ANALES DE DOCUMENTACION, Nº 11, 2008, PÁGS. 197-218

ESTUDIO DE LA SIMILITUD DE LA RESPUESTA DE LOS PRINCIPALES MOTORES DE BÚSQUEDA EN LA WEB

José Vicente Rodríguez Cáceres*

Grupo de Tecnologías de la Información. Universidad de Murcia.

Francisco Javier Martínez Méndez

Grupo de Tecnologías de la Información. Universidad de Murcia.

José Vicente Rodríguez Muñoz

Grupo de Tecnologías de la Información. Universidad de Murcia.

Resumen: El aumento de la información en la web y su empleo como fuente principal para la recuperación de información en Internet propicia que el uso de los sistemas de recuperación de información en la web (los motores de búsqueda), cobre mayor auge. Si bien estas herramientas realizan una labor encomiable, no resulta menos necesaria la tarea de evaluar su rendimiento y analizar la información que proporcionan. Nuestro estudio propone llevar a cabo un análisis de la similitud de los resultados ofrecidos por los principales motores de búsqueda. Para ello se ha construido un *metabuscador* que nos va a permitir realizar los experimentos de búsqueda y los cálculos de estos valores de una manera rápida e interactiva. Finalmente se intentará extraer un conjunto de conclusiones válidas de estos resultados, contrastándolas con datos obtenidos en otros trabajos y entre los mismos ofrecidos por nuestro experimento, que viene a demostrar el amplio grado de divergencia entre las respuestas de estos sistemas de recuperación de información, tratándose éste de un tema de relevante interés a la par que no definitivamente consensuado en su solución.

Palabras clave: Buscadores web; evaluación de la recuperación de información; sistemas de recuperación de información.

Title: A STUDY ON THE SIMILARITY IN THE RESULTS OFFERED BY THE MAIN SEARCH ENGINES ON THE WEB.

Abstract: The huge increment of the available information on the web added to a progressive augment of its use as main source of data has caused that the utilization of information retrieval systems -IRS- in the web (also known as search engines), takes on a higher relevance. Even though these tools carry out an essential work, no less necessary is the task of evaluating them and studying the data that they provide us. This research proposes to analyze the main search engines of the market through a direct comparison of the URLs returned from a heterogeneous set of queries to intend to find out the current degree of similarity of the responses. For this purpose we have developed a metasearch engine which will let us make our search experiment and calculate the distances in a quick and interactive way. Finally, we will try to get valid conclusions of the results contrasting them with data from other researches and from our own experiment, what will demonstrate the high degree of divergence in the responses of the IRS.

* gti@um.es Grupo de Tecnologías de la Información

Keywords: Web search engines; information retrieval evaluation; information retrieval systems.

"¿Cómo podría el mundo trazar una trayectoria hasta tu puerta cuando dicha trayectoria se encuentra inexplorada, descatalogada, y sólo se podría descubrir por una hermosa casualidad?".

Paul Gister

1. INTRODUCCIÓN

Los sistemas de recuperación de información en la web han sido evaluados prácticamente desde su implantación. Esta necesidad de evaluar siempre ha acompañado al desarrollo de los sistemas de recuperación de información, por lo que no es de extrañar esta circunstancia. Son muy numerosos los aspectos evaluados, tal como podemos consultar en Oppenheim (2000) y Martínez y Rodríguez (2003). Los análisis de efectividad de la recuperación de la información son los trabajos más citados, destacando sobremanera el trabajo de Chu y Rosenthal (1996) como el que más características analizó; las sucesivas aportaciones de Leighton y Svristava (1999) y finalmente el trabajo de Gordon y Pathak (1999), el que más sistemas evaluó.

Es un hecho cierto que estos sistemas almacenan vastas colecciones de documentos en sus índices y que su parecido es bastante pequeño. Tradicionalmente se ha asumido que este parecido es del 15%, dato que de ser cierto (normalmente este tipo de comentarios no suele venir apoyado por estudios que los avalen), supondría una enorme diversidad en la composición de estos índices (algo que más o menos todo el mundo piensa cuando hace una misma búsqueda en distintos sistemas, acción cada vez menos frecuente, por cierto). A este factor debemos unir otro quizá más influyente y decisivo, el alineamiento o ranking que emplea el sistema de recuperación de la información para presentar la respuesta al usuario. Cada sistema implementa un algoritmo diferente y por tanto, el grado de divergencia de la respuesta que podemos apreciar los usuarios es aún mucho mayor.

¿Cuál de los dos grados de similitud/divergencia es el que nos interesa analizar? Desde el punto de vista del usuario final elegiríamos el de la respuesta ofrecida, desde el punto de vista del administrador de un sistema de recuperación de información en la web también tendría interés analizar este factor sobre la composición de los índices pero para los usuarios de la web su importancia es secundaria. Todo ello sin olvidar que difícilmente se podrá acceder a todas estas colecciones para analizarlas.

Datos sobre esta concordancia sólo aparecen en los trabajos de Losjland (2000a), (2000b) y Martínez (2002), además de los datos sobre el solapamiento de los sistemas de recuperación de información en la web que calculaba Notess (2002) para la web *searchengineshowdown.com*, pero hace tiempo que no disponemos de una fuente que nos aporte información actualizada al respecto. Es por ello que consideramos interesante aportar un método que permita llevar a cabo análisis de forma periódica de este factor y aplicarlo sobre las respuestas de los sistemas de recuperación de información en la web más utilizados en la web actual.

2. PANORAMA ACTUAL Y OBJETO DEL ESTUDIO

La práctica totalidad de usuarios de la web realizan sus tareas de búsqueda en no más de media docena de entre la plétora de sistemas de recuperación de información disponibles. Los motivos que impulsan a un internauta a elegir un determinado ingenio de búsqueda se deben más a razones subjetivas (familiaridad con el mismo, pleno convencimiento de su calidad, desconocimiento del resto, costumbre de uso, etcétera) que a una evaluación crítica y razonada de las alternativas.

El ranking realizado por la consultora especializada Nielsen/NetRatings (2007) a partir de los datos del mes de julio de 2007 respecto al número de usuarios y búsquedas realizadas sobre los principales buscadores web no deja lugar a dudas, el número de búsquedas de tres motores de búsqueda -Google (53%), Yahoo! (20%) y Windows Live (14%)- totalizan el 87% del total. Tan sólo dos competidores -AOL y Ask.com- parecen mantener muy de lejos el ritmo mientras que el resto se reparten un escaso 5% del mercado. Esta tremenda concentración, mayor en otros estudios como es el caso del realizado en diciembre de 2007 por la consultora Comscore (Burns, 2007) disipa cualquier duda sobre qué buscadores seleccionar para nuestro estudio porque analizando la respuesta de estos tres motores estaríamos diseñando un estudio que cubre casi el noventa por ciento de las búsquedas realizadas diariamente por la comunidad de usuarios de la red.

Además, estos tres grandes sistemas de búsqueda han recopilado (o dicen haber recopilado) unas extensísimas colecciones documentales cada uno de ellos, quedando pendiente de conocer cuál es el verdadero tamaño de la web y también por qué se diferencian tanto los resultados que ofrecen los motores. Una posible vía de cálculo de esa diferencia puede ser comparar los resultados que ofrece cada motor sobre una misma colección documental de prueba (por ejemplo *TREC Web* o *Terabyte*[1]), lo que permitiría valorar los distintos métodos de alineamiento independientemente de las características de la colección externa y de las respuestas patrocinadas de cada motor. Pero en realidad esta vía de estudio no ofrecería datos propios del contexto de la web, hábitat del cual algunos de estos sistemas extraen información para sus sistemas de alineamientos -*Pagerank* (Page, 98), el algoritmo de alineamiento de Google, precisa de información procedente de la misma web- sino que aportaría datos de una versión muy reducida de la web. Es por ello que son varios los proyectos que prefieren emplear a la propia web como corpus documental sobre el que realizar los experimentos (Palfrey, 2006).

Antes de abordar la toma de muestras para el posterior análisis de evaluación de los motores seleccionados, se ha realizado un somero estudio de los sistemas de alineamiento que emplean cada uno de ellos. Todo ello, con objeto de dar una mejor visión y claridad a la hora de justificar la elección de este conjunto de sistemas de recuperación de información.

Google

El ranking de páginas web de Google se basa en el sistema *PageRank* implementado por los fundadores de la compañía Larry Page y Sergey Brin. Desde Google (2007) se afirma que este sistema se basa únicamente en la "naturaleza democrática de la web", su

[1] Más información sobre estas colecciones documentales de prueba en la web del *Grupo de Investigación en Recuperación de Información de la Universidad de Glasgow*, <http://ir.dcs.gla.ac.uk/test_collections/>. [Consulta: 12 de febrero de 2008].

vasta estructura de enlaces, como un indicador del valor de una página. Además del número de enlaces que recibe una página, Google analiza el valor de la página desde la que se enlaza por lo que cada enlace será ponderado por el valor de la propia página que lo contiene. El sistema *PageRank* se combina con sofisticadas técnicas de búsqueda de texto, tanto en la página enlazada como en la que contiene los enlaces, de forma que cada *query* encaje en mejor medida con los resultados ofrecidos. Finalmente, Google aclara que además del número de veces que un término aparece en una página, se analizan docenas de aspectos del contenido de la misma, entre las que no destaca el uso de las metaetiquetas, cuestión esta que cuando menos resulta curiosa dada la inmensa bibliografía que sobre las mismas hay desarrollada.

Yahoo!

Yahoo! ofrece una serie de consejos para mejorar la posición de un sitio web dentro de su ranking de resultados. En primer lugar se recomienda usar aquellos términos que se consideren clave en la construcción de la web. Además, se destaca el uso de las metaetiquetas '*description*' y '*keyword*' y se recomienda el uso de enlaces en HTML y de texto alternativo para imágenes. Sólo al final, se encomienda al desarrollador el uso de enlaces con páginas relevantes.

Windows Live

Este buscador enumera una serie de prácticas que pueden hacer mejorar la posición de un sitio web en el ranking que se encarga de realizar su web-spider MSNBot. Especial atención merece la construcción HTML well-formed que favorece la correcta indexación por parte de MSNBot y que este motor aprecia en positivo. Adicionalmente, se recomienda el uso de un fichero robots.txt o bien de metaetiquetas para ayudar al indexador. Es importante destacar que en los tres casos se trata de disuadir al desarrollador web del uso de técnicas fraudulentas de mejora en los rankings también conocidas como *Web Spam* (Hunt, 2005).

3. PROCESO DE RECUPERACIÓN DE DATOS

Seleccionados los buscadores sobre los que se realizará el estudio, procede ahora abordar la fase de toma de muestras, de tal forma que nos permita a posteriori el cálculo de similitud de los resultados obtenidos y en consecuencia una comparación de los buscadores en cuanto al conjunto de la respuesta emitido por cada uno de ellos.

El primer paso consiste en el análisis de la APIs[2] para búsqueda que cada buscador proporciona. El resultado de este análisis no hizo sino confirmar lo expuesto anteriormente, en el camino hacia la supremacía en el mercado de la recuperación de información en la red cada motor de búsqueda establece sus propias reglas de juego, limitando duramente, como en el caso de Google, el número de resultados obtenidos (hasta hace unas semanas ocho como máximo[3]) u obligando a la adopción de lenguajes de programación propietarios como el caso de Windows Live. Por el momento, es Yahoo! el buscador que permite

[2] API son las siglas de *Application Programming Interface*, que consisten en una serie de funciones que están disponibles para realizar programas para un cierto entorno, en este caso el de la búsqueda en la web.

[3] El límite de 8 documentos por respuesta se modificó a 32 documentos por respuesta hacia finales de enero de 2008, cuando la parte de toma de datos de este estudio estaba finalizada.

un uso de su índice menos limitado. Sin duda alguna, no podemos más que concluir que cualquier decisión tomada por los propietarios de estas herramientas respecto a la apertura al público desarrollador, y por ende a sus principales competidores, de sus índices de búsquedas y sus tan reservadamente establecidos criterios de ranking se deben siempre a razones de negocio y en ningún caso a limitaciones de la tecnología.

Como se ha expuesto, la API de Google restringía a ocho el número de resultados devueltos para cada consulta, lo que ha sesgado parte del alcance de nuestro estudio, obligando a tomar tan solo ocho muestras de cada motor para cada *query*. Sin embargo, no creemos que este factor menosprecie los resultados obtenidos al final del estudio ya que, en cualquier caso, habremos logrado una comparación de casi la totalidad de la primera página de resultados devuelta por cada motor, la única parte de la respuesta de un motor que consultan detenidamente más del 80% de sus usuarios.

El siguiente paso del proceso requiere la implementación del código necesario para interrogar a cada ingenio y recuperar los conjuntos de resultados (*resultsets*) proporcionados por cada uno, con el objetivo de crear la base de datos de muestras que finalmente será estudiada y analizada. Como resultado de este proceso se desarrolla un *metabuscador* capaz de consultar cualquiera de los tres motores de búsqueda y de almacenar la información en una base de datos. Debido a la heterogeneidad de las APIs dispuestas por cada buscador, se plantea la necesidad de, en la medida de lo posible, establecer un criterio de programación sobre el cual disponer el núcleo de nuestro *metabuscador* y, a partir de este, diseñar el resto de piezas de esta herramienta. Lo obtenido al final es una herramienta en la que se fragmenta el código dentro de la misma según su funcionalidad. En resumen se dispone de:

- **Núcleo de navegación** basado en los lenguajes de programación PHP y Javascript, y en código HTML.
- **Motor de búsqueda**. Se diferencian los fragmentos para cada buscador:
  - **Google**. Realiza la comunicación mediante funciones Javascript usando la tecnología AJAX.
  - **Yahoo!** Utiliza comunicación vía REST queries mediante código PHP.
  - **Windows Live**. Dispone de un Web Service al que se accede gracias a un cliente SOAP escrito en PHP.
- **Interfaz de base de datos**. Escrita en PHP, se encarga de la comunicación con la base de datos MySQL.

Figura 1. Interfaz del metabuscador construido.

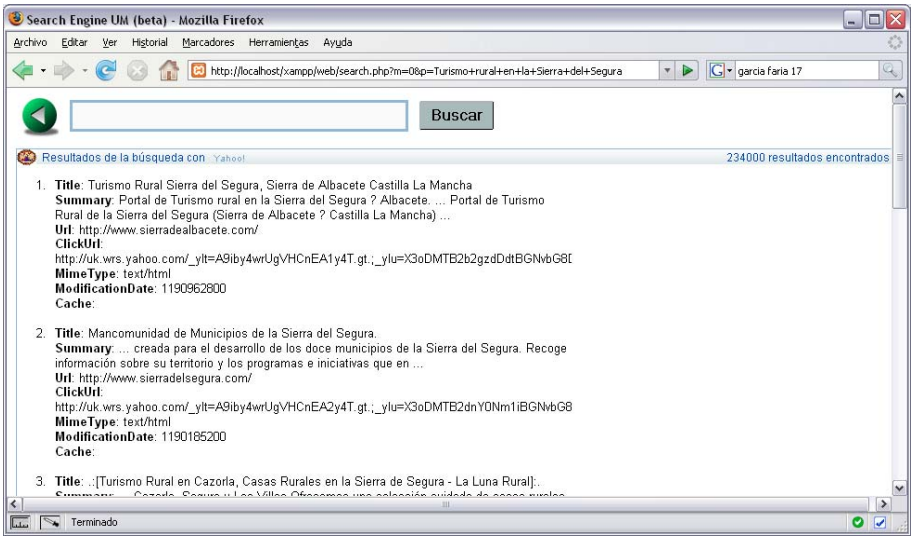

Figura 2. Conjunto de la respuesta de una búsqueda del metabuscador.

Por su parte, el diseño de la base de datos permite el almacenamiento de los resultados de búsqueda para cada consulta. Más concretamente, recopila para cada resultado la URL, el título, la descripción y su posición en el ranking, así como su relación con la dupla

query-motor. Además, se dispone de una tabla para almacenar las estadísticas referentes a cada consulta.

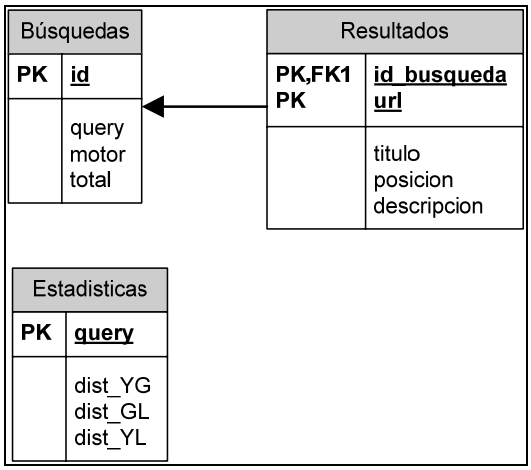

Figura 3. Tablas para el almacenamiento de resultados de cada búsqueda.

Para el último paso, es necesaria la selección de las *queries* sobre las que recoger todas las muestras del estudio. En total hemos trabajado con 80 *queries* distintas, divididas en tres grupos bien diferenciados. Por una parte se han tomado las treinta consultas empleadas en el estudio de Martínez Méndez (2002), así dispondremos de una experiencia previa con la cual podremos comparar nuestros resultados.

Queries	
Turismo rural en la Sierra del Segura	Alquiler de apartamentos en Málaga
Historia del Camino de Santiago	Curso a distancia de Programación en PHP
Principio de incertidumbre de Heisenberg	Diseño de sistemas multimedia para el aprendizaje
Academias de idiomas en Valencia	Discurso del Método de Descartes
Diseño accesible a páginas Web	Recetas de cocina y dieta mediterránea
Teoría de la Evolución de Darwin	Semana Santa en Murcia
Bibliografía de Miguel de Unamuno	Estrategias de Representación del Conocimiento
Galerías de Arte en Murcia	Empresas de fabricación de calzado en Alicante
Influencia de la televisión en los niños	Apuntes de Sistemas Digitales
Apuntes de Estadística Descriptiva	Modelos pedagógicos para la educación a distancia
Principio de Conservación de la Energía	Librerías de antiguo en España
Apuntes de Historia del Arte Barroco	Temario de Oposiciones de Matemáticas en Secundaria
Recopilación de Legislación en Derecho Civil	Historia de la ciudad de Ceuta
Compra-Venta de automóviles de segunda mano en Madrid	Evaluación de la calidad de la enseñanza universitaria
Literatura Española en el Siglo de Oro	Plan de Estudios de Licenciado en Comunicación Audiovisual

Tabla I. Consultas aplicadas en Martínez Méndez (2002).

En segundo lugar se investigó el *Top 10* de búsquedas durante 2006 en Estados Unidos en los principales sistemas de búsqueda (Kopytoff, 2007), lo que nos proporcionó otro grupo de treinta consultas. De esta forma podremos comparar precisamente entre un buscador y otro sus *queries* más demandadas.

Google	***Yahoo!***	***Ask.com***
Bebo	Britney Spears	MySpace
MySpace	WWE	Dictionary
World Cup	Shakira	Games
Metacafe	Jessica Simpson	Cars
Radioblog	Paris Hilton	Food
Wikipedia	American Idol	Song lyrics
Video	Beyonce Knowles	Poems
Rebelde	Chris Brown	New York
Mininota	Pamela Anderson	Baby names
Wiki	Lindsay Lohan	Music

Tabla II. *Top 10* de los buscadores estudiados en Estados Unidos.

Finalmente, se seleccionaron al azar términos en español menos específicos que en el primer grupo, intentando así contrastar los resultados de ambos.

Google	***Yahoo!***	***Ask.com***
Agencias de viajes	Cocina mediterránea	Politonos
Alquiler en Valencia	Coches de ocasión	Recetas de cocina
Baloncesto	Comida a domicilio	Restaurantes japoneses
Barcelona	Conciertos en Madrid	TDT
Bibliotecas municipales	China	Videojuegos
Boda real	Energía renovable	Vinos rosados
Cámaras de fotos	Mecí	
Cartelera de teatro	Móviles	
Cine y televisión	Ordenadores portátiles	

Tabla III. Términos de contraste para el análisis.

4. METODOLOGÍA PARA CALCULAR LA SIMILITUD DE LOS RESULTADOS DE LA RESPUESTA DE LOS MOTORES DE BÚSQUEDA

El objeto del trabajo se centra en la comparación literal de una pequeña muestra de las URLs devueltas por cada ingenio de búsqueda ante una determinada *query*. Si consideramos estas muestras como vectores de datos podemos aseverar que estamos ante un claro ejemplo de cálculo de similitud de vectores. Se dispone de varias alternativas de funciones (*Coseno, Dice, Jaccard,* etc.) aunque todas presentan un problema, todas operan con resultados únicos y en nuestro caso disponemos de resultados duales (el binomio formado por un motor y la respuesta). Para poder aplicar una función de las anteriores hemos de reducir el número de dimensiones. Para conseguirlo seguimos lo propuesto en Martínez Méndez (2002), transformando la información a resultados individuales de cada motor

con respecto a un único espacio de documentos y posteriormente para determinar la similitud se empleará la *función coseno,* basada Modelo del Espacio Vectorial (Salton, 1975). Esta función ha demostrado empíricamente ser la que mejores resultados ofrece en otros experimentos similares realizados en el contexto de la web (Spertus, 2005) y además resulta mucho más intuitiva y cercana para los usuarios que otras.

4.1 Procedimiento de cálculo de similitud de dos vectores de resultados en el experimento

Dentro de un estudio algo más amplio cuyo objetivo general estaba dirigido a establecer el grado de efectividad de la respuesta de varios motores de búsqueda (Martinez, 2002) se calculó también el nivel de similitud de la respuesta ofrecida por estos sistemas (Google, Altavista, All the Web, Terra, MSN y Wisenut). Para ello se realizaron 30 consultas en cada uno de estos sistemas y se analizó el grado de similitud existente entre los 10, 20 y 30 primeros documentos de la respuesta. Inspirados en Lojslund (1999), se introdujeron una serie de adaptaciones a los vectores para hacer posible aplicar la función de similitud.

$$Cos(P,N) = \frac{\sum_{i=1}^{n}(p_i \times n_i)}{\sqrt{\sum_{i=1}^{n} p_i^2} \times \sqrt{\sum_{i=1}^{n} n_i^2}}$$

Ilustración 1: función de similitud del coseno.

Al mismo tiempo se introdujo una ponderación del peso de cada documento devuelto en función de su posición dentro del vector respuesta (ponderación denominada *factor de relevancia*). Este factor permitió añadir un cierto grado de precisión a la hora de discriminar la distancia de los resultados para cada sistema, intentando calcular no sólo la coincidencia de documentos en la respuesta (el solapamiento, en definitiva) sino intentar medir cómo se parece la respuesta en cuanto al orden de cómo son entregados estos documentos. Para ilustrar nuestra propuesta, se dispone del siguiente ejemplo: se ha realizado una búsqueda sobre los sistemas A y B, cuyos resultados se representan en la siguiente tabla (se resaltan en negrita aquellas URLs coincidentes); en la columna de la derecha aparecen los pesos asignados a cada URL en función de su posición (*factor de relevancia*):

Sistema A		**Sistema B**	
	Peso		**Peso**
web.first.com	1	**web.first.com**	1
web.second.com	0.99	web.two.com	0.99
web.third.com	0.98	web.three.com	0.98
web.fourth.com	0.97	**web.second.com**	0.97
web.fifth.com	0.96	**web.fifth.com**	0.96
web.sixth.com	0.95	**web.sixth.com**	0.95
web.seventh.com	0.94	**web.third.com**	0.94
web.eighth.com	0.93	web.eight.com	0.93

Tabla IV. Ejemplo de cálculo de pesos para medir el factor de relevancia. Si el documento es no relevante su peso es igual a cero.

En este punto se dispone de una distribución de elementos con dos atributos: URL y peso del objeto. Para aplicar la función de similitud debemos reducir la distribución obtenida a un espacio n-dimensional de elementos con un único atributo, de manera que *n* será el número de URLs coincidentes en la respuesta de ambos motores de búsqueda más el número de URLs aportadas por uno sólo de los buscadores. Se necesita diseñar un método que transforme los iniciales vectores de resultados duales en vectores de naturaleza simple, cuyos componentes sean los pesos que presenta cada URL en el buscador original con respecto a espacio común de documentos. Así, la matriz "Sistema A" (respuesta del motor A) se transforma en el vector "VMotorA" y la matriz "Sistema B" se transforma en "VMotorB", cuyos valores refleja la siguiente tabla, que representa todas las URLs localizadas (el vector resultado global) y los vectores VMotor_A y VMotor_B que en realidad muestran la respuesta de cada uno de los sistemas por separado al vector resultado global:

Vector Resultado	**VMotor_A**	**VMotor_B**
web.first.com	1	1
web.second.com	0.99	0.97
web.third.com	0.98	0.94
web.fourth.com	0.97	0
web.fifth.com	0.96	0.96
web.sixth.com	0.95	0.95
web.seventh.com	0.94	0
web.eighth.com	0.93	0
web.two.com	0	0.99
web.three.com	0	0.98
web.eight.com	0	0.93

Tabla V. Ejemplo de representación vectorial de los buscadores A y B.

Ahora que se manejan vectores con resultados individuales, sí se puede calcular la función de similitud *Coseno (*VMotor_A, VMotor_B*)*. El resultado en este caso es un valor de 0.63, que equivale a una similitud del 63%.

4.2 Análisis de los resultados

Una vez realizadas todas las consultas expuestas anteriormente mediante nuestro *metabuscador*, disponemos de toda la información necesaria para llevar a cabo el experimento propuesto. Por tanto, siguiendo el procedimiento ya descrito lo que corresponde es calcular las distancias entre cada vector de resultados devueltos por cada ingenio de búsqueda.

En la tabla VI se muestran las similitudes calculadas para las consultas del trabajo original de Martínez Méndez (2002). Se resaltan en negrita los valores máximos para cada columna de similitudes que como se puede observar en la tabla coincide en los tres casos para la misma consulta.

Query	*Similitud Google-Yahoo!*	*Similitud Yahoo!-Windows Live*	*Similitud Google-Windows Live*
Academias de idiomas en Valencia	0,2462	0,3828	0,2513
Alquiler de apartamentos en Málaga	0,0000	0,2591	0,0000
Apuntes de Estadística Descriptiva	0,2603	0,1315	0,1249
Apuntes de Historia del Arte Barroco	0,1342	0,3735	0,0000
Apuntes de Sistemas Digitales	0,1342	0,0000	0,0000
Bibliografía de Miguel de Unamuno	0,2460	0,2564	0,2550
Compra-Venta de automóviles de segunda mano en Madrid	0,1342	0,2536	0,0000
Curso a distancia de Programación en PHP	0,1328	0,0000	0,1315
Discurso del Método de Descartes	0,1315	0,1275	0,0000
Diseño accesible a páginas web	0,2564	0,3865	0,1315
Diseño de sistemas multimedia para el aprendizaje	0,1248	0,1248	0,0000
Empresas de fabricación de calzado en Alicante	0,0000	0,1160	0,0000
Estrategias de Representación del Conocimiento	0,0000	0,1315	0,0000
Evaluación de la calidad de la enseñanza universitaria	0,0000	0,0000	0,0000
Galerías de Arte en Murcia	0,2550	0,2537	0,2564
Historia de la ciudad de Ceuta	0,1274	0,0000	0,0000
Historia del Camino de Santiago	0,5013	0,2409	0,4944
Influencia de la televisión en los niños	0,2576	0,1275	0,2590
Librerías de antiguo en España	0,1275	0,0000	0,0000
Literatura Española en el Siglo de Oro	0,2371	0,0000	0,1235
Modelos pedagógicos para la educación a distancia	0,1236	0,0000	0,0000
Plan de Estudios de Licenciado en Comunicación Audiovisual	0,0000	0,0000	0,0000
Principio de Conservación de la Energía	0,2459	0,1328	0,1288
Principio de incertidumbre de Heisenberg	0,5038	0,2499	0,2538
Recetas de cocina y dieta mediterránea	0,0000	0,1262	0,0000
Recopilación de Legislación en Derecho Civil	0,1185	0,1342	0,1211
Semana Santa en Murcia	0,0000	0,0000	0,1249
Temario de Oposiciones de Matemáticas en Secundaria	0,1315	0,0000	0,0000
Teoría de la Evolución de Darwin	0,2474	0,2643	0,2499
Turismo rural en la Sierra del Segura	**0,5128**	**0,5155**	**0,5027**
Media	***0,1730***	***0,1529***	***0,1136***

Tabla VI. Resultados para el primer bloque de preguntas comparando las URLs completas.

En el trabajo original de Martínez Méndez la media obtenida al analizar los diez primeros resultados de seis motores de búsqueda era 0.134 y la similitud media calculada para Google y Microsoft Network (antecesor de Windows Live) fue de 0.15. Estos valores coinciden prácticamente con los obtenidos en el presente trabajo (0.149 y 0.11), de lo que

se deduce que la similitud media ha variado muy poco a pesar del increíble aumento del tamaño de los índices de estos buscadores (en 2002 Google rozaba los dos mil millones de documentos, MSN tenía un tamaño de trescientos millones; actualmente ambos motores superan los quince mil millones de documentos). En cuanto a los valores medios calculados para la similitud entre cada buscador en este bloque de búsquedas, podemos observar que, en general, los valores son bastante homogéneos, si bien los más cercanos resultan ser Google y Yahoo!

Al mismo tiempo hemos calculado la similitud considerando como documentos iguales a dos páginas alojadas en un mismo host o servidor web (nuestra idea es conocer ahora el grado de coincidencia de servidores en los primeros lugares de la respuesta, además no podemos olvidar que un amplio porcentaje de documentos que se localizan en la web suelen localizarse gracias a la navegación, así que entregar el mismo servidor en dos lugares de la respuesta puede parecer redundante). En este caso, los resultados (ver tabla VII), resultan superiores a los anteriores, de hecho los valores medios se duplican.

Query	*Similitud Google-Yahoo!*	*Similitud Yahoo!-Windows Live*	*Similitud Google-Windows Live*
Academias de idiomas en Valencia	**0,7603**	0,5090	0,5053
Alquiler de apartamentos en Málaga	0,0000	0,2591	0,0000
Apuntes de Estadística Descriptiva	0,5192	0,1315	0,1249
Apuntes de Historia del Arte Barroco	0,2643	0,6219	0,2510
Apuntes de Sistemas Digitales	0,3762	0,1160	0,0000
Bibliografía de Miguel de Unamuno	0,3735	0,6364	0,2550
Compra-Venta de automóviles de segunda mano en Madrid	0,2589	0,2536	0,0000
Curso a distancia de Programación en PHP	0,3824	0,3683	0,1315
Discurso del Método de Descartes	0,2629	0,2577	0,1275
Diseño accesible a páginas web	0,3892	0,5166	0,2616
Diseño de sistemas multimedia para el aprendizaje	0,2458	0,1248	0,1210
Empresas de fabricación de calzado en Alicante	0,2434	0,1160	0,0000
Estrategias de Representación del Conocimiento	0,0000	0,1315	0,0000
Evaluación de la calidad de la enseñanza universitaria	0,0000	0,3801	0,0000
Galerías de Arte en Murcia	0,5206	0,5180	0,5181
Historia de la ciudad de Ceuta	0,1274	0,1223	0,1288
Historia del Camino de Santiago	0,5025	0,6156	**0,8690**
Influencia de la televisión en los niños	0,2576	0,1275	0,2590
Librerías de antiguo en España	0,2564	0,0000	0,0000
Literatura Española en el Siglo de Oro	0,6262	0,3707	0,2458
Modelos pedagógicos para la educación a distancia	0,1236	0,0000	0,0000
Plan de Estudios de Licenciado en Comunicación Audiovisual	0,0000	0,1274	0,0000
Principio de Conservación de la Energía	0,6337	0,5127	0,5156
Principio de incertidumbre de Heisenberg	0,6393	0,5169	0,6444
Recetas de cocina y dieta mediterránea	0,2536	0,2422	0,2458
Recopilación de Legislación en Derecho Civil	0,4944	0,2656	0,1211
Semana Santa en Murcia	0,0000	0,1223	0,2485
Temario de Oposiciones de Matemáticas en Secundaria	0,5014	0,1236	0,0000
Teoría de la Evolución de Darwin	0,5090	0,3918	0,6392
Turismo rural en la Sierra del Segura	0,6376	**0,6402**	0,5027
Media	***0,3387***	***0,3040***	***0,2239***

Tabla VII. Resultados del primer bloque comparando solamente los nombres de *hosts*.

La similitud entre las respuestas de los tres motores a las consultas de términos en castellano de carácter más general que los del bloque anterior ofrece resultados ligeramente superiores que los obtenidos en el caso anterior. Igualmente, se verifica que la distancia entre Google y Yahoo! resulta la más próxima. (Ver tabla VIII).

Query	***Distancia Google-Yahoo!***	***Distancia Yahoo!-Windows Live***	***Distancia Google-Windows Live***
Agencias de viajes	0,3827	0,3812	0,2616
Alquiler en Valencia	0,1262	0,2550	0,0000
Baloncesto	0,5076	0,3814	0,3774
Barcelona	0,2576	0,0000	0,2562
Bibliotecas municipales	0,2537	0,1249	0,0000
Boda real	0,5012	0,2552	0,2590
Cámaras de fotos	0,0000	0,0000	0,0000
Cartelera de teatro	0,2485	0,3892	0,1275
Cine y televisión	0,0000	0,1261	0,3697
Cocina mediterránea	0,1236	0,2498	0,1223
Coches de ocasión	0,2536	0,1185	0,2511
Comida a domicilio	0,3645	0,2498	0,3760
Conciertos en Madrid	0,6350	0,1262	0,2538
China	0,3853	0,2510	0,2549
Energía renovable	0,0000	0,2590	0,0000
Fútbol	0,1185	0,2486	0,0000
Historia	0,1328	0,2525	0,0000
Imperio Romano	0,3893	**0,5091**	0,3853
Madrid	**0,7563**	0,2577	0,2590
Matemáticas	0,1328	0,2512	0,0000
Medicina	0,1342	0,1223	0,1235
Messi	0,5102	0,1288	0,1301
Móviles	0,3879	0,1302	0,1235
Ordenadores portátiles	0,2552	0,1342	0,3918
Politonos	0,0000	0,1198	0,1342
Recetas de cocina	0,2526	0,1288	0,3774
Restaurantes japoneses	0,0000	0,2576	0,0000
TDT	0,1342	0,1342	**0,5115**
Videojuegos	0,3852	0,1262	0,2603
Vinos rosados	0,2537	0,1236	0,2577
Media	***0,2627***	***0,2031***	***0,1955***

Tabla VIII. Resultado para los términos en castellano comparando las URLs completas.

Asimismo se ha calculado la similitud entre la respuesta de los motores tomando en consideración únicamente la dirección del servidor web. Al igual que pasaba en el primer bloque de preguntas, vuelve a aumentar considerablemente el parecido de la respuesta al reducirse las distancias, aunque no se llega a duplicar.

Query	***Similitud Google-Yahoo!***	***Similitud Yahoo!-Windows Live***	***Similitud Google-Windows Live***
Agencias de viajes	0,3827	0,3812	0,2616
Alquiler en Valencia	0,1262	0,6156	0,0000
Baloncesto	0,8657	0,5062	0,3774
Barcelona	0,3878	0,2656	0,5099
Bibliotecas municipales	0,6325	0,2511	0,1262
Boda real	0,6300	0,5027	0,3775
Cámaras de fotos	0,2458	0,0000	0,0000
Cartelera de teatro	0,5088	0,3905	0,2590
Cine y televisión	0,0000	0,1261	0,3697
Cocina mediterránea	0,2409	0,2511	0,2510
Coches de ocasión	0,3877	0,3842	**0,5154**
Comida a domicilio	0,3645	0,2498	0,3760
Conciertos en Madrid	0,7625	0,2577	0,3801
China	0,3853	0,2510	0,2549
Energía renovable	0,3931	0,3852	0,2616
Fútbol	0,2514	0,2486	0,0000
Historia	0,3826	0,2525	0,0000
Imperio Romano	0,6442	**0,7627**	0,5142
Madrid	**0,8825**	0,2577	0,2590
Matemáticas	0,3824	0,3826	0,2538
Medicina	0,1342	0,1223	0,1235
Messi	0,5102	0,3787	0,2590
Móviles	0,3879	0,2549	0,2483
Ordenadores portátiles	0,3763	0,1342	0,3918
Politonos	0,3892	0,2539	0,2670
Recetas de cocina	0,2526	0,1288	0,3774
Restaurantes japoneses	0,0000	0,6312	0,0000
tdt	0,1342	0,1342	0,5115
Videojuegos	0,5153	0,1262	0,2603
Vinos rosados	0,2537	0,1236	0,3852
Media	***0,3937***	***0,3003***	***0,2724***

Tabla IX. Resultado para términos en castellano comparando solamente los nombres de *hosts*.

Continuando con la presentación de resultados, mostramos la similitud de la respuesta de cada uno de los motores analizados para las consultas del *Top 10* norteamericano. En esta ocasión los valores medios vuelven a descender a los niveles ofrecidos por el primer bloque de preguntas y se repite el hecho de que los motores más cercanos son Google y Yahoo! (Ver tabla X).

Query	***Similitud Google-Yahoo!***	***Similitud Yahoo!-Windows Live***	***Similitud Google-Windows Live***
American Idol	0,2590	0,0000	0,2604
Baby names	0,0000	0,0000	**0,5060**
Bebo	0,1262	0,3800	0,1328
Beyonce Knowles	0,2512	0,1315	0,1288
Britney Spears	0,1315	0,1235	0,3773
Cars	0,2383	0,0000	0,0000
Chris Brown	0,0000	0,0000	0,0000
David Beckham	0,2643	0,3827	0,1328
Dictionary	0,1249	0,2563	0,1301
Food	0,1301	0,1173	0,0000
Games	0,2446	0,0000	0,0000
Jessica Simpson	0,1315	0,0000	0,2578
Lindsay Lohan	0,2576	0,1315	0,1315
Metacafe	0,0000	0,0000	0,2604
Mininova	0,0000	0,0000	0,1342
Music	0,0000	0,2434	0,0000
MySpace	0,1315	0,1315	0,1288
New York	0,0000	0,0000	0,0000
Pamela Anderson	0,2525	0,0000	0,1315
Paris Hilton	0,1342	0,2590	0,2564
Poems	0,0000	0,0000	0,1211
Radioblog	0,2510	0,3786	0,0000
Rebelde	0,3787	0,0000	0,0000
Shakira	0,3699	0,1211	0,3827
Song lyrics	0,0000	0,0000	0,2526
Video	0,0000	0,0000	0,0000
Wiki	**0,3892**	0,1315	0,2500
Wikipedia	0,2656	**0,3945**	0,3906
World Cup	0,1274	0,1342	0,1185
WWE	0,0000	0,1302	0,2656
Media	***0,1486***	***0,1149***	***0,1583***

Tabla X. Resultado para los términos Top 10 comparando las URLs completas.

Igual que en los casos anteriores se han obtenido similitudes medias considerando únicamente la dirección del servidor web ofrecido en la respuesta de cada motor (ver tabla XI). En este último caso, los ascensos han sido menos significativos que en los dos bloques de preguntas anteriores, aunque se ha mantenido la tendencia positiva.

Query	***Similitud Google-Yahoo!***	***Similitud Yahoo!-Windows Live***	***Similitud Google-Windows Live***
American Idol	0,3878	0,0000	0,2604
Baby names	0,2460	0,0000	**0,5060**
Bebo	0,2577	0,4998	0,1328
Beyonce Knowles	0,2512	0,1315	0,1288
Britney Spears	0,1315	0,2510	0,3773
Cars	0,3725	0,0000	0,0000
Chris Brown	0,0000	0,0000	0,0000
David Beckham	0,3945	0,3827	0,2616
Dictionary	0,1249	0,3774	0,1301
Food	0,1301	0,1173	0,0000
Games	0,2446	0,0000	0,0000
Jessica Simpson	0,2603	0,0000	0,2578
Lindsay Lohan	0,2576	0,1315	0,2630
Metacafe	0,2670	0,2670	0,3932
Mininova	0,0000	0,1223	0,2670
Music	0,0000	0,2434	0,0000
MySpace	0,2629	0,2564	0,3825
New York	0,1328	0,0000	0,0000
Pamela Anderson	0,2525	0,0000	0,1315
Paris Hilton	0,3878	0,4999	0,3812
Poems	0,2460	0,0000	0,1211
Radioblog	0,3772	0,5047	0,1198
Rebelde	**0,6391**	0,0000	0,0000
Shakira	0,3699	0,1211	0,3827
Song lyrics	0,0000	0,0000	0,2526
Video	0,2643	0,1315	0,0000
Wiki	0,3904	0,2577	0,3802
Wikipedia	0,2670	**0,6418**	0,3919
World Cup	0,2562	0,1342	0,1185
WWE	0,2550	0,1302	0,2670
Media	***0,2476***	***0,1734***	***0,1969***

Tabla XI. Resultado para los términos Top 10 comparando solamente los nombres de *hosts.*

Para finalizar nuestro análisis de resultados, se muestran tres gráficas en las que se comparan las similitud entre un buscador respecto de los otros dos estudiados. En concreto, se han dispuesto en el eje de abscisas las distancias obtenidas para todo el rango de consultas analizado sin atender a más orden que el alfabético para situarlas a lo largo del

eje. Además, se han representado las líneas de tendencia para cada grupo de datos tratando así de visualizar más fácilmente la diferencia entre ambas curvas.

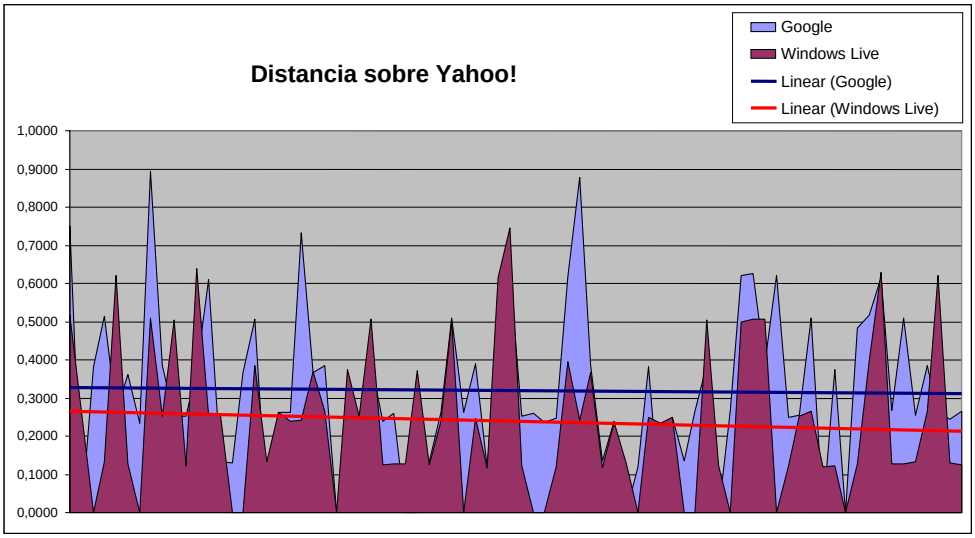

Figura 3. Distancia sobre Yahoo!

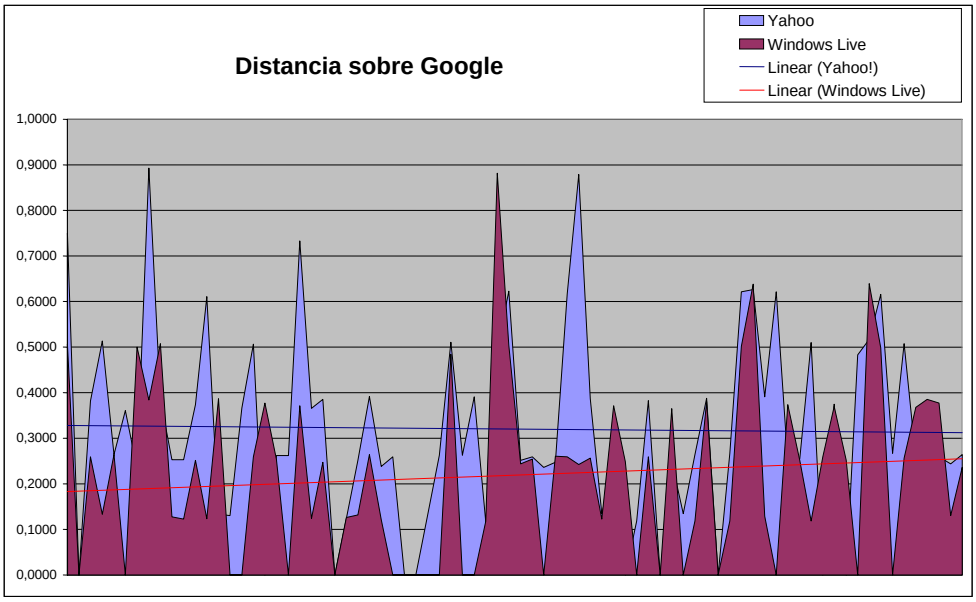

Figura 4. Distancia sobre Google.

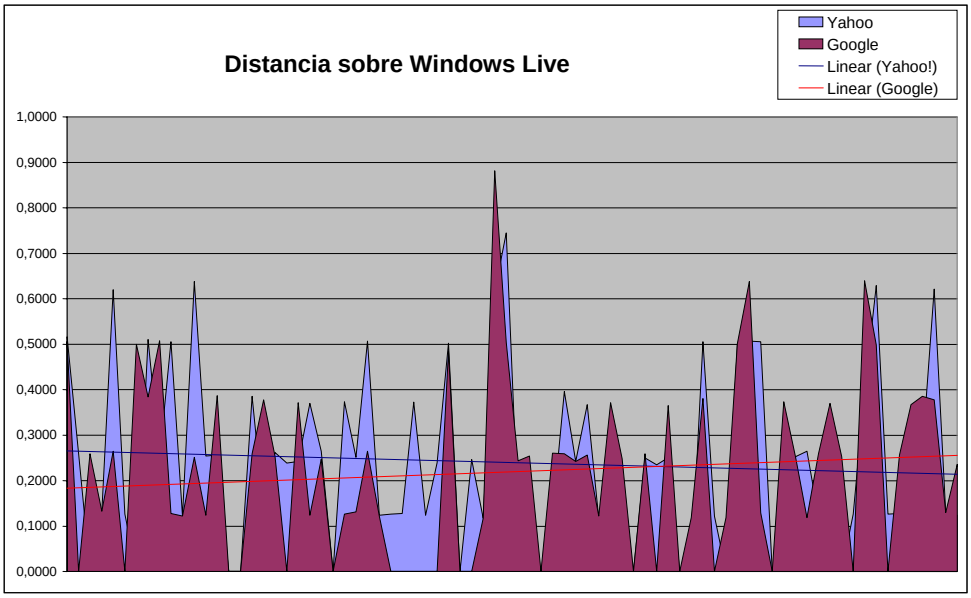

Figura 5. Distancia sobre Windows Live.

Como era de esperar en vista de los resultados anteriores, las gráficas demuestran que los resultados de Google y Yahoo! se encuentran más próximos comparándolos con los devueltos por Windows Live. Sin embargo, respecto a este último la distancia sobre Google y Yahoo! es muy similar. Más concretamente, los valores medios de todo nuestro estudio revelan los que se muestran en las tablas XII y XIII.

Similitud Google-Yahoo!	***Similitud Yahoo!-Windows Live***	***Similitud Google-Windows Live***	***Similitud Media***
0,1948	**0,1570**	**0,1558**	**0.1696**

Tabla XII. Valores medios de todo el muestreo comparando las URLs completas.

Similitud Google-Yahoo!	***Similitud Yahoo!-Windows Live***	***Similitud Google-Windows Live***	***Similitud Media***
0,3275	**0,2621**	**0,2336**	**0.274**

Tabla XIII. Valores medios de todo el muestreo comparando solamente los *hosts*.

5. CONCLUSIONES

A la vista de los resultados obtenidos por este estudio, aún estando algo lejos de pretender ser un análisis exhaustivo de la respuesta de los principales sistemas de recuperación de información en la web, el hecho de que la información que se ha obtenido coincide con la que la mayoría de los usuarios de estos sistemas utilizan, consideramos que sí se puede aspirar a proporcionar a la comunidad investigadora una visión objetiva del rendi-

miento ofrecido por los principales motores de búsqueda en el último tercio de 2007, aportando las siguientes conclusiones:

1. Los buscadores que se encuentran más próximos en su respuesta son Google y Yahoo!, si bien es cierto que como cabía esperar, las diferencias respecto al tercer competidor no son muy significativas. La similitud media entre Yahoo! y Windows Live supera en apenas una décima a la cercanía entre este último y Google.
2. Otro hecho significativo es el distinto comportamiento de los buscadores respecto a las *queries* específicas y a las *queries* de términos genéricos. Ha quedado reflejado en el experimento que los motores tienden a devolver resultados más parecidos en el caso de búsquedas de expresiones menos específicas en castellano que para queries más concretas.
3. Comparando los resultados de nuestro estudio con el de Martínez Méndez (2002) se verifica que el grado de la similitud en la respuesta de los motores de búsqueda apenas ha variado, a pesar de haber cambiado muy sustancialmente el entorno de la web en estos cinco años. El mayor tamaño de la respuesta que entregan hoy en día los buscadores debería implicar una mayor diversidad en el orden de la respuesta de cada uno de ellos (hay más documentos donde elegir para situar entre los diez primeros) y debería ser aún menor el parecido entre las respuestas, suposición que no parece cumplirse.
4. No es extraño que aumenten las similitudes de la respuesta al considerar idénticas dos direcciones de un mismo servidor, entre otras razones porque es una en modo alguno desatinada, ya que la mayoría de las veces acceder a una dirección lleva ineludiblemente a la otra. Pueden considerarse similares dentro de la óptica de un análisis coherente.
5. El dato quizá algo más desconcertante es la dispersión en los resultados ofrecidos por el conjunto de términos del *Top 10* de consultas en Estados Unidos. Centrándonos por ejemplo en el caso de Yahoo! y Google, los términos genéricos en Español muestran una similitud mucho mayor que para los términos del *Top 10*. Para justificar este resultado se podría barajar la hipótesis de que la programación de nuestro metabuscador se ha realizado ajustando las llamadas a cada motor con los parámetros de búsqueda en español. Sin embargo está hipótesis pierde su valor si tenemos en cuenta que gran parte de los términos hacen referencia a nombres propios anglosajones y vocablos ingleses aceptados en castellano.

6. REFLEXIÓN FINAL

Para ofrecer una mejor visión de la situación, se espera poder ampliar las prestaciones de nuestro metabuscador tanto en el número de documentos analizados como de motores de búsqueda (aunque no son muchos los que proporcionan APIs de forma óptima para su implementación). Si se consigue expandir el alcance del análisis, se estará en condiciones de mostrar una visión más completa de la situación actual en cuanto al parecido de la respuesta de estos ingenios de búsqueda.

7. BIBLIOGRAFÍA

BRIN, S. y PAGE, L. *The PageRank Citation Ranking: Bringing Order to the Web*. [En línea]. Computer Science Department, Stanford University, 1999 <http://dbpubs.stanford.edu/pub/1999-66>. [Consulta: 3 de agosto de 2007].

BURNS, E. *Top 10 Search Providers.* [En línea]. The Clicz Network, 2007. <http://searchenginewatch.com/showPage.html?page=3626903>. [Consulta: 3 de septiembre de 2007].

BURNS, E. *U. S. Search Engine Rankings, December 2007*. [En línea] searchenginewatch.com: 2007. <http://searchenginewatch.com/showPage.html?page=3628341>. [Consulta: 12 de febrero de 2008].

CHU, H. y ROSENTHAL, M. Search engines for the World Wide Web: a comparative study and evaluation methodology. *ASIS 1996 Annual Conference*, October 19-24, 1996. <http://www.asis.org/annual-96/ElectronicProceedings/chu.html>. [Consulta: 12 de enero de 2008].

GOOGLE. *Our Search: Google Technology* [En línea] Google: Moutain View, CA, 2007. <http://web.google.com/technology/>. [Consulta: 29 de septiembre de 2007].

GORDON, M. y PATHAK, P. Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management* 35, 1999. p. 141-180

HUNT, B. *What, Exactly, is Search Engine Spam?* [En línea] searchenginewatch.com: 2005. <http://searchenginewatch.com/showPage.html?page=3483601>. [Consulta: 30 de septiembre de 2007].

KOPYTOFF, V. *Year's top search terms* [En línea] San Francisco: Chronicle, 2007. <http://sfgate.com/cgi-bin/article.cgi?f=/c/a/2006/12/25/BUGOBN387R1.DTL>. [Consulta: 4 de agosto de 2007].

LEIGHTON, H. V. y SRIVASTAVA, J. First 20 precision among World Wide Web search services (search engines). *Journal of the American Society for Information Science* 50 (10), 1999, p. 870-881.

LJOSLAND, M. (2000b) *Evaluation of twenty Web search engines on ten rare words ranking algorithms.* Trondheim and Sør-Trøndelag: University, 2000. [En línea] <http://www.aitel.hist.no/~mildrid/dring/paper/Comp20.doc>. [Consulta: 11 de enero de 2008].

LJOSLAND, M. Evaluation of Web search engines and the search for better ranking algorithms. *SIGIR99 Workshop on Evaluation of Web Retrieval* August 19, 1999. [En línea] Trondheim and Sør-Trøndelag: University, 2000. <http://www.aitel.hist.no/~mildrid/dring/paper/SIGIR.html>. [Consulta: 18 de enero de 2008].

MARTÍNEZ MÉNDEZ, F. J. *Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en Internet*. [Tesis Doctoral] [En línea] Alicante: Biblioteca Virtual Miguel de Cervantes, 2002. <http://web.cervantesvirtual.com/FichaObra.html?Ref=10010>. [Consulta: 8 de julio de 2007].

MARTÍNEZ MÉNDEZ, F. J. y RODRÍGUEZ MUÑOZ, J. V. Síntesis y crítica de las evaluaciones de la efectividad de los motores de búsqueda en la web. *Information Re-*

search, 8(2), paper no. 148, 2003. <http://InformationR.net/ir/8-2/paper148.html>. [Consulta: 18 de noviembre de 2007].

NIELSEN/NETRATINGS. *Nielsen Online Reports Topline U.S. Data for November 2007*. [En línea] Nielsen: 2007. <http://www.nielsen-netratings.com/press.jsp>. [Consulta: 21 de octubre de 2007].

NOTESS, G. R. *Search engine statistics*. [En línea] Nottes.com: 2007 <http://www.searchengineshowdown.com/stats/>. [Consulta: 11 de enero de 2008].

OPPENHEIM, C., MORRIS, A., MCKNIGHT, C. y LOWLEY, S. The evaluation of WWW search engines. *Journal of Documentation*, 56(2), 2000, p. 190-211.

PAGE, L.; BRIN, S.; MOTWANI, R. y WINOGRAD, T. *The PageRank Citation Ranking: Bringing Order to the Web*. Manuscript in progress. [En línea] Stanford: University, 1998. <http://google.stanford.edu/~backrub/pageranksub.ps>. [Consulta: 18 de octubre de 2007].

PALFREY, D. Experiments with Search Engine Distance Measures. [En línea] *Digital History at Western wiki*. digitalhistory.uwo.ca/wiki: 2006. <http://digitalhistory.uwo.ca/wiki/index.php/Experiments_with_Search_Engine_Distance_Measures>. [Consulta: 8 de enero de 2008].

SALTON, G.; WONG, A. y YANG, C. A vector space model for automatic indexing. *Communications of the ACM*, Volume 18, Issue 11, Nov. 1975, p. 613-620.

SPERTUS, E.; SAHAMI, M. y BUYUKKOKTEN, O. Evaluating similarity measures: a large-scale study in the orkut social network *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2005)*, p. 678-684.