Procesamiento del lenguaje Natural, nº 41 (2008), pp. 165-172 recibido 7-05-2008; aceptado 16-06-2008

Sistema de Recomendación para la Recuperación Automática de Enlaces Web Rotos *

Recommendation System for Automatic Recovering Broken Web Links

Juan Martinez-Romo
UNED
juaner@lsi.uned.es

Lourdes Araujo
UNED
lurdes@lsi.uned.es

Resumen: Tanto en las páginas *Web* a las que accedemos cuando navegamos por *Internet*, como en las nuestras propias, a veces encontramos enlaces que han dejado de ser válidos. A menudo la búsqueda de la página que correspondía a dichos enlaces no es sencilla. En este trabajo investigamos distintas formas de recuperar automáticamente dichas páginas, de manera que le podamos ofrecer al usuario una lista de direcciones *Web* candidatas para sustituir el enlace roto. Concretamente utilizamos de forma alternativa o combinada, dependiendo de las características de la página y del enlace, el texto del ancla e información extraída de la *Web* en la que se encuentra el enlace roto. La información extraída de estas fuentes se utiliza para realizar una consulta con un motor de búsqueda usual, como *Google* o *Yahoo*. El sistema ordena posteriormente las páginas recuperadas en base a su contenido, utilizando técnicas de recuperación de información, y finalmente el resultado es presentado al usuario. Presentamos los resultados del análisis realizado sobre numerosos enlaces seleccionados aleatoriamente, los cuales nos han permitido decidir en qué condiciones es posible hacer una recomendación con un alto grado de fiabilidad.
Palabras clave: recuperación de información, *World Wide Web*, enlaces rotos

Abstract: In the *Web* pages accessed when navigating through *Internet* or even in our own *Web* pages, we sometimes find links which are not valid any more. The search of the right *Web* pages which correspond to those links is often hard. In this work we have analyzed different sources of information to automatically recover broken *Web* links so that the user can be offered a list of possible pages to substitute that link. Specifically, we have used either the anchor text or the *Web* page containing the link, or a combination of both. The information extracted is then used to perform a search with some of the usual search engines, such as *Google* or *Yahoo*. The candidate pages are then ranked applying information retrieval techniques on their content. Finally, the user is presented the pages resulting from this process. We report the analysis of a number of issues on a set of links randomly chosen, what has allowed us to decide the conditions under which the system can make a reliable recommendation.
Keywords: information retrieval, *World Wide Web*, broken links, link integrity

1. Introducción

La *Web* es un sistema altamente dinámico en el que constantemente desaparecen, se crean o se mueven las páginas de información. Esto provoca que algunos de los enlaces a los que apuntan dichas páginas se rompan un tiempo más o menos largo después de su creación. Nos encontramos esta situación frecuentemente en *Internet*. También nos obliga a revisar periódicamente nuestros sitios *Web* para comprobar que todos sus enlaces siguen siendo válidos. Encontrar la nueva ubicación de la página a la que apuntaba un enlace roto no siempre es trivial. La recuperación de enlaces en nuestras propias páginas debería ser fácil, aunque puede resultar tediosa.

Existen algunos trabajos enfocados a la re-

* Trabajo financiado por el proyecto TIN2007-67581-C02-01

ISSN: 1135-5948

cuperación de enlaces, aunque se basan en información anotada por anticipado en el enlace. El sistema *Webvise* (Grønbæk, Sloth, y Ørbæk, 1999), integrado con software de *Microsoft*, permite cierto grado de recuperación de enlaces *Web* rotos utilizando información redundante sobre los enlaces almacenada en bases de datos de servidores de *Internet*. La información se almacena al crearse o modificarse el enlace. Davis (Davis, 2000) analiza las causas del problema de los enlaces rotos y propone soluciones enfocadas a la recopilación de información sobre la estructura de la red de enlaces. Nakamizo y colaboradores (Nakamizo et al., 2005) han desarrollado un sistema de recuperación de enlaces basado en lo que denominan "enlaces con autoridad" de una página que son otras páginas que enlazan a la primera con enlaces que siempre se actualizan cuando la página *Web* se mueve. Para ello utilizan servidores de este tipo de páginas. Shimada y Futakata (Shimada y Futakata, 1998) propusieron la creación de una base de datos de enlaces, *SEDB*, en la que son posibles ciertas operaciones de reparación de los enlaces almacenados. *SEDB* maneja los documentos usando enlaces con tipos entre ellos. Sólo los enlaces se almacenan de una forma centralizada, mientras que los documentos quedan en sus localizaciones originales. Este sistema aplica una reparación automática de enlaces diseñada para preservar la topología de la red de enlaces.

También se han desarrollado trabajos que, aunque con propósitos diferentes de la recuperación de enlaces rotos, han investigado mecanismos de extracción de información a partir de los enlaces y sus contextos. Algunos de los mecanismos utilizados en estos trabajos han sido investigados en nuestro sistema de recuperación de enlaces. McBryan (McBryan, 1994) propuso el uso del texto del ancla como una ayuda para la búsqueda. En este trabajo se describe *WWWW*, una herramienta de localización de recursos. Este programa explora *Internet* localizando todo tipo de recursos *Web* con los que construye una base de datos. El motor de búsqueda de *WWWW* se ejecuta cuando un usuario accede a la página de este servicio y rellena un formulario de búsqueda. A partir de esta información, que puede ser de distintos tipos, incluyendo Urls y anclas se hacen búsquedas de patrones de cadenas. Chakrabarti y colaboradores (Chakrabarti et al., 1998) han desarrollado un algoritmo para la compilación automática de recursos *Web* con autoridad en cualquier temática suficientemente amplia. Dicho algoritmo se basa en una combinación de la información extraída de un análisis local de los textos y de los enlaces de las páginas.

Nuestro trabajo difiere de los anteriores ya que no presupone la existencia de ninguna información almacenada de antemano sobre los enlaces y es aplicable a cualquier página de *Internet*.

Cuando se trata de páginas a las que hemos llegado navegando por *Internet*, a veces podemos recuperarlas utilizando un buscador *Web* con los términos del ancla del enlace roto. Sin embargo, en muchos casos, el texto del ancla no es suficientemente informativo para recuperar la página deseada. Entonces podemos realizar consultas complementando la información del ancla con datos extraídos de otras fuentes: la página *Web* en la que se encontraba el enlace, la página almacenada por el buscador en su última indexación, la Url, etc.

En este trabajo hemos diseñado un sistema para automatizar este proceso. Nuestro sistema comprueba los enlaces de la página que se le indica. Si alguno de ellos está roto, hace una propuesta al usuario de una serie de páginas candidatas para sustituir el enlace roto. Las páginas candidatas se obtienen mediante búsquedas en *Internet* compuestas de términos extraídos de distintas fuentes. A las páginas recuperadas con la búsqueda *Web* se les aplica un proceso de ordenación que refina los resultados antes de hacer la recomendación al usuario. La figura 1 presenta un esquema del sistema propuesto.

Hemos comenzado este trabajo analizando numerosas páginas *Web* y sus enlaces para determinar qué fuentes de información y qué combinaciones de ellas son más apropiadas en cada caso. Este análisis nos ha permitido extraer criterios para determinar cuando tiene sentido hacer una recomendación al usuario, y cuando la información disponible es insuficiente para llevar a cabo la recuperación. En este caso, se informa al usuario de la situación. Si la información es suficiente se hace una recomendación de páginas candidatas ordenadas por relevancia.

El resto del artículo se organiza de la siguiente forma: en la sección 2 se describe la metodología seguida para estudiar la utili-

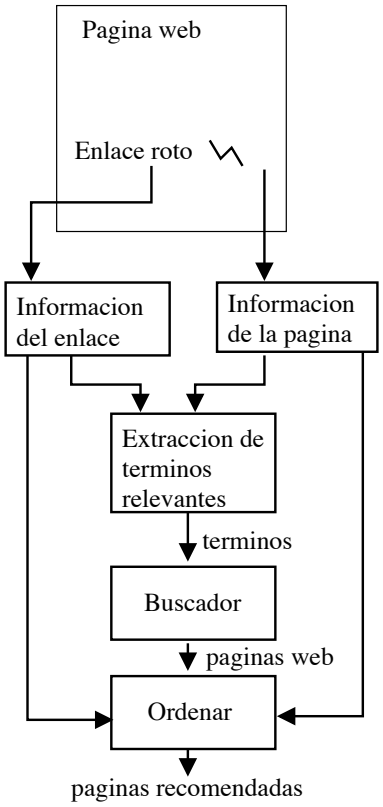

Figura 1: Esquema del funcionamiento del sistema de recomendación para la recuperación de enlaces rotos.

dad de la información de las distintas fuentes de información consideradas, mientras que la sección 3 presenta los resultados de dichos estudios y analiza su utilidad para nuestro problema. La sección 4 describe el proceso de ordenación de los documentos obtenidos. La sección 5 presenta el esquema resultante de los análisis anteriores y su evaluación sobre un conjunto de enlaces rotos, y finalmente, la sección 6 resume las conclusiones del trabajo.

2. Metodología

Si analizamos la utilidad de las distintas fuentes de información utilizadas directamente sobre enlaces rotos, es muy difícil evaluar la calidad de las páginas candidatas a sustituir el enlace. Por ello, en esta fase de análisis trabajamos con enlaces *Web* tomados de forma aleatoria, que no están realmente rotos, y que denominamos *supuestamente rotos*. De esta forma disponemos de la página a la que apuntan y podemos evaluar la recomendación que hacemos utilizando cada fuente de información.

2.1. Selección de los enlaces a recuperar

Para realizar el análisis, tomamos los enlaces de páginas seleccionadas aleatoriamente mediante peticiones sucesivas a *www.randomwebsite.com*, un sitio que proporciona páginas *Web* aleatorias.

Hemos impuesto ciertos requisitos a nuestras páginas de prueba:

- Intentamos restringir el idioma al inglés, considerando los siguientes dominios: “.com”, “.org”, “.net”, “.gov” y ”.edu”.
- Buscamos páginas con al menos 250 palabras, con el objetivo de utilizar este texto para caracterizar la página. Además, el texto deberá contener al menos diez términos que no sean palabras vacías, es decir, palabras comunes como artículos, pronombres, etc., que no sirven para discriminar.

También exigimos que la página contenga al menos cinco enlaces potencialmente analizables, lo que significa:

- Estar en línea si buscamos enlaces activos o roto si estamos buscando enlaces rotos. Este parámetro cambia según el tipo de estudio.
- El sistema analiza enlaces externos, por lo tanto los enlaces que apunten al mismo sitio son descartados.
- El texto del ancla no debe estar vacío o ser un número o una Url.
- En el caso de que el texto del ancla sea un solo carácter y además coincida con un signo de puntuación, este enlace será descartado.

Las páginas que no poseen estas características son descartadas, y el proceso de selección no finaliza hasta que se han reunido un total de cien páginas, lo que supone al menos 500 enlaces a estudiar. Algunos experimentos preliminares nos indicaron que es frecuente encontrar páginas en las que la mayoría de los enlaces son correctos y otras en las que la mayoría de los enlaces son incorrectos. Cuando estas páginas tienen muchos enlaces, sesgan los resultados en uno u otro sentido. Por ello decidimos limitar el número de enlaces tomado de cada página a diez. La elección de este subconjunto de enlaces se realiza por

cada página de prueba y siguiendo una distribución aleatoria y uniforme sobre el conjunto total de sus "enlaces analizables".

3. Fuentes de Información

En esta sección analizamos cada una de las fuentes de información consideradas, extrayendo estadísticas de su utilidad para la recuperación de enlaces cuando se aplican por separado o combinadas.

3.1. Texto del ancla de los enlaces

En muchos casos las palabras que componen el texto del ancla de un enlace son la principal fuente de información para identificar la página apuntada. Para verificar esta teoría, hemos realizado un estudio, que se muestra en el cuadro 1, del número de casos en los que los enlaces rotos se han recuperado buscando en *Google* el texto del ancla entrecomillado.

Para considerar que un enlace se ha recuperado, se ha utilizado una combinación de distintos mecanismos. En primer lugar se comprueba si la Url de la página candidata a sustituir el enlace coincide con la del enlace analizado (que recordemos, en esta fase de análisis no está roto en realidad). Sin embargo, hemos encontrado casos en los que la página que se recupera tiene el mismo contenido que la del enlace supuestamente roto, pero distinta Url. Por ello si las Urls no coinciden, comprobamos si el contenido de las páginas es el mismo. También hemos encontrado varios casos en los que el contenido de las páginas no es exactamente el mismo, pero es muy similar: cambia algún anuncio, la fecha, etc. Por ello, si el contenido no coincide, aplicamos el modelo de espacio vectorial (Manning, Raghavan, y Schütze, 2008), representando cada una de las páginas a comparar por un vector de términos, y hayamos la distancia dada por el coseno entre ellos. Si este valor es mayor de 0.9, consideramos la página recuperada. Para valores menores que este umbral, como un 0.8, aunque en la mayoría de los casos se trata de la misma página con pequeños cambios como los mencionados, hemos encontrado algún caso en que se trataba de páginas distintas, aunque del mismo sitio *Web*.

El cuadro 1 muestra el número de enlaces supuestamente rotos que se ha conseguido recuperar entre las diez primeras posiciones de los documentos devueltos por el buscador. Podemos observar que utilizando un umbral de similitud de 0.9 se ha conseguido recuperar un 41 % de los enlaces entre las diez primeras posiciones (*Google*). Además un 66 % de los enlaces recuperados han logrado encontrarse en la primera posición. Estos datos demuestran que el texto del ancla de un enlace es una gran fuente de información de cara a recuperar un enlace roto. Los resultados para las filas correspondientes a grados de similitud menores de 0.9 muestran que el número de enlaces adicionales que se conseguiría bajando el umbral, es muy pequeño. Por ello, y dados los casos erróneos que se podrían incluir con otros umbrales, hemos utilizado un umbral de 0.9.

Grado Sim.	1 pos.	1-10 pos.	E.N.R.
0.9	253	380	536
0.8	3	4	529
0.7	2	6	521
0.6	4	13	504
0.5	4	22	478

Cuadro 1: Valores agregados (descontando los del nivel anterior) de la búsqueda del texto del ancla en *Google* según el grado de similitud utilizado. La primera columna indica el grado de similitud requerido para calcular estos valores, entendidos como el incremento que se conseguiría al pasar de 0.9 a 0.8, etc. 1 pos. representa el número de enlaces "supuestamente rotos" que se han recuperado en primera posición entre los resultados del buscador, y 1-10 pos. el número de los recuperados entre las 10 primeras posiciones. E.N.R. representa los enlaces que no se han conseguido recuperar.

Sin embargo, hay ocasiones en las que los términos del ancla pueden ser poco o nada descriptivos. Imaginemos un enlace cuyo texto de anclaje es "pincha aqui". En este caso, el encontrar el enlace roto podría calificarse como imposible. Por este motivo también es muy importante analizar estos términos para poder decidir qué tareas realizar dependiendo de su número y calidad.

En este trabajo hemos optado por realizar un reconocimiento de entidades nombradas (nombres de personas, organizaciones o lugares) sobre el texto del ancla, para poder extraer determinados términos cuya importancia sea mayor que la del resto. Para tal fin, existen varias soluciones software como *LingPipe*, *Gate*, *FreeLing*, etc. También exis-

ten múltiples recursos en forma de *gazetteers*, pero el amplio dominio sobre el que trabajamos ha impedido conseguir resultados precisos. Estamos en un entorno en el que analizamos páginas aleatorias cuyo único factor común es el idioma (inglés). Además, el hecho de que el texto de las anclas sean conjuntos muy reducidos de palabras y/o números, hace que los sistemas usuales de reconocimiento de entidades proporcionen resultados muy pobres.

Por estos motivos, hemos decidido emplear la estrategia opuesta. En lugar de encontrar entidades nombradas, hemos optado por recopilar un conjunto de diccionarios y descartar las palabras comunes y números, suponiendo que el resto de palabras son entidades nombradas, además trabajamos con unitérminos. Aunque hemos encontrado algunos *falsos negativos*, como por ejemplo la compañía *"Apple"*, en el caso de las anclas hemos obtenido mejores resultados con esta técnica.

El cuadro 2 muestra los resultados de recuperación de los enlaces "supuestamente rotos" en función del contenido de entidades nombradas de las anclas. Podemos ver que cuando el ancla no contiene ninguna entidad nombrada, el número de enlaces para los que no se consigue recuperar la página es mucho mayor que el número de los que se recupera, mientras que cuando hay entidades nombradas ambas cantidades son similares. Esto demuestra que la presencia de entidades nombradas en el ancla favorece la recuperación del enlace.

Tipo de ancla	E. N. R.	E. R.
Ent. Nomb.	240	232
No Ent.	296	148

Cuadro 2: Análisis del tipo de ancla de los enlaces no recuperados (E.N.R.) y recuperados (E.R.). Ent. Nomb. representa a las anclas con una o más entidades nombradas, y No Ent, a las que no contienen ninguna entidad nombrada.

También hemos analizado los resultados de recuperación en función del número de términos del ancla. El cuadro 3 muestra este estudio. El resultado más claro es que cuando el ancla consta de un sólo término y este no es una entidad nombrada, el número de casos en los que se consigue recuperar el documento correcto es realmente muy pequeño. Cuando hay entidades nombradas, aunque haya un solo término, el número de casos recuperados es importante. Otro dato que podemos observar es que a partir de dos términos, el número de términos del ancla no representa una gran variación en los resultados.

Tipo de ancla	Términos	E. N. R.	E. R.
Ent.Nomb.	1 term.	102	67
	2 term.	52	75
	3 term.	32	29
	4+ term.	57	61
No Ent.	1 term.	145	7
	2 term.	91	49
	3 term.	27	45
	4+ term.	33	47

Cuadro 3: Análisis de los enlaces no recuperados (E.N.R.) y recuperados (E.R.) en función del tipo de ancla, con (Ent. Nomb.) y sin (No Ent.) entidades nombradas, y del número de términos del ancla. *4+ term.* se refiere a anclas con cuatro o más términos.

3.2. El texto de la página

Los términos más frecuentes encontrados en una página *Web* son una forma de caracterizar el tema principal de dicha página. Esta técnica requiere que el contenido de la página sea suficientemente grande. Un ejemplo claro de utilidad de esta información son los enlaces a páginas personales. Es muy frecuente que el ancla de un enlace a una página personal esté formada por el nombre de la persona a la que corresponde la página. Sin embargo, en muchos casos los nombres, incluido el apellido, no identifican a una persona de forma unívoca. Por ejemplo, si buscamos en *Google* por el nombre *"Juan Martínez"*, el nombre de uno de los autores de este trabajo, nos aparecen numerosas entradas (99.900 aprox. en el momento de escribir este artículo). La primera respuesta del buscador que corresponde a *Juan Martínez Romo* ocupa la décima posición. Sin embargo, si añadimos algún término de los que aparecen en su página *Web*, como "Web search", entonces la entrada a su página pasa a ser la primera. Este ejemplo nos muestra la utilidad del uso de una selección adecuada de términos de la página que contiene el enlace.

Hemos aplicado técnicas clásicas de recuperación de información para extraer los

términos más representativos de la página. Una vez eliminadas las palabras vacías, generamos un índice de términos ordenado por frecuencias. Los diez primeros términos de este índice se utilizan, uno a uno, para expandir la consulta formada por el texto del ancla. Es decir, se expande con cada uno de ellos y se toman los diez primeros documentos recuperados en cada caso.

En el cuadro 4 se puede observar como la expansión mejora globalmente los resultados aumentando el número de enlaces recuperados en las diez primeras posiciones y por tanto reduciendo los enlaces no recuperados. A pesar de esto, el número de enlaces recuperados en primera posición se ve reducido. El

Análisis.	1 pos.	1-10 pos.	E.N.R.
No EXP	253	380	536
EXP	213	418	498

Cuadro 4: Análisis del número de documentos recuperados en primera posición (1 pos.), entre las diez primeras posiciones (1-10 pos.) o no recuperados (E.N.R.) en función de utilizar (EXP) o no (No EXP), el método de expansión de la consulta.

cuadro 5 muestra el número de casos en los que la expansión ha mejorado los resultados, y en los que los ha empeorado. Podemos ver que aunque el número de casos en que mejora es bastante mayor, casi el doble (90 frente a 52), el número de casos en los que empeora no es despreciable. Por ello consideramos que lo más adecuado es aplicar ambas formas de recuperación, y ordenar después los resultados para presentar al usuario los más relevantes en primer lugar.

Resultado expansión	Num. Casos
Mejora	90
Empeora	52

Cuadro 5: Número de casos en los que la expansión mejora y empeora los resultados.

Analizando los casos en los que se consigue recuperar la página correcta con y sin entidades nombradas (cuadro 6) y en función del número de términos del ancla (cuadro 7) vemos que las proporciones obtenidas recuperando sin expandir la consulta se mantienen. Es decir, los mejores resultados se obtienen cuando hay entidades nombradas y cuando hay dos o más términos. Sin embargo, en este caso, es decir con expansión, el número de enlaces recuperados cuando el ancla consta de un único término y no es una entidad nombrada es 25, que ya puede ser una cantidad significativa. Esto sugiere intentar recuperar con expansión también en este caso, siempre que sea posible comprobar la validez de los resultados, como se explica después en la sección 5.

Tipo de ancla	E. N. R.	E. R.
Ent. Nomb.	248	224
No Ent.	250	194

Cuadro 6: Análisis, cuando se aplica el método de expansión de la consulta, de los enlaces no recuperados (E.N.R.) y recuperados (E.R.) en función del tipo de ancla, con (Ent. Nomb.) y sin (No Ent.) entidades nombradas.

Tipo de ancla	Términos	E. N. R.	E. R.
Ent.Nomb.	1 term.	104	65
	2 term.	55	72
	3 term.	30	28
	4+ term.	59	59
No Ent.	1 term.	127	25
	2 term.	70	70
	3 term.	22	50
	4+ term.	31	49

Cuadro 7: Análisis, cuando se aplica el método de expansión de la consulta, de los enlaces no recuperados (E.N.R.) y recuperados (E.R.) en función del tipo de ancla, con (Ent. Nomb.) y sin (No Ent.) entidades nombradas, y del número de términos del ancla. *4+ term.* se refiere a anclas con 4 o más términos.

4. Ordenación de los enlaces a recomendar

En este momento hemos recuperado un conjunto de enlaces candidatos a sustituir al enlace roto, procedentes de la búsqueda con el ancla y con el ancla expandida con cada uno de los diez primeros términos que representan a la página padre. Ahora queremos ordenarlos por relevancia para presentarlos al usuario. Para calcular esta relevancia hemos considerado dos fuentes de información. En primer lugar, si existe, la página a la que apuntaba el enlace roto almacenada en la caché del buscador, en nuestro

caso de *Google*. Si esta información no existe, entonces utilizamos la página padre que contiene el enlace roto. La idea es que la página enlazada tratará en general sobre una temática relacionada con la página en la que se encuentra el enlace.

De nuevo hemos aplicado el modelo de espacio vectorial (Manning, Raghavan, y Schütze, 2008) para estudiar la similitud entre la página que contenía el enlace roto y las páginas recuperadas. Con esta técnica calculamos la similitud o bien con la caché o bien con la página padre. El cuadro 8 muestra los resultados obtenidos ordenando por similitud con la caché, mientras que el cuadro 9 muestra los resultados ordenando por similitud con la página padre. En el primer caso, la mayoría de los documentos correctos recuperados se presentan entre los diez primeros documentos, con lo que si se dispone de la caché, podremos hacer recomendaciones muy fiables. En el caso de la similitud con la página padre, el orden de los resultados es peor. Por lo que sólo recurriremos a esta información si no se dispone de la caché.

N primeros docs. seleccionados	Apariciones mejor candidato
10	301
20	305
30	306
50	307
80	310
100	312
110	313

Cuadro 8: Número de apariciones de páginas correctas en el ranking elaborado, seleccionando los N mejores candidatos según la similitud con la caché.

N primeros docs. seleccionados	Apariciones mejor candidato
10	47
20	105
30	132
50	191
80	263
100	305
110	313

Cuadro 9: Número de apariciones de páginas correctas en el ranking elaborado, seleccionando los N mejores candidatos según la similitud con la página padre.

5. *Algoritmo de Recuperación Automática de enlaces*

```
si long(ancla) = 1 y NoEN(ancla) ent
  si EnCache(pagina) ent
    docs = busqueda_Web(ancla + info_pagina)
    ordenar(docs,cache)
    si similitud(docs, cache(pagina) > 0.9) ent
      propuesta_usuario(docs)
    sino
      No_se_recupera
  sino
    No_se_recupera
sino
  docs = busqueda_Web(ancla)
  docs = docs +
       busqueda_Web(ancla + info_pagina)
  si EnCache(pagina) ent
    ordenar(docs,cache)
  sino
    ordenar(docs,pagina_padre)
  propuesta_usuario(docs)
```

Figura 2: Algoritmo de recomendación de enlaces sustitutos de uno roto.

Los resultados del análisis descrito en las secciones anteriores sugieren criterios para decidir en qué casos hay información suficiente para intentar la recuperación del enlace y qué fuentes de información utilizar. De acuerdo con ellos proponemos el procedimiento de recuperación que aparece en la figura 2. En primer lugar se comprueba si el número de términos del ancla es sólo uno (long(ancla) = 1) y si no contiene entidades nombradas (NoEN(ancla)). En este caso sólo se intenta recuperar si la página desaparecida está en la cache y por tanto tenemos información que nos permita comprobar que la propuesta que hagamos al usuario sea relevante. Si no es así, se informa al usuario de la imposibilidad de hacer la recomendación. Si la página está en la cache, entonces se recupera, expandiendo la consulta de los términos del ancla con los extraídos de la pagina padre, se ordenan los resultados y sólo si hay alguno suficientemente próximo al contenido de la cache se hace la recomendación al usuario. En los casos restantes, es decir anclas con más de un término o que contienen alguna entidad nombrada, se recupera con los términos del ancla, también expandiendo con términos de la página padre y se juntan y ordenan todos los documentos. Si la cache de la página desaparecida está disponible se utiliza para la

ordenación, y si no se utiliza la página padre.

Hemos aplicado este algoritmo a enlaces que están realmente rotos, pero sólamente de los que se dispone de caché, para poder evaluar los resultados. El cuadro 10 muestra los resultados de la posición de los documentos más relevantes en una ordenación por similitud con la página padre. La relevancia se mide por similitud con la caché. Hemos comprobado que en unos casos se trata de la página original, que ha cambiado de Url, y en otros casos de páginas con contenido muy relacionado en una localización diferente. Podemos observar, que aún si no contamos con la caché y ordenamos por similitud con la página padre, el sistema es capaz de presentar documentos sustitutos relevantes entre las 10 primeras posiciones en un 48 % de los casos y entre las 20 primeras en un 76 %.

N primeros	E.R
1-10	12
10-20	7
20-50	6

Cuadro 10: Número de apariciones de páginas sustitutas (de acuerdo con su similitud con el contenido de la caché) entre los N primeros documentos ordenados por similitud con la página padre.

6. Conclusiones y Futuros trabajos

En este trabajo hemos analizado distintas fuentes de información que podemos utilizar para hacer una recuperación automática de enlaces *Web* que han dejado de ser válidos. Los resultados indican que los términos del ancla pueden ser muy útiles, especialmente si hay más de uno y si contienen alguna entidad nombrada. Hemos estudiado también el efecto de añadir términos procedentes de la página que contiene el enlace, con el fin de reducir la ambigüedad que puede conllevar la cantidad limitada de términos del ancla. Este estudio ha mostrado que los resultados mejoran a los obtenidos utilizando sólo los términos del ancla. Sin embargo, como hay casos en los que la expansión empeora el resultado de la recuperación, hemos decidido combinar ambos métodos, ordenando después los documentos obtenidos por relevancia, para presentar al usuario las mejores páginas candidatas en primer lugar. El resultado de este análisis ha sido un algoritmo que ha conseguido recuperar una página muy cercana a la desaparecida entre las diez primeras posiciones de los documentos candidatos en un 48 % de los casos, y entre las 20 primeras en un 76 %.

En este momento trabajamos en analizar otras fuentes de información que pueden ser útiles para la recuperación, como las Urls o las páginas apuntadas por otros enlaces de la página que contiene el enlace roto.

Bibliografía

Chakrabarti, S., B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, y S. Rajagopalan. 1998. Automatic resource list compilation by analyzing hyperlink structure and associated text. En *Proceedings of the 7th International World Wide Web Conference*.

Davis, H.C. 2000. Hypertext link integrity. *ACM Computing Surveys Electronic Symposium on Hypertext and Hypermedia*, 31(4).

Grønbæk, Kaj, Lennert Sloth, y Peter Ørbæk. 1999. Webvise: Browser and proxy support for open hypermedia structuring mechanisms on the world wide web. *Computer Networks*, 31(11-16):1331–1345.

Manning, Christopher D., Prabhakar Raghavan, y Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

McBryan, Oliver A. 1994. GENVL and WWWW: Tools for Taming the Web. En O.Ñierstarsz, editor, *Proceedings of the first International World Wide Web Conference*, página 15, CERN, Geneva.

Nakamizo, A., T. Iida, A. Morishima, S. Sugimoto, , y H. Kitagawa. 2005. A tool to compute reliable web links and its applications. En *SWOD '05: Proc. International Special Workshop on Databases for Next Generation Researchers*, páginas 146–149. IEEE Computer Society.

Shimada, Takehiro y Atsushi Futakata. 1998. Automatic link generation and repair mechanism for document management. En *HICSS '98: Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences-Volume 2*, página 226, Washington, DC, USA. IEEE Computer Society.