Ranking Global de Páginas Web Basado en Atributos de los Enlaces

Ricardo Baeza-Yates Emilio Davis

Centro de Investigación de la Web
Depto. de Ciencias de la Computación
Universidad de Chile
Blanco Encalada 2120
Santiago 6511224, Chile
E-mail: {rbaeza,edavis}@dcc.uchile.cl

Abstract

Presentamos una variante de Pagerank, el algoritmo más conocido para realizar ranking de páginas Web usando enlaces, que considera distintos atributos de cada enlace para dar distinta importancia a los mismos. Nuestros resultados muestran que la precisión de las respuestas mejora en más de un 10 %.

1. Introdución

Desde que los creadores de Google publicaran el trabajo que menciona por primera vez el algoritmo PageRank [BP98], se ha observado una revolución en el ámbito de la recuperación de información en lo que se refiere a los motores de búsqueda Web. Esta revolución posiblemente ha llevado a la mayoría de los motores a utilizar el algoritmo PageRank o una variación de éste dentro del cálculo de su *ranking*, debido a la efectividad del uso de enlaces.

Por otro lado, han aparecido otros algoritmos que si bien a primera vista no son similares a PageRank, por ejemplo HITS [Kle99], que depende de la consulta del usuario, todos pertenecen a la familia de los rankings normalizados, como se ve en [DHH+02].

Un aspecto que llama la atención, es el hecho que ningún algoritmo del tipo PageRank hace diferencia en los enlaces contenidos en una página determinada. Estos algoritmos "reparten" la misma cantidad de peso en cada enlace, sin tomar en cuenta la relevancia que pueden tener en el resultado de una búsqueda e ignorando los esfuerzos del creador de la página por destacar ciertos enlaces que considere importantes (porque llevan a recursos que él considera más relevantes que otros). Por ejemplo, en la figura 1 las páginas más importantes están en un tono más oscuro y eso muchas veces puede ser inferido a partir de los enlaces.

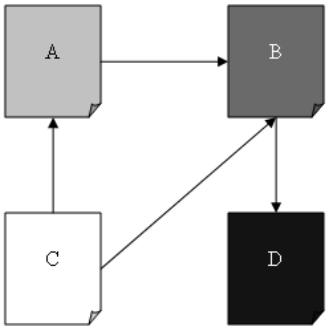

Figura 1: Ejemplo de enlaces entre páginas de distinta importancia.

linkedWebPages.ps

Tomando en cuenta lo explicado anteriormente, se puede concluir que un sistema de *ranking* que pertenezca a la familia de PageRank y que utilice la información que entreguen los generadores de contenido sobre la importancia relativa de los enlaces dentro de una página, superaría la eficacia de PageRank.

En este trabajo presentamos una variante de PageRank que da distinta importancia a los enlaces basados en tres atributos: posición relativa en la página, etiqueta (*tag*) donde se usa el enlace y largo del texto que describe al enlace (*anchor text*). Nuestros resultados muestran que todos estos atributos mejoran el algoritmo original de PageRank, en particular la combinación de ellos.

Comenzamos presentando PageRank, seguido de nuestra variante, WLRank. Luego entregamos la evaluación de la calidad de las respuestas de WLRank y las conclusiones de nuestro trabajo.

2. PageRank

La idea detrás de PageRank es que las *"buenas"* páginas Web son referenciadas mediante enlaces (*links*) por otras páginas, y esto es bastante claro. Si una página contiene información interesante para una persona es problable que en sus páginas la referencie.

El tema de las referencias o citas, no es nuevo, la gran mayoría de publicaciones científicas las utiliza y hay variados enfoques en análisis de citaciones académicas [Gar95] y [Gof71] por ejemplo, y desde un buen tiempo se ha utilizado para jerarquizar tanto las publicaciones como sus autores. Por ejemplo ISI [ISI] o CiteSeer [Cit].

Si bien, tanto páginas Web como publicaciones científicas utilizan un sistema similar para enlazarse, hay una diferencia fundamental. Para que un material académico se publique, éste debe pasar por concienzudos y escrupulosos análisis, mientras, por otro lado, las páginas Web pueden publicarse con una facilidad asombrosa, permitiendo incluso, que programas generen y publiquen grandes volúmenes de páginas con el propósito de aumentar el conteo de referencias para una página en particular. Por esta razón existe *spamming* de enlaces (por ejemplo el fenómeno llamado *Google bombing*) y por ende los algoritmos exactos de uso de enlaces no son públicos.

PageRank [BP98] en su versión simplificada se define así:

$$R(u) = c \sum_{v \epsilon B_u} \frac{R(v)}{N_v}$$

donde u es una página Web, B_u el conjunto de páginas que referencia a u, N_u el número de páginas apuntadas por u y c una constante de normalización. Nótese que el *ranking* de una página es dividido por el número de páginas que apunta para contribuir al *ranking* de éstas. La ecuación es recursiva pero convergente, es decir, si se comienza con un *ranking* arbitrario para cada página y se itera sobre todas las páginas calculando su nuevo *ranking*, al cabo de algunas iteraciones los valores se habrán estabilizado. Más aún, lo que importa es el orden y no los valores exactos, así que basta iterar hasta que no puedan producirse más cambios de orden.

Hay un pequeño problema en esta versión que se ilustra en el siguiente ejemplo: dos páginas se apuntan mutuamente y una tercera apunta a alguna de ellas, entonces durante las iteraciones esas páginas acumularán *ranking* que no distribuirán pues no apuntan a otras páginas. Lo mismo ocurre con autoreferencias o secuencias de enlaces que vuelven a su origen. Para evitar esto se introduce una fuente externa de *ranking* para cada página ($E(u)$) que resuelve el problema, así, PageRank se define como [PBMW98]:

$$R(u) = c \sum_{v \epsilon B_u} \frac{R(v)}{N_v} + E(u)$$

Si vemos PageRank como una navegación aleatoria, donde se escoge un enlace al azar para continuar o se salta a una página al azar de la Web en cualquier momento con probabilidad q, podemos usar $E(u) = q/T$ y $c = 1 - q$, donde T es el número de páginas en la colección. De esta manera la suma de valores de $R(u)$ valdrá siempre 1. Esta es la versión que usamos para explicar nuestra variante.

3. Algoritmo WLRank

WLRank (Weighted Links Rank) asigna el ranking $R(i)$ a la página i de la colección, según las siguientes ecuaciones:

$$R(i) = \frac{q}{T} + (1 - q) \sum_j \frac{W(j,i) R(j)}{\sum_k W(j,k)}$$

$$W(j,i) = L(j,i)(c + T(j,i) + AL(j,i) + RP(j,i))$$

Donde:

- $L(j,i)$ es 1 si existe un enlace desde página j a la página i ($j \in B_i$) y 0 si no existe.

- c es una constante que entrega un peso base a cada enlace.
- $T(j,i)$ es un valor que depende del *tag* en que está inserto el enlace entre la página j y la página i.
- $AL(j,i)$ es el largo del texto del enlace entre la página j y la página i ponderado por una constante d, que es la proporción entre una constante de peso base y una estimación del promedio de los largos de textos de anclaje.
- $RP(j,i)$ es la posición relativa del enlace dentro de la página ponderada por una constante b.

La figura 2 muestra un ejemplo concreto de la importancia que tendrían distintos enlaces de acuerdo a los atributos definidos.

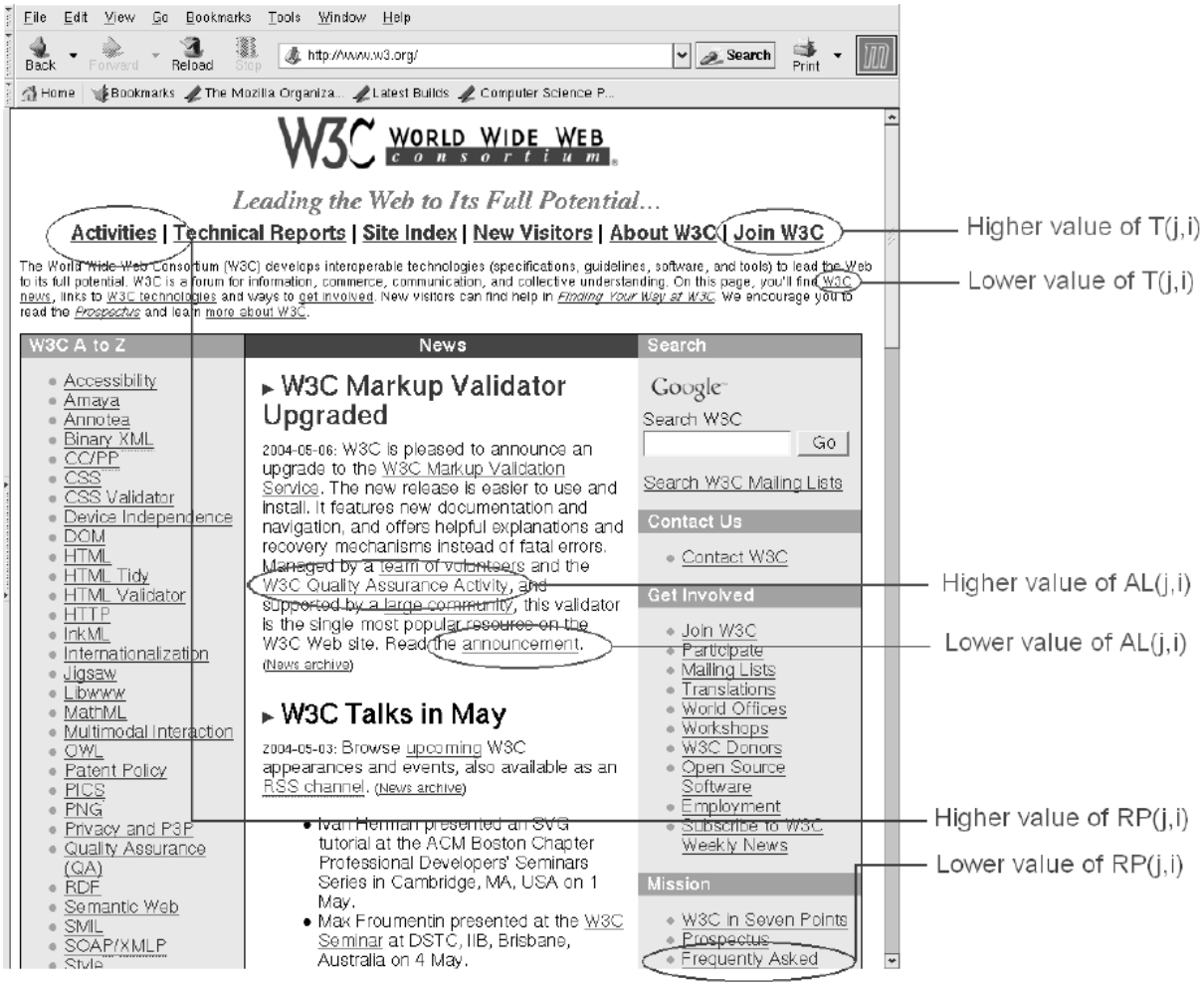

Figura 2: Ejemplo de distintos atributos de enlaces y su importancia relativa.

El *ranking* $R(i)$ corresponde a la probabilidad de un usuario de llegar a la página i, y se compone de dos términos. El primero corresponde al "salto" aleatorio desde cualquier página a ésta (distribución uniforme con probabilidad q/T) y el segundo, al peso ponderado con que aporta cada página que apunta a i.

Como se ve, si el término $W(j,i)$ fuese igual a $L(j,i)$ se obtiene PageRank. A continuación se explican las modificaciones.

El término $T(j,i)$ será una serie de constantes dependiendo del *tag* en que se encuentre el enlace, por ejemplo, un enlace que se encuentre entre $tags < h1 >$, tendrá un valor $T(j,i)$ alto, un poco más bajo si está entre $tags < h2 >$, etc. Lo mismo para otros *tags* de énfasis como $< strong >$, $< b >$, etc.

El término $AL(j,i)$ dará más valor a los enlaces en que el creador de la página haya explicado con más detalle el recurso que está siendo apuntado.

Por último, el término $RP(j,i)$ valorará más los enlaces que se ubiquen más hacia el principio de la página que hacia el final.

El algoritmo para calcular WLRank es el siguiente, basado en el de PageRank:

- Inicializar WLRank para cada página en $1/TotalPaginas$.
- Calcular la suma de los pesos de los enlaces salientes para cada documento.
- Iterar N veces (con N suficientemente grande para asegurar convergencia).
  - Para cada página i en la colección
    - Para cada página j que es apuntada por i
      - `WLRank(j) += WLRank(i) * PesoEnlace(i,j) / SumaPesosEnlacesDesde(i)`
  - Para cada página i en la colección
    - `WLRank(i) = q/TotalPaginas + (1-q)WLRank(i)`

Según [PBMW98] $N = 50$ asegura convergencia para colecciones de 300 millones de enlaces. Aunque WLRank es más complicado, en nuestros experimentos hemos necesitado valores menores de N para obtener convergencia.

4. Evaluación

En el área de recuperación de la información, existen dos parámetros importantes para evaluar la calidad de las respuestas de un motor de búsqueda.

- Recuperación (*recall*): Razón entre el número de documentos relevantes recuperados y el número total de documentos relevantes. Para poder calcular este indicador es necesario contar con una colección de documentos, un conjunto de consultas y las respuestas relevantes para cada consulta, lo que en la Web es difícil de generar.
- Precisión: razón entre el número de documentos relevantes recuperados y el número total de documentos recuperados. A diferencia del indicador anterior, para calcular éste sólo se necesita una colección de documentos y usuarios con la capacidad de realizar consultas y determinar si un resultado es relevante o no, dentro de un cierto número de respuestas.

Para probar WLRank se realizó una colecta sobre la Web chilena y usuarios de prueba que hicieron consultas al sistema y evaluaron si las respuestas que entrega son pertinentes o no, dentro de la primera página de resultados (cada página contiene diez respuestas). Cada persona evaluaba los resultados sin conocer cual caso era el que correspondía en cada conjunto de resultados. Se buscó un conjunto de usuarios de prueba de dominios de conocimiento variados de manera de simular un universo representativo de los usuarios de la Web.

La colecta sobre la Web chilena utiliza como punto de partida el conjunto de URLs que mantiene NIC Chile[1]. Nuestro sistema de búsqueda fue alimentado con este conjunto de URLs y se realizó el ciclo de recolección hasta contar con un conjunto de 460 mil páginas indexadas. Este índice se utilizó para realizar todas las pruebas.

En esta colecta se calculó WLRank usando peso unitario para el peso base ($c = 1$), el factor de posición relativa ($b = 1$), el peso de los tags <b> y <h1> (sin considerar otros tags), y el factor del largo del texto de anclaje considera largo promedio del texto de anclaje de 100 caracteres ($d = 1/100$). Además se calculó WLRank usando cada una de las modificaciones a PageRank, es decir, un WLRank agregando sólo la posición relativa, otro sólo agregando los *tags* y finalmente uno utilizando sólo el largo del texto de anclaje, usando todos ellos sobre la misma colección de páginas para que la comparación sea válida.

A cada usuario se le asignó como tarea realizar tres consultas de su elección al sistema (se supone que será un usuario experto para cada uno de esas consultas), utilizando PageRank y WLRank (en sus cuatro formas) como sistema de *ranking*. El usuario debió determinar, dentro de la primera página de resultados, cuáles respuestas fueron pertinentes a la consulta que realizó. En el cuadro 1 se observan los resultados de esta prueba utilizando PageRank y WLRank completo. En los cuadros 3 y 4 del apéndice se muestran los resultados utilizando los WLRank parciales.

Con estos datos se calculó la precisión de PageRank y WLRank (completo y parciales), los resultados se observan en la figura 3.

Para determinar si WLRank es una mejora a PageRank se pueden comparar ambos con un *ranking* "perfecto". Definimos como un *ranking* "perfecto" a un sistema de ordenamiento que logre ubicar dentro de las primeras páginas de resultados sólo respuestas relevantes. Con esto vemos, en el cuadro 2, el desempeño de PageRank y WLRank contra un *ranking* "perfecto". Como se observa, WLRank mejora en aproximadamente un 12 % a PageRank.

[1] Entidad encargada de la administración del dominio .cl, http://www.nic.cl

Consulta	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**
aspirina	×	√	×	×	√	√	×	×	√	×
educacion	√	√	×	×	×	×	×	×	×	×
bicicleta	√	√	√	×	×	√	×	×	×	√
computacion	√	√	√	√	√	√	√	×	×	√
mascota	×	×	×	×	×	√	√	√	√	√
starwars	√	×	√	√	√	√	√	√	√	√
todocl	√	×	×	×	×	√	×	√	×	×
cosmeticos	×	√	×	×	×	√	√	×	√	×
hulk	√	√	√	√	√	√	√	√	√	√
quake	√	√	×	×	×	×	×	×	×	×
mozilla	×	×	×	×	×	×	×	×	×	×
cine	×	√	×	√	×	×	×	√	√	×
“andres bello”	×	√	√	×	×	×	×	×	×	√
citibank	×	√	×	×	×	×	×	×	×	×
musica	√	√	×	√	√	√	√	√	√	√
futbol	×	√	√	√	√	√	√	√	√	√
caballo corralero	×	×	√	×	√	√	√	×	√	√
biotecnologia	√	×	×	√	√	√	√	√	√	√
cuadernos	√	√	×	×	×	×	×	√	×	√
viajes	×	×	√	√	√	√	√	√	√	×
aspirina	√	×	×	×	√	×	×	√	×	×
educacion	√	×	√	×	×	×	×	×	×	√
bicicleta	√	√	√	×	×	√	√	×	×	×
computacion	√	√	√	√	√	√	√	√	√	×
mascota	×	×	×	×	×	√	√	√	√	√
starwars	√	×	√	√	√	√	√	√	√	√
todocl	√	×	×	×	×	√	√	×	×	×
cosmeticos	×	√	√	√	×	√	√	√	×	×
hulk	√	√	√	√	√	√	√	√	√	√
quake	√	√	×	×	×	×	×	×	×	√
mozilla	×	×	×	×	×	×	×	×	×	×
cine	×	√	√	√	√	×	×	√	×	√
“andres bello”	×	√	√	√	×	×	×	×	×	×
citibank	×	√	×	×	×	×	×	×	×	×
musica	√	√	√	√	√	√	√	√	√	√
futbol	√	×	√	√	√	√	√	√	√	√
caballo corralero	×	×	×	√	√	√	√	×	√	√
biotecnologia	×	√	√	√	×	√	√	√	√	√
cuadernos	√	√	×	×	×	×	√	×	√	√
viajes	√	×	√	√	√	√	√	√	√	√

Cuadro 1: Resultados de las pruebas usando PageRank (arriba) y WLRank completo (abajo).

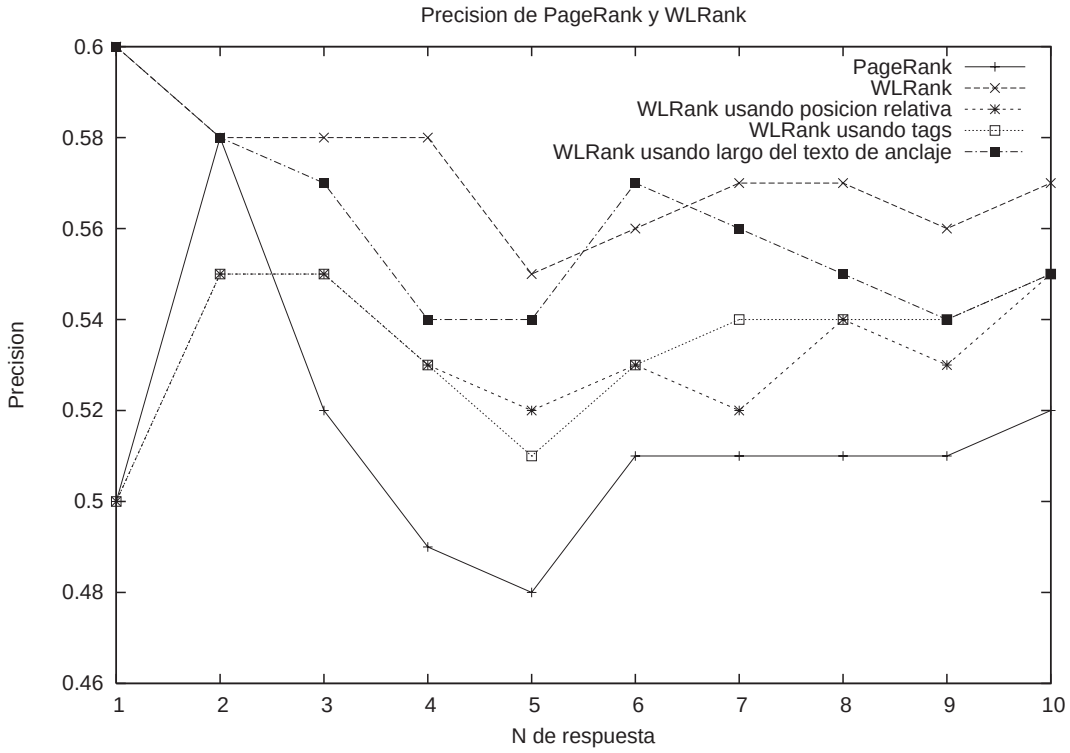

Figura 3: Precisión de PageRank y WLRank.

5. Conclusiones

Nuestros resultados muestran que el uso de enlaces de peso variable puede mejorar la precisión en más de un 10 %. Esto indica que usar estos atributos u otros similares puede mejorar el ranking. En nuestro caso la posición relativa no fue tan efectiva, lo que puede indicar que dado el diseño actual de páginas Web, un enlace puede estar lógicamente arriba pero no al comienzo del HTML asociado.

Trabajo futuro incluye una evaluación con más usuarios y un análisis detallado de los factores de importancia de cada atributo, lo que permitirá mejorar aún más WLRank. Cabe hacer notar que los valores dependen de los creadores de páginas Web y no de los que navegan por aquellas páginas. Por otro lado, estos factores deben ser confidenciales y posiblemente modificados permanentemente para evitar enlaces que intentan engañar al buscador (*spamming* de enlaces).

Respuesta	**Perfecto**	**Perfecto - Pagerank**	**Perfecto - WLRank**	**Mejora (%)**
1	1	0.5	0.4	-25
2	1	0.43	0.43	0
3	1	0.48	0.42	9
4	1	0.51	0.43	8
5	1	0.52	0.45	9
6	1	0.49	0.44	10
7	1	0.49	0.43	12
8	1	0.49	0.43	12
9	1	0.49	0.44	11
10	1	0.49	0.44	11
Error total	**0**	**4.89**	**4.29**	12

Cuadro 2: Comparación de PageRank y WLRank contra un ranking “perfecto”.

Referencias

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *7th WWW Conference*, Brisbane, Australia, April 1998.

[Cit] CiteSeer. http://citeseer.com.

[DHH+02] Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha, and Horst Simon. Pagerank, hits and a unified framework for link analysis. *LBNL Tech Report 49372*, 2001-2002.

[Gar95] Eugene Garfield. New international professional society signals the maturing of scientometrics and informetrics. *The Scientist*, 9(16), 1995.

[Gof71] William Goffman. A mathematical method for analyzing the growth of the scientific discipline. *Journal of the ACM*, 18(2):173–185, 1971.

[ISI] ISI. Citation index, http://isinet.com.

[Kle99] J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 48:604–632, 1999.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

Apéndice: Resultados de las Pruebas Parciales

Consulta	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**
aspirina	×	√	×	×	√	×	×	√	×	×
educacion	√	×	√	×	×	×	×	×	×	×
bicicleta	√	√	√	×	×	√	×	×	×	√
computacion	√	√	√	√	√	√	√	√	×	√
mascota	×	×	×	×	×	√	√	√	√	√
starwars	√	×	√	√	√	√	√	√	√	√
todocl	√	×	×	×	×	×	×	×	×	×
cosmeticos	×	√	√	√	×	√	√	√	×	√
hulk	√	√	√	√	√	√	√	√	√	√
quake	√	√	×	×	×	×	×	×	×	×
mozilla	×	×	×	×	×	×	×	×	×	×
cine	×	√	×	√	√	×	×	√	√	√
“andres bello”	×	√	√	×	×	×	×	√	×	×
citibank	×	√	×	×	×	×	×	×	×	×
musica	√	√	√	√	√	√	√	√	√	√
futbol	√	×	√	√	√	√	√	√	√	√
caballo corralero	×	×	×	√	√	√	√	×	√	√
biotecnologia	×	√	√	×	√	√	√	√	√	√
cuadernos	√	√	×	×	×	×	×	√	√	√
viajes	×	×	√	√	√	√	√	√	√	√

Cuadro 3: Resultados de las pruebas usando WLRank con posición relativa.

Consulta	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**
aspirina	×	√	×	×	√	√	×	×	√	×
educacion	√	×	√	×	×	×	×	×	×	×
bicicleta	√	√	√	×	×	×	√	×	×	×
computacion	√	√	√	√	√	√	√	×	×	√
mascota	×	×	×	×	×	√	√	√	√	√
starwars	√	×	√	√	√	√	√	√	√	√
todocl	√	×	×	×	×	√	√	×	×	×
cosmeticos	×	√	√	×	×	×	√	√	√	√
hulk	√	√	√	√	√	√	√	√	√	√
quake	√	√	×	×	×	×	×	×	×	×
mozilla	×	×	×	×	×	×	×	×	×	×
cine	×	√	√	√	×	√	×	√	√	√
“andres bello”	×	√	√	×	×	×	×	×	×	×
citibank	×	√	×	×	×	×	√	×	×	×
musica	√	√	√	√	√	√	√	√	√	√
futbol	×	√	√	√	√	√	√	√	√	√
caballo corralero	×	×	×	√	√	√	√	×	√	√
biotecnologia	√	×	×	√	√	√	√	√	√	√
cuadernos	√	√	×	×	×	×	×	√	×	√
viajes	×	×	√	√	√	√	√	√	√	√

aspirina	√	×	×	×	√	√	×	×	×	√
educacion	√	×	√	×	×	×	×	×	√	×
bicicleta	√	√	×	√	×	√	×	√	×	√
computacion	√	√	√	√	√	√	√	×	×	√
mascota	×	×	×	×	×	√	√	√	√	√
starwars	√	×	√	√	√	√	√	√	√	√
todocl	√	×	×	×	×	√	×	×	×	×
cosmeticos	√	√	√	×	×	√	√	√	×	√
hulk	√	√	√	√	√	√	√	√	√	√
quake	√	√	×	×	×	×	×	×	×	×
mozilla	×	×	×	×	×	×	×	×	×	×
cine	×	√	√	×	√	√	×	√	×	×
“andres bello”	×	√	√	×	×	×	×	×	×	×
citibank	×	√	×	√	×	×	×	×	×	×
musica	√	√	√	√	√	√	√	√	√	√
futbol	×	√	√	√	√	√	√	√	√	√
caballo corralero	×	×	×	√	√	√	√	×	√	√
biotecnologia	√	×	√	×	√	√	√	√	√	√
cuadernos	√	√	×	×	√	×	×	×	×	√
viajes	×	×	√	√	√	√	√	√	√	√

Cuadro 4: Resultados de las pruebas usando WLRank con *tags* (arriba) y con largo del texto de anclaje (abajo).