrar, ordenar y refinar las búsquedas, se encuentran basado en siete factores:

- Número de páginas que enlazan a cada resultado. Cuantos más enlaces de terceros sitios cuente una página, más arriba dentro de los resultados estará. Se basa en la presunción de que si muchas páginas enlazan a una en concreto, es que ésta será de calidad. Esto forma el ranking de página o como ellos lo denominan "PageRank".

- Palabras empleadas en los enlaces. Se refiere a que los texlinks de los sitios enlazantes coincidan con la palabra buscada. Por ejemplo si otra página enlaza a DiarioRed.com con

[60] Para estar al corriente de todas las nuevas tendencias e innovaciones del equipo de Google se puede visitar el siguiente enlace http://labs.google.com/.

el enlace: Noticias de Internet, será un punto a favor para estar más arriba bajo esa determinada búsqueda.

- Peso de las páginas enlazantes. A su vez, valora estar enlazado por una tercera página con más puntos positivos, es decir mejor situada en el ranking. En consecuencia, no es igual un enlace de Yahoo que de una página personal.

- Que en la url o dirección aparezca las palabras claves buscadas.

- Estar o no listado en el directorio Dmoz. Google utiliza Dmoz, un directorio construido por el esfuerzo de editores voluntarios y sin ánimo de lucro, como base de información ampliada.

- Palabras en el título, y secundariamente, en las etiquetas o metatags.

- Que las palabras buscadas aparezcan en el contenido de la página en relación a su densidad. Si la página esta formada por mil palabras y las palabras claves aparecen una sola vez tendrá menos éxito que si aparece 2 veces entre cien palabras de contenido.

Ninguno de estos factores son por si solos garantía de posicionamiento, sino que es necesaria la combinación positiva de varios para alcanzar un buen lugar. Por otro lado, Google hace uso de dos métodos de búsqueda el básico y avanzado.

La Búsqueda Básica:

En esta búsqueda, se introduce una consulta (palabras descriptivas de su búsqueda) y se da clic en el botón Búsqueda en Google para obtener su lista de resultados pertinentes.

Google sólo busca páginas que coinciden exactamente con sus términos de búsqueda, de modo que puede intentar usar distintas versiones de sus términos de búsqueda. De igual forma el cambiar la frase de consulta en distinto orden muy seguramente devuelve distintos conjuntos de resultados.

Google agrega automáticamente "and" entre las palabras que escriba, de modo que devuelve sólo las páginas que incluyan todos sus términos de búsqueda. Para restringir una búsqueda aún más, sólo incluya más términos. Google también prefiere las páginas

en las cuales los términos de consulta relacionados están cerca uno del otro, es decir Google respeta la ubicación de sus términos de búsqueda dentro de una página.

No sólo los resultados de Google contienen todos sus términos de búsqueda, sino que éste también analiza la proximidad de esos términos dentro de una página. A diferencia de muchos otros motores de búsqueda, Google da prioridad a los resultados según la proximidad de los términos de búsqueda.

Por otro lado, algunas veces, Google realiza una búsqueda que está en el área correcta, pero entrega demasiados resultados. Para reducir el número de resultados de la búsqueda, puede que desee realizar una nueva búsqueda que sólo considere las direcciones URL devueltas por su primera consulta de búsqueda. Esto con frecuencia se denomina "limitar una búsqueda" o "buscar en los resultados de la búsqueda actual."

Google facilita este proceso. Puesto que Google sólo devuelve páginas Web que contengan todas las palabras de su consulta, lo único que debe hacer es agregar más palabras de consulta a los términos que ya ha escrito. Esta nueva consulta devolverá un subconjunto específico de las páginas devueltas para su consulta original "demasiado amplia".

Búsqueda avanzada:

Google, ofrece esta página para permitir refinar las búsquedas es bastante sencilla e intuitiva. Entre las características en la cual se encuentra estructurada esta página tenemos:

Frase exacta. Escogencia del idioma, tipo de archivo, fechas de la página e incluso permite buscar páginas similares, entre otras opciones; Permite también buscar páginas que están enlazadas a otras.

Es importante hacer notar que los resultados principales provistos por Google son provistos por su propia base de datos y algunos resultados secundarios de Open Directory.

Para ampliar aún mas las prestaciones que este poderoso motor de búsqueda ofrece se presenta a continuación la tabla24.

Google Búsqueda	**TABLA24. PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BUSQUEDA (www.google.com)**
Información general	
Tipo de Servicio (Motor/Directorio):	Motor de búsqueda pero tambien directorio.
Acceso(libre,comercial):	Libre
Nro de URLs:	8 billones de paginas web reportados.
Origen:	Google se comenzó a gestar en 1996 en la Universidad de Stanford (Estados Unidos), por los entonces estudiantes Sergey Brin y Larry Page. En 1998, se creó la empresa Google Inc.
Frecuencia de actualizacion de la base de datos(mensual, semanal, quincenal,diaria):	mensual.
Nº de paginas que recoge por dia	No mencionado.
Tiempo que dura en dar las paginas web registradas:	2-6 semanas.
Versión analizada:	Español. Pero existe la version en Inglés.
Portal:	No.
Obedece el protocolo de exclusión(Sirven para indicarle a los robots que la página que los contiene no debe ser indexada y/o que los enlaces de tal página no deben ser seguidos).	Se puede hacer uso del estandar robot.txt en el servidor para para indicarle a Google que explore algunas o ninguna de las partes de algun sitio específico.
Recolección	
Robot:	Spider(Robot explorador llamado Googlebot).
Método(Humano/Automático):	Automático
Primero en profundidad:	No aplica
Primero en Anchura:	Si aplica
Tipo de cobertura (www, gopher,WAIS, ftp,telnet,IRC, UseNet News, productos multimedia, etc.,):	Aparte la búsqueda por defecto en páginas web de Internet, Google ofrece otros tipos de búsquedas como: buscador de imágenes, newsgroups, buscador de noticias, buscador de información de productos online(Froogle), Buscador de productos, dentro de los catálogos de venta por correo de cientos de empresas...etc.,
Cobertura geográfica:	cobertura Internacional.
Objeto de Cobertura (general o contenido especializado):	General.
Frecuencia de actualizacion para visitar los mismos sitios/documentos:	diariamente.
Otras bases de datos:	Utiliza la tenología de Google para buscar dentro de las categorías del Open Directory (DMOZ).

Google Búsqueda	PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BUSQUEDA (www.google.com)
Indexación	
Que indexa (titulos, cabeceras, información de cabecera, Links(URLs),etiquetas HTML, resumenes, texto completo, etc.,):	Google indexa automáticamente todas las paginas que encuentra y lo hace teniendo en cuenta el texto completo, los metatags no son tenidos encuenta(también indexa las imágenes -- bueno los nombres de las imágenes – de la pagina web para sus búsquedas por imágenes). Por tanto, la importancia que la pagina que indexa, trate temas muy concretos o tenga combinaciones de palabras poco
Que no indexa:	Spam, URLs con caracteres especiales(&, ?, =, %, $). No indexa palabras vacías(Stopwords, marcos, mapas de imagen, etiquetas de meta (excepto robot etiqueta de meta), comentarios.
Sistema de Recuperación	
Herramienta de Búsqueda:	Motor de búsqueda.
Tipo de Recuperación (Booleana, Vectorial, Probabilística, otro):	Probabilística.
Estructuras de consulta y operaciones soportadas:	Consignada en la tabla 25.
Criterio de Búsqueda:	Número de veces que la palabra aparece en el documento, en qué campos aparece (título, texto), Número de veces que este documento está referenciado en otros.
Sistema busca *por default/por campos seleccionables*(URLs, titulos,Resumen, Texto completo, otros.,)	Por defecto en el texto completo.
Mejora en la Busqueda (reducción de resultados, reformulación, vocabulario controlado, mediante ejemplo,otros.,):	No permite refinar los resultados de la búsqueda.
Incluye textos alternativos(inf. similar):	Si aplica
Ordenamiento:	Google presenta una tecnología desarrollada por PageRankTM que aprovecha la estructura del web, donde cada página puede vincularse con otra, de manera instantánea, directa y sin terceros. De cierta forma, esta estructura, soportada en el vínculo, elimina la jerarquía y permite que la información transite sin dificultades por la red. Google analiza cada vínculo conectado a otras páginas y permite que la naturaleza vasta y abierta de Internet nos provea las búsquedas más relevantes.Es decir, Una vez que haya otros vínculos apuntando a su página, será indexada.
Despliegue de resultados: (Título, descripción,URL, tamaño,fecha de alta, Nº total de saltos, correspondencia de términos,orden de relevancia,valor de relevancia,Nº de saltos por pagina, formato variable, detecta novedades, permite traducir,otros.,):	Permite ver título con el enlace, la URL, las primeras palabras a modo de resumen, correspondencia de términos, Nro total de saltos, la salida se compone por defecto de 10 resultados por pagina, el usuario puede elegir los criterios de relevancia en el orden de presentación de los resultados. Ademas cuenta con información de la fecha, el tamaño del archivo, titulares de noticias, traducción de páginas web, archivos PDF, vínculos en caché, vinculo para páginas similares. Restricción de dominios, Voy a tener suerte.

PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BUSQUEDA (www.google.com)

Interfaz de Usuario	
Descripción de la Interfaz	Menus estructurados, desplegables.
Claridad de la interfaz y de la pagina de búsqueda:	Interfaz sencilla e intuitiva.
Ofrece ayuda en cuanto a la documentacion de búsqueda:	La ayuda que brinda es muy completa y detalla la forma de realizar las búsquedas, los operadores admitidos y los criterios de relevancia a aplicar en la presentación de resultados y mucho más.
Busqueda Simple/Avanzada	Ambas(formato:formulario)
Establece preferencias	Si permite establecer preferencias y guardar la configuración.
Idiomas Interfaz:	35 idiomas.
Servicios:	Buscador de noticias, servicio de correo electrónico(Gmail), ofrece la posibilidad de la Integración de una barra de búsqueda de Google dentro del navegador web (Google Toolbar, disponible para MS IExplorer sobre MS Windows y en linux se dispone de Googlebar), ofrece serivio de traductor, entre otros.
URL Pagina de Ayuda:	http://www.google.com/intl/es/help/faq_images.html
URL Pagina deTraduccion:	http://www.google.com.co/language_tools?q=salud&hl=es&lr=
Ranking:	Trabaja con ranking basado en popularidad.Considera los enlaces que un sitio tiene dentro de la web y "rankea" los resultados tomando en cuenta que tan "citados" son, de esta forma determina como medida para evaluar su calidad informativa.

TABLA25. ESTRUCTURAS DE CONSULTA Y OPERACIONES SOPORTADAS

Google

Búsqueda

Operador default:	realiza consulta automática con AND
Operador de Existencia: (exclusión/inclusion)	-/+
Operadores Booleanos:	AND(Y), OR(O), no admite el Not, And Not.
Operadores de Proximidad:	No aplica
Búsqueda por campos:	link:, allinurl:, inurl: , intitle:, allintitle:, site:
Operadores de exactitud:	Si aplica. Frase literal o uso comillas " ".
Lenguaje Natural:	Aplica
Truncamiento:	No. *, ?, %: como comodín, no es admitido.
Agrupacion de términos y operadores:	No aplica
Sencible mayúsculas/minúsculas:	No aplica
Reconoce sinónimos:	No aplica. Google busca las palabras del modo en que fueron escritas, conviene incluir en la misma búsqueda o en otra los sinónimos o plurales.
Caso sensitivo a acentos, puntuación:	No aplica.
Filtro por idioma:	Si Aplica
Filtro por formato:	Si Aplica: permite restringir la búsqueda a un tipo de archivo o en su defecto permitir cualquiera(entre los formatos que maneja: .pdf, .xls, .ppt, .doc, .ps, .rtf)
Filtro por Ubicación:	Si Aplica. Permite producir resultados del dominio o sitio Web.
Filtro por numero de resultados de páginas:	Si Aplica
Filtro por fecha:	Si Aplica. Permite filtrado por intervalo de tiempo.
Creador de consultas:	Si aplica. Puede crear fácilmente potentes consultas introduciendo palabras en una combinación de estos cuadros de texto(todas las palabras, frase exacta, cualquier de estas palabras, ninguna de estas palabras).
Permite bloqueo de contenido ofensivo:	No aplica.

Por otro lado, Google ofrece interesantes opciones, como la calculadora (prueba a introducir fórmulas matemáticas en el buscador); una sinopsis pertinente de cada devolución. En vez de resúmenes de páginas Web que nunca cambian, Google extrae el texto que coincide con su consulta - con sus términos de búsqueda resaltados - justo en los resultados de búsqueda. Esta personalización le ahorra el tiempo y la frustración de cargar una página Web inútil; ¡Google puede hacerlo sentir afortunado!

Google se distingue por producir primero el resultado correcto para consultas comunes como nombres de compañías. Han instalado un botón denominado "Me siento afortunadoTM", que lo lleva directamente al sitio Web del primer resultado de búsqueda. La característica "Me siento afortunado" de Google tiene el objetivo de llevarlo rápidamente a una información útil.

Google almacena páginas Web en memoria caché, con el fin de recuperarlas para los usuarios como una copia de seguridad, en caso de que el servidor de la página falle temporalmente. Si el servidor no está disponible, la memoria caché de Google de la página que usted necesita puede ser una salvación. Con frecuencia, este material en caché puede ser mucho más rápido que seguir el vínculo normal, aunque la información que usted reciba puede estar menos actualizada. El Directorio Google otra opción de gran utilidad permite realizar búsquedas restringiendo el significado de la palabra al área de conocimiento en el que se realiza la búsqueda.

C.3 MOTOR DE BÚSQUEDA " YAHOO! "

Como muchos otros aspectos de la era digital, Yahoo! comenzó siendo una idea, creció hasta convertirse en una afición, y acabó siendo una pasión y negocio a tiempo completo. Ese negocio trata de hacer de Internet un lugar más agradable en el que estar, que facilita la búsqueda de información y la realización de actividades en Internet. Ha tenido tanto éxito que en la actualidad millones de personas utilizan Yahoo! regularmente para abrirse camino por la Red.

Yahoo! (http://www..yahoo.com, http://es.yahoo.com, su versión en castellano) continua siendo el directorio general internacional más conocido y utilizado por los usuarios. Se caracteriza por recoger información ubicada en servidores de todo el mundo que cubren una vasta extensión de materias.

Yahoo, propietaria del portal más popular de la Internet, ha optado por una estrategia más multidimensional que pretende darle a sus usuarios un paquete completo: entretenimiento, herramientas para el trabajo y para la investigación. Aunque Yahoo! incorpora un buscador en el que se introducen las palabras clave, lo que ha hecho indiscutiblemente famoso este servidor es el directorio.

C.3.1. ORIGEN Y DESCRIPCION

David Filo y el Dr. Jerry Yang, los fundadores de Yahoo!, eran estudiantes de doctorado de Ingeniería Eléctrica en la Universidad de Stanford cuando decidieron iniciar una recopilación de sus intereses en Internet creando la guía Yahoo! a principios de 1994.

Durante 1994 convirtieron Yahoo! en una base de datos personalizada diseñada para cubrir las necesidades de miles de usuarios que comenzaron a utilizar el servicio a través de la muy limitada comunidad de Internet. Desarrollaron un software personalizado para localizar, identificar y editar de forma eficaz el material almacenado en Internet.

Según David A. Kaplan autor de varios libros de Sillicon Valley, la principal atracción del directorio en expansión de Yang y Filo consistía en que nadie más había compilado otro. El secreto de su posterior éxito, como tan a menudo lo ha sido en Sillicon Valley, residió

en haberlo compilado en el momento preciso. Un par de años antes, simplemente no había muchas cosas en la Web como para tener que clasificarlas, un par de años después, seguramente Microsoft hubiera querido intervenir. Yang y Filo se iniciaron en el momento justo cuando el navegador Mosaic comenzaba a despegar.

Varios fueron los nombres que se dieron inicialmente a este Directorio temático que hoy en día también tiene funcionalidad de motor de búsqueda: "Vía rápida de Jerry a Mosaic" se convirtió en un nombre tan conocido como el de Wal-Mart y sus sustitutos siguientes que no fueron mucho mejores: "Guía de Jerry Yang a la WWW" y "Guía de Jerry y Dave a la World Wide Web". Yang y Filo reemplazaron todas ellas con algo más adecuado para su directorio: Yahoo! sin duda ésta palabra debía tener algún significado. De modo que Yang y Filo salieron con el Yet another Hierarchical Officius Oracle, una especie de parodia informática. "Yet another" formaba parte de la jerga de los programadores de software y el Hierarchical venía por el orden de su clasificación, pero el nombre completo era la práctica de galimatías, y Yang y Filo siempre siguieron adelante con Yahoo!.

Mientras el tiempo transcurría, Jerry y David pronto se dieron cuenta de que no eran los únicos que estaban interesados en un sitio donde se pudiera encontrar una base de datos con las páginas más útiles e interesantes. Cientos de personas accedían a esa información, incluso fuera de Stanford. Empezó a correr la voz y rápidamente consiguieron un número significante de visitas para lo que entonces era la comunidad de internet. A finales de 1994, se celebraba el millón de visitas y los casi 100.000 visitantes únicos. La gran cantidad de tráfico (la red informática de Stanford sufría sus consecuencias) y la entusiasta acogida que tuvo Yahoo! hizo ver a sus fundadores que tenían entre manos un negocio con un enorme potencial.

En marzo de 1995 se constituyó como empresa y empezaron a buscar socios capitalistas entre los inversores de Silicon Valley y es apartir de abril de 1995, que se fundó Yahoo! con un capital inicial de casi 2 millones de dólares aportado por Sequoia Capital, una reconocida empresa que había invertido en Apple Computer, Atari, Oracle y Cisco Systems que se da inicio a este importante proyecto.

Meses mas tarde fue Marc Andreessen, co-fundador de Netscape Communications en Mountain View, California (que desarrolló dos de los navegadores más populares, Mosaic y el mencionado Netscape) quién invitó a Filo y a Yang a traspasar sus ficheros a unos ordenadores de mayor tamaño alojados en Netscape. Como resultado, la red de ordenadores de Stanford volvió a la normalidad, después del volumen de tráfico que recibía su red, y ambas partes salieron beneficiadas.

A finales de 1995 consiguieron nuevos inversores como Reuters Ltd. y Softbank. En abril de 1996, cuando Yahoo! contaba con 49 empleados, empezó a cotizar en Bolsa. Es a partir del 12 este mes, que Yahoo! Se listó como empresa publica.

Conforme la popularidad de Yahoo! aumentaba, también la gama de servicios crecía. Esto convirtió a Yahoo! en "el único lugar a donde alguien tiene que ir para encontrar cualquier cosa, comunicarse con cualquier persona o comprar cualquier cosa".

Entre oct. y enero de 1996: Se lanzan Yahoo! Alemania, Francia y Reino Unido, en septiembre de 1997: Francia, Alemania y Reino Unido cuentan con más de 2 millones de páginas vistas al día, en noviembre de 1997: Yahoo! Suecia, Dinamarca y Noruega (Asia, Australia y Nueva Zelandia y Corea). En noviembre de 1998: En Europa, Yahoo! registra una media de 6,7 millones de páginas vistas al día. Se lanza Yahoo! España.

Es a mediados del 2001, cuando Yahoo realizó un cambio importante al movilizar sus resultados al modo "motor de búsqueda". Los criterios de búsqueda son los mismos que utiliza Google, ya que yahoo! toma los resultados provenientes de la misma base de datos.

Con el objeto de retomar sus esfuerzos en torno a búsqueda, que fue lo que lo hizo grande, Yahoo! ha hecho adquisiciones como la de Inktomi en Diciembre de 2002 y Overture en Julio de 2003 (junto con sus filiales Altavista, AllTheWeb), prescindiendo en parte de los servicios que hasta entonces, su rival Google le estaba ofreciendo como principal motor de sus búsquedas. Así mismo, por esa misma temporada llevó a cabo una alianza en México con una empresa llamada Ideas Interactivas para lanzar un directorio telefónico impreso bajo la marca Yahoo! Páginas Útiles, rompiendo paradigmas y

complementando su estrategia de búsqueda local en el mundo físico. Esto le permite competir fuertemente con Google y la nueva funcionalidad de búsqueda de MSN eficientemente.

En diciembre del 2004 El portal Yahoo! incorporó un nuevo buscador específico para localizar vídeos en fase de pruebas. Permite localizar los archivos que habitualmente pueden reproducirse en los programas más conocidos como son QuickTime, Real Media y Windows Media. De igual forma, en ese mismo año Yahoo! utiliza su propio sistema de búsquedas con algoritmo propio, terminando en forma definitiva con los servicios prestados por Google.

Yahoo! ha pasado por diferentes fases durante su vida en el web. Desde proponer, siendo el primer directorio temático de la red hasta ser una empresa reactiva, contestando a la competencia con servicios similares, cosa que ha cambiado con una serie de anuncios y compras interesantes, siendo precisamente en marzo de este año 2005, el Creative Commons Search su última propuesta. El valor de este servicio es único y su finalidad se basa en que miles de personas pueden publicar sus contenidos con licencias que permiten usarlas, siempre dando el crédito respectivo para proyectos comerciales o sin fines de lucro dependiendo del tipo de licencia, permitiendo de esta manera, beneficiar a las publicaciones digitales independientes o empresas medianas y pequeñas basadas en Internet.

Hoy en día, Yahoo! contiene información organizada de decenas de miles de ordenadores conectados a la Red. El objetivo del desarrollo del buscador Yahoo! fue incrementar la capacidad de alcance del mismo, a través de contenidos locales y dirigidos a una audiencia con intereses específicos, para aunar por un lado el potencial internacional y globalizador de la red con la especifidad necesaria que requiere cada cultura y sociedad. Con equipos basados en cada país, Yahoo! consigue enfrentarse mejor a las necesidades de los usuarios de la Red que buscan información que enfatice su herencia nacional.

Es así, que Yahoo! ofrece un diseño más claro y limpio e incorpora nuevas opciones de búsqueda. Cabe hacer notar, que Yahoo! con respecto a los resultados de sus búsquedas, éstas son provistas principalmente de sus propias bases de datos que están formadas por directorios hechos a mano, sin embargo por ser un proceso que no va al

paso del crecimiento de Internet, se provee de sus resultados secundarios de la base de datos de Inktomi.

C.3.2. FUNCIONAMIENTO Y PRESTACIONES

El motor de búsqueda interno de Yahoo!, permite plantear consultas introduciendo directamente las palabras clave para buscar en toda la base de datos o en algunas de sus categorías, si bien es en las búsquedas con términos generales en las que se obtienen mejores resultados. Para ser principalmente un directorio, sorprende la amplitud de opciones para personalizar las búsquedas.

Yahoo!, ofrece en las páginas de resultados enlaces a otros tipos de búsquedas, justo encima de los resultados:

- Web: Permite la búsqueda de sitios relevantes provistos por el motor de búsqueda de Yahoo! Search, mezclados con sitios listados en el Directorio Yahoo!
- Directorio: Encuentra sitios listados en la guía organizada por los editores de Yahoo!.
- Imágenes: Mira fotos, ilustraciones e iconos extraídos de toda la Red.
- Noticias: Busca en Yahoo! Noticias de actualidad relacionada con el tema de búsqueda.
- Compras: Encuentra tiendas, ofreciendo la posibilidad de comparar millones de artículos con los miles de comerciantes disponibles.

YAHOO! Es un directorio temático con buscador:

❒ Directorio:

La forma de recorrer «el árbol jerárquico» de las categorías de Yahoo! consiste en ir marcando, en forma sucesiva, la palabra en la que se supone a priori, se encuentra el tema a buscar.

El funcionamiento es sencillo. Como lo expresa el documento de ayuda, las categorías aparecen en negrita, mientras que las páginas web aparecen en tipología normal. Para bajar en la jerarquía, simplemente pulse la siguiente categoría y se le mostrará el siguiente

nivel. Para subir en la jerarquía, simplemente pulse la categoría adecuada en el encabezamiento de la página.

La manera como es estructurado el Directorio, es producto de la manera como los empleados de Yahoo! examinan uno a uno los servidores de Internet con el objetivo de crear un "índice de Internet". De acuerdo con Ibáñez (1998) "una vez examinado el material, incorporan cada una de las páginas en una categoría predeterminada (con referencias cruzadas a otras si es necesario), hacen un pequeño resumen de su contenido y lo publican en el catálogo general, un árbol que recoge "todo lo que pueda existir" completamente organizado en más de 20.000 categorías y cientos de miles de páginas documentadas".

❒ Buscador:

Yahoo! Por otra parte, también dispone de un buscador por palabra, con el cual se pueda buscar a través de toda la base de datos o sólo en una categoría o nivel seleccionados.

Yahoo, ofrece la Búsqueda avanzada, que ayuda a encontrar sitios que encajan con un criterio muy específico, incluyendo combinaciones de palabras y la fecha en que el sitio fue actualizado por última vez, restringe las búsquedas a sitios con un dominio específico, a un país de interés entre otras opciones que serán expuestas en la tabla26. Las opciones que se elijan serán aplicadas sólo a la búsqueda actual.

Entre otras formas de búsqueda Yahoo!, ofrece los Atajos que posibilita cuando se use determinadas palabras de búsqueda, la obtención de información precisa que se necesita encima de los resultados de sitios web, por otro lado también se cuenta con la opción de Preferencias de búsqueda, donde se configura las pautas personales de búsqueda de modo que sean aplicadas cada vez que buscas en la web usando Yahoo!.

YAHOO! SEARCH Search

TABLA26. PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.yahoo.com)

Información general	
Tipo de Servicio (Motor/Directorio):	Motor de búsqueda pero tambien directorio.
Acceso(libre,comercial):	Libre
Nro de URLs:	5 billones de paginas web reportados.
Origen:	David Filo y el Dr. Jerry Yang, los fundadores de Yahoo!, eran estudiantes de doctorado de Ingeniería Eléctrica en la Universidad de Stanford cuando decidieron iniciar una recopilación de sus intereses en Internet creando la guía Yahoo! a principios de 1994. Y es apartir de Abril de 1995 que se fundo Yahoo!, como un negocio de enorme potencial.
Frecuencia de actualizacion de la base de datos(mensual, semanal, quincenal,diaria):	mensualmente.
Nº de paginas que recoge por dia	
Tiempo que dura en dar las paginas web registradas:	4-8 semanas.
Versión analizada:	Ingles. Pero existe la version en castellano.
Portal:	Si.
Obedece el protocolo de exclusión(Sirven para indicarle a los robots que la página que los contiene no debe ser indexada y/o que los enlaces de tal página no deben ser seguidos).	Soporta el uso del estandar robot.txt. Esto lo lee mediante el meta-tag noindex: <META NAME="robot.txt" CONTENT="noindex"> .
Recolección	
Robot:	Spider
Método(Humano/Automático):	Directorio: Humano y Buscador:Automático
Primero en profundidad:	No aplica
Primero en Anchura:	No Aplica
Tipo de cobertura (www, gopher,WAIS, ftp,telnet,IRC, UseNet News, productos multimedia, etc.,):	Web, Gopher, FTP, Usenet (News Group), Telnet.
Cobertura geográfica:	cobertura Internacional.
Objeto de Cobertura (general o contenido especializado):	General.
Frecuencia de actualizacion para visitar los mismos sitios/documentos:	periodicamente.
Otras bases de datos:	Cuenta ademas de su Base de datos manual(Directorio) Tambien por las bases de datos proporcionadas por Inktomi y Overture.

YAHOO! SEARCH Search	PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.yahoo.com)
Indexación	
Que indexa (titulos, cabeceras, información de cabecera, Links(URLs),etiquetas HTML, resumenes, texto completo, etc.,):	No indiza el contenido del documento, solo realiza una descripción del mismo. La información está organizada y tematizada en forma jerárquica, en categorías principales (14) y subcategorías. Además indexa los primeros 500 KB de una pagina web.
Que no indexa:	Spam.
Sistema de Recuperación	
Herramienta de Búsqueda:	Motor de búsqueda.
Tipo de Recuperación (Booleana, Vectorial, Probabilística, otro):	Tipo de recuperacion por "The best matched".
Estructuras de consulta y operaciones soportadas:	Consignada en la tabla 27.
Sistema búsca *por default/por campos seleccionables(*URLs, titulos,Resumen, Texto completo, otros.,)	Título, url, descripción o resumen y asi como el origen de los links asociados.
Mejora en la Busqueda (reducción de resultados, reformulación, vocabulario controlado, mediante ejemplo,otros.,):	Si permite refinar el resultado de la búsqueda.
Incluye textos alternativos:	Si aplica.
Ordenamiento:	Los resultados se obtienen ordenados de acuerdo a su relevancia en cantidad de palabras y a criterios como coincidencias con la palabra clave, relevancia de la sección en el documento(Documentos que se relacionan con la palabra buscada, que se encuentran en el título son ordenados con un rango mayor que aquellas que se encuentran en el cuerpo del documento o en la dirección URL), Generalidad de la categoría(Palabras que se encuentran relacionadas con las categorías más altas del árbol de jerarquía de la estructura de búsquedas de Yahoo).
Despliegue de resultados: (Título, descripción,URL, tamaño,fecha de alta, Nº total de saltos, correspondencia de términos,orden de relevancia,valor de relevancia,Nº de saltos por pagina, formato variable, detecta novedades, permite traducir,otros.,):	Permite ver título con el enlace, la URL, las primeras palabras a modo de resumen, correspondencia de términos, Nro total de saltos, la salida se compone por defecto de 10 resultados por pagina, el usuario puede elegir los criterios de relevancia en el orden de presentación de los resultados. Ademas cuenta con información del tamaño del archivo, archivos PDF, vínculos en caché, vinculo para visualizar mas paginas del sitio mediante el enlace *More from this site*.

YAHOO! SEARCH

Search

PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.yahoo.com)

Interfaz de Usuario	
Descripción de la Interfaz	Menus estructurados, desplegables.
Claridad de la interfaz y de la pagina de búsqueda:	Buena y muy intuitiva en cuanto al buscador. Por otro lado el directorio es muy bueno si se tiene una idea general de qué es lo que se quiere por estar organizado por categorias.
Ofrece ayuda en cuanto a la documentacion de búsqueda:	La ayuda que brinda es muy buena y detalla la forma de realizar las búsquedas, los operadores admitidos y los criterios de relevancia a aplicar en la presentación de resultados y mucho más.
Busqueda Simple/Avanzada	Ambas
Establece preferencias	Si permite establecer preferencias y guardar la configuración.
Idiomas Interfaz:	37idiomas.
Servicios:	Servicios para usuarios: información en internet (noticias, viajes, deportes, juegos, meteorología); servicios comerciales (compras, subastas, finanzas, anuncios clasificados); servicios de comunicación (e-mail, Yahoo! Messenger, Yahoo! Groups) y servicios wireless para PDAs, teléfonos móviles y otros dispositivos. Entre otros servicios para empresas y negocios, etc.,
URL Pagina de Ayuda:	http://help.yahoo.com/help/us/ysearch/
URL Pagina deTraduccion:	No aplica.
Ranking:	Rankea basado en relevancia(ocurrencia de términos).

TABLA27. ESTRUCTURAS DE CONSULTA Y OPERACIONES SOPORTADAS

YAHOO! SEARCH

Search

Operador default:	Por defecto establece la relación AND entre los términos.
Operador de Existencia: (exclusión/inclusion)	-/+
Operadores Booleanos:	AND(Y), OR(O),NOT or AND NOT.
Operadores de Proximidad:	No aplica
Búsqueda por campos:	site:, hostname:, link:, url: , inurl:, intitle:. De igual manera restringe la búsqueda por dominio(Site/Domain:búsqueda avanzada).
Operadores de exactitud:	Si aplica. Usa la frase literal con " ".
Lenguaje Natural:	
Truncamiento:	No aplica en el buscador web. Sin embargo en el directoro de Yahoo! El truncamiento * es admitido.
Agrupacion de términos y operadores:	Si aplica. Admite el uso de los paréntesis "( )".
Sencible mayúsculas/minúsculas:	No aplica
Reconoce sinónimos:	
Caso sensitivo a acentos, puntuación:	No aplica.
Filtro por idioma:	Si Aplica
Filtro por formato:	Si Aplica: permite restringir la búsqueda a un tipo de archivo o en su defecto permitir cualquiera(entre los formatos que maneja: .pdf, .htm, .htmL, .pdf, .xls, .ppt, .doc, .xml, .rdf, .rss, .txt.)
Filtro por Country:	Si Aplica. Permite producir resultados de un país especifico.
Filtro por numero de resultados de páginas:	Si Aplica
Filtro por fecha:	Si aplica. Permite obtener los documentos más recientes, mediante tres opciones específicas dentro de los pasados 3, 6 meses o el último año.
Creador de consultas:	Si aplica. Puede crear fácilmente potentes consultas introduciendo palabras en una combinación de estos cuadros de texto(todas las palabras, frase exacta, cualquier de estas palabras, ninguna de estas palabras).
Permite bloqueo de contenido ofensivo:	Si aplica. (Busqueda Avanzada) y en el sitio de preferencias.

Es importante hacer notar que en Yahoo!, que sus búsquedas se recupera una lista de páginas web que contienen las palabras que se introdujeron a priori. Estos resultados son una combinación de sitios de toda la Web (obtenidos por motor de búsqueda, Yahoo! Search) y del Directorio Yahoo!. Por otra parte, el Directorio Yahoo! es una "biblioteca" de sitios web organizada por catalogadores en categorías y subcategorías. Ellos revisan personalmente estos sitios y eligen el mejor lugar para ubicarlos.

Cabe señalar, la existencia en las categorías de algunos símbolos cuyo significado se expresa a continuación:

* "@" La existencia de otra categoría superior de ese mismo tema.
* Un número entre paréntesis que indica en número de opciones contenidas.
* XTRA!, indica que existe algún artículo o novedad sobre ese tema.

C.4 MOTOR DE BÚSQUEDA " ALTAVISTA "

AltaVista, empresa de Overture Services, Inc., es un destacado proveedor de tecnología y servicios de búsqueda. AltaVista continúa desarrollando la búsqueda a través de Internet con nuevas tecnologías y herramientas diseñadas para mejorar las búsquedas de los usuarios. AltaVista, con sede en Sunnyvale, Calif., cuenta con una larga trayectoria de innovación, reflejada en sus 61 patentes de búsqueda.

Altavista es un Motor de Búsqueda de origen internacional que posee versiones regionales para los diferentes idiomas. Su nombre que significa "una visión desde las alturas", se vio inspirada por la creación de grandes ideas de un equipo de expertos fascinados con el seguimiento de la información.

Hoy en día Altavista, se basa en un objetivo o principio básico que consiste en proporcionar a la comunidad global acceso a la información y fijar el estándar de la tecnología de búsqueda y la forma en que la gente encuentra la información.

C.4.1 ORIGEN Y DESCRIPCION

Durante la primavera de 1995, los científicos del Laboratorio de investigaciones de Digital Equipment Corporation en Palo Alto, California, con Louis Mornier a la cabeza, se pusieron manos a la obra en su creación; tras numerosas pruebas se diseño una araña que buscaba y recuperaba automáticamente información de las páginas web y que fue bautizada como Scooter. Asimismo, decidieron crear un índice de toda la Web, tarea que hasta ese momento se pensaba irrealizable, por el vasto volumen que ya entonces estaba alcanzando.

Para conseguir este ambicioso objetivo crearon un programa de indización capaz de indicar a texto completo las páginas con la misma velocidad de Scooter podía recuperarlas. Según Mornier, los laboratorios de Digital eran el único lugar del mundo donde podía realizarse este trabajo con tanta rapidez, ya que pocas empresas tenían el personal investigador necesario y ninguna universidad podría haberse permitido tal inversión en equipamiento.

Altavista fue el nombre del código de ese proyecto, que acabo manteniéndose finalmente de manera definitiva, convirtiéndose en la primera base de datos de texto completo en la que se podían realizar búsquedas en la World Wide Web.

En 1997, la empresa de Altavista, añadió búsquedas multilingües con compatibilidad para 25 idiomas.

En 1999, lanzó compatibilidad con búsqueda de archivos multimedia (audio, video, imágenes). Entre 1999 y el 2001 da apertura a 20 sitios locales en diferentes países entre los mas conocidos motor de búsqueda para España (www.altavista.es).

Fue el primer motor de búsqueda importante que introdujo la búsqueda gratuita de noticias en Internet en 2001. Por otro lado en marzo de este mismo año Altavista incluye la capacidad de búsqueda multilingüe de Internet y la primera tecnología de búsqueda que admitió los idiomas chino, japonés y coreano. Babel Fish, fue el primer servicio de traducción mecanizada de la Web que puede traducir palabras, frases o sitios web completos hacia y desde el inglés, el español, el francés, el alemán, el portugués, el italiano y el ruso.

En febrero del 2002 AltaVista anuncia el lanzamiento oficial de "AltaVista Shortcuts". Esto es una tentativa de proporcionar el contenido " del Web Invisible " proporcionando de esta manera los links escogidos para búsquedas mas comunes en la Web. Por otro lado, se presentó Prisma™, su poderosa herramienta de búsqueda asistida, en ese mismo año, es básicamente una nueva forma de ayudar a refinar una pregunta de búsqueda y producir los resultados mas relevantes. Este escanea el texto de los 50 top de resultados de búsqueda regular y luego selecciona los resultados subdivididos en 12 subgrupos junto con una lista separada con todas las páginas web que contienen la palabra solicitada.

Desde abril del 2003, AltaVista forma parte de Overture. Altavista es una de las propiedades Internet que más ha cambiado de dueños. Primero perteneció a DEC, luego a Compaq, tuvo después un breve período de independencia y fue a partir de este año que adquirió Overture (antes conocido como GoTo.com), famoso por ser el primer buscador exitoso de "Pay Per Click". Fue en este mismo año, que AltaVista se consolidó como la mejor opción de búsqueda para contenidos multimedia: imágenes, vídeos o audio.

En marzo del 2004, La base de datos de Web de AltaVista ha sido substituida en gran medida por la base de datos de búsqueda de Yahoo!.
Su directorio temático se encuentra potencializado Open Directory, los archivos de imágenes, audio y video y paginas de noticias es potencializado por su propia base de datos. Sin embargo mucho de los resultados web proviene de la Base de datos de Yahoo.

C.4.2 FUNCIONAMIENTO

Altavista maneja tanto la Búsqueda Simple como la avanzada.

Búsqueda Simple: Se introduce dentro del formulario presente en su pagina principal, la o las palabras claves de búsqueda y se da click en el botón encontrar. En esta búsqueda se hace uso normalmente del operador por defecto “AND” o de los operadores de inclusión; sin embargo, la búsqueda por reconocimiento de frase literal es llevada a cabo en forma también en forma automática.

Búsqueda Avanzada: Altavista soporta completamente la búsqueda booleana, Ofrece también la opción de Display site collapse(on/off). Que si se encuentra activada, permite agrupar y desplegar uno o dos paginas por sitio; de lo contrario, recupera todo las páginas de resultados sin ningún tipo de agrupamiento. Entre otra de las opciones que dispone tenemos:

- Related Pages , paginas relacionadas con el tema de búsqueda.
- Translate, permite la traducción atomática de la pagina.
- More pages from . . . , Se obtiene más páginas del sitio obtenido en la pagina de resultados.

A continuación se presenta información mas detallada de este particular Motor de Búsqueda.

altavista ENCONTRAR	**TABLA28. PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.altavista.com)**
	Información general
Tipo de Servicio (Motor/Directorio):	Motor de búsqueda pero tambien cuenta con el nivel de directorio.
Acceso(libre,comercial):	Libre
Nro de URLs:	1,1 billion
Origen:	Fue creado en 1995 por Digital Equipment Corporation en los laboratorios de investigación de Palo Alto (Estados Unidos).
Frecuencia de actualizacion de la base de datos(mensual, semanal, quincenal,diaria):	mensualmente.
Nº de paginas que recoge por dia	No mencionado.
Tiempo que dura en dar las paginas web registradas:	1-2 dias
Versión analizada:	Ingles. Pero existe la version en castellano.
Portal:	No.
Obedece el protocolo de exclusión(Sirven para indicarle a los robots que la página que los contiene no debe ser indexada y/o que los enlaces de tal página no deben ser seguidos).	Hace uso del archivo Robots.txt: es un archivo de texto que permite indicar que páginas no deben ser indexadas en el sitio.
	Recolección
Robot:	Spider(Robot explorador llamado Scooter).
Método(Humano/Automático):	Automático
Primero en profundidad:	No aplica.
Primero en Anchura:	Si aplica.
Tipo de cobertura (www, gopher,WAIS, ftp,telnet,IRC, UseNet News, productos multimedia, etc.,):	Web, Usenet, productos multimedia(audio, imágenes y video).
Cobertura geográfica:	cobertura Internacional.
Objeto de Cobertura (general o contenido especializado):	General.
Frecuencia de actualizacion para visitar los mismos sitios/documentos:	2 a 3 días. Y a nivel de directorio por ser estáticos son recalculados una vez a la semana.
Otras bases de datos:	Utiliza la tenología de Overture, para la búsqueda de paginas web y para buscar dentro de las categorías(Directory) del Open Directory (DMOZ). Entre otras bases de datos a mencionar tenemos Images, MP3/Audio, Video, News.

altavista ENCONTRAR	**PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.altavista.com)**
Indexación	
Que indexa (titulos, cabeceras, información de cabecera, Links(URLs),etiquetas HTML, resumenes, texto completo, etc.,):	El sistema es capaz de reconocer las etiquetas META de título, descripción y palabras-clave y extraer la información que contienen.
Que no indexa:	Spam.
Sistema de Recuperación	
Herramienta de Búsqueda:	Motor de búsqueda.
Tipo de Recuperación (Booleana, Vectorial, Probabilística, otro):	Booleana, Probabilística.
Estructuras de consulta y operaciones soportadas:	Consignada en la tabla 29.
Criterio de Búsqueda:	*Rastreo por enlaces de popularidad:* donde la popularidad de una página se detecta analizando cuantos enlaces existen hacia otra página.
Sistema búsca *por default/por campos seleccionables(*URLs, titulos,Resumen, Texto completo, otros.,)	Por defecto en el texto completo.
Mejora en la Busqueda (reducción de resultados, reformulación, vocabulario controlado, mediante ejemplo,otros.,):	Contemplan la posibilidad de 'Afinar' la búsqueda mdiante la búsqueda Avanzada.
Incluye textos alternativos(inf. similar):	Si aplica
Ordenamiento:	Es susceptible de tener un gran número de páginas de resultado por tanto el estímulo de clasificación por enlaces de popularidad: mediante el número de enlaces que existen a ella desde otras páginas. El nivel de directorio donde se encuentra la página. Los más altos son considerados como más importantes. Si una página está muy al fondo, el robot o spider no irá tan abajo y nunca la encontrará.
Despliegue de resultados: (Título, descripción,URL, tamaño,fecha de alta, Nº total de saltos, correspondencia de términos,orden de relevancia,valor de relevancia,Nº de saltos por pagina, formato variable, detecta novedades, permite traducir,otros.,):	Permite ver título con el enlace, la URL, las primeras palabras a modo de resumen, correspondencia de términos, la salida se compone por defecto de 10 resultados por pagina, visualiza el número total de documentos encontrados que coinciden con su criterio de búsqueda, No de saltos por página, y en preferencias da la opcion de activar la visualizacion del tamaño de la pagina y el lenguaje de la pagina. Permite la búsqueda por paginas relacionadas y permite con el link More Pages from Site: visualizar mas paginas del sitio listado.

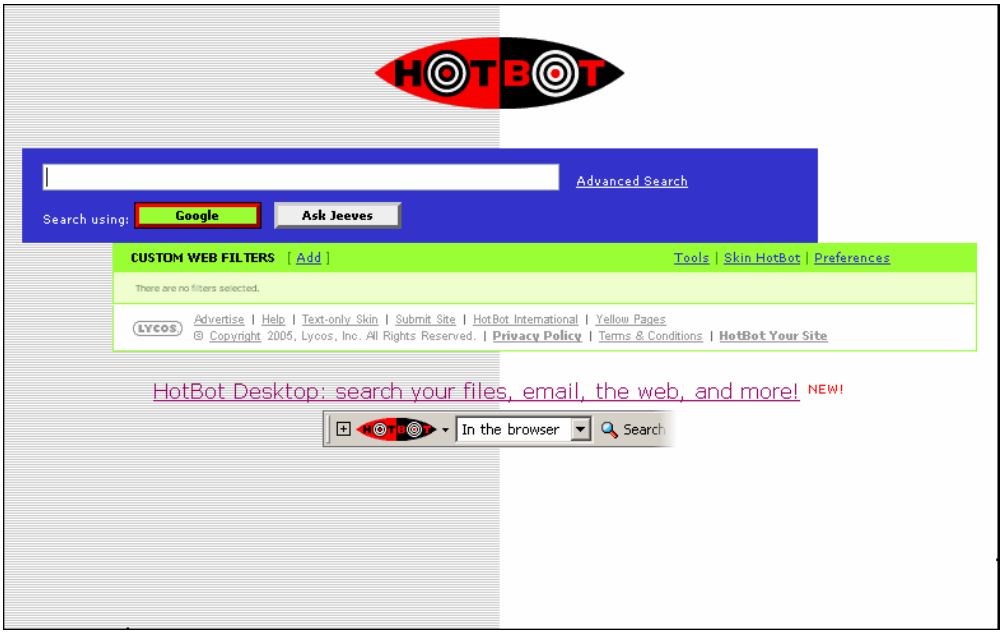
	**PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.altavista.com)**
Interfaz de Usuario	
Descripción de la Interfaz	Menus estructurados, desplegables(Busqueda Avanzada).
Claridad de la interfaz y de la pagina de búsqueda:	Interfaz clara con instrucciones detalladas.
Ofrece ayuda en cuanto a la documentacion de búsqueda:	La ayuda detalla la forma de realizar las búsqueda avanzada, los operadores a utilizar y los criterios de relevancia que se aplican en la presentación de resultados entre otras características.
Busqueda Simple/Avanzada	Ambas.
Establece preferencias	Si permite establecer preferencias y guardar la configuración.
Idiomas Interfaz:	26 idiomas.
Servicios:	Hay opciones de búsqueda por newsgroups, personas y empresas. Posee un sistema de traducción automática que traduce textos y documentos web, desde y hacia diferentes idiomas.
URL Pagina de Ayuda:	http://www.altavista.com/help/search/help_adv
URL Pagina deTraduccion:	http://babelfish.altavista.com/. Con la posibilidad de traducir paginas web en 30 lenguajes diferentes.
Ranking:	Trabaja con ranking basado en la cantidad de enlaces, que se han hecho desde otras páginas. De igual forma tambien tiene encuenta las páginas largas con mucho texto significativo, y páginas con un buen sistema de navegación, con un montón de vínculos a páginas con contenido relacionado.

TABLA29. ESTRUCTURAS DE CONSULTA Y OPERACIONES SOPORTADAS

Operador default:	AND
Operador de Existencia: (exclusión/inclusion)	Incluye el termino (+ y -)
Operadores Booleanos:	AND, OR, NOT, AND NOT
Operadores de Proximidad:	Si aplica. NEAR(10 palabras)
Búsqueda por campos:	domain, host, link, title, url. habilidad para restringir la búsqueda a ciertas partes o a un tipo de documentos, imagen: image:, Java applets: applet:, y enlaces, link: entre otros.
Operadores de exactitud:	Maneja la frase literal haciendo uso de las comillas " ".
Lenguaje Natural:	No mencionado
Truncamiento:	Maneja truncación derecha e izquierda con '*'.
Agrupacion de términos y operadores:	Si mediante el parentesis operador de paréntesis ( ).
Sencible mayúsculas/minúsculas:	No aplica
Reconoce sinónimos:	Si aplica.
Caso sensitivo a acentos, puntuación:	No aplica.
Filtro por idioma:	Si aplica.
Filtro por formato:	Si Aplica: permite restringir la búsqueda a un tipo de archivo, entre los formatos que maneja: .pdf, .html, .pdf, .xls, .ppt, .doc, txt.).
Filtro por Ubicación:	Por dominio y URL.
Filtro por numero de resultados de páginas:	Si aplica. Permite recuperar de 10, 20, 30, 40 o 50 resultados al tiempo.
Filtro por fecha:	Si aplica. Dispone de bien filtra por rango o por semanas, meses o un año.
Creador de consultas:	Puede crear fácilmente consultas introduciendo palabras en una combinación de estos cuadros de texto(todas las palabras, frase exacta, cualquier de estas palabras, ninguna de estas palabras).
Permite bloqueo de contenido ofensivo:	Si aplica. Mediante el uso de activacion de la opción Family Filter.

C.5 MOTOR DE BÚSQUEDA " HOTBOT "

Hotbot es el motor de búsqueda más colorido y el que goza de un gran éxito crítico. Se caracteriza por ser un buscador internacional de origen americano. Su sistema de búsqueda inteligente y complejo con variedad de herramientas, permite ayudar a los usuarios a realizar mejor sus búsquedas y obtener resultados importantes de forma rápida y fácil.



HotWired e Inktomi crearon esta herramienta de búsqueda capaz de indexar el World Wide Web entera. "La misión de HotBot consiste fundamentalmente en ayudar a la gente en la permanencia de economía de de la información de hoy, estando a la vanguardia del panorama de crecimiento que se desarrolla tan rápidamente en la Web.

HotBot está en una posición única siendo un motor de los más exhaustivos en la búsqueda combinando con un conjunto de sofisticados e intuitivos instrumentos para las consultas mas precisas.

C.5.1. ORIGEN Y DESCRIPCION

Hotbot (www.hotbot.com), se lanzó en mayo de 1996 como la entrada en el mercado de los motores de búsqueda en Internet de Wired Digital. Fue desarrollada por los editores de la revista Wired (revista de excelente reconocimiento) junto con la contribución de Nicholas Negroponte y Esther Dyson. Es a partir de este año, que Hotbot adquirió un fuerte reconocimiento por su calidad y exhaustividad en sus resultados proporcionados por Inktomi, en ese entonces. De igual forma llamo la atención por la interfaz web que manejaba, por sus colores inusuales y su imagen tan divertida.

Durante octubre de 1998, El grupo Wired fue adquirido en su totalidad por Lycos (ahora Terra Lycos) . A partir de ahí, Lycos no pudo hacer de la búsqueda una prioridad en su

sitio, por tanto, se centró en la adición de características "portal" descuidando de alguna manera a Hotbot.

En 1999 HotBot ganó más notoriedad cuando emprendió un cambio frente a los resultados proporcionados para sus listados principales que eran llanamente proporcionados por Direct Hit . Direct Hit era en ese tiempo un de los motores de búsqueda más novedosos que habían aparecido recientemente. Infortunadamente la calidad de los resultados ofrecidos no superaba al también reciente buscador que debutaba en el mismo tiempo, Google, por tanto el renombre de Hotbot comenzó a decaer en importancia. En ese mismo año a mediado de Abril, implementó su propia version de directorio temático del Open Directory.

A finales del 2002, HotBot cuenta con cuatro base de datos Inktomi, Google, FAST, and Teoma, posibilitando al usuario mayor control en sus búsquedas permitiendo fácilmente consultas por cuatro diferentes motores de búsqueda desde una sola interfaz. Lográndose en ese entonces búsquedas relacionadas muy poderosas entre ellos y búsqueda avanzada mediante operaciones booleanas.

A principio del 2003, La versión de HotBot, Teoma finalmente ha añadido algunos características avanzadas de búsqueda que están disponibles en el sitio de Teoma de las cuales se pueden destacar la búsqueda por lenguaje, fecha, región, filtro para contenido ofensivo, entre otros.

En junio del 2003, HotBot lanza la barra de herramienta “desk bar”. Lo interesante de esta barra de herramientas que la diferencia de Dogpile y Google es que esta no requirió ser instalada dentro del browser, sino por el contrario trabajaba en la barra de tareas del explorador de Windows y bajo cualquier browser por defecto. En julio del 2003 Hotbot que se encuentra potencializado por cuatro motores de búsqueda en su sitio, tres de los cuatro fueron modificados en sus nombres: Inktomi que fue denominado "HotBot", FAST denominado "Lycos" y Teoma llamado "Ask Jeeves".

En marzo del 2004 Hotbot elimina una de sus bases de datos, originalmente denominada FAST y posteriormente cambiada a Lycos. Lo interesante en esa fecha, es que siendo Hotbot propiedad de Lycos eliminó la base de datos que potencializa a su propio

propietario. En abril de ese mismo año Hotbot ofrece una nueva utilidad de búsqueda de escritorio llamada “Desktop” no sólo busca en el web, sino que este también permite indexar archivos y correo electrónico sobre su ordenador, haciéndolo mucho mas manejable.

Actualmente, Hotbot ofrece en su sitio únicamente dos bases de datos Google y Ask caracterizados por ofrecer una búsqueda bastante depurado. Incluso el sistema nos proporciona la opción de bajarnos de Internet dos software que se instala en nuestro ordenador para realizar búsquedas desde nuestro escritorio, el cual puede instalarse pulsando el enlace "tools". Estas herramientas mencionadas son HotBot Desktop y HotBot Deskbar. Es importante aclarar que Hotbot no actúa como metabuscador, para realizar las búsquedas se requiere que elija una de las bases de datos (Google o ASK) para que se recupere resultados propios del motor elegido dentro de su interfaz. Independiente de manejar dos bases de datos potencializados por otros proveedores este tiene su propio índice encargado de recuperar la información pertinente a la necesidad informativa especificada.

C.5.2 FUNCIONAMIENTO

Hotbot es un tipo robot, un servidor que va almacenando y catalogando infinidad de páginas de la red. En general los buscadores de este tipo revisan todo Internet y van agregando las direcciones de las páginas que encuentran a su base de datos. Utiliza sistemas de contabilización de palabras, así como, revisa los metatags de cada página para poder asignar una puntuación a cada web, luego cuando alguien le pregunta por una palabra clave, el servidor responde con las direcciones de las páginas que las contienen organizadas según la puntuación que les ha conferido previamente.

Este buscador utiliza una opciones llamadas filtros que permiten escoger a través de que motor de búsqueda quiere buscar un término, estos dos filtros son: Google y Ask Jeeves.

HotBot ha desarrollado, a través de innumerables pruebas, una intuitiva y eficaz interfaz de búsqueda, con el que tanto los usuarios inexpertos como los más avanzados encuentran siempre el mejor resultado de manera rápida y transparente. Filtros

específicos de búsqueda El usuario puede elegir de entre varias opciones: Búsqueda Rápida, Avanzada.

La Búsqueda Rápida ofrece una búsqueda muy rápida y sencilla dentro de la enorme bases de datos que conforma HotBot. La opción de que el usuario pueda utilizar filtros, uno de los valores añadidos y específicos de HotBot, se encuentra dentro de la Búsqueda Avanzada, que ayuda a hacer la búsqueda más eficaz; por ejemplo, puede predeterminarse el idioma de la búsqueda, especificarse la fecha de publicación del documento.

En la tabla30, se expone con mas amplitud las características destacables por este interesante motor de búsqueda.

Cabe mencionar que Hotbot cuenta con la capacidad “skin”, es decir la posibilidad de poder personalizar el look y ambiente de su interfaz permitiendo modificar el fondo, los colores en áreas específicas así, como dar un formato de fuente a su gusto. Además ofrece dentro de su pagina el link Yellow Pages, sitio propio de Terra Lycos su propietario, y es un lugar desarrollado para proveer un amplio portafolio de temas ordenados por categorías tales como información de páginas web, titulares de noticias, avisos clasificados, buscador exclusivo de contenido de los sitios de las bibliotecas, etc. Por otro lado, Hotbot no reconoce en la URL caracteres especiales como ?, =, % and &. Entre las asistencias que ofrecen sus barras de herramientas podemos identificar:

HotBot Desktop...

- Búsqueda en el Web
- Búsqueda en computador local
- Subscripción a RSS News Feeds
- Bloquea Pop-ups.
- Permite el acceso fácil a las búsquedas del sitio desde cualquier sitio en el Internet

HotBot Deskbar...

- Permite la búsqueda desde escritorio de su computador
- Búsqueda en el Web
- Ofrece funciones y herramientas (e.g. alarmas, calculadora)
- Acceso a su calendario personal en línea
- Crea búsqueda Shortcuts

TABLA30. PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.hotbot.com)

Información general	
Tipo de Servicio (Motor/Directorio):	Motor de búsqueda.
Acceso(libre,comercial):	Libre
Nro de URLs:	2,2 billones de paginas web reportados.
Origen:	En mayo de 1996 fue lanzado Hotbot al mercado de los motores de búsqueda. Desarrollado por Wired Magazine, pero posteriormente la Wired Digital fue comprado por Lycos(Actualmente llamado Terra Lycos) . Sus resultados de búsqueda fueron proporcionados inicialmente por la base de datos de Inktomi.
Frecuencia de actualizacion de la base de datos(mensual, semanal, quincenal,diaria):	1 mes.
Nº de paginas que recoge por dia	No mencionado.
Tiempo que dura en dar las paginas web registradas:	2-4 semanas.
Versión analizada:	Ingles. Pero existe la version en castellano (http://www.hotbot.lycos.es)
Portal:	No.
Obedece el protocolo de exclusión(Sirven para indicarle a los robots que la página que los contiene no debe ser indexada y/o que los enlaces de tal página no deben ser seguidos).	Soporta el uso del estandar robot.txt.
Recolección	
Robot:	Crawler, basado en los Motores de Búsqueda Google y Ask Jeeves.
Método(Humano/Automático):	Buscador:Automático
Primero en profundidad:	No aplica
Primero en Anchura:	No Aplica
Tipo de cobertura (www, gopher,WAIS, ftp,telnet,IRC, UseNet News, productos multimedia, etc.,):	Especificamente Web.
Cobertura geográfica:	cobertura Internacional.
Objeto de Cobertura (general o contenido especializado):	General.
Frecuencia de actualizacion para visitar los mismos sitios/documentos:	1 y 3 días .
Otras bases de datos:	Actualmente cuenta con las bases de datos proporcionadas por Google y Ask Jeeves.

	PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.hotbot.com)
Indexación	
Que indexa (titulos, cabeceras, información de cabecera, Links(URLs),etiquetas HTML, resumenes, texto completo, etc.,):	Dependiendo de la base de datos bien puede indexar texto completo o bien, indexar titulos, meta tags, resumenes.
Que no indexa:	Spam, URLs con caracteres especiales. Hotbot detecta el spoofing, es decir el reconoce técnicas que incluyen la repetición de palabras, la inserción de meta tags con contenidos no relacionados con el documento o el uso de palabras que no puede ser leido debido al taamaño tan pequeño o al color.
Sistema de Recuperación	
Herramienta de Búsqueda:	Motor de búsqueda.
Tipo de Recuperación (Booleana, Vectorial, Probabilística, otro):	Booleana y probabilística.
Estructuras de consulta y operaciones soportadas:	Consignada en la tabla 31.
Sistema busca *por default/por campos seleccionables(*URLs, titulos,Resumen, Texto completo, otros.,)	words appear frequently in a short document
Mejora en la Busqueda (reducción de resultados, reformulación, vocabulario controlado, mediante ejemplo,otros.,):	Permite refinar los resultados de la búsqueda. (Búsqueda Avanzada).
Ordenamiento:	Por defecto, ordena los sitios por orden de relevancia, es decir, utiliza algoritmos de contabilización de palabras, así como, revisa los metatags. Dependiendo de la base de datos el despliegue del ordenamieto se basa en la ubicación de información, es decir en base al numero de veces que el término aparece, si este aparece en el titulo o al comienzo del documento o bien en la relevancia basados en N° de links desde páginas web más consultadas.
Despliegue de resultados: (Título, descripción,URL, tamaño,fecha de alta, Nº total de saltos, correspondencia de términos,orden de relevancia,valor de relevancia,Nº de saltos por pagina, formato variable, detecta novedades, permite traducir,otros.,):	Despliega, titulo, la descripcion, Url, tamaño(bytes), fecha, Nº total de saltos, correspondencia de términos, el orden de relevancia y la salida se compone por defecto de 10 resultados por pagina.

PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.hotbot.com)

Interfaz de Usuario	
Descripción de la Interfaz	Menus estructurados, desplegables.
Claridad de la interfaz y de la pagina de búsqueda:	Buena e intuitiva. Y lo mejor permite la búsqueda en dos bases de datos en un mismo sitio.
Ofrece ayuda en cuanto a la documentacion de búsqueda:	La ayuda que brinda da claridad en algunos aspectos, se basa en preguntas mas frecuentes que los usuarios comunmente tienenacerca del uso de la interfaz. Sin embargo la informacion no esta del todo actualizada.
Busqueda Simple/Avanzada	Ambas
Establece preferencias	Si permite establecer preferencias y guardar la configuración.
Idiomas Interfaz:	34 idiomas.
Servicios:	Hotbot ofrece dentro de su sitio link yellow pages de Terra lycos, que es actualmente su propietario, el cual posibilita tener opciones de búsqueda por newsgroups, negocios, productos multimedia y mucho más. Entre la barras de herramientas destacadas en su sitio esta HotBot Desktop y HotBot Deskbar, las cuales ofrecen muy buenas opciones no solo en la web sino en el ordenador local.
URL Pagina de Ayuda:	http://help.lycos.com/hotbot/hotbot_1_help.asp
URL Pagina deTraduccion:	No aplica.
Ranking:	El Ranking de relevancia está basado en una combinación de frecuencia de aparición del término y ubicación dentro del documento.

TABLA31. ESTRUCTURAS DE CONSULTA Y OPERACIONES SOPORTADAS

Operador default:	Multiples términos de búsqueda son procesados como una operación AND por defecto.
Operador de Existencia: (exclusión/inclusion)	-/+
Operadores Booleanos:	AND, OR, NOT .
Operadores de Proximidad:	No aplica.
Búsqueda por campos:	title: , domain: , site: ,son aplicables directamente al formulario incial. (la búsqueda por campos en Search Advance no esta disponible).
Operadores de exactitud:	Si aplica. Usa la frase literal con " ".
Lenguaje Natural:	No es mecionado
Truncamiento:	No aplica.
Agrupacion de términos y operadores:	Si aplica. Admite el uso de los paréntesis "( )".
Sencible mayúsculas/minúsculas:	No aplica.
Reconoce sinónimos:	No es mecionado.
Caso sensitivo a acentos, puntuación:	No aplica.
Filtro por idioma:	Si Aplica.
Filtro por formato:	No aplica.
Filtro por Domain/Site:	Si Aplica. Pemite la inclusion o la exclusion de un dominio o sitio específico.
Filtro por región:	Aplica si la búsqueda avanzada se filtra mediante el uso de la base de datos As jeeves.
Filtro por numero de resultados de páginas:	No Aplica
Filtro por fecha:	Si aplica. Limita los resultados de la página dentro de un específico periodo de tiempo.
Creador de consultas:	Filtro por exclusión/inclusión de palabras o a una parte específica de la pagina. Aplica solo si la búsqueda se filtra mediante la base de datos Google.
Permite bloqueo de contenido ofensivo:	Si aplica. (Busqueda Avanzada) en ambas Bases de datos(Google y Ask Jeeves).

C.6 MOTOR DE BÚSQUEDA " EXCITE "

Excite es un destacado portal de Internet con un motor de búsqueda incluido de ahí que sea considerado un motor de búsqueda híbrido, que cumple con la definición tradicional de motor de búsqueda(permitiendo la visita a sitios Web y clasificándolos usando un programa de software) pero también cuenta con un directorio, con una variedad de sitios ordenados por categorías.

Excite es uno de los más pequeños motores de búsqueda. Pero es muy reconocido, no solo por la variedad de contenidos sino también por las funcionalidades que ofrece: proporciona la personalización sofisticada, ofrece resultados relevantes excelentes para preguntas muy populares, incluye web-mail y su búsqueda de Noticias proporciona el acceso importante a las versiones Web de periódicos, revistas, y cables de noticias (el índice ubicado a la izquierda del sitio, permite navegar este ejemplar).

C.6.1. ORIGEN Y DESCRIPCION

Excite fue inicialmente desarrollado y administrado en California en febrero de 1993 por un grupo de amigos que comenzaron a trabajar en un programa que gestionase la información existente en la Web.

Fundado por Mark Van Haren, Ryan McIntyre, Ben Lutch, Joe Kraus, Graham Spencer y Martin Reinfried. Estas personas (cinco hackers y un experto en ciencias políticas), investigaron para la Biblioteca de la Universidad de Stanford, cual sería la mejor forma de buscar y recuperar información para solucionar el problema de dicha biblioteca

Esta idea condujo a la creación de Architext, software que mas tarde se incorporó a Excite. Su particularidad estribaba en que ofrecían una gran base de datos, acceso a la

información mediante categorías temáticas y, como primicia la evaluación y comentario – Una breve sinopsis del documento realizada por expertos de Excite- de algunas paginas web de interés. De esta forma se podría comprobar rápidamente si esos recursos son relevantes o no a la búsqueda y si en efecto resultan relevantes.

En Diciembre de 1994, Kleiner, Perkins, Caulfield, Byers y una empresa constituida por capitales de riesgo invirtieron en Excite! USD 4000 para la compra de los primeros equipos. Su lanzamiento en Internet fue un año después (Diciembre de 1995).

A mediados de 1996 adquieren al Motor de Búsqueda Magellan y a fines del mismo año adquieren WebCrawler[61]. En ese mismo año ganaron acuerdos de distribución exclusivos con empresas como Netscape, Microsoft y Apple Computer. En ese entonces Excite se caracterizaba por ofrecer búsquedas basadas en palabras claves o basadas en conceptos (no sólo buscando los términos deseados por el usuario sino también los similares).

En 1999, Excite combinado con el servicio de alta velocidad de internet @Home.com llego hacer Excite@Home, pero su fusión afectó de manera considerable las perspectivas de la empresa, debido a intereses de proyección un tanto dispares entre ellas.

Hacia el 2001, había tensión entre los inversionistas principales y se inicio un incremento de pérdida de dinero en efectivo en la empresa. En octubre de ese mismo año, Excite@Home presenta una solicitud de declaración de quiebra. A partir de ahí, iWon.com, empresa de New York basada en Internet dejó caer todos los proyectos en desarrollo y comenzó a construir completamente un nuevo pero familiar sitio web "Excite". Unas semanas mas tarde, la empresa hizo una oferta conjunta con Infospace, (una empresa de Seattle) para la compra del dominio y la marca Excite y fue en noviembre de ese mismo año que el tribunal acepta la oferta y dió iWon la oportunidad de lanzar el nuevo portal Excite.

[61] Actualmente, Webcrawler y Excite se encuentran fusionados por la misma Base de datos. De modo que los resultados de cada motor son idénticos aun cuando su interfaz y el sentido de cada sitio sean diferentes. Por tanto Excite y Webcrawler se encuentran catalogados con las mismas clasificaciones en cuanto a resultados de búsqueda se trata.

A principios del 2002 Excite mantiene sus características y capacidades de portal invariables pero substituye su motor original de búsqueda por Dogpile de InfoSpace. A partir de ese momento, Excite es considerado un meta buscador debido al uso de la misma tecnología subyacente de otros motores de búsqueda provenientes del meta de InfoSpace, cuyos resultados eran recibidos desde Google, LookSmart, Inktomi, Ask Jeeves, About, Overture, FindWhat y eventualmente FAST.

La red Excite, Irvington, New York, fue adquirido por Ask Jeeves, Inc. en el 2004. Excite ha sido un sitio predominante de propiedad interactiva de búsqueda, cuyas marcas y filiales incluye Excite, iWon, Inc., My Way, My Search, My Web Search and the MaxOnline. Al inicio del 2004, Excite fue clasificado y considerado el noveno sitio Web mas visitado en los Estados Unidos. Mientras los sitios combinados de su entonces Propiedad paternal, Interactiva de búsqueda, alcanzada fue de tan solo el 17% de todos los usuarios estadounidenses De Internet.

Actualmente Excite cuenta con tecnología metasearch que retorna automáticamente una combinación de resultados de los principales motores de búsqueda del Web como Google, Yahoo! y Ask Jeeves. Es decir ofrece un espectro más amplio y más relevante de resultados que se puede conseguir en comparación con la utilización de un solo motor de búsqueda. El portal Excite no solo tiene cobertura en los Estados unidos sino también tiene entradas en todos los principales mercados europeos entre ellos tenemos Italia, Alemania, Inglaterra, Francia, España, Japón, Austria entre otros.

Excite cuenta con diferentes tecnologías de búsqueda entre ellos tenemos:

- Motores de búsqueda que incluyen Google, Ask Jeeves, Yahoo y Teoma.
- Directorio Web normalmente seleccionados y clasificados en forma manual. Provenientes de About, Looksmart and Open Directory.
- Pay-For-Placement: Son motores de Búsqueda que devuelven listados relevantes de patrocinadores pagados. Motores pagados, por los que Excite realiza la búsqueda incluye Overture and FindWhat.

C.6.2. FUNCIONAMIENTO

Excite presenta en su sitio tres modalidades de búsqueda: Excite Search (Búsqueda), la cual realiza búsquedas sobre los listados generados por su motor de búsqueda en la forma tradicional de recorrer el Web y obtener información de las páginas visitadas. Channels (Canales), el cual se orienta a entregar información por temas y finalmente Excite News Tracker (Noticias), que permite generar listados sobre sitios dedicados especialmente a entregar información de noticias.

- Directorio: Ofrece revisiones del Web de sitios clasificados y organizados temáticamente, considerado como los más altos en la red en una variedad de categorías o en algunos casos denominados canales permitiendo acceder a una variada gama de servicios. El directorio esta formado por 22 canales de acceso a diferentes tipos de informaciones (ubicados en "Explore Excite" de los cuales encontramos Salud, Noticias, informática, educación etc.,). Estas categorías se subdividen en otras más específicas.

- Motor de Búsqueda:

Búsqueda simple: Se ingresan las palabras separadas por espacios en blanco dentro del formulario y se da click en el botón search. Excite localiza los documentos que contengan las palabras introducidas o la mayor parte de ellas. Sin embargo se tiene la opción de colocar el símbolo (+/-) para la que obligue a que se encuentre o se excluya los documentos que contengan dicha palabra.

Búsqueda avanzada: Excite tiene una página especial búsqueda avanzada que le permite aumentar la exactitud de la consultas. Permite filtrar por el dominio, por idioma o por la fecha. En la tabla32, se especifica con más detalle las prestaciones de este gran metabuscador.

Entre otras características que Excite proporciona en la pagina de resultados, podemos destacar:

Are You Looking For? Que despliega los resultados más probables para la búsqueda basada en una colección exclusiva de datos de búsqueda y análisis. Website Match: Identifica la dirección de un sitio web (URL) que se relaciona con el término de búsqueda. Related Content : Despliega una colección de links que se encuentran relacionados con el

termino de búsqueda. Y por último Recent Searches : Guarda el rastro de las 15 búsquedas más recientes.

excite Search	**TABLA32. PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.excite.com)**
	Información general
Tipo de Servicio (Motor/Directorio):	Motor de búsqueda pero tambien directorio. - Motores de búsqueda que incluyen Google, Ask Jeeves, Yahoo y Teoma. -Directorio provenientes de About, Looksmart and Open Directory.
Acceso(libre,comercial):	Libre
Nro de URLs:	No mencionado.
Origen:	Excite fue inicialmente desarrollado y administrado en California en febrero de 1993 y denominado en ese momento Architext. Fundado por Mark Van Haren, Ryan McIntyre, Ben Lutch, Joe Kraus, Graham Spencer y Martin Reinfried(cinco hackers y un experto en ciencias políticas). Su lanzamiento en Internet tuvo lugar en Diciembre de 1995.
Frecuencia de actualizacion de la base de datos(mensual, semanal, quincenal,diaria):	Una vez al mes.
Nº de paginas que recoge por dia	No mencionado.
Tiempo que dura en dar las paginas web registradas:	2-4 semanas.
Versión analizada:	Ingles. Pero existe la version en castellano (http://www.excite.es).
Portal:	Si.
Obedece el protocolo de exclusión(Sirven para indicarle a los robots que la página que los contiene no debe ser indexada y/o que los enlaces de tal página no deben ser seguidos).	Si hace uso del estandar robot.txt en el servidor para permitir explorar algunas o ninguna de las partes de algun sitio específico.
	Recolección
Robot:	Spider
Método(Humano/Automático):	Automático y manual
Primero en profundidad:	No aplica
Primero en Anchura:	No aplica
Tipo de cobertura (www, gopher,WAIS, ftp,telnet,IRC, UseNet News, productos multimedia, etc.,):	Permite la búsqueda sobre Noticias, Web, Audio, Imagenes y video.
Cobertura geográfica:	cobertura Internacional.
Objeto de Cobertura (general o contenido especializado):	General.
Frecuencia de actualizacion para visitar los mismos sitios/documentos:	Diariamente.
Otras bases de datos:	Búsqueda de imagenes son suministradas por Yahoo! Images y Ditto. Búsqueda de Audio Search proporcionada por Yahoo! Audio y Singingfish. Búsqueda de Video proveido por Yahoo! Multimedia y Singingfish. Búsqueda de Noticias suministrada por Yahoo! News, Topix, FoxNews y ABC News.

	**PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.excite.com)**
Indexación	
Que indexa (titulos, cabeceras, información de cabecera, Links(URLs),etiquetas HTML, resumenes, texto completo, etc.,):	Excite indexa los documentos que tengan una mayor concordancia con las palabras de su consulta. Igualmente buscará documentos que traten sobre los mismos conceptos que describe su consulta, por lo que algunas veces Excite devuelve artículos que no mencionen ninguna de las palabras de su consulta original.
Que no indexa:	Si los motores de búsqueda detecta una técnica spamming, ellos pueden degradar la clasificación de una página o lo excluyen de los listados totalmente.
Sistema de Recuperación	
Herramienta de Búsqueda:	Motores de búsqueda e indice temático.
Tipo de Recuperación (Booleana, Vectorial, Probabilística, otro):	Manual, Booleana(e Inteligencia artificial).
Estructuras de consulta y operaciones soportadas:	Consignada en la tabla 33.
Criterio de Búsqueda:	Número de veces que la palabra aparece en el documento, en qué campos aparece (título, texto), Número de veces que este documento está referenciado en otros.
Sistema busca *por default/por campos seleccionables(*URLs, titulos,Resumen, Texto completo, otros.,)	Esto es dependiente de las reglas de cada motor al cual Excite se encuentra conformado.
Mejora en la Busqueda (reducción de resultados, reformulación, vocabulario controlado, mediante ejemplo,otros.,):	Si en una búsqueda se consiguen demasiados resultados vagos, Excite permite refinar los resultados gracias a su tecnología de clasificación automática. Esta especial caraterística, alinea automáticamente los resultados en categorías basadas en palabras y frases contenidas en las búsquedas.
Ordenamiento:	Excite alinea los resultados de dos formas: La manera tradicional, donde agrupa los mas altos resultados subyacentes de cada uno de los motores de búsqueda y el otro modo, permite clasificar los resultados por relevancia; en esencia, resultados por las cuales la mayor parte de los motores de búsqueda han votado como los mas altos de la lista.
Despliegue de resultados: (Título, descripción,URL, tamaño,fecha de alta, Nº total de saltos, correspondencia de términos,orden de relevancia,valor de relevancia,Nº de saltos por pagina, formato variable, detecta novedades, permite traducir,otros.,):	Permite ver título con el enlace, la URL especificando de que motor se recuperó, las primeras palabras a modo de resumen, correspondencia de términos, Nro total de saltos, la salida se compone por defecto de 20 resultados por pagina, en caso de elegir por relevancia; si se desea que se agrupe por motor de búsqueda éste se encuentra determinado por los resultados que cada motor retorne. Igualmente el usuario puede elegir los criterios de relevancia en el orden de presentación de los resultados.

PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (www.excite.com)

Interfaz de Usuario	
Descripción de la Interfaz	Menus estructurados, desplegables.
Claridad de la interfaz y de la pagina de búsqueda:	Interfaz muy llamativa con muchas posibilidades de búsqueda sobre cualquier tema en particular. Y ofrece personalizacion de búsquedas de las mas avanzadas entre otros motores de búsqueda hoy disponibles.
Ofrece ayuda en cuanto a la documentacion de búsqueda:	La ayuda que brinda es muy completa y detalla, en la que se viusaliza una serie de links que direccionan a paginas donde se explica la manera de como encontrar exactamente lo que se quiere buscar en Excite. Sin embargo se encuetra en ingles.
Busqueda Simple/Avanzada	Ambas
Establece preferencias	Si permite personalizar preferencias y guardar la configuración.
Idiomas Interfaz:	10 idiomas.
Servicios:	Proporciona servicios de búsqueda, directorio, información y de diversión, imágenes, audio/MP3, vídeo, noticias, foros, informacion sobre el tiempo entre otros.
URL Pagina de Ayuda:	Motor de Búsqueda: http://www.infospace.com/info.xcite/search/help/faq.htm. Directorio:http://help.excite.com/
URL Pagina deTraduccion:	No aplica.
Ranking:	El ranking de Excite es basado sobre la relevancia y el performance. Si un particular motor de búsqueda no procesa una consulta rapidamente, sus resultados no seran incluidos en él.

TABLA33. ESTRUCTURAS DE CONSULTA Y OPERACIONES SOPORTADAS

Operador default:	Usa por defecto el operador OR.
Operador de Existencia: (exclusión/inclusion)	-/+
Operadores Booleanos:	AND, OR, NOT y AND NOT
Operadores de Proximidad:	
Búsqueda por campos:	Los campos admitidos site:, url:
Operadores de exactitud:	Permite el uso de comillas " ". Pero tambien se cuenta con la opcion para habilitar frase exacta ubicada en la parte inferior del formulario de introduccion de los términos de búsqueda después de que la consulta es ejecutada por primera vez.
Lenguaje Natural:	No mencionado.
Truncamiento:	No admitido.
Agrupacion de términos y operadores:	Si aplica
Sencible mayúsculas/minúsculas:	No Aplica.
Reconoce sinónimos:	Si aplica.
Caso sensitivo a acentos, puntuación:	No Aplica.
Filtro por idioma:	Si Aplica.
Filtro por formato:	No Aplica.
Filtro por Dominio:	Si Aplica. Permite el filtro por extenciones de dominio como .com, .gov and .edu, y/o la inclusión de un dominio especifico.
Filtro por numero de resultados de páginas:	Si Aplica. Por defecto 20, sin embargo se tiene opciones de 10,15,18,20,30 y hasta 40 resultados por pagina.
Filtro por fecha:	Si Aplica. Permite filtrado por intervalo de tiempo.
Creador de consultas:	Si aplica. Puede crear fácilmente potentes consultas introduciendo palabras en una combinación de estos cuadros de texto(todas las palabras, frase exacta, cualquier de estas palabras, ninguna de estas palabras).
Permite bloqueo de contenido ofensivo:	Si aplica. Mediante el filtro "Adulter Filter".

C.7 MOTOR DE BÚSQUEDA " MSN SEARCH "

MSN es líder mundial en servicios Web para usuarios y ventas de publicidad para negocios en todo el mundo. Como el servicio en línea más útil e innovador actualmente, MSN proporciona a los clientes todo lo que necesitan de la Web para aprovechar al máximo su tiempo en línea.

MSN Search! Se basa no sólo en proporcionar a los usuarios una larga lista de enlaces Web, sino en proporcionarles acceso a las respuestas e información que buscan. Hoy en día MSN Search cuenta con un nuevo motor que incluye, índice y rastreador de búsqueda, todo ello creado desde cero con tecnología Microsoft. Este nuevo servicio, es capaz de encontrar lo que estás buscando de una manera más rápida.

"la red de servicios de MSN actualmente se concentra en 25 mercados y en 10 idiomas con el servicio de búsqueda MSN Search".

C.7.1. ORIGEN Y DESCRIPCION

MSN Search fue desarrollado por un equipo de trabajo de Microsoft en Octubre de 1998, en ese entonces se destacaba su directorio que era analizado y organizado de manera manual. Sin embargo para la recuperación de resultados más precisos ante una consulta dada, MSN Search hacía uso de algoritmos de búsqueda que eran impulsados principalmente por Looksmart y por Inktomi, obteniendo de esta forma resultados de sitios más relevantes.

En julio de 1999 Microsoft adiciona Direct Hit a MSN Search, que provee los resultados en base a la popularidad, es decir priorizaba los enlaces que los usuarios mas usaban. De este modo, se lograba dar una mejor experiencia en la búsqueda con sitios más precisos y relevantes.

En octubre del 2000, MSN Search lanza su nueva versión e Inktomi es reemplazado. Por tanto, el directorio LookSmart y el motor de búsqueda AltaVista son en ese entonces sus bases de datos principales. Sin embargo a finales de este año, MSN search regresa nuevamente con la base de datos InKtomi, convirtiéndose en un competidor fuerte frente a su rival Altavista. Su decisión se fundamenta en la mejorada funcionalidad en las búsquedas al integrar nuevos métodos interpretativos de búsqueda y etiquetas que lo hacen más fácil y más rápido encontrar la información de muchas maneras, Incluso, a su amplio contenido como portal en el que se incluyen una riqueza de noticias actualizadas, música y clips vídeo entre otros.

En el 2003, MSN Search hace uso de la base de datos Overture para anuncios publicitarios, sus principales resultados son impulsados por LookSmart, e Inktomi para resultados secundarios que ha sido utilizado durante años. Adicionalmente Microsoft inicia con la creación de una nueva base de datos construida con su propio robot. El nuevo MSNBOT constituye en un avance lento pero promisorio en el Web, sin embargo por ser un prototipo, ninguno de sus documentos recuperados directamente provenían de la base de datos MSN Search.

En enero del 2004 LookSmart deja de ofrecer su resultados sobre MSN Search y sus resultados principales fueron por tanto, suministrados por Inktomi. En noviembre, MSN Search finalmente comienza a añadir una búsqueda de noticias. Sin embargo, es sólo en la beta y está sólo disponible en los sitios de MSN Search para el Reino Unido, Francia, Italia, y España. El cual era impulsado por Moreover, que solicitaba información de más de 4,000 fuentes. En abril, MSN Search proporciona el acceso a listados de Yahoo, pero no con tanta funcionalidad en términos de otros tipos de búsquedas que Yahoo ofrece por sí mismo. Sin embargo, MSN desarrolla y madura su propia tecnología basado en robot y planea otros cambios que permitirán revivificar el servicio en el posterior año al 2004.

El 11 de noviembre del 2004, MSN Search lanzó su nuevo, única base de datos de su motor de motor de búsqueda beta.search.msn.com. A diferencia de anteriores pruebas, esta versión ofrecía características de búsqueda avanzada bajo el enlace de "Search Builder" incluyendo filtro de resultados por sitio, país y región, idioma y permitía ajustar la clasificación de los resultados, el manejo de operaciones booleanas, búsqueda por campos entre otras opciones.

En Febrero de este año, la empresa estadounidense Microsoft lanza formalmente su buscador en Internet 'MSN Search' en 25 países, varios meses después de pasar por un periodo de pruebas, este nuevo motor algorítmico reemplaza la tecnología de Yahoo, mejora la personalización, resuelve operaciones matemáticas y busca en fuentes externas. El portal MSN también cambió para acelerar la carga y dar más relevancia al buscador. No obstante, el buscador de Yahoo! se sigue usando para las listas pagadas por compañías como anuncios.

Actualmente Estados Unidos cuenta plenamente con todas las herramientas de búsqueda de MSN Search, incluido el servicio de respuesta directa (Direct Answers), el de búsqueda local “Cerca de mí” (Near to me), el acceso a los más de 40.000 artículos de MSN Encarta, la integración con MSN Music, o la búsqueda de archivos en el disco duro, entre otros. Sin embargo en los próximos meses se incorporarán plenamente estos servicios en todos los mercados, incluido España.

C.7.2. FUNCIONAMIENTO

MSN Search es el motor de búsqueda para el sitio portal MSN. Durante años había usado bases de datos de otros vendedores en estos, se incluye Inktomi, LookSmart, and Direct Hit. Desde el 1 de Febrero del 2005, este comenzó a usar su base de datos propia y única.

Entre varias características nuevas que ofrece el motor de búsqueda MSN Search y que permite encontrar lo que se busca en un tiempo relativamente corto, se debe a la tecnología de Microsoft creada desde la base para poner a disposición lo mejor del Web. Entre nuevos servicios de búsqueda incluyen:

- Resultados rápidos y relevantes, donde se puede encontrar información en areas: como Web(donde se escribe uno o varios términos de búsqueda y, a continuación, se da click en el botón Search); MSN Newsbot (Se da click en News y examina las noticias y los titulares más recientes); imágenes (donde se escribe los términos que describen la fotografía de búsqueda y, a continuación, se da click en Images; de igual forma esta el area de Music y Desktop para encontrar información de su computador local.

- Más opciones de idiomas y cobertura ampliada en todo el mundo, éstas características facilitan el acceso al contenido más relevante según su ubicación. También se puede encontrar contenido en uno o varios idiomas o de uno o varios países o regiones.
- Generador de búsquedas(Search Builder), que permite hacer uso palabras clave especiales con operadores booleanos, manejo de operadores de campo y demás opciones que ayudan a crear búsquedas más precisas.

En seguida se muestra en la tabla 34, las prestaciones que ofrece el motor de búsqueda MSN Search:

TABLA34. PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BUSQUEDA (search.msn.com)	
Información general	
Tipo de Servicio (Motor/Directorio):	Motor de búsqueda. Usa su propio motor para impulsar los resultados de búsqueda.
Acceso(libre,comercial):	Libre
Nro de URLs:	5 billion
Origen:	Fue desarrollado por Microsoft en octubre de 1998.
Frecuencia de actualizacion de la base de datos(mensual, semanal, quincenal,diaria):	mensual.
Nº de paginas que recoge por dia	No mencionado.
Tiempo que dura en dar las paginas web registradas:	1-2 semanas aprox.
Versión analizada:	Ingles. Pero existe la version en castellano.
Portal:	Sí.
Obedece el protocolo de exclusión(Sirven para indicarle a los robots que la página que los contiene no debe ser indexada y/o que los enlaces de tal página no deben ser seguidos).	El registro del servidor contiene el mensaje "Archivo no encontrado" para robots.txt o etiquetas META para controlar cómo indizan un sitio MSNBot y otros rastreadores. El archivo robots.txt informa a los rastreadores Web qué archivos o carpetas no pueden rastrear.
Recolección	
Robot:	Robot o agente de búsqueda llamado MSNBot.
Método(Humano/Automático):	Automático
Primero en profundidad:	No aplica.
Primero en Anchura:	No aplica.
Tipo de cobertura (www, gopher,WAIS, ftp,telnet,IRC, UseNet News, productos multimedia, etc.,):	Web, Usenet News, productos multimedia (imágenes, musica, etc,.)
Cobertura geográfica:	cobertura Internacional.
Objeto de Cobertura (general o contenido especializado):	General.
Frecuencia de actualizacion para visitar los mismos sitios/documentos:	Peridicamente.
Otras bases de datos:	Utiliza a Yahoo! Search Marketing Solutions (formalmente conocido como Overture). Para listas pagadas para efectos publicitarios. Cabe mensionar que MSN Newsbot, busca más de 4.800 fuentes nuevas en todo el mundo para ofrecer la cobertura más actualizada.
Indexación	
Que indexa (titulos, cabeceras, información de cabecera, Links(URLs),etiquetas HTML, resumenes, texto completo, etc.,):	Indiza el contenido completo de las paginas.
Que no indexa:	Frames, Spam, es decir MSN Search no indiza páginas con palabras irrelevantes en un intento de aumentar la densidad de palabras clave de una página. Tampoco paginas que utilicen texto o vínculos ocultos.

msn Search Search	**PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BUSQUEDA (search.msn.com)**
Sistema de Recuperación	
Herramienta de Búsqueda:	Motor de búsqueda.
Tipo de Recuperación (Booleana, Vectorial, Probabilística, otro):	Usa la recuperación Booleana.
Estructuras de consulta y operaciones soportadas:	Consignada en la tabla 35.
Criterio de Búsqueda:	El algoritmo de clasificación de MSN Search analiza el contenido de las páginas, la cantidad y calidad de los sitios que tienen vínculos con tus páginas y la relevancia del contenido del sitio con respecto a palabras clave. Se trata de un algoritmo complejo y sin intervención humana. Además utiliza un conjunto único de instrucciones para asignar a cada página una clasificación dentro del índice. La clasificación viene determinada por muchos factores, pero los más importantes son, entre otros: El idioma en el que el sitio está escrito, la calidad y la cantidad del contenido.
Sistema búsca *por default/por campos seleccionables*(URLs, titulos,Resumen, Texto completo, otros.,)	Urls, titulos, Descripción, contenido de las paginas.
Mejora en la Busqueda (reducción de resultados, reformulación, vocabulario controlado, mediante ejemplo,otros.,):	Contemplan la posibilidad de 'Afinar' la búsqueda mdiante la búsqueda Avanzada.
Incluye textos alternativos(inf. similar):	Si aplica. Cuando no se logra encontrar lo que se busca, MSN Search muestra algunos sitios Web relacionados que pueden ayudar a llegar hasta la información que se esta buscando.
Ordenamiento:	MSN Search permite cambiar el orden de los resultados de la búsqueda con clasificación de resultados. Por medio del control deslizante del ecualizador, se puede modificar la búsqueda y dar mayor prioridad a los sitios que se han agregado recientemente al índice de búsqueda y/o dar prioridad a los sitios en función del número de sitios que incluyen vínculos a ellos y/o dar la máxima prioridad a la coincidencia entre las palabras de búsqueda exactas y los resultados. La dirección que se personalice reducirá la importancia de los otros dos criterios de clasificación del control deslizante. (Ver Search Builder)
Despliegue de resultados: (Título, descripción,URL, tamaño,fecha de alta, Nº total de saltos, correspondencia de términos,orden de relevancia,valor de relevancia,Nº de saltos por pagina, formato variable, detecta novedades, permite traducir,otros.,):	Despliega títulos descriptivos que reflejan los términos de búsqueda utilizados, las direcciones Web (direcciones URL) en las que se encuentran los sitios, Fecha que indica cuándo indizamos por última vez el sitio Web, Vínculos a una página almacenada en la memoria caché. Vínculos numerados para navegar entre las páginas de resultados, situados en la parte superior de cada página. Resalta tus términos de búsqueda en los resultados, el número de resultados que despliega por página es 10.

PRESTACIONES E INFORMACION GENERAL DEL MOTOR DE BÚSQUEDA (search.msn.com)

Interfaz de Usuario	
Descripción de la Interfaz	Menus estructurados, desplegables(Busqueda Avanzada).
Claridad de la interfaz y de la pagina de búsqueda:	Interfaz es agradable visualmente y facil de usar.
Ofrece ayuda en cuanto a la documentacion de búsqueda:	La ayuda detalla la forma de realizar las búsqueda avanzada, los operadores a utilizar y los criterios de relevancia que se aplican en la presentación de resultados entre otra información correspondiente a su funcionamiento interno como motor de búsqueda, servicios disponibles, etc.,
Busqueda Simple/Avanzada	Ambas.
Establece preferencias	Si permite establecer preferencias y guardar la configuración, en la opción "Settings" .
Idiomas Interfaz:	12 idiomas.
Servicios:	MSN Hotmail, MSN Premium/ MSN Plus, Barra de Herramientas MSN, MSN Health, MSN Messenger, MSN Mobile, MSN Dinero, MSN Wallet, MSN Direct, MSN Zone.
URL Pagina de Ayuda:	http://search.msn.com/docs/default.aspx?FORM=HLHP
URL Pagina deTraduccion:	No aplica.
Ranking:	El ranking de MSN Search, esta determinado en base a la clasificación por relevancia. Es decir, los sitios que se actualizan periódicamente, que tienen una gran cantidad de contenido y a los que muchos otros sitios tienen vínculos obtienen una clasificación más alta que los demás.

TABLA35. ESTRUCTURAS DE CONSULTA Y OPERACIONES SOPORTADAS

msn Search

Operador default:	AND
Operador de Existencia: (exclusión/inclusion)	Incluye los terminos (+ / -) respectivamente.
Operadores Booleanos:	AND (o el simbolo ampersand &), OR (o elsimbolo pipe |), y NOT. Cualquiera de los dos, AND NOT o NOT pueden ser usados.
Operadores de Proximidad:	No aplica.
Búsqueda por campos:	Si aplica. site:, language:, loc: or location:, link:, url:
Operadores de exactitud:	Permite el uso de las comillas, " " alrededor de la frase de búsqueda, para acotar resultados.
Lenguaje Natural:	No mencionado.
Truncamiento:	No admitido.
Agrupacion de términos y operadores:	Permite términos de búsqueda entre paréntesis, si hay otro término de búsqueda que les afecta a todos. Los términos que se encuentran dentro del paréntesis se tratan como una unidad.
Sencible mayúsculas/minúsculas:	No aplica. Haciendo uso de mayúsculas y minúsculas se obtienen los mismos resultados.
Reconoce sinónimos:	No aplica.
Caso sensitivo a acentos, puntuación:	No aplica.
Filtro por idioma:	Si aplica.
Filtro por formato:	No aplica.
Filtro por vinculos:	Permite buscar páginas Web que contengan vínculos a una dirección URL (dirección Web específica).
Filtro por Sitio/dominio:	Permite buscar sólo páginas Web de este sitio o dominio o Excluir páginas Web de este sitio o dominio.
Filtro por numero de resultados de páginas:	Si aplica. Permite mostrar 10, 15, 30 o 50 resultados por pagina. Además permite agrupar resultados del mismo sitio. Mostrando los primeros 1, 2 o 3 resultados.
Filtro por Región/País	Encontrar páginas Web en una ubicación específica. Se tiene para seleccionar 22 posibles paises y/o regiones.
Filtro por fecha:	Si aplica. Dispone de bien filtra por rango o por semanas, meses o un año.
Creador de consultas:	Puede crear fácilmente consultas introduciendo téminos en una combinación de cualquiera de estas opciones (todas las palabras, frase exacta, cualquier de estas palabras, ninguna de estas palabras).
Clasificador de resultados:	Mediante un control deslizante del ecualizador puede determinar la clasificación en función a la actualización mas reciente, mas popular o a la coincidencia mas exacta.
Permite bloqueo de contenido ofensivo:	Si Aplica. En el enlace "Settings page" que comunmente conocemos como preferencias de configuración, permite bloquear sitios para adultos y el contenido sexual explícito en los resultados de búsqueda (Estricta, Moderada, Desactivada).

CONCLUSIONES

La propuesta empleada, posibilita y permite analizar la eficacia en el funcionamiento de los motores de búsqueda de Internet generales e internacionales, los cuales fueron objeto de estudio siete específicamente. En la misma se pudo observar resultados bastante razonables, en relación con otras informaciones recientes, ofrecidas por algunos reconocidos sitios de Internet (ver sitios Search Engine Showdown y Search Engine Watch), facultadas en ésta línea de investigación. Por tanto, estas técnicas vislumbran y demuestran la viabilidad y facilidad de adaptar técnicas ya existentes de evaluación de la recuperación de información, a los servicios de búsqueda en Internet.

Es importante hacer énfasis que este experimento es en realidad, una muestra a pequeña escala en cuanto a resultados comparativos se refiere. Esto no quiere decir, que carezca de importancia, confiabilidad y eficiencia al momento de aplicar los criterios preestablecidos, sino que su trabajo es bastante arduo y repetitivo, lo cual requiere de más personal para abarcar mayores consultas, más motores de búsqueda, mayores volúmenes de documentos a evaluar, que permitan dar un análisis menos subjetivo y más absoluto, con respuestas mas certeras sobre los alcances en cuanto recuperación útil se refiere y sobre todo aplicar infinidad de pruebas que por tiempo quedaron latentes pero que de igual forma resulta interesante abarcar en futuras investigaciones.

Como todo proyecto siempre debe haber una serie de etapas que faciliten obtener alcances de un objetivo propuesto; por eso, en este experimento empírico, se llevó a cabo una metodología conformada y organizada, donde inicialmente se plantea y detalla la forma como se llevará a cabo la propuesta de evaluación y termina con la puesta en práctica de los criterios previamente preestablecidos. La serie de criterios propuestos, dio lugar a un enfoque multidimensional para dar análisis integral sobre estos sistemas de recuperación de información.

Como se vió, en los apéndices se da cierta información sobre qué son los sistemas de recuperación de información, documentación clave para poder interpretar el origen del tema de esta monografía; seguidamente se hace un paralelo entre los sistemas de información de datos y los sistemas de información documental, que aunque claramente tienen diferencias en su funcionamiento los segundos sirvieron de base para el desarrollo de los sistemas de recuperación Web. También se dió una visión global sobre los sistemas de recuperación de información de Internet hoy a nuestro alcance, logrando profundizar un poco, a cerca de sus particularidades, funcionamiento y cuando hacer uso de ellos. Posteriormente después de una investigación ardua, se pudo dar un análisis evolutivo de los motores de búsqueda elegido en este estudio, para lograr con mas detalle la dinámica llevada desde su creación hasta nuestros días, lo que reitera el constante avance y variabilidad en lo que respecta al entorno Web, que les exige estar a la Vanguardia de suplir en mayor medida las necesidades de los usuarios como crear nuevos servicios que les permita ser mas competitivos; en esta información también se resume las prestaciones propias de cada buscador, características claves para entender más su forma de indexación y la manera como logran recuperar resultados relevantes producto de una búsqueda realizada.

En lo que respecta a la metodología de evaluación de la recuperación de información de los motores de búsqueda en el entorno de la World Wide Web, hay que recordar que las consideraciones principales llevadas a cabo para el desarrollo del mismo:

a) Planteamiento e incorporación de preguntas reales, que permitieran exigir a los diferentes buscadores ante la ingente y heterogénea información disponibles en su base de datos, la recuperación de documentos en dos idiomas (Ingles y Español), manejo de temas tanto específicos como generales y desempeño frente a sintaxis de consulta variada (frase literal, booleanas, etc.,) y disponibles en todos los buscadores para asegurar la homogeneidad y viabilidad en los resultados obtenidos.
b) El uso de las medidas de relevancia, exhaustividad y precisión para evaluar los Sistemas de recuperación de información.

c) Manejo de niveles de relevancia que permitieran facilitar separación de lo que corresponde al ruido documental a lo que realmente es información útil, a una necesidad informativa dada.

Aunque la relevancia tiende a considerarse por algunos autores como un concepto afectado de una gran dosis de subjetividad que puede ser explicado de múltiples maneras por distintas personas y, por tanto, dentro del contexto de una búsqueda en un sistema de recuperación de información, este concepto puede ser precisado de muchas maneras distintas por todos aquellos que realicen búsquedas en el mismo. No obstante, cabe mencionar que en este estudio fue especificado el grado de relevancia, determinándose como una medida en términos de función continua (óptima, relevante, escasamente relevante) y en la que se especifica con un ejemplo el significado al que se categorizaba la relevancia, permitiendo delimitar una necesidad de información en un contexto más específico para su respectivo análisis.

d) Manejo del método de Salton y MeGill para su aplicación a los buscadores web, como medida para lograr resultados a los pares de valores Exhaustividad-Precisión que permitiera lograr el indicador de rendimiento de estos sistemas de recuperación. Se estableció para este criterio el análisis de los 10 primeros documentos recuperados y de acuerdo a su alineamiento determinar la Exhaustividad (Respuesta) y Precisión (Pertinencia). De igual forma para dar una aproximación más representativa de los valores E-P, se usó la regresión logarítmica, por ofrecer los mejores resultados de ajuste en este criterio. También se corroboró que a causa de la gran cantidad de información que se acumula en las bases de datos, y a la intervención de múltiples factores que hacen de la indización un proceso muy complejo, los procedimientos de recuperación no pueden ser nunca íntegramente exhaustivos y precisos.

e) Sea trabajado el criterio de similitud de resultados en los diez primeros documentos. Se deja para a objeto de estudio, trabajar con un mayor número de documentos analizados con el fin de obtener un valor de similitud mayor en los resultados. No obstante en la prueba se llegó a resultados donde la similitud media más alta era obtenida por los pares de motores Yahoo-Altavista y Google-Hotbot. Además con la similitud obtenidas se pudo hallar la distancia (1- Similitud), permitiendo de manera hipotética establecer un centro considerado como el núcleo de la mejor colección de documentos web y mediante una distribución radial establecer cual esta más cercano a ese centro ideal. Finalmente se

utilizó una técnica de agrupamiento, que permitiera visualizar los motores de búsqueda mas afines, lográndose 5 agrupamientos que permiten establecer el porcentaje de sus índices los hacen absolutamente diferentes entre sí.

En resumen con base a los resultados obtenidos, en este estudio, se reitera la tendencia de los motores de búsqueda a ser sistemas de recuperación de información mas exhaustivos que precisos sin embargo se notó un mayor grado de precisión en motores como Yahoo, Altavista, Google y Hotbot y en algunos casos Excite poniendo de manifiesto, mayores posibilidades de encontrar información útil durante sus búsquedas. Los resultados manifiestan la coherencia frente a esta afirmación, gracias al método utilizado.

En términos generales podemos sugerir el ranking de rendimiento de los motores de Búsqueda de mayor a menor nivel de la siguiente manera: Google, Hotbot, Altavista, Yahoo, Excite, MSN Search y Lycos. Debido a que algunos motores potencializan parte de su base de datos a otros motores con el fin de mejorar su servicio en el mercado, se puede suponer que de alguna manera esto ha influido en la similitud en sus resultados (Ej: Yahoo → Altavista, Google → Hotbot). En este estudio, Google, Hotbot, Yahoo y Altavista, se identifican como los mejores motores, en cuanto al comportamiento relacionado con la efectividad de recuperación de la información, lo cual justifica que estén entre los 10 motores más populares y exitosos en la Internet.

Excite no se destacó entre los mejores ni entre los peores, en conjunción con todos los criterios analizados, lo que hace suponer que el algoritmo interno (Spider) no es del todo apropiado, sin embargo, su reducido índice de alguna forma también influye un poco en los resultados al no poder indexar una mayor cantidad de documentos en la Red. MSN Search también no tuvo una influencia predominante, pero tampoco negativa del todo. Sin embargo hay que resaltar su nuevo índice que fue puesto al público a principios de este año; el algoritmo que maneja es propio de Microsoft y este fue creado desde cero, por tanto apenas está en etapa de maduración pero con muchas posibilidades de ser un gran competidor como herramienta de búsqueda Web. Lycos en comparación con los demás buscadores, fue el que se vió mas confinado en cuanto a calidad de sus resultados.

En cuanto al ruido documental precedente (enlaces inactivos y duplicados) se percibió la existencia prevaleciente de este problema en todos los motores de búsqueda. Aunque en este caso vimos que Lycos, Excite y MSN Search presentaban los mayores valores promedios de ruido documental en relación a los demás, superando mas del 10% este aspecto sugiere dificultad o cierto descuido, para mantener perfectamente actualizado sus índices existentes en sus bases de datos, debido a los porcentajes considerados altos sobre los treinta primeros documentos recuperados.

Finalmente puedo concluir, que el trabajo aquí expreso, aborda con claridad y cumplimiento los objetivos propuestos al inicio del mismo. De igual forma es importante señalar que el propósito principal de este estudio no es determinar en sí cuál es el mejor motor de búsqueda presente hoy en día en la Internet, puesto que las conclusiones a las que se pudieron llegar son poco duraderas, sino mas bien establecer una propuesta que pueda aplicarse a la evaluación de los resultados por éstos ofrecidos, permitiendo así la realización de estudios periódicos que determinen y analicen la evolución de estas herramientas de búsqueda tan populares por los internautas.

PROPUESTAS DE EXPLORACION FUTURAS

En torno a este tema, hay diferentes posibilidades que pueden originar útiles e interesantes trabajos de investigación.

- Aunque este estudio se basaba en presentar un seguimiento detallado sobre algunos posibles parámetros útiles para estimar la evaluación de los sistemas de recuperación en Internet, queda abierto la posibilidad en próximas investigaciones abarcar esta metodología en los 20 y 30 primeros resultados, ya que como se vió el énfasis fue básicamente en los 10 primeros documentos. Por otro lado, también se establece como medida de estudio, en cuanto a resultados relevantes, realizar una serie de pruebas para los diferentes grados de relevancia ya establecidos, sin duplicados para los diez, veinte o treinta primeros documentos recuperados, atendiendo a los diferentes criterios establecidos en esta monografía.

- Se deja también como inquietud, realizar el cálculo del tiempo de respuesta que un sistema de búsqueda puede alcanzar cuando se evalúa en primer lugar el retardo obtenido a partir de la ruta de comunicación adoptada para acceder al servidor cuando la pregunta es inmediatamente enviada; en segundo lugar, determinar el retardo conseguido en producir y entregar los resultados al usuario. Por otro lado, la capacidad del entorno y el horario en la cual se realiza la conexión con estos sistemas, es un factor determinante a evaluar de manera reiterativa el comportamiento de respuesta de los mismos, ante una necesidad informativa dada. Por tanto, sería ideal realizar pruebas en diferentes localidades (hogar, oficina y/o universidad) en especial en horas de congestión, debido a que éste enfoque da garantía de percibir un tiempo de respuesta más real de cada sistema.

- Para analizar los servicios de búsqueda a gran escala se deben realizar estudios donde la evaluación sea más meticulosa, es decir donde la cantidad de expresiones de búsqueda y páginas a examinar sean mayores; de tal forma, viabilizar con mas certeza la efectividad del funcionamiento de estos sistemas de recuperación.

- El método aquí aplicado a servicios de búsqueda generales e internacionales podría aplicarse a servicios de búsqueda con cobertura más limitada en razón de su contenido (especializados) o de su cobertura geográfica (nacionales).

- Ciertamente, los resultados de cualquier estudio de la precisión de cualesquiera de los sistemas de recuperación de la información en Internet pierden vigencia con rapidez – debido al carácter tan dinámico de este entorno- Por tanto la necesidad de repetir estudios similares con periodicidad que permitan evidenciar los alcances del funcionamiento y efectividad sobre recuperación de información que prestan estos sistemas.

- Aplicar más parámetros de evaluación existentes, aplicar y comparar métodos empleados para un mismo criterio para confirmar la analogía en los resultados, proponer otros nuevos que permitan evidenciar con mayor firmeza los rasgos diferenciales o rasgos semejantes entre los sistemas de búsqueda a ser comparados.

BIBLIOGRAFIAS

[1] R. BAEZA-YATES AND B. RIBEIRO NETO, Modern Information Retrieval, ACM Press, 1999, ISBN 0-201-39829.
<http://www.sims.berkeley.edu/~hearst/irbook/1/node1.html>

[2] BLAIR, D.C. Language and representation in information retrieval. Amsterdam: Elsevier Science Publishers, 324 p. 1990. ISBN: 0-444-88437-8.

[3] BLAIR, D. C.; MARON, M. E. An evaluation of retrieval effectiveness for a full-text document retrieval system, Communications of the ACM, 28(3), 281-299, 1985.

[4] CACHEDA, F., PUENTES, F. AND CARNEIRO, V. A new performance evaluation technique for web information retrieval systems.
<http://www.tic.udc.es/~fidel/docs/publications/iadis2004.pdf>

[5] CHANG, G., HEALEY, M.J., MCHUGH, J.A.M. y WANG, J.T.L. (2001) Mining the World Wide Web: an information search approach". Norwell, MA: Kluwer Academic Publishers, 2001. <http://www.ler.pucpr.br/~roose/docs/b2.ouksel.pdf>

[6] CHU, H. and ROSENTHAL, M. Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology [En línea] Asis 1996 Annual Conference Proceedings, 1996. <http://www.asis.org/annual- 96/ElectronicProceedings/chu.html>.

[7] C|NET. KEIZER, GREGG. Search engine CNET compares the top 5 engines. <http://gti1.edu.um.es:8080/javima/CNET_com-CNET's-Ultimate-Guide-to-Search.htm>

[8] COOPER, W.S. 'On selecting a Measure of Retrieval Effectiveness'. Journal of the American Society for Information Science, v. 24, March-April 1973. p.87-92

[9] COURTOIS, MARTIN, BAER, WILLIAM, STARK, MARCELLA. Cool tools for searching the Web - a performance evaluation, 1995 pp.14-32.
<http://www.amherst.edu/~seedelbe/search.html>

[10] DELGADO, ADELAIDA M. Mecanismos de Recuperación de Información en la WWW, 1998. < http://servidorti.uib.es/adelaida/tice/modul6/memfin.pdf>

[11] DAVIS, E.T. A comparison of seven search engines. Columbus, OH: Kent State University, 1996. < http://www.iwaynet.net/~lsci/Search/paper.htm>

[12] DESAI, B. C. Supporting discovery in virtual libraries, Journal of the American Society for Information Science, 48(3), 190-204, 1997.
<http://www.cs.concordia.ca/~bcdesai/web-publ/test-of-index-systems-revisited.html>

[13] DING, W.; MARCHIONINI, G. A comparative study of web seareh serviee performance, Proceedings qf the AS'IS Annual Conference, 33. 136-142, 1996.

[14] ENGR. KHALIL AHMED. Computer & Information Center, Royal Saudi Naval Forces, Riyadh. How to find Information On the Internet. IEP-SAC Journal 1999-2000. <http://www.iepsac.org/downloads/p06b.pdf>

[15] FRANTS , V.I. et al. Automated information retrieval : theory and methods. San Diego [etc.] : Academic Press, cop.1997. XIV, 365 p.

[16] GAUCH. S: WANG, G. Information Fusion with ProFusion. Webnet 96 Conference, San Francisco, CA, Oetober 15-19, 1996.
<http://www.csbs.utsa.edu:80/info/webnet96/httnt/155.htm>.

[17] GERARD SALTON, Professor PhD Harvard University, 1958 .
<http://www.cs.cornell.edu/Info/Department/Annual95/Faculty/Salton.html>

[18] GORDON, M., PATHAK, P. "Finding information on the World Wide Web: the retrieval effectiveness of search engines". Information Processing and Management 35, 1999. p. 141-180, <http://www.cindoc.csic.es/cybermetrics/pdf/60.pdf>

[19] GWIZDKA, J. & CHIGNELL, M. Towards information retrieval measures for evaluation of Web search engines. Toronto: University of Toronto, Mechanical & Industrial Engineering Department. (Unpublished manuscript), 1999. <http://www.imedia.mie.utoronto.ca/people/jacek/pubs/webIR_eval1_99.pdf>

[20] HARMAN, D. K. Overview of the Second Test REtrieval Conference (TREC-2), Information Processing and Management, 31(3), 271-289, 1995. <http://trec.nist.gov/pubs/trec2/t2_proceedings.html>

[21] HARMAN, D. K. The TREC Conferences. En: Sparck Jones, K.; Willett, P. (ed.), Readings in infórmation retrieval, San Francisco: Morgan Kaufmann, 1997. p. 247-256, ISBN 1558604545. <http://trec.nist.gov/pubs/trec6/t6_proceedings.html>

[22] JAIMES LUIS GABRIEL. Uso de técnicas de clasificación en conglomerados para describir perfiles en grandes bases de datos educativas, Universidad de Puerto Rico Mayagüez Campus, 2004. http://grad.uprm.edu/tesis/jaimes.pdf

[23] LANCASTER. F. W. Evaluation of the Medlars demand search service. Bethesda, Md.: National Library of Medicine, 1968; e Infórmation retrieval systems: characteristics, testing arrd evaluation. 2.ª ed. New York: Willey, 1979.

[24] LANCASTER, F. W. AND WARNER, A.J. Information Retrieval Today. Arlington, Virginia : Information Resources, 1993.

[25] LEBEDEV, A. Best search engines for finding scientific information in the Net. Moscow: State University, 1996. <http://www.pharm.unito.it/itcrs/comparis.html>

[26] LEIGHTON, V. H.; SRIVASTAVA, J. Precision arnong World Wide Web Search Services (Search Engines): Altavista, Excite, Hotbot Infoseek, Lyco,s. act. 10 jul 97. Disponible en: http://www.winona.edu/library/webind2/webind2.htm, <http://cybermetrics.cindoc.csic.es/cybermetrics/pdf/96.pdf>.

[27] LEIGHTON, V.H.; SRIVASTAVA, J. First 20 Precision among World Wide Web Search Services (Search Engines), Journal of the American Society for Information Science, 50 (10), 870-881, 1999. <http://www.asis.org/Publications/JASIS/v50n1099.html>

[28] LEONARD, A. Search engine: where to find anything en the net. 1996 <http://www.cnet.com/Content/Reviews/Compwc/Seareh>.

[29] LJOSLAND, M. A comparison between twenty Web search engines on ten rare words [En línea]. Trondheim: Dept. of Computer and Information Science, Norwegian University of Science and Technology, 2000. <http://www.aitel.hist.no/~mildrid/dring/paper/Comp20.doc>

[30] LJOSLAND, M. Comparison between two and one well established Web search engines [En línea]. Trondheim: Dept. of Computer and Information Science, Norwegian University of Science and Technology, 2000. http://citeseer.ist.psu.edu/cache/papers/cs/9908/http:zSzzSzwww.dei.unipd.itzSz~imszSzsigir99z<SzpaperszSz4-ljosland.pdf/ljosland99evaluation.pdf>

[31] MARCHIONINI, G.; BARLOW. D.; HILL, L. Comparing WAIS and boolean capabilities. Journal of the American Society for Information Science, 45(8), 561-564, 1994.< http://pdf.aiaa.org/preview/1993/PV1993_4707.pdf>

[32] MARTÍNEZ MÉNDEZ, F.J. "Aproximación general a la evaluación de la recuperación de información por medio de los motores de búsqueda en Internet", 2001. <http://www.um.es/gtiweb/fjmm/ibersid2000.PDF>

[33] MARTÍNEZ M. FRANCISCO J.; RODRÍGUEZ M. JOSE V. Reflexiones sobre la evaluación de los sistemas de recuperación de información: necesidad, utilidad y

viabilidad, Anales de documentacion, nº 7, 2004, págs. 153-170. < http://www.um.es/fccd/anales/ad07/ad0710.pdf>

[34] MARTÍNEZ MÉNDEZ, F.J. Aplicación de un modelo de evaluación de los SRI en la web. Universidad de Murcia, 2002. <http://descargas.cervantesvirtual.com/servlet/SirveObras/02472741989036164198835/010010_7.pdf>

[35] MING, H. Comparison of Three Search Engines. Toronto: University, <http://vered.rose.utoronto.ca/people/ming/Three_Internet.pdf>

[36] NOTESS, G. R. Search Engines Statistics: Dead Links Report, Database Change Over Time, Unique Hits Report. Bozeman, MT: Notes.com, 2002. <http://www.searchengineshowdown.com/stats/size.shtml>.

[37] OLVERA M.D. Rendimiento de los sistemas de recuperación de información en la web: evalución de servicios de búsqueda (search engines). Pag. 302 - Revista Española de Documentación Científica Vol. 23, Nº 3, 2000, ISSN 0210-0614, Pag. 302

[38] PETER INGWERSEN. Information and Information Science. taken from : 'Enclyclopaedie of Library and Information Science', Vol. 56, supplement 19, p. 137-174, publ. Marcel Dekker, Inc. New York. <http://webhost.ua.ac.be/ibw/artikels/IIS_Ingwersen.doc>

[39] PETERSON, R.E. "Eight internet search engines compared.", 1997 <http://www.firstmonday.dk/issues/issue2_2/peterson/>

[40] SALTON, G. The state of retrieval system evaluation, Information Processing and Measurement, pp. 441-449. 1992.

[41] SALTON, G., MCGILL, M. (1983). Introduction to Modern Information Retrieval. New York: McGraw-Hill Publishing Company, 1983, p.166-167. ISBN 0070544840.

[42] SPARCK JONES, K. Information retrieval experiment. London: Butterworths. ed. 1981 <http://www.itl.nist.gov/iad/894.02/projects/irlib/pubs/ire/ire_toc.html>

[43] SCOVILLE, RICHARD. "Special Report: Find it On The Net!." PC WORLD online, 1996 <http://www.pcworld.com/reprints/lycos.htm>.

[44] STOBART, S. and KERRIDGE, S. WWW search engine study. Sunderland: University, 1996. <http://www.ariadne.ac.uk/issue6/survey/>

[45] SULLIVAN, D. StatMarket search engine ratings. Jupitermedia Corporation, 2002. <http://www.searchenginewatch.com/reports/statmarket.html/>

[46] SULLIVAN, D. Search engine sizes. Jupitermedia Corporation, 2002. <http://www.searchenginewatch.com/reports/sizes.html>

[47] SULLIVAN, D. Jupiter Media Metrix search engine ratings. Jupitermedia Corporation, 2002. <http://searchenginewatch.com/reports/mediametrix.html>

[48] SULLIVAN , D. Media Metrix Search Engine Ratings, 2005. <ttp://searchenginewatch.com/reports/article.php/2156431>.

[49] SULLIVAN , D. Nielsen//NetRatings search engine ratings, 2005 <http://searchenginewatch.com/reports/netratings.html>

[50] TRAMULLAS J.; OLVERA D. Mª. Recuperación de la Información en Internet, edit. Rama, 2001. ISBN84-7897-458-X

[51] JESÚS TRAMULLAS Y KRONOS Introducción a la Documática, 1997, 2000. <http://tramullas.com/documatica/8-6.html>

[52] WESTERA, G. Robot-drive search engine evaluation: overview, 1997. <http://www.curtin.edu.au/curlin/library/ staffpages/gwpersonal/senginestudy/ >

[53] WESTERA, G. *Comparison of Search Engine User Interface Capabilities* [En línea]. Curtin: University of Technology, 2000. <http://library.curtin.edu.au/staff/personal/gwpersonal/compare.html>

[54] WISHARD, L. Precision Among Internet Search Engines: An Earth Sciences Case Study. [En línea]. Pennsyilvania: State University, 1998. <http://gti1.edu.um.es:8080/javima/Precision-Among-Internet-Search-Engines.htm>

[55] WINSHIP, I. "World Wide Web searching tools - an evaluation". VINE (99), 1995. p.49-54 También disponible en http://gti1.edu.um.es:8080/javima/World-Wide-Web-searching-tools-an-evaluation.htm

[56] VAN RIJSBERGEN, C.J. (1979) *Information retrieval*, 2d ed. London: Butterworths., 1979. <http://www.dcs.gla.ac.uk/Keith/Preface.html.>

[57] VAN SLYPE, G., Conception et gestion des systèmes documentaires, Paris, Editions d'organisation, 1977.

[58] ZAZO R. ANGEL; FIGUEROLA P. CARLOS; ALONSO B. JOSE; GÓMEZ RAQUEL. Recuperación de información utilizando el modelo vectorial, Cap4, pp. 11-4, 2001. http://reina.usal.es/trabajos/zazo02recuperacion.pdf

[59] SHERMAN, C. How To Choose The Best General-purpose Search Tool [En línea]. New York: About.com Inc. 2000. <http://webpages.charter.net/michaelsullivan/best_search_engine.html >

[60] Oppenheim, C., Morris, A., McKnight, C. & Lowley, S. (2000) "The evaluation of WWW search engines". Journal of Documentation, 56(2), 190-211. http://irsweb.blogspot.com/2005_02_01_irsweb_archive.html

[61] LEIGHTON, V. H. Performance of Four World Wiáe Web (WWW) Index Services: Infoseek, Lvcos, Webcrawler and WWWWorm. 1995. Disponible en: <http://www.winona.edu/library/webind.htm>.

[62] TOMAIUOLO, N. G.; PACKER. J. G. Results of 200 subject searches in Altavista, Infbseek Lycos, Magellan and Point, performed Oct. to De(. 1995. 20 may 1996. <http://portal.acm.org/citation.cfm?id=231446&coll=Portal&dl=GUIDE&CFID=46616613&CFTOKEN=58199338>http://www.bstc.az/html/websearch.html

[63] TRAMULLAS JESUS. Docunautica le ofrece el texto del manual "Introducción a la Documática. Ed. Kronos, 1997. < http://docunautica.com/>

[64] DAVID A. GROSSMAN, OPHIR FRIEDER Information Retrieval : Algorithms and Heuristics. Publisher: Springer, 276p. 1998. ISBN: 0792382714.

[65] K. SPARCK JONES AND P. WILLETT. Readings in Information Retrieval. Morgan Kaufmann Publishers, San Francisco, California. 1997.

[66] C. J. VAN RIJSBERGEN. Information Retrieval. Butterworths, London, second edition, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/index.htm>

[67] MARTI HEARST. User Interfaces and Visualization in Modern Information Retrieval, 1999.
< http://www.sims.berkeley.edu/~hearst/irbook/10/node1.html>

[68] WATSTEIN S.B AND KESSELMAN M. End user searching services and providers. American Library Association, Chicago, Estados Unidos. 1998.

[69] PRIETO-DIAZ, R. and ARANGO, G. Domain Analysys: Acquisition of Reusable Information for Software Construction. New York: IEEE Press, 1991.

[70] WITTEN I. H, MOFFAT ALISTAIR AND BELL T.C. Managing gigabytes:Compressing and Indexing Documents and Images. Morgan Kaufmann Publishers. Inc. San Francisco, California, 1999. ISBN: 1558605703.
<http://newdownline.com/Computers/Software/Databases/Managing%20Gigabytes%20Compressing%20and%20Indexing%20Documents%20and%20Images.aspx?userid=1>

[71] PEMBERTON, J. MICHAEL. Telecommunication: Technology and Devices. Records Management Quarterly, 23, 46-48, 1989.

OTRAS FUENTES DE CONSULTA: COFERENCIAS Y PUBLICACIONES DE EXPOSICIÓN RETROSPECTIVA COMO ACTUAL SOBRE SISTEMAS DE RECUPERACION INFORMACION.

Publicaciones:

- Information Retrieval. A new journal, edited by Paul Kantor and Stephen Robertson, whose stated aims are to publish high quality technical work in IR. The first number is due to appear any day now.
- Journal of the American Society for Information Science. A good, standard publication source for information science, with much good work in IR. Experiment and research in IR.
- Journal of Documentation. Published by Aslib. Long one the standard and most important journals in IR, which avoids the US bias of JASIS. Experiment and research in IR.
- ACM Transactions on Information Systems. This is a standard journal for substantial, archival work in IR, with a computer science emphasis. Experiment and research in IR.

Para publicaciones, comparaciones e información acerca de search engines:

- Search Engine Watch, www.searchenginewatch.com. Look at the section "Reviews, ratings and tests" which will take you to details of the size of the different engines, how frequently they update, how they index, popularity etc.
- Search Engine Showdown, www.searchengineshowdown.com also gives access to reviews and analysis.

Para detalles de search engines(motores de búsqueda) especializados en un tema particular o que abarcan una región/país específico:

- All SearchEngines.com, www.allsearchengines.com

- The Search Engine Guide, www.searchengineguide.com, and click on "Internet search engines".
- Search Engine Colossus, www.searchenginecolossus.com - excellent for finding search engines to do with any country.

Conferencias:

TREC-n. Proceedings of the nth Text REtrieval Conference. It has become a standard place for publication of high-quality IR evaluation papers. The most important new results in computer-oriented IR are now first published in this forum. The entire set of proceedings is available on-line, at http://trec.nist.gov/

National Online Meeting. Most papers in these proceedings are not terribly high-level, but there are always a few of interest. This is probably the best conference to report on new work in operational IR systems. <http://www.servitel.es/scripts/acid/acid.dll?accion=lista>

ANEXOS

ANEXO1. FORMATO DE CONSULTAS REALIZADAS

FORMULACION DE PREGUNTAS		CRITERIOS						
1. Find pages about dyslexia —>dyslexia		LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA			X	X			X	
EXCITE			X	X			X	
GOOGLE			X	X			X	
HOTBOT			X				X	
SN SEARC			X	X			X	
LYCOS			X				X	
YAHOO			X	X			X	
2. Inform.Seguridad informatica: Criptografia o certificados d		LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA	>>> "certificados digitales" OR criptografia		X		X	X		X
EXCITE	AND "seguridad informatica"		X		X	X		X
GOOGLE			X		X	X		X
HOTBOT			X		X	X		X
SN SEARC			X		X	X		X
LYCOS			X		X	X		X
YAHOO			X		X	X		X
3. Who is Elkin Patarroyo?.		LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA	>>> Who is Elkin Patarroyo?	X					X	
EXCITE		X					X	
GOOGLE		X					X	
HOTBOT		X					X	
SN SEARC		X					X	
LYCOS		X					X	
YAHOO		X					X	
4. Informacion acerca de guerras mundiales		LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA	>>> guerras AND mundiales			X	X	X		X
EXCITE				X	X	X		X
GOOGLE				X	X	X		X
HOTBOT					X	X		X
SN SEARC				X	X	X		X
LYCOS					X	X		X
YAHOO				X	X	X		X
5. Find pages that inform you, about Milky way(via lactea).		LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA	>>> +milky +way				X			X
EXCITE					X			X
GOOGLE					X			X
HOTBOT					X			X
SN SEARC					X			X
LYCOS					X			X
YAHOO					X			X
6. Información acerca de redes virtuales privadas		LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA	>>> "red privada virtual"		X				X	
EXCITE			X				X	
GOOGLE			X				X	
HOTBOT			X				X	
SN SEARC			X				X	
LYCOS			X				X	
YAHOO			X				X	

7. Information about which drugs are used to treat depress	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> drugs AND treat AND depression			X		X		X
EXCITE			X		X		X
GOOGLE			X		X		X
HOTBOT					X		X
SN SEARC			X		X		X
LYCOS					X		X
YAHOO			X		X		X

8. Tratado de libre comercio —> "tratado de libre comercio"	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA		X	X				X
EXCITE		X	X				X
GOOGLE		X	X				X
HOTBOT		X					X
SN SEARC		X	X				X
LYCOS		X					X
YAHOO		X	X				X

9. Information about Theory Relativity.	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> "theory" AND "relativity"		X			X	X	
EXCITE		X			X	X	
GOOGLE		X			X	X	
HOTBOT		X			X	X	
SN SEARC		X			X	X	
LYCOS		X			X	X	
YAHOO		X			X	X	

10. Informacion sobre como jugar Golf de campo.	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> +reglas +golf		X	X	X			X
EXCITE		X	X	X			X
GOOGLE		X	X	X			X
HOTBOT		X		X			X
SN SEARC		X	X	X			X
LYCOS		X		X			X
YAHOO		X	X	X			X

11. Albert Einstein.	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >> "The Earth's Atmosphere"		X				X	
EXCITE		X				X	
GOOGLE		X				X	
HOTBOT		X				X	
SN SEARC		X				X	
LYCOS		X				X	
YAHOO		X				X	

12. Violencia Doméstica	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> violencia AND doméstica			X		X	X	
EXCITE			X		X	X	
GOOGLE			X		X	X	
HOTBOT					X	X	
SN SEARC			X		X	X	
LYCOS					X	X	
YAHOO			X		X	X	

13. Information abaut class of aves.	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> "class aves"		X					X
EXCITE		X					X
GOOGLE		X					X
HOTBOT		X					X
SN SEARC		X					X
LYCOS		X					X
YAHOO		X					X

14. Informacion acerca de la tabla periodica de los elemento	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> "tabla periodica" AND elementos		X	X	X	X	X	
EXCITE		X	X	X	X	X	
GOOGLE		X	X	X	X	X	
HOTBOT		X		X	X	X	
SN SEARC		X	X	X	X	X	
LYCOS		X		X	X	X	
YAHOO		X	X	X	X	X	

15. Hitoria de las pirámides de Egipto.	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> historia AND piramides AND Egipto					X	X	
EXCITE					X	X	
GOOGLE					X	X	
HOTBOT					X	X	
SN SEARC					X	X	
LYCOS					X	X	
YAHOO					X	X	

16. Find pages that discuss about yellow fever	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> "yellow fever"		X	X			X	
EXCITE		X	X			X	
GOOGLE		X	X			X	
HOTBOT		X				X	
SN SEARC		X	X			X	
LYCOS		X				X	
YAHOO		X	X			X	

17. Información general acerca de las causas de los marem	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> +causas AND maremotos OR tsunamis		X		X	X	X	
EXCITE		X		X	X	X	
GOOGLE		X		X	X	X	
HOTBOT		X		X	X	X	
SN SEARC		X		X	X	X	
LYCOS		X		X	X	X	
YAHOO		X		X	X	X	

18. Information about scientific names of plants.	LN	FL	BA	I/E	OB	ES	GE
ALTAVISTA >>> "scientific names" AND plants		X			X		X
EXCITE		X			X		X
GOOGLE		X			X		X
HOTBOT		X			X		X
SN SEARC		X			X		X
LYCOS		X			X		X
YAHOO		X			X		X

BA: Búsqueda avanzada, **I/E:** Inclusión/Exclusión, **OB**:Operaciones Booleanas,
FL: Frase literal, **LN:** Lenguaje Natural, **ES:** Preguntas específicas, **GE:**Preguntas generales.

ANEXO2. TABLA DE DATOS-CRITERIO DE INDICE

RESULTADOS BASADO EN EL TAMAÑO DE INDICE PROMEDIO
(Febrero 21 del 2005)

	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
P1	1.500.000	92	2.150.000	239.000	717.366	514.400	1.520.000
P2	167.000	62	26.300	479	48.903	16.100	165.000
P3	1.670	63	4.530	887	2.395	1.340	1.750
P4	126.000	63	119.000	21.800	81.766	26.800	126.000
P5	700.000	69	436.000	95.700	1.286.301	184.700	710.000
P6	27.000	66	169.200	3.600	20.133	4.050	27.300
P7	2.130.000	70	2.300.000	333.000	3.258.712	421.700	2.150.000
P8	344.000	60	354.000	63.300	121.192	78.400	314.200
P9	1.570.000	73	1.110.000	208.000	295.847	498.100	1.580.000
P10	144.000	63	123.000	13.100	48.500	44.400	142.000
P11	538.000	63	350.000	64.900	246.842	165.800	544.000
P12	364.000	65	360.000	49.100	195.344	68.500	358.000
P13	60.400	63	82.800	7.980	19.612	90.800	60.200
P14	907.000	95	541.000	12.000	16.055	408.400	924.000
P15	53.000	67	29.500	739	20.488	7.910	51.800
P16	730.000	65	984.000	94.900	427.571	368.200	730.000
P17	2.370.000	65	24.400	2.810	512.645	140.200	2.340.000
P18	110.000	66	122.000	27.900	47.385	45.600	113.000
PROM	**657.893**	**68,3**	**515.874**	**68.844**	**409.281**	**171.411**	**658.736**

RESULTADOS BASADO EN EL TAMAÑO DE INDICE PROMEDIO
(Abril 30 del 2005)

	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
P1	1.850.000	89	2.020.000	537.000	771.534	544.500	1.920.000
P2	153.000	62	44.100	526	44.969	173	162.000
P3	3.090	63	5.800	968	2.860	128	2.910
P4	111.000	69	127.000	23.100	88.689	28.600	116.000
P5	2.390.000	76	2.100.000	388.000	1.427.144	887.900	2.580.000
P6	27.200	64	22.800	3.440	19.849	197	26.700
P7	4.170.000	94	1.820.000	523.000	3.733.174	156	3.860.000
P8	330.000	68	529.000	86.500	158.579	176	326.000
P9	1.400.000	73	2.570.000	479.000	275.517	568.300	1.430.000
P10	96.300	64	87.300	14.000	56.360	173	100.000
P11	525.000	74	434.000	99.800	249.830	186	559.000
P12	284.000	66	471.000	48.000	125.957	166	358.000
P13	41.800	67	42.000	7.070	20.770	151	42.200
P14	32.800	70	32.200	534	16.361	169	32.800
P15	37.200	68	35.700	581	22.247	125	41.000
P16	714.000	75	581.000	154.000	426.070	378.800	742.000
P17	8.620.000	64	16.100	2.670	535.157	908.600	8.990.000
P18	107.000	77	117.000	28.400	50.385	177	109.000
PROM	**1.160.688**	**71,3**	**614.167**	**133.144**	**445.858**	**184.371**	**1.188.756**

ANEXO3. TABLA DE DATOS-CRITERIO DE TIEMPO DE RESPUESTA

RESULTADOS BASADO EN EL TIEMPO PROMEDIO (Segundos)
(Febrero 21 del 2005)

	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
P1	2,2	3,9	1,4	2	2	4,2	2,8
P2	3,4	3,8	1,4	2	1,8	4,05	2,8
P3	4	4,3	1,5	1,6	2,1	3,8	2,6
P4	2,5	3,9	1,2	2,2	1,9	3,4	3,4
P5	2,2	3,2	1	2	1,7	3,9	2,4
P6	2,3	3,3	1,25	1,6	1,85	3,3	2,3
P7	2,2	3,75	1,4	2,5	1,7	5	2,4
P8	2,3	3,5	1,1	1,8	1,8	3,6	2,25
P9	2,4	2,8	3,4	2,3	2,15	2,8	2,7
P10	2,9	4,2	1,8	1,6	1,9	3,9	2,7
P11	2,7	3,6	1,3	1,9	1,85	3,8	2,8
P12	2,1	4,4	2	2,1	1,85	3,3	2,3
P13	2,3	3	1,1	1,4	2,1	3,5	2,5
P14	2,3	2,8	1,2	1,9	1,9	3,3	2,1
P15	2,1	3,5	1,5	1,7	1,8	3,95	1,7
P16	2,2	3,8	1,1	1,9	1,75	2,7	2,3
P17	2,4	3,2	1,6	1,4	1,9	4,2	2,5
P18	2,2	3,5	1,3	1,7	1,8	2,8	2,6
PROM	**2,48**	**3,58**	**1,48**	**1,87**	**1,88**	**3,64**	**2,51**

RESULTADOS BASADO EN EL TIEMPO PROMEDIO (Segundos)
(Abril 30 del 2005)

	Altavista	Excite	Google	Hotbot	MSN Search	Lycos	Yahoo
P1	2,8	2,4	1,4	2,1	1,7	3,6	2
P2	1,9	2,7	1	1,4	1,65	2,95	1,8
P3	1,8	1,2	1,45	1,3	1,4	3,1	2,1
P4	1,75	2,5	1,1	1,4	1,4	3,3	2
P5	1,65	2,6	1,15	1,85	1,7	3,4	2,2
P6	1,6	2,7	0,95	1,3	1,45	3	1,8
P7	1,7	2,6	1,6	1,2	1,5	3,85	1,9
P8	1,9	2,5	1,15	1,5	1,4	3,2	2,15
P9	2,1	2,9	1,25	1,8	1,7	3,25	2,3
P10	1,7	2,5	0,9	1,35	1,6	3	1,7
P11	1,85	2,7	1	1,5	1,5	3,3	2
P12	1,8	2,2	1,05	2,2	1,4	3,6	2,2
P13	1,9	2,65	1	1,7	1,5	3,4	2
P14	2	2,4	0,9	1,6	1,4	3,5	2,25
P15	1,55	2,55	1	1,5	1,6	3,6	2,3
P16	1,8	2,2	1,05	1,3	1,6	3,4	1,9
P17	1,9	2,6	0,8	1,35	1,45	3	1,95
P18	1,6	2,7	0,95	1,4	1,5	3,3	1,7
PROM	**1,85**	**2,48**	**1,09**	**1,54**	**1,53**	**3,32**	**2,01**

ANEXO4. TABLA DE DATOS-CRITERIO DE RUIDO DOCUMENTAL

RESULTADOS BASADO EN EL RUIDO DOCUMENTAL
(Análisis tomado sobre los 30 primeros documetos recuperados)

	Altavista			Excite			Google			Hotbot			MSN Search			Lycos			Yahoo		
	IN	D	IR	IN	D	IR	IN	D	IR	IN	D	IR	IN	D	IR	IN	D	IR	IN	D	IR
P1	0	8	1	2	7	8	0	3	3	0	5	2	0	4	1	0	7	4	0	6	2
P2	0	7	5	0	3	6	0	5	4	0	5	10	1	3	5	1	6	6	0	5	2
P3	0	4	1	3	3	2	4	2	3	3	4	2	0	3	8	3	4	2	1	3	4
P4	2	3	5	2	4	3	1	4	2	4	3	1	1	3	10	2	6	5	3	4	3
5	0	5	4	1	6	5	2	3	6	1	4	7	0	2	6	0	5	4	2	5	3
P6	4	4	3	0	7	4	0	4	3	0	3	1	0	4	10	1	6	4	3	2	3
P7	0	6	2	0	8	1	0	2	4	0	3	2	0	1	4	0	7	5	0	5	4
P8	0	4	4	3	5	3	1	5	1	0	4	1	0	3	1	3	4	3	0	4	4
P9	0	6	0	1	4	3	0	3	3	0	3	2	0	6	1	1	8	4	0	6	1
P10	1	8	2	1	7	5	0	3	4	0	2	5	0	5	4	2	7	6	1	4	6
P11	2	4	2	1	7	1	1	4	1	1	5	1	0	4	2	2	9	5	4	2	3
P12	2	1	1	1	2	4	1	3	4	2	2	3	0	5	2	2	1	6	1	1	3
P13	1	1	3	2	4	2	1	5	5	2	4	4	1	5	7	2	2	5	1	1	4
P14	1	4	0	2	6	0	2	5	1	1	3	1	0	4	4	6	3	0	2	3	0
P15	2	4	4	1	3	5	0	3	5	1	4	4	1	5	6	2	5	5	0	5	2
P16	0	1	6	0	9	4	0	2	4	0	2	6	0	3	4	2	3	9	0	1	4
P17	0	3	0	1	3	5	2	4	2	2	6	3	2	4	7	1	6	5	0	1	9
P18	0	4	2	1	7	6	0	3	0	0	1	2	2	3	6	1	1	2	0	4	2
Total	15	77	45	22	95	67	15	63	55	17	63	57	8	67	88	31	90	80	18	62	59

ANEXO5. TABLA DE DATOS (1-10, 11-20 Y 21-30 DOCUMENTOS RECUPERADOS)

CRITERIO BASADO EN LA EXAUTIVIDAD-PRECISIÓN
(Analisis tomado sobre 1-10 primeros documetos recuperados)

	Altavista				Excite				Google				Hotbot				MSN Search				Lycos				Yahoo			
	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
P1	1	1	6	2	1	2	5	2	0	1	7	2	0	1	7	2	0	0	7	3	1	2	5	2	0	1	7	2
P2	1	0	6	3	3	1	3	3	2	0	4	4	2	1	7	0	1	1	6	2	2	0	7	1	0	1	7	2
P3	0	1	8	1	2	0	4	4	2	1	5	2	2	1	4	3	0	6	4	0	1	2	6	1	2	0	5	3
P4	3	1	1	2	2	0	8	0	2	2	5	1	1	3	5	1	2	3	5	0	4	2	3	1	2	3	5	0
P5	1	1	5	3	2	0	6	2	3	0	5	2	3	0	5	2	0	2	8	0	0	0	7	3	2	1	5	2
P6	3	1	5	1	1	1	8	0	1	1	8	0	1	1	8	0	2	1	7	0	3	2	5	0	1	0	8	1
P7	0	0	9	1	0	1	9	0	0	1	9	0	0	1	9	0	0	1	8	1	2	1	7	0	0	0	10	0
P8	0	0	10	0	2	1	7	0	0	1	9	0	0	1	9	0	0	3	7	0	3	2	5	0	2	0	8	0
P9	0	0	10	0	0	0	9	1	1	0	9	0	1	0	9	0	0	1	8	1	1	1	8	0	0	0	9	1
P10	0	1	8	1	0	1	9	0	0	0	10	0	1	1	8	0	1	0	6	3	0	2	8	0	1	1	6	0
P11	1	1	7	1	1	1	7	1	1	0	7	2	1	0	8	1	0	1	9	0	2	1	6	1	1	0	8	1
P12	1	1	7	1	1	0	6	3	2	2	6	0	2	1	6	1	1	1	7	1	1	1	6	2	1	0	8	1
P13	1	1	8	0	0	0	10	0	1	1	8	0	2	0	8	0	1	0	9	0	2	0	8	0	0	1	9	0
P14	1	0	9	0	0	0	10	0	1	0	9	0	0	0	10	0	1	0	6	3	2	0	8	0	0	0	10	0
P15	3	0	7	0	2	1	6	1	3	0	6	1	2	0	6	2	1	1	5	1	4	1	5	0	0	0	7	3
P16	0	1	9	0	2	2	6	0	0	3	6	0	1	2	7	0	1	1	5	1	3	1	6	0	0	0	10	0
P17	0	1	8	1	1	0	9	0	2	0	6	2	2	0	7	1	2	0	8	0	2	3	5	0	2	0	6	2
P18	1	2	6	1	2	0	8	0	0	0	10	0	0	1	9	0	3	0	7	0	2	0	7	1	1	2	6	1

CRITERIO BASADO EN EL RUIDO EN LA RECUPERACION DE LA INFORMACION
(Analisis tomado sobre 11-30 primeros documetos recuperados)

	Altavista				Excite				Google				Hotbot				MSN Search				Lycos				Yahoo			
	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
P1	0	0	8	2	4	0	5	1	2	0	6	2	0	1	6	3	0	1	8	1	2	0	4	4	0	0	7	3
P2	1	1	6	2	1	0	6	3	1	0	6	3	4	0	3	3	1	3	3	3	2	2	4	2	1	2	7	0
P3	0	0	8	2	2	0	8	0	2	2	6	0	1	0	7	2	4	2	4	0	2	1	7	0	2	1	5	2
P4	2	1	7	0	1	1	8	0	0	2	7	1	2	2	5	1	3	3	4	0	3	1	5	1	2	1	5	2
P5	1	1	6	2	2	1	6	1	2	0	8	0	3	0	7	0	2	2	5	1	1	1	6	2	1	2	7	0
P6	2	2	5	1	2	2	5	1	2	0	8	0	0	1	9	0	1	0	5	1	1	1	7	1	3	0	6	1
P7	0	1	8	1	2	2	5	1	2	0	8	0	0	1	9	0	1	1	7	1	1	1	7	1	3	0	6	1
P8	1	0	9	0	1	1	7	1	0	2	8	0	0	1	9	0	0	2	7	1	2	2	6	0	0	1	9	0
P9	0	1	8	1	2	0	8	0	1	1	7	1	0	0	9	1	0	0	10	0	2	0	8	0	1	0	8	1
P10	2	3	5	0	2	1	7	0	2	0	8	0	2	0	8	0	2	0	7	1	3	1	5	1	2	2	6	0
P11	1	0	8	1	1	0	7	2	1	0	7	2	0	0	8	2	1	2	7	0	2	0	6	2	3	0	7	0
P12	2	1	6	1	1	1	7	1	1	2	7	0	1	1	6	2	1	2	7	0	3	0	4	3	1	0	7	2
P13	2	0	8	0	1	0	9	0	2	2	6	0	3	0	7	0	0	0	10	0	2	0	8	0	1	1	8	0
P14	0	1	9	0	1	1	8	0	1	0	9	0	1	0	9	0	3	1	6	0	1	0	8	1	1	0	9	0
P15	2	2	6	0	2	0	6	2	1	0	7	2	2	0	5	3	1	2	7	0	1	1	6	2	0	0	8	2
P16	3	3	5	1	1	3	6	1	1	1	8	1	2	1	7	0	1	2	7	0	4	0	6	0	2	0	7	1
P17	0	2	8	0	2	0	8	0	0	1	8	1	0	0	8	2	3	3	5	0	3	1	5	1	3	0	6	1
P18	0	1	6	2	1	0	9	0	0	1	9	0	1	1	8	0	5	0	5	0	0	1	8	1	0	3	7	0

CRITERIO BASADO EN EL RUIDO EN LA RECUPERACION DE LA INFORMACION
(Analisis tomado sobre 21-30 primeros documetos recuperados)

	Altavista				Excite				Google				Hotbot				MSN Search				Lycos				Yahoo			
	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
P1	0	0	8	2	5	1	2	1	1	3	6	0	2	2	5	1	1	2	6	1	1	2	5	2	2	0	7	1
P2	3	1	3	3	2	1	5	2	1	0	6	3	4	2	1	3	4	0	5	1	3	1	6	0	1	1	3	5
P3	1	2	7	0	1	2	7	0	3	1	6	0	2	1	5	2	4	3	3	0	2	2	6	0	2	2	5	1
P4	1	3	6	0	2	2	6	0	1	3	6	0	2	1	6	1	5	3	2	0	3	2	5	0	2	3	4	1
P5	2	0	5	3	2	1	5	2	3	1	5	1	2	1	7	0	3	1	6	0	3	0	7	1	2	3	5	0
P6	2	1	6	1	1	2	7	0	1	1	7	1	0	2	8	0	4	2	4	0	1	2	6	1	2	2	5	1
P7	2	2	6	0	1	0	9	0	2	2	6	0	1	1	8	0	3	1	6	0	2	1	7	0	2	0	7	1
P8	3	0	6	1	4	1	5	0	2	0	8	0	1	2	7	0	1	1	7	1	1	1	8	0	2	1	7	0
P9	0	2	7	0	2	0	7	1	1	1	8	0	1	1	8	0	0	2	7	1	2	1	6	1	0	2	6	2
P10	1	2	6	1	4	0	6	0	2	1	7	0	2	1	6	1	1	0	8	1	5	1	4	0	4	1	5	0
P11	2	1	5	2	0	1	8	1	0	3	6	1	1	1	8	0	1	1	8	0	3	1	5	1	3	0	7	0
P12	0	2	8	0	3	1	5	1	2	2	6	0	2	0	8	0	2	2	6	0	4	0	4	2	2	0	8	0
P13	1	1	7	1	3	0	7	0	3	1	6	0	2	0	7	1	1	0	8	1	3	1	6	0	4	0	6	0
P14	0	3	7	0	1	1	7	1	1	0	8	1	1	1	8	0	4	0	6	0	3	0	6	1	1	0	8	1
P15	1	2	6	1	2	1	6	1	2	1	7	0	1	1	7	1	2	2	6	0	2	2	6	0	2	0	7	1
P16	3	2	7	0	1	2	7	0	3	1	6	0	3	1	5	1	2	2	6	0	4	1	5	0	2	0	8	0
P17	0	3	7	0	3	2	5	0	2	1	5	2	3	1	6	0	2	3	5	0	1	2	6	1	3	0	6	1
P18	2	2	6	0	4	0	6	0	0	1	8	1	1	0	9	1	1	1	8	0	1	1	8	0	1	1	7	1

ANEXO6. REPRESENTACION GRAFICA, RENDIMIENTO PROMEDIO EXHAUSTIVIDAD PRECISIÓN

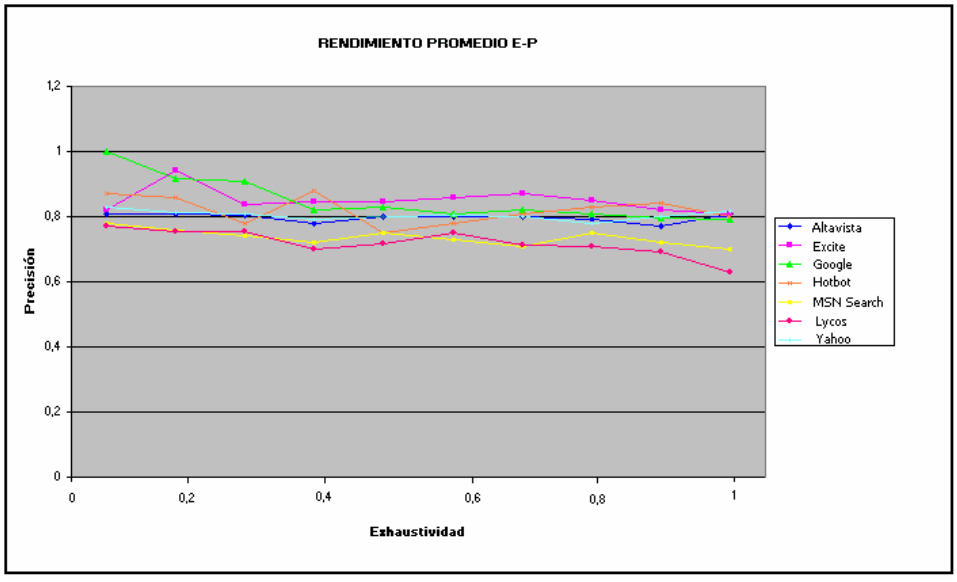


TABLA DE DATOS EXHAUSTIVIDAD-PRECISIÓN (E-P),
(10 PRIMEROS DOCUMENTOS RECUPERADOS)

Altavista		Excite		Google		Hotbot		MSN Search		Lycos		Yahoo	
E	P	E	P	E	P	E	P	E	P	E	P	E	P
0,09	0,81	0,1	0,82	0,12	1	0,09	0,87	0,08	0,78	0,1	0,77	0,1	0,83
0,21	0,808	0,2	0,94	0,22	0,915	0,21	0,86	0,18	0,757	0,2	0,755	0,22	0,812
0,3	0,805	0,28	0,839	0,32	0,91	0,29	0,78	0,23	0,74	0,32	0,753	0,33	0,81
0,43	0,78	0,41	0,845	0,42	0,82	0,41	0,88	0,4	0,72	0,42	0,7	0,46	0,79
0,52	0,802	0,49	0,844	0,53	0,83	0,5	0,75	0,53	0,75	0,52	0,715	0,55	0,8
0,64	0,8	0,64	0,86	0,63	0,81	0,58	0,78	0,65	0,73	0,68	0,75	0,68	0,805
0,69	0,798	0,71	0,87	0,71	0,82	0,71	0,81	0,76	0,71	0,76	0,713	0,78	0,799
0,91	0,792	0,86	0,85	0,81	0,81	0,83	0,83	0,87	0,75	0,85	0,71	0,88	0,78
0,95	0,77	1	0,82	0,9	0,795	0,95	0,84	0,93	0,72	0,94	0,69	0,91	0,8
1	0,81	1	0,81	1	0,79	1	0,8	1	0,7	1	0,63	1	0,817

ANEXO7. PROCEDIMIENTO DE AGRUPAMIENTO
(TÉCNICA DE PROMEDIO ARITMÉTICO)

	B	C	D	E	F	G
A	0,21	0,346	0,294	0,311	0,428	0,078
B		0,393	0,388	0,496	0,318	0,206
C			0,109	0,405	0,433	0,356
D				0,417	0,44	0,381
E					0,441	0,455
F						0,447

En base a las distancias medias de motor a motor, se prosigue a la aplicación del algoritmo de agrupamiento, siempre escogiendo la menor distancia entre ellos.

	C	D	E	F	A+G
B	0,393	0,388	0,496	0,318	0,206
C		0,109	0,405	0,433	0,356
D			0,417	0,44	0,381
E				0,441	0,455
F					0,447

Agrup.	Motores	Distancia
1	Altavista-Yahoo	0,078

	C+D	D	E	F	A+G
B	0,39	0,388	0,496	0,318	0,206
C+D		0,109	0,411	0,436	0,368
E				0,441	0,455
F					0,447

0,39= (0,393+0,388)/2
0,411=(0,405+0,417)/2
0,436=(0,433+0,44)/2
0,368=(0,356+0,381)/2

Agrup.	Motores	Distancia
1	Altavista-Yahoo	0,078
2	Google-Hotbot	0,109

	E	F	A+G+B
C+D	0,411	0,436	0,368
E		0,441	0,455
F			0,447

Agrup.		Motores	Distancia
1		Altavista-Yahoo	0,078
2		Google-Hotbot	0,109
3		Altavista-Yahoo-Excite	0,206

	E	F	A+G+B+C+D
E		0,441	0,455
F			0,447

Agrup.		Motores	Distancia
1		Altavista-Yahoo	0,078
2		Google-Hotbot	0,109
3		Altavista-Yahoo-Excite	0,206
4		Altavista-Yahoo-Excite-Google-Hotbot	0,368

	A+G+B+C+D+E+F
E+F	0,447

Agrup.		Motores	Distancia
1		Altavista-Yahoo	0,078
2		Google-Hotbot	0,109
3		Altavista-Yahoo-Excite	0,206
4		Altavista-Yahoo-Excite-Google-Hotbot	0,368
5		Altavista-Yahoo-Excite-Google-Hotbot-MSN Search-Lycos	0,447

ANEXO8. REPRESENTACION DE LA TOMA DE DATOS Y CÁLCULO DE SIMILITUD (PAR DE MOTORES DE BÚSQUEDA ALTAVISTA-EXCITE)

CRITERIO DE SIMILITUD DE RESULTADOS ENTRE PARES DE MOTORES DE BUSQUEDA

		M_A	M_E	$M_A * M_E$	$(M_A)^2$	$(M_B)^2$
P1	http://www.dyslexia.com/	1	1	1	1	1
	http://www.dislexia.net/	1	0	0	1	0
	http://www.interdys.org/	1	0,9	0,9	1	0,81
	http://www.dyslexia-teacher.com/	1	1	1	1	1
	http://www.bda-dyslexia.org.uk/main/home/index.asp	1	0	0	1	0
	http://www.audiblox2000.com/dyslexia_dyslexic/dyslexia.htm	1	0,9	0,9	1	0,81
	http://www.bonnieterrylearning.com/	0	1	0	1	0
	http://www.readingupgrade.com/html/rudyslexia.htm	0	1	0	1	0
	http://www.gow.org/?src=overture	0	1	0	1	0
	http://www.schwablearning.org/articles.asp?r=43&g=1	0	1	0	1	0
	http://infoscouts.com/health/dyslexia.htm	0	1	0	0	1
P2	http://www.criptored.upm.es/guiateoria/gt_m001a.htm	1	0,8	0,8	1	0,64
	http://www.htmlweb.net/seguridad/seguridad.html	1	1	1	1	1
	http://www.abcdatos.com/tutoriales/tutorial/l6091.html	1	0	0	1	0
	http://html.rincondelvago.com/criptografia_seguridad-informatica.html	1	1	1	1	1
	http://www.conocimientosweb.net/dt/article1035.html	1	1	1	1	1
	http://www.hispasec.com/unaaldia/1957	1	0,9	0,9	1	0,81
	http://www.kriptopolis.com/more.php?id=A72_0_1_0_M	1	0,8	0,8	1	0,64
	http://dmoz.org/World/Espa%c3%b1ol/Computadoras/Seguridad/	0,9	1	0,9	0,81	1
	http://www.incywincy.com/default?p=139409	0,8	1	0,8	0,64	1
P3	http://buscabiografias.com/cgi-bin/verbio.cgi?id=5358	1	0,9	0,9	1	0,81
	http://www.biografiasyvidas.com/biografia/p/patarroyo.htm	1	0,8	0,8	1	0,64
	http://www.fonendo.com/noticias/43/2001/03/1.shtml	1	0	0	1	0
	http://www.campusred.net/campusdiario/20020723/entrev.htm	1	0	0	1	0
	http://www.diariomedico.com/	1	1	1	1	1
	http://es.wikipedia.org/wiki/Manuel_Elkin_Patarroyo	1	0,9	0,9	1	0,81
	http://www.sextocontinente.org/apoyohumano/Patarroyo.htm	1	0	0	1	0
	http://www.artehistoria.com/historia/personajes/7851.htm	1	1	1	1	1
	http://web.li.gatech.edu/~rdrury/special/colombia/colombians/manuel.htm	0	1	0	0	1
	http://www.banrep.gov.co/blaavirtual/credencial/113manuel.htm	0,8	1	0,8	0,64	1
	http://www.educarchile.cl/ntg/sitios_educativos/1618/article-62875.html	0	1	0	0	1
	http://www.thirdworldtraveler.com/Heroes/Manuel_Patarroyo.html	0	1	0	0	1
	http://www2.rnw.nl/rnw/es/temas/cienciaysalud/salud/archivo_cienciaysalu	0	1	0	0	1
P4	http://www.monografias.com/trabajos/guemun/guemun.shtml	1	1	1	1	1
	http://www.galeon.com/guerras/	1	0,8	0,8	1	0,64
	http://karasu.iespana.es/karasu/ww/ww.html	1	1	1	1	1
	http://www.libreriaelastillero.com/index.php?mnu=materia&materia=guerr	1	1	1	1	1
	http://www.mgr.org/wwdatesEsp.html	1	1	1	1	1
	http://clio.rediris.es/articulos/memoria_guerras.htm	1	0	0	1	0
	http://www.itacor.com.ar/detodo/guerras.html	1	1	1	1	1
	http://www.aunmas.com/ataque/parte_046.htm	1	1	1	1	1
	http://pastranec.net/historia/contemporanea/gm.htm	0	1	0	0	1
P5	http://www.seds.org/messier/more/mw.html	1	1	1	1	1
	http://adc.gsfc.nasa.gov/mw/milkyway.html	1	1	1	1	1
	http://antwrp.gsfc.nasa.gov/apod/lib/milky_way.html	1	1	1	1	1
	http://www.star.le.ac.uk/edu/mway/	1	1	1	1	1
	http://www.damtp.cam.ac.uk/user/gr/public/gal_milky.html	1	1	1	1	1
	http://map.gsfc.nasa.gov/m_uni/uni_101mw.html	1	0,9	0,9	1	0,81
	http://csep10.phys.utk.edu/guidry/violence/ginfo1.html	1	1	1	1	1
	http://starchild.gsfc.nasa.gov/docs/StarChild/universe_level2/milky_way.h	1	0	0	1	0
	http://www.sciam.com/search/index_combo.cfm?sc=F&ec=overture03&Q	1	0	0	1	0
P6	http://inicio.tiendapc.com/internet/SInicio?j=11462&f=40258	1	0,8	0,8	1	0,64
	http://max.uc3m.es/docs/vpn/	1	0	0	1	0
	http://www.uv.es/~montanan/redes/trabajos/VirtualPrivateNetwork.PDF	1	0,8	0,8	1	0,64
	http://www.microsoft.com/latam/technet/articulos/200008/art04/	1	1	1	1	1
	http://criterios.raudo.com/red-privada-virtual.html	1	0,8	0,8	1	0,64
	http://publib.boulder.ibm.com/html/as400/v5r1/ic2931/info/rzaja/rzaja.pdf	1	1	1	1	1
	http://www.qsl.net/lw2dtq/vpn.htm	0	1	0	0	1
	http://www.um.es/atica/rpv/	0	1	0	0	1
	http://www.windowstimag.com/atrasados/2001/49_ene01/articulos/suplem	0	1	0	0	1
	http://support.microsoft.com/default.aspx?scid=kb%3Bes%3B314076	0	1	0	0	1

P7	http://home.blarg.net/~charlatn/depression/DepMeds.html	1	1	1	1	1
	http://home.blarg.net/~charlatn/Depression.html	1	0,9	0,9	1	0,81
	http://www.carolandjacque.com/depression1/depression-drugs.php	1	0,9	0,9	1	0,81
	http://www.low-cost-drugs.net/paxil-0042.html	1	0,9	0,9	1	0,81
	http://www.thermoflowproducts.com/fibromyalgia.html	1	0,9	0,9	1	0,81
	http://www.depressed-no-more.com/	1	0	0	1	0
	http://www.umich.edu/~pharm660/	1	0	0	1	0
	http://biz.yahoo.com/ibd/050211/health_1.html	0	1	0	0	1
	http://www.coreynahman.com/antidepressantdrugsdatabase.html	0,9	1	0,9	0,81	1
	http://www.feelserenity.com/?b=1658&ovmkt=L7GFEB26GS6RDN35NV4	0	1	0	0	1
	http://www.remedyfind.com/type.asp?id=13&TYPE_ID=4	0	1	0	0	1
	http://www.ahrq.gov/news/press/pr2002/adhdpr.htm	0	1	0	0	1
	http://www.mercola.com/2001/mar/31/depression.htm	0,9	1	0,9	0,81	1
	http://www.keepmedia.com/pubs/HealthDay/2004/06/16/489152?extID=10	0	1	0	0	1
	http://www.healthyplace.com/communities/depression/treatment/antidepre	0	1	0	0	1
P8	http://es.kelkoo.com/search.jsp?popups=no&siteSearchQuery=tratado+de	1	0	1	0	1
	http://www.monografias.com/trabajos11/eltlcfin/eltlcfin.shtml	1	0,9	0,9	1	0,81
	http://www.hoy.com.ec/temas/tlc/tlc.htm	1	0	1	0	1
	http://tlcecuador.blogspot.com/	1	1	1	1	1
	http://www.elcolombiano.com/proyectos/serieselcolombiano/textos/tlc/Ho	1	0	1	0	1
	http://www.tlcperu-eeuu.gob.pe/index.php	1	1	1	1	1
	http://www.udlap.mx/~tesis/lri/garcia_p_ay/	1	0	1	0	1
	http://www.mincomercio.gov.co/VBeContent/NewsDetail.asp?ID=1445&ID	1	0,9	0,9	1	0,81
	http://www.comex.go.cr/acuerdos/comerciales/CAFTA/texto/	1	1	1	1	1
	http://www.continents.com/nuevasfronteras.html	0	1	0	0	1
	http://www.nafta-sec-alena.org/DefaultSite/index.html	0,9	1	0,9	0,81	1
	http://www.rmalc.org.mx/					
	http://www.causa.sieca.org.gt/	0	1	0	0	1
P9	http://www.bartleby.com/173/	1	1	1	1	1
	http://www-groups.dcs.st-and.ac.uk/~history/HistTopics/General_relativity	1	0,9	0,9	1	0,81
	http://www2.slac.stanford.edu/vvc/theory/relativity.html	1	1	1	1	1
	http://www.drphysics.com/relativity.html	1	1	1	1	1
	http://www.damtp.cam.ac.uk/user/gr/public/qg_ss.html	1	0	0	1	0
	http://en.wikipedia.org/wiki/Theory_of_relativity	1	1	1	1	1
	http://www.pbs.org/wgbh/nova/einstein/relativity/	1	1	1	1	1
	http://archive.ncsa.uiuc.edu/Cyberia/NumRel/GenRelativity.html	1	0	0	1	0
	http://www.amazon.com/exec/obidos/ASIN/0517884410/104-1671034-374	1	0,9	0,9	1	0,81
	http://www.sciam.com/search/index_combo.cfm?sc=F&ec=overture03&Q	0	1	0	0	1
	http://www.questia.com/Index.jsp?CRID=theory_of_relativity&OFFID=se2	0	1	0	0	1
	http://www.muppetlabs.com/~breadbox/txt/al.html	0,9	1	0,9	0,81	1
	http://csep10.phys.utk.edu/astr161/lect/history/einstein.html	0,9	1	0,9	0,81	1
P10	http://www.golfspainfederacion.com/page/reglas.asp	1	1	1	1	1
	http://www.fvg.org/reglasgolf.pdf	1	0,8	0,8	1	0,64
	http://www.basegolf.com/reglas/rules.htm	1	1	1	1	1
	http://www.golfmagazine.com.mx/reglas.htm	1	0	0	1	0
	http://www.golfaventura.com/Reglas/pagreglasindex.htm	1	1	1	1	1
	http://www.aug.com.uy/reglasdegolf2004.pdf	1	0	0	1	0
	http://www.federacioncolombianadegolf.com/modulo.php3?modulo=9	1	1	1	1	1
	http://www.golfspain.com/esp/golfmania_reglas.asp	0,9	1	0,9	0,81	1
	http://www.amcgolf.com/	0,9	1	0,9	0,81	1
	http://www.golfdamadenoche.com/index.swf	0	1	0	0	1
P11	http://csep10.phys.utk.edu/astr161/lect/earth/atmosphere.html	1	1	1	1	1
	http://www.enchantedlearning.com/subjects/astronomy/planets/earth/Atm	1	1	1	1	1
	http://www.windows.ucar.edu/tour/link=/earth/Atmosphere/overview.html	1	1	1	1	1
	http://www.rcn27.dial.pipex.com/cloudsrus/atmosphere.html	1	0,9	0,9	1	0,81
	http://www.usd.edu/esci/exams/atmosph.html	1	0,9	0,9	1	0,81
	http://archive.ncsa.uiuc.edu/Edu/RSE/RSEred/WeatherLesson1.html	1	0,9	0,9	1	0,81
	http://en.wikipedia.org/wiki/Earth%27s_atmosphere	1	0,9	0,9	1	0,81
	http://www.ux1.eiu.edu/~cfjps/1400/atmos_origin.html	0	1	0	0	1
	http://curriculum.calstatela.edu/courses/builders/lessons/less/les3/layers.h	0	1	0	0	1
	http://zebu.uoregon.edu/internet/l2.html	0	1	0	0	1
P12	http://www.afscme.org/spanish/abuso.htm	1	1	1	1	1
	http://www.vidahumana.org/vidafam/violence/violence_index.html	1	1	1	1	1
	http://www.servicioslegales.org/violencia_domestica.html	1	1	1	1	1
	http://www.mujersanjuan.com/violencia.html	1	0,9	0,9	1	0,81
	http://www.nlm.nih.gov/medlineplus/spanish/domesticviolence.html	1	1	1	1	1
	http://www.psicologia-online.com/colaboradores/paola/violencia/index2.sh	1	0	0	1	0
	http://www.psiqweb.med.br/infantil/violdome.html	1	0	0	1	0
	http://www.idph.state.il.us/about/womenshealth/spfactsheets/dv.htm	1	1	1	1	1
	http://www.dvalianza.org/	0	1	0	0	1
	http://dmoz.org/World/Espa%C3%B1ol/Sociedad/Problem%C3%A1ticas/	0,8	1	0,8	0,64	1
	http://www.guardiaciv il.org/mujer/index.jsp	0,9	1	0,9	0,81	1
	http://www.justicewomen.com/help_know_your_rights_sp.html	0	1	0	0	1
	http://dir.foromarbella.com/Top/World/Espa%C3%B1ol/Sociedad/Problem	0	1	0	0	1

P13	http://www.ucmp.berkeley.edu/diapsids/birds/birdintro.html	1	1	1	1	1
	http://www.sidwell.edu/us/science/vlb5/Labs/Classification_Lab/Eukarya/A	1	1	1	1	1
	http://faculty.evansville.edu/de3/b10802/PPoint/Aves/sld001.htm	1	0,9	0,9	1	0,81
	http://www.lions.odu.edu/~kkilburn/209_lectures/aves1.pdf	1	0,9	0,9	1	0,81
	http://www.okc.cc.ok.us/biologylabs/Documents/Animals/Aves.htm	1	0,9	0,9	1	0,81
	http://www.mscarolinesci.com/MarineBirds.pdf	1	0,9	0,9	1	0,81
	http://www.thefreedictionary.com/class+Aves	1	0,9	0,9	1	0,81
	http://www.birdnature.com/borderintro.html	1	0,8	0,8	1	0,64
	http://animaldiversity.ummz.umich.edu/site/accounts/information/Aves.htı	0,8	1	0,8	0,64	1
	http://animals.about.com/od/birds/p/aves.htm	0	1	0	0	1
	http://www.inhs.uiuc.edu/cbd/ilspecies/birdsplist.html	0	1	0	0	1
P14	http://www.lenntech.com/espanol/tabla-periodica.htm	1	1	1	1	1
	http://www.mcgraw-hill.es/bcv/tabla_periodica/mc.html	1	1	1	1	1
	http://www.geocities.com/erkflores/Tabla.htm	1	1	1	1	1
	http://galilei.iespana.es/galilei/qui/tablaperiodica0.htm	1	1	1	1	1
	http://site.ifrance.com/okapi/quimica.htm	1	1	1	1	1
	http://www.webelements.com/	1	0,9	0,9	1	0,81
	http://www.prodigyweb.net.mx/degcorp/Quimica/Tabla_Periodica.htm	1	1	1	1	1
	http://www.aldeaeducativa.com/aldea/elementos.asp	1	1	1	1	1
	http://www.visionlearning.com/library/module_viewer.php?mid=52&l=s&c	0,9	1	0,9	0,81	1
	http://inicia.es/de/sistemaperiodico/	0,9	1	0,9	0,81	1
P15	http://www.liceus.com/cgi-bin/aco/his/02/01/0100.asp	1	1	1	1	1
	http://icarito.tercera.cl/icarito/2001/821/	1	0,9	0,9	1	0,81
	http://www.geocities.com/athens/marble/1732/	1	0,9	0,9	1	0,81
	http://www.egiptomania.com/	1	0,8	0,8	1	0,64
	http://www.geocities.com/serbino/piramide.html	1	1	1	1	1
	http://www.monografias.com/cgi-bin/search.cgi?query=egipto	1	0	0	1	0
	http://egipto.com/arte/articulos/34.html	1	0	0	1	0
	http://www.tubreveespacio.com/remar-piramides%20de%20egipto.htm	0	1	0	0	1
	http://www.casadellibro.com/fichas/fichabiblio/0,1094,2900000577639,00.	0	1	0	0	1
	http://www.step.es/personales/jms/egipto/egipto2.html	0	1	0	0	1
P16	http://www.cdc.gov/travel/diseases/yellowfever.htm	1	0,9	0,9	1	0,81
	http://www.cdc.gov/ncidod/dvbid/yellowfever/	1	1	1	1	1
	http://www.astdhpphe.org/infect/yellow.html	1	1	1	1	1
	http://www.who.int/csr/disease/yellowfev/en/	1	1	1	1	1
	http://www.health.state.ny.us/nysdoh/communicable_diseases/en/yellow.h	1	1	1	1	1
	http://www.emedicine.com/emerg/topic645.htm	1	1	1	1	1
	http://www.who.int/mediacentre/factsheets/	1	0,9	0,9	1	0,81
	http://www.netdoctor.co.uk/travel/diseases/yellowfever.htm	1	0,9	0,9	1	0,81
	http://www.shopping.com/xGS-Yellow_Fever~NS-1~linkin_id-3062971~	0	1	0	0	1
P17	http://www.sinaproc.gob.pa/maremotos.htm	1	1	1	1	1
	http://www.angelfire.com/nt/tsunamis/	1	1	1	1	1
	http://www.maristas.com.ar/champagnat/poli/terremoto.htm	1	1	1	1	1
	http://es.wikipedia.org/wiki/Tsunami	1	0,8	0,8	1	0,64
	http://nees.orst.edu/IT/info/InformationES.htm	1	1	1	1	1
	http://www.familia.cl/Framearea.asp?p=c&c=779	1	0	0	1	0
	http://www.ineter.gob.ni/geofisica/sis/dep-sis.html	1	1	1	1	1
	http://www.epoca.es/node_463.jsp?id_producto=8199&id_categoria=142	1	0	0	1	0
	http://www.sierramaestra.cu/sismico.htm	1	0	0	1	0
	http://www.cisas.org.ni/Gestion/zonas/maremotos.htm	0	1	0	0	1
	http://tsunami174.tripod.com/id2.html	0	1	0	0	1
	http://www.latinoseguridad.com/LatinoSeguridad/Fenat/Tsunami.shtml	0	1	0	0	1
	http://www.escalofrio.com/n/Catastrofes/Maremotos_Tsunamis/Maremoto	0	1	0	0	1
P18	http://www.lib.umd.edu/MCK/GUIDES/plant_names.html	1	1	1	1	1
	http://www.plantsall.com/commonscientificnamesforplants/	1	0,9	0,9	1	0,81
	http://www.desert-tropicals.com/Plants/sci_names_A.html	1	1	1	1	1
	http://elib.cs.berkeley.edu/photos/flora/sci-A.html	1	0	0	1	0
	http://www.tpwd.state.tx.us/gis/vegetation_types/appendix/	1	0,8	0,8	1	0,64
	http://botanicallatin.org/latinhandout.pdf	1	0	0	1	0
	http://members.aol.com/magarland/botlat/testhand.htm	0	1	0	0	1
	http://www.citygardening.net/comname/	0,9	1	0,9	0,81	1
	http://www.backyardnature.net/namelatn.htm	0,8	1	0,8	0,64	1
	http://www.coepark.org/plantlist.html	0	1	0	0	1
	http://www.geocities.com/RainForest/7109/latin.htm	0,9	1	0,9	0,81	1

PONDERACION TOTAL:	6,79	8,37	8,82
		2,89	2,97
			8,59

FUNCION DEL COSENO: **0,7899** **78,9%**