UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

TÉCNICAS DE RECUPERACIÓN DE INFORMACIÓN PARA LA RESOLUCIÓN DE PROBLEMAS EN LA WEB

TESIS DOCTORAL

Juan Martínez Romo
Ingeniero Informático
2010

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

TÉCNICAS DE RECUPERACIÓN DE INFORMACIÓN PARA LA RESOLUCIÓN DE PROBLEMAS EN LA WEB

Juan Martínez Romo
Ingeniero en Informática por la Universidad Rey Juan Carlos

Director

Lourdes Araujo Serna
Profesora Titular de Universidad del Departamento de Lenguajes y Sistemas Informáticos de la Universidad Nacional de Educación a Distancia

A mi abuela María;
a mis padres, Juan y Pepi;
a mi hermana Carolina;
a Cristina.

Resumen

En esta tesis, se abordan dos de los problemas más importantes que afectan a la Web en la actualidad. El crecimiento vertiginoso de esta red mundial, ha propiciado la conexión en esta tesis de uno de sus principales problemas desde el origen en 1989, los enlaces rotos, con una reciente preocupación de los motores de búsqueda, el web spam. El vínculo entre el problema de los enlaces rotos en las páginas web y el spam de buscadores, se ha establecido mediante el uso común de un conjunto de técnicas de recuperación de información, en forma de sistema de recuperación de información web.

El inconveniente que genera la desaparición de una página web, ha sido afrontado mediante el diseño de un Sistema de Recuperación de Enlaces Rotos (SRER). Este sistema analiza la información disponible acerca de una página desaparecida, y recomienda al usuario un conjunto de documentos candidatos para reemplazar el enlace obsoleto. El SRER propuesto en esta tesis, a diferencia del resto de sistemas con objetivos similares, no necesita del almacenamiento previo de ningún tipo de información acerca de la página desaparecida, para poder realizar una recomendación. El diseño de este sistema se compone de cuatro etapas, en las que se aplican diferentes técnicas de recuperación de información y procesamiento del lenguaje natural, para obtener el mejor rendimiento.

La primera etapa consiste en un proceso de selección de información, en el cual se analiza en primer lugar, el texto del ancla del hiperenlace que ha dejado de funcionar. Los términos que componen el ancla son una pieza fundamental en el buen funcionamiento del sistema, y de esta forma se realiza un reconocimiento de entidades nombradas, con el objetivo de determinar aquellos términos con un valor descriptivo superior. En segundo lugar, se extrae información del contexto del hiperenlace para conseguir un mayor grado de precisión. Cuando una página web desaparece, durante un periodo de tiempo variable, es posible encontrar datos acerca de dicha página en la infraestructura web. Teniendo en cuenta la presencia de esta información, en tercer lugar se propone el uso de varios recursos disponibles

en la Web, con el objetivo de seguir el rastro que ha dejado la página desaparecida. Entre estos recursos se encuentran aplicaciones proporcionadas por los principales motores de búsqueda, librerías digitales, servicios web y redes sociales.

La segunda etapa se centra en las fuentes de información obtenidas a partir del contexto del enlace y de los recursos online disponibles. En algunos casos, el tamaño de dichas fuentes es demasiado grande como para discriminar la información relevante de la que no lo es. Por este motivo se lleva a cabo un proceso de extracción de terminología a fin de sintetizar la información. Con el objetivo de optimizar la extracción de los términos más relevantes en cada caso, se han analizado diferentes técnicas de recuperación de información.

En la tercera etapa, el SRER analiza la información obtenida y establece un conjunto de consultas, que posteriormente serán ejecutadas en un motor de búsqueda. En esta fase se parte de los datos obtenidos del texto del ancla y a continuación se realiza un proceso de expansión de consultas. Por cada una de las consultas, el sistema recupera los primeros resultados devueltos por el buscador.

Una vez finalizada la etapa de expansión de consultas y recuperados las páginas candidatas a reemplazar al enlace roto, se lleva a cabo una ordenación por relevancia, para mostrar al usuario un conjunto de resultados en orden decreciente. Para establecer el orden de aparición, se han analizado algunas funciones de ranking. Estas funciones utilizan la información disponible en la primera etapa para otorgar un valor de relevancia a cada documento. Finalmente, el sistema presenta al usuario una lista de resultados ordenados según su relevancia.

Las cuatro etapas en las que se divide el SRER, se encuentran dirigidas por un algoritmo que analiza la información disponible en cada caso, y toma una decisión, con el objetivo de optimizar por un lado los resultados mostrados al usuario y por otro lado el tiempo de respuesta del sistema.

Entre las aportaciones de esta tesis, también se encuentra el desarrollo de una metodología de evaluación, que evita el juicio de humanos a fin de ofrecer unos resultados más objetivos.

Por último, el SRER, representado a su vez por el algoritmo de recuperación de enlaces rotos, ha sido integrado en una aplicación web denominada Detective Brooklynk.

La recuperación de un enlace, es decir, encontrar una página en Internet en función de la información relativa a ella disponible en la página que la apunta, está basada en la hipótesis de que dicha información es coherente. Existen casos es los que los autores de páginas web manipulan la información relativa a una determinada página, con el objetivo de obtener algún beneficio. En esta tesis, analizamos los casos en los que una página web inserta información incoherente acerca de una segunda página apuntada, con el objetivo de promocionarla en un buscador.

En la segunda parte de esta tesis, enmarcada dentro del área de la detección de web spam, se parte del concepto de recuperación de enlaces para detectar aquellos de naturaleza fraudulenta. En esta ocasión, el motor del sistema de recuperación de enlaces rotos es modificado para la recuperación de enlaces activos. El objetivo de

dicha adaptación es localizar los enlaces cuya información acerca del recurso apuntado es voluntariamente incoherente y por tanto resulta imposible su recuperación. El sistema resultante es capaz de proporcionar un conjunto de indicadores por cada página analizada, empleados para una etapa posterior de clasificación automática.

El web spam se divide principalmente en dos grupos de técnicas: aquellas que inciden sobre los enlaces de las páginas web, y las que emplean el contenido para promocionarlas. De esta forma, si mediante el sistema de recuperación de enlaces se consiguen detectar los enlaces fraudulentos, en esta tesis se ha decidido completar la detección de spam de contenido. Para ello, se ha llevado a cabo un análisis de la divergencia entre el contenido de dos páginas enlazadas.

El resultado de esta segunda parte de la tesis dedicada a la detección de web spam, es la propuesta de utilización de dos nuevos conjuntos de indicadores. Además, la combinación de ambas características da lugar a un sistema ortogonal que mejora los resultados de detección de ambos conjuntos por separado.

Juan Martínez Romo
juaner@lsi.uned.es
NLP & IR Group, http://nlp.uned.es
DEPARTAMENTO DE LENGUAJES Y SISTEMAS INFORMÁTICOS
UNED, MADRID, 7 DE MAYO DE 2010.

Abstract

This thesis addresses two of the most important problems affecting the Web today. The rapid growth of this global network has facilitated the connection in this thesis of one of its main problems from the beginning in 1989, broken links, with a recent worry of the search engines, web spam. The problem of broken links on web pages and search engine spam have been linked by means of the use of a common web information retrieval system.

The problem which generates a missing web page has been addressed through the design of a Recovery System of Broken Links (SRER). This system analyses the available information about a missing page and proposes to the user a set of documents to replace the outdated link. SRER, unlike other systems with similar objectives, does not need any previous stored data about the missing page in order to propose a recommendation. The design of this system consists of four stages where various information retrieval and natural language processing techniques are applied in order to obtain the best performance.

First, a process of selection of information is carried out, in which the anchor text is analysed. Anchor terms are a key element in the performance of the system, so a named entity recognition is executed in order to identify those terms with a better descriptive value. Second, information in the context of a hyperlink is extracted in order to obtain a higher accuracy. When a web site disappears, for a variable slot of time, it is possible to find information about that page in the web infrastructure. Taking into account the presence of this information, it is proposed the use of various resources available on the Web with the aim of following the trail left by the missing page. These resources include applications provided by major search engines, digital libraries, web services and social networks.

Second phase focuses on the sources of information obtained from the context of a link and the available online resources. In some cases, the size of these sources is too large to discriminate relevant information. For this reason, a term extraction process is carried out to synthesise that information. In order to optimise the extrac-

tion of the most relevant terms in each case, various information retrieval techniques have been analysed.

In the third stage, SRER analyses the recovered information and establishes a set of queries which will be executed in a search engine. In this phase, the anchor text is the base of the query and after that, a query expansion process is performed. For each query, the system retrieves the top results returned by the search engine.

When the query expansion stage has finished and the top results have been recovered, the system performs a ranking with these candidate pages. Some ranking functions have been analysed with the aim of set the position of every page. These functions use the information available in the first stage to establish a relevance score to each document. Finally, the system presents the user with a list of ranked results.

Four stages in which SRER is divided are managed by an algorithm. This algorithm analyses the information available in each case, and makes a decision with the aim of optimising, in one hand the results displayed to the user, and on the other hand the response time.

Among the contributions of this thesis, a new evaluation methodology has been developed which avoids human judgements, providing a set of results more objective. Finally, the recovery broken links algorithm, has been integrated into a web application called "Detective Brooklynk".

The recovery of a link, that is, to find a web page in Internet just using the information about it, available in the pointed page, is based on the assumption that such information is consistent. There are cases where web pages authors manipulate that information about a particular site with the aim of obtaining some benefit. In this thesis, the cases, in which a web page contains inconsistent information about a second pointed page with the aim of promoting it in a search engine, have been analysed.

In the second part of this thesis, framed in the area of web spam detection, it is used the concept of recovery of links to detect the fraudulent ones. The recovery system of broken links is modified for the recovery of active links. The purpose of this adaptation is to find links which contain inconsistent information about the pointed resource, and therefore are impossible to recover. Through this technique, the system provides a set of features for each page. Those features will be used in a later automatic classification.

Web spam focuses basically on two types of techniques, those that affect the web page links, and those that use content to promote them. Thus, this thesis has completed the previous study by means of the detection of spam content. For that, it was carried out an analysis of the divergence among different parts of two linked pages.

The main result of this second part of the thesis, devoted to the detection of web spam, is the proposal of two new sets of features. Furthermore, the combination of both set of features gives rise to an orthogonal system that improves detection results of both sets in a separate way.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Parte I

Preámbulo

CAPÍTULO 1

Introducción

1.1. Carencias en la Integridad Referencial de las Páginas Web

Cuando en 1989, Tim Berners-Lee creó la base de la tecnología necesaria para el funcionamiento de la Web, como el lenguaje HTML, el protocolo HTTP y el sistema de localización de objetos URL, probablemente no imaginaba de que una de las principales ventajas respecto a los anteriores sistemas gestores de información iba a convertirse en un problema al que no se le podría dar solución incluso 20 años después. La Web no necesitaba enlaces bidireccionales, y de esta forma se creó un sistema con enlaces unidireccionales, lo que simplificaba la implementación de servidores y navegadores. Pero con el paso del tiempo, muchos recursos web enlazados por hiperenlaces desaparecen, cambian su localización, o simplemente son reemplazados con distinto contenido. En estos casos y debido a que las páginas que contienen hiperenlaces a estos recursos no conocen dichos eventos, se produce un fenómeno que se denomina "enlaces rotos".

Desde 1989, ha habido momentos en los que el problema de los "enlaces rotos" era considerado el segundo problema más serio de la Web[MB02], y es que es muy frecuente encontrar algún enlace que no funciona correctamente cuando navegamos por Internet. Durante 10 años el grupo de investigación "GVU's WWW Survey Team" del instituto de tecnología de Georgia[1] dirigió una serie de estudios[2] preguntando a usuarios de la Web, en los que analizaban entre otros asuntos los principales problemas que encontraban en la Web. En la Figura 1.1, extraída de uno de estos estudios, se puede apreciar el porcentaje de usuarios que clasifican cada uno de los problemas propuestos como el más importante en la Web. Un dato muy significativo es el hecho de que los enlaces rotos se sitúan como el segundo ma-

[1] http://gvu.cc.gatech.edu/

[2] http://www.cc.gatech.edu/gvu/user_surveys/

yor problema, tan solo superado por la velocidad, habiendo aumentado esta última considerablemente desde la realización de este estudio.

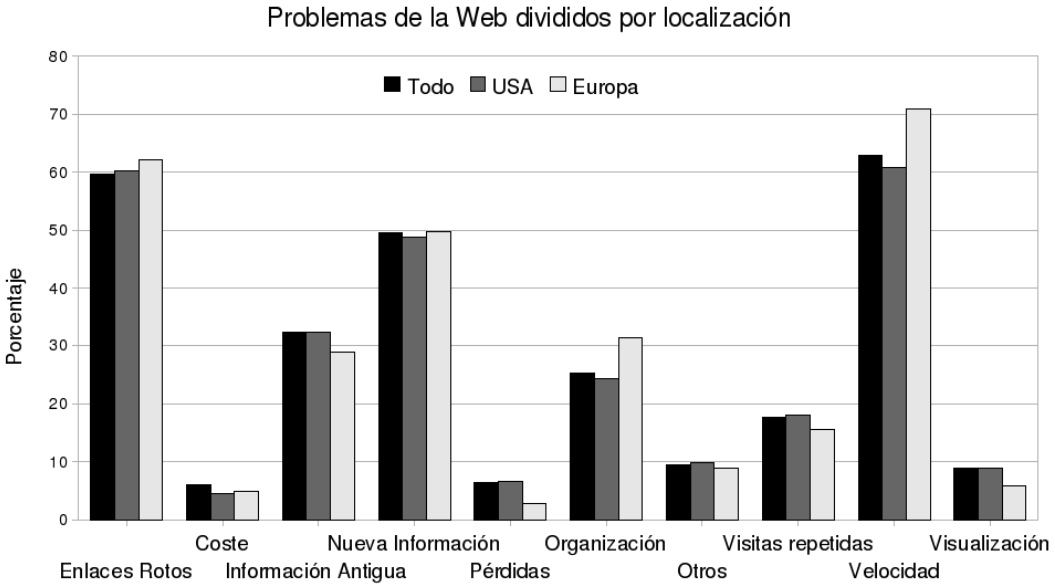

Figura 1.1: *Estudio que describe los principales problemas de la Web por localización, según los usuarios. Fuente:GVU's Eight User Survey*

En este tiempo, han aparecido numerosos trabajos[Kah97, Koe99, Koe02, Koe04, MB02, MCNB05, AMM08] que han mostrado la evolución de la forma en la que las páginas web desaparecen con el paso del tiempo. Y es que la mayoría de los usuarios de Internet probablemente confirmarán que el mensaje de error HTTP 404 "Page Not Found" forma parte del día a día en la navegación web. Sin embargo creemos que las páginas Web no desaparecen sin dejar ningún rastro sino que además de la replicación de información en Internet, las páginas que dejan de estar activas, a menudo lo hacen para moverse a otro lugar. De esta forma, el uso de algunas de las técnicas habituales de recuperación de información podrían ayudar a encontrar la nueva ubicación de muchas páginas. En relación al frecuente cambio en el mapeo de las URLs (Uniform Resource Locator) de un sistema, existen cuatro escenarios generales:

- La misma URL apunta al mismo contenido o un contenido muy similar antes o después.
- La misma URL apunta a un contenido diferente antes o después.
- Una URL diferente apunta al mismo contenido o un contenido muy similar antes o después.
- El contenido deja estar apuntado por ninguna URL.

La existencia de enlaces rotos en un sitio o página web tiene implicaciones tanto para los potenciales usuarios como para los motores de búsqueda. Es evidente que de cara a un usuario, el hecho de encontrar este tipo de problemas dificulta y entorpece sus búsquedas de información, pero además esto repercute en una perdida de fiabilidad y prestigio. Estas consecuencias, en definitiva, significan un deterioro en las sensaciones del usuario, pero además existen otra serie de repercusiones como son la penalización por parte de los motores de busqueda[II08] a la hora de ofrecer una lista de resultados ante una consulta. Incluso han aparecido algoritmos[EMT04] que proponen la modificación del *PageRank*[MRS08] para tener en cuenta este tipo de errores, penalizando las páginas con enlaces rotos debido a que es considerado como un problema en crecimiento y que podría tener un gran impacto en los rankings generados por los motores de busqueda. De este ultimo trabajo[EMT04] se ha extraido la Figura 1.2, que muestra como más del 6 % de los enlaces seguidos durante un crawling en el año 2004 de mil doscientos millones de páginas, fueron enlaces rotos.

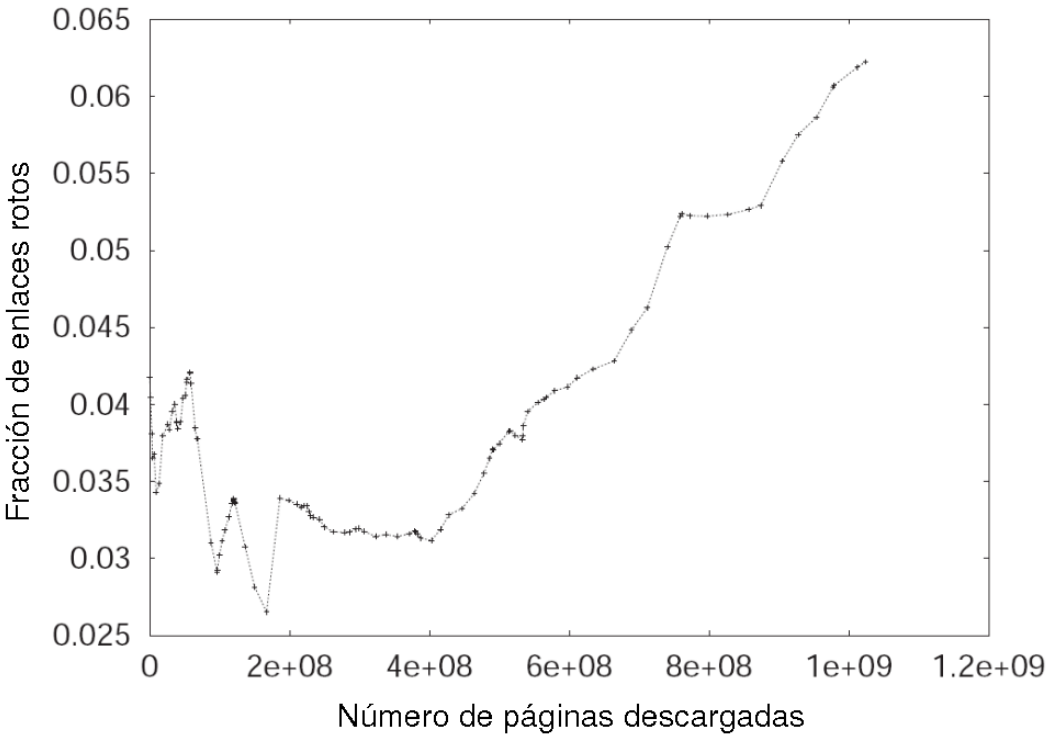

Figura 1.2: *Enlaces rotos encontrados (Errores 403 y 404) durante un crawling de mil doscientos millones de páginas. Fuente:Ranking the Web Frontier, WWW'04*

1.2. Promoción Fraudulenta de los Recursos en un Motor de Búsqueda

Desde hace unos años, los motores de búsqueda se han convertido en uno de los principales recursos ante la necesidad de información. De hecho, un estudio[3] realizado en 2009 por la agencia Reuters, señala a Internet como la principal fuente de información (56 %) por delante de la televisión (21 %), prensa escrita (10 %) y radio (10 %). Esta popularidad y prestigio se ha convertido en uno de los principales objetivos del marketing y la publicidad. A día de hoy, la publicidad en Internet forma parte del modelo de negocio de muchas compañías, y esto ha repercutido en que haya empresas que se dediquen a analizar el funcionamiento interno de los motores de búsqueda, con el objetivo de adquirir el conocimiento necesario para dotar a una determinada página o sitio web de una mayor importancia y/o relevancia ante las consultas de los usuarios. El *Web Spam* o *Spamdexing*[GGM05] podría definirse como cualquier acción destinada a mejorar el ranking en un buscador por encima de lo que se merece. Esto comprende una serie de técnicas relacionadas con el contenido de las páginas y los enlaces entrantes y salientes de cada una de ellas. Al mismo tiempo, los motores de búsqueda están invirtiendo recursos con el objetivo de detectar este tipo de páginas fraudulentas y así eliminarlas de sus índices y penalizar todos aquellos enlaces relacionados. Estas acciones a su vez provocan en los *spammers* la eliminación y reubicación de estás páginas detectadas, provocando finalmente una gran cantidad de enlaces rotos en la Web. De esta forma, existe una relación directa entre los enlaces rotos de una página web y la posibilidad de que dicha página esté relacionada con estas actividades ilícitas o lo haya estado en el pasado.

En la Figura 1.3 puede observarse un ejemplo típico de spam en el cual un enlace que debería llevarnos a una página con libros online, según indica el texto de su ancla (“online books”), enlaza realmente a una página de venta de productos farmacéuticos. Si anteriormente citábamos los enlaces rotos como una presunta característica de spam, en esta ocasión es la divergencia entre el contenido de las dos páginas enlazadas, el indicio que nos puede llevar a detectar un caso de enlaces fraudulentos o nepotismo.

Por los motivos expuestos anteriormente, en este trabajo proponemos el desarrollo de un sistema de recuperación de enlaces rotos, que por un lado ofrezca a un potencial usuario un listado ordenado de páginas candidatas a sustituir un determinado enlace roto, y que por otro lado este sistema sea aplicable también a la extracción de un conjunto de rasgos de páginas web para la detección de spam.

[3] http://www.reuters.com/article/idUSTRE55G4XA20090617

Figura 1.3: *Ejemplo de web spam en el cual la página origen enlaza a otra que supuestamente tiene libros online, pero que realmente vende Viagra y otros productos farmaceúticos.*

1.3. Objetivos del Trabajo

Desde el punto de vista de la recuperación de páginas desaparecidas, proponemos el desarrollo de un sistema que recomiende al usuario una serie de páginas candidatas a sustituir a un enlace roto, teniendo en cuenta la información contextual que se puede obtener acerca de ese enlace.

En una primera fase, tras detectar los enlaces rotos de una página o sitio web, el sistema deberá recuperar toda la información disponible en el contexto de un determinado enlace roto. Esta información consistirá en el texto del ancla, el párrafo donde se encuentre el enlace y el contenido o el título de dicha página. Como parte de esta primera fase, el sistema debería utilizar los recursos disponibles en la infraestructura web para recolectar la mayor cantidad de información posible, relativa a la página desaparecida. Entre estos recursos disponibles, se consultará el *Internet Archive*[4] que almacena diferentes versiones de una gran cantidad de páginas web desde 1996. Los motores de búsqueda, también podrían ser una fuente de información muy importante debido a que también almacenan una versión relativamente reciente de muchas páginas. Por último, desde hace un tiempo vienen tomando una gran relevancia los sistemas de redes sociales. Algunos de estos sistemas se han desarrollado en forma de etiquetado social, en los cuales los usuarios asignan una o varias etiquetas (palabras) a un recurso. Estas etiquetas podrían ser datos significativos a la hora de obtener más información acerca de la temática o peculiaridades del enlace roto.

En una segunda fase, el sistema deberá filtrar la información de la que dispone y generar una serie de consultas a un buscador; con todos los términos de la fuente de información en algunos casos, o con los términos mas relevantes en el caso

[4]http://www.archive.org

de las fuentes con un texto mayor. Después, para cada consulta, el sistema deberá recuperar los primeros resultados devueltos por el buscador, reuniendo un conjunto con las páginas más relevantes. Para la extracción de terminología de las fuentes de información más amplias, sería necesario seleccionar la técnica de recuperación de información[MRS08] más apropiada (*Tf*[5], *Tf-Idf*[6], modelos de lenguaje, etc.)

En tercer lugar, el sistema deberá mostrar este conjunto de páginas relevantes de una forma ordenada por relevancia, mostrando la primera página candidata al principio. Para la elaboración de esta lista ordenada, sería necesario analizar las funciones de ranking[MRS08] que pudieran aplicarse a esta tarea (*Tf*, *Tf-Idf*, co-ocurrencia, modelos de lenguaje, etc.), teniendo en cuenta la información de la que se pudiera disponer.

Finalmente deberá establecerse un marco de evaluación objetivo para ofrecer una serie de datos acerca del funcionamiento del sistema.

Detección de Web Spam

Por otro lado y teniendo en cuenta la estrecha relación entre los enlaces rotos y el web spam, proponemos la adaptación del sistema de recuperación de enlaces, anteriormente descrito, para la extracción de una serie de rasgos que pudieran ser útiles a un clasificador en la difícil tarea de detección de Web Spam.

Además, como complemento a estos rasgos relacionados con la tipología del enlace (activo o roto) y teniendo en cuenta las características del ejemplo de spam mostrado en la Figura 1.3, proponemos el uso de técnicas de recuperación de información para analizar la divergencia entre los modelos de lenguaje de dos páginas enlazadas. De esta forma se podría detectar una posible discrepancia entre el tema del enlace expuesto y el contenido real de la página apuntada.

1.4. Estructura de la Tesis

Este documento se ha dividido de la siguiente manera:

- Capítulo 1: En este primer capítulo se ha expuesto la introducción de esta tesis, describiendo por un lado las carencias en la integridad referencial de las páginas web y la promoción fraudulenta de recursos en los motores de búsqueda. En este capítulo también se presentan los objetivos de la tesis.

- Capítulo 2: Incluye la definición del problema de la integridad referencial de las páginas web junto a una revisión del estado del arte. El capítulo 2 también se compone de una serie de definiciones que permitirán comprender mejor el marco conceptual de este trabajo, una revisión bibliográfica sobre los

[5]Frecuencia de términos

[6]Frecuencia de términos - frecuencia inversa en documentos

temas relevantes de nuestra investigación, así como una investigación sobre las técnicas empleadas en esta tesis.

- Capítulo 3: Contiene la definición del problema del web spam y se hace un breve repaso a los principales tipos de web spam, junto a los principales trabajos de cada área. Este capítulo cuenta también con un estado del arte que engloba los trabajos más estrechamente relacionados con esta tesis. Finalmente se presentan un conjunto de conceptos, líneas de investigación y métodos que permitirán asimilar mejor las ideas presentadas en el capítulo 5.

- Capítulo 4: Exhibe los estudios empíricos realizados durante el desarrollo del sistema de recuperación de enlaces rotos. Incluye el diseño del sistema, un estudio de las etapas en las que se divide el proceso de recuperación, un análisis de las principales fuentes de información empleadas, dos estudios comparativos sobre técnicas de extracción de terminología y métodos de ranking y un análisis del rendimiento del sistema. En este capítulo también se muestran los resultados más destacados así como unas conclusiones relativas a este capítulo.

- Capítulo 5: Engloba dos tipos de subsistemas que extraen rasgos acerca de un conjunto de páginas web con el objetivo global de componer un sistema de detección de web spam. El primer subsistema se centra en el análisis de la calidad de los enlaces, mientras que el segundo subsistema analiza la divergencia de contenido entre dos páginas enlazadas. En este capítulo también se describirá el proceso de desarrollo del sistema y la metodología empleada. También se muestran los resultados del sistema de detección de web spam en el que se integran los subsistemas anteriores, así como unas conclusiones relativas a este capítulo.

- Capítulo 6: Este último capítulo se centra en el desarrollo de las conclusiones y la exposición de futuras líneas de investigación que se proponen a raíz de esta tesis.

- Apéndice A: Este apéndice muestra el funcionamiento de una aplicación web que implementa el sistema de recuperación de enlaces rotos presentado en el capítulo 4.

Parte II

Antecedentes

CAPÍTULO **2**

Integridad Referencial en Hiperenlaces

En este capítulo se va a realizar una introducción al problema de los enlaces rotos en la Web. En líneas generales se van a tratar cuatro temas; en primer lugar se muestra el origen del problema y cómo las URLs tienen una fecha de caducidad relativamente pequeña; a continuación se presentan los principales trabajos de investigación que han intentado de resolver el problema y los foros científicos que han tratado activamente de solicitar trabajos en esta línea; el siguiente tema introduce los dos pilares sobre los que se asentará el futuro sistema de recuperación de enlaces rotos, como son la infraestructura web y la extracción de terminología; finalmente se exponen las principales medidas de evaluación en el campo de la recuperación de información y que deberán ser aplicadas al sistema de recuperación, con el objetivo de demostrar sus características de eficacia y eficiencia.

2.1. Definición del Problema

En esta sección analizaremos el origen del problema de la pérdidad de integridad referencial en la Web y cuales son los cambios que se están produciendo en el diseño de los sitios web que perjudican la solución de estos errores.

2.1.1. Longevidad de las URLs

El origen del problema de los enlaces rotos está compartido tanto por el diseño original de los enlaces unidireccionales como por la persistencia de las páginas web. Por este motivo, en esta primera sección vamos a presentar el origen del problema para poder entender la dimensión del mismo y de esta manera analizar las posibles soluciones.

A pesar de las guías de estilo que muestran consejos para crear URLs duraderas[BL98], la mayoría de usuarios de Internet probablemente confirmarán que el mensaje de error HTTP 404 “Page Not Found” forma parte del día a día en

la navegación Web. La pérdida de integridad de los enlaces (link integrity) en la Web ha sido estudiada por numerosos investigadores[Dav00a, Dav98, ADWC98, Ash00].

En este tiempo, han aparecido trabajos que han mostrado la evolución de cómo las páginas web desaparecen con el paso del tiempo. En 1997, Brewster Kahle (fundador del Internet Archive[1]) empezó a invertir tiempo y recursos para intentar preservar la Internet ante la desaparición continua y definitiva de páginas, estableciendo el tiempo de vida medio[Kah97] de una página web en 44 días. Más tarde, y después de diferentes estudios[Koe99, Koe02, Koe04], Koehler estableció la persistencia de una página en algo menos de dos años, aunque podría ser variable en función del tipo de recurso[NA02]. Como parte de este estudio, Koehler hizo un seguimiento a un conjunto aleatorio de páginas durante un periodo de 4 años, suficiente como para considerarlo un periodo de tiempo estable, tras el cual un 67 % de estas URLs habían dejado de estar activas. Otro estudio[MB02] monitorizó los recursos de tres cursos online durante 14 meses y pudo comprobar como el 16.5 % de los enlaces habían desaparecido o no estaban disponibles. Después de esto, salió a la luz un nuevo trabajo en el que se confirmaba el hecho de que cada semana desaparecía un enlace de cada 200 en Internet. Recientemente[AMM08] se ha intentado demostrar la influencia de este tipo de errores en los servicios web disponibles en Internet, arrojando la significativa cifra de que el 16.44 % de los servicios web tienen algún enlace roto o recurso no disponible.

También en el ámbito de los artículos científicos se han realizado análisis de este tipo[MCNB05], descubriendo que era imposible acceder a la mitad de las URLs citadas en la revista "D-Lib Magazine" 10 años después de su publicación. Lawrence et al.[LPF+01] descubrieron en el año 2000 que entre el 23 % y el 53 % de todas las URLs que se encontraban en artículos relacionados con las ciencias de la computación entre 1994 y 1999 habían desaparecido. En este mismo trabajo, los autores mediante búsquedas manuales, fueron capaces de reducir esta cantidad de enlaces rotos a un 3 %, buscando réplicas en la red o la nueva ubicación de estos recursos. Este último hecho confirma nuestra hipótesis de que la mayoría de enlaces rotos que aparecen, son generados por el movimiento del recurso apuntado. Spinellis[Spi03] condujo un estudio muy similar investigando la accesibilidad de las URLs presentes en los artículos científicos publicados en "Communications of the ACM" y en "IEEE Computer Socierty". Spinellis descubrió que el 28 % de todas las URLs habían desaparecido después de 5 años y esta cifra ascendía al 41 % después de 7 años. También muestra el dato relevante de que para el 60 % de las URLs no accesibles, el error devuelto por el servidor web era el 404. Al mismo tiempo, estimó que el tiempo de vida medio de las URLs era de 4 años desde su fecha de publicación. Nelson y Allen[NA02] estudiaron la disponibilidad de los recursos en bibliotecas digitales, concluyendo que el 3 % de las URLs desaparecían después de un año. Dellavalle et al.[DRP03] examinaron las referencias en Internet de artículos publi-

[1] http://www.archive.org

cados en revistas con un gran factor de impacto, proporcionado por el Institute for Scientific Information (ISI). Los autores revelan que el 30 % de los artículos contienen referencias a recursos en la Web, y que estas referencias se convierten en inaccesibles un mes después de que la revista que los contiene alcanza el factor de impacto más alto (1 % de los artículos) en las áreas de ciencia y medicina. A su vez, descubrieron que el porcentaje de referencias inactivas se vio incrementado desde el 3.8 % (3 primeros meses) al 10 % después de 15 meses, y llegando al 13 % al cabo de 27 meses. En este estudio la mayoría de referencias inactivas correspondían al dominio *.com* (46 %) y al dominio *.org* (5 %), aunque buscando manualmente en el *Internet Archive* fueron capaces de recuperar el 50 % de estas referencias. Markwell y Brooks[MB02] también analizaron el dominio de los enlaces inactivos, considerando el dominio *.gov* como el más estable, y el dominio *.edu* como el más provisional.

La Tabla 2.1 muestra un breve resumen de la información más relevante de los artículos previamente citados desde el punto de vista de la longevidad de las URLs.

Publicación	**Año**	**Objeto de Interés**	**Enlaces Rotos**	**Información adicional**
Lawrence et al.[LPF+01]	2000	URLs en artículos del área de Ciencias de la computación publicados entre 1994 y 1999	23-53 %	Búsqueda manual redujo los enlaces rotos al 3 %
Koehler[Koe02]	2002	Colección de URLs aleatorias	67 % después de 4 años	Vida media estimada de 2 años
Markwell y Brooks[MB02]	2002	URLs en cursos online	16.5 %	Dominio *.gov* el más estable
Nelson y Allen[NA02]	2002	Recursos de bibliotecas digitales	3 %	Realizaron una búsqueda manual posterior
Dellavalle et al.[DRP03]	2003	URLs de revistas con alto factor de impacto	3.8 % tras 3 meses y 13 % tras 27 meses	Recuperaron manualmente el 50 % de los enlaces
Spinellis[Spi03]	2003	URLs publicadas en artículos de ACM e IEEE	28 % tras 5 años y 41 % tras 7 años	Tiempo de vida medio 4 años
McCown et al.[MCNB05]	2005	URLs en artículos de D-Lib en 1995-2004	30 %	Tiempo de vida medio 10 años

Tabla 2.1: *Resumen de la persistencia de URLs según diferentes artículos de investigación.*

2.1.2. Deterioro del Diseño Original de la Web

La Web evoluciona en la mayoría de los casos de manera positiva, pero en algunas situaciones se crean tecnologías y se implementan modificaciones en el diseño original de la red que provocan daños colaterales. En este trabajo se pone de manifiesto el problema latente de los enlaces rotos. Pues bien, si la resolución de este problema ha sido una tarea difícil desde el comienzo, en la actualidad se han realizado algunos cambios en la respuesta de los servidores que dificultan aún más esta labor.

Bar-Yossef et al.[BYBKT04] introdujeron una medida formal de la degradación de la Web. En su trabajo muestran el sorprendente hecho de que no solo páginas individuales muestran un deterioro significativo, sino que también se produce en grupos de páginas e incluso en zonas enteras de la Web. En su trabajo, también describen otro de los problemas que provocan la degradación de la red como son la respuesta a los errores 404 (Page Not Found). Indiscutiblemente, detectar los errores 404 "puros" (casos en los que el error 404 es realmente devuelto por el servidor) y detectar URLs sintácticamente incorrectas es una tarea trival, pero ellos describen un algoritmo para detectar los también llamados "errores 404 suaves" (softs 404) - que son errores 404 que no devuelven el código de respuesta HTTP correspondiente sino el código 200 cuyo significado es respuesta correcta. Realmente la petición del usuario genera un error 404 interno en el servidor debido a su configuración pero el mismo servidor genera un mensaje de respuesta configurado previamente por el administrador al mismo tiempo que se devuelve el código 200 al navegador. De cara a los usuarios, incluso esta respuesta generada por el servidor podría ser más agradable, ya que se podría mostrar el logotipo de la empresa, un mensaje de disculpa e incluso la oportunidad de utilizar un buscador interno de la propia compañía. Pero el problema de este tipo de respuestas es que el navegador interpreta que la página se ha encontrado correctamente y por lo tanto el integrar soluciones para la recuperación de páginas no encontradas es aún más difícil en estos casos.

En la Figura 2.1 se muestra un ejemplo de un "error 404 suave" donde la URL solicitada era `http://pages.ebay.com/information-retrieval`, que a pesar de estar bien formada no existía en la realidad. En esta imagen además de mostrarse un texto adaptado al usuario, ofrecen una serie de opciones de navegación para intentar que el usuario continúe navegando por el sitio.

Bar-Yossef et al.[BYBKT04] presentaron un método para detectar si un servidor web produce errores 404 suaves. Este método envía dos peticiones diferentes al servidor sospechoso. La primera petición consiste en una página que previamente se ha comprobado que existe, y en la segunda petición se genera automáticamente una dirección que con una alta probabilidad no existe. Posteriormente se comparan las dos respuestas del servidor teniendo en cuenta si existe algún tipo de redirección. Además, el contenido de las páginas devueltas es comparado usando *shingles*[BDGM95], lo que consiste en tomar un conjunto de términos contiguos de ambos documentos y comparar el número de coincidencias. Mediante esta técnica,

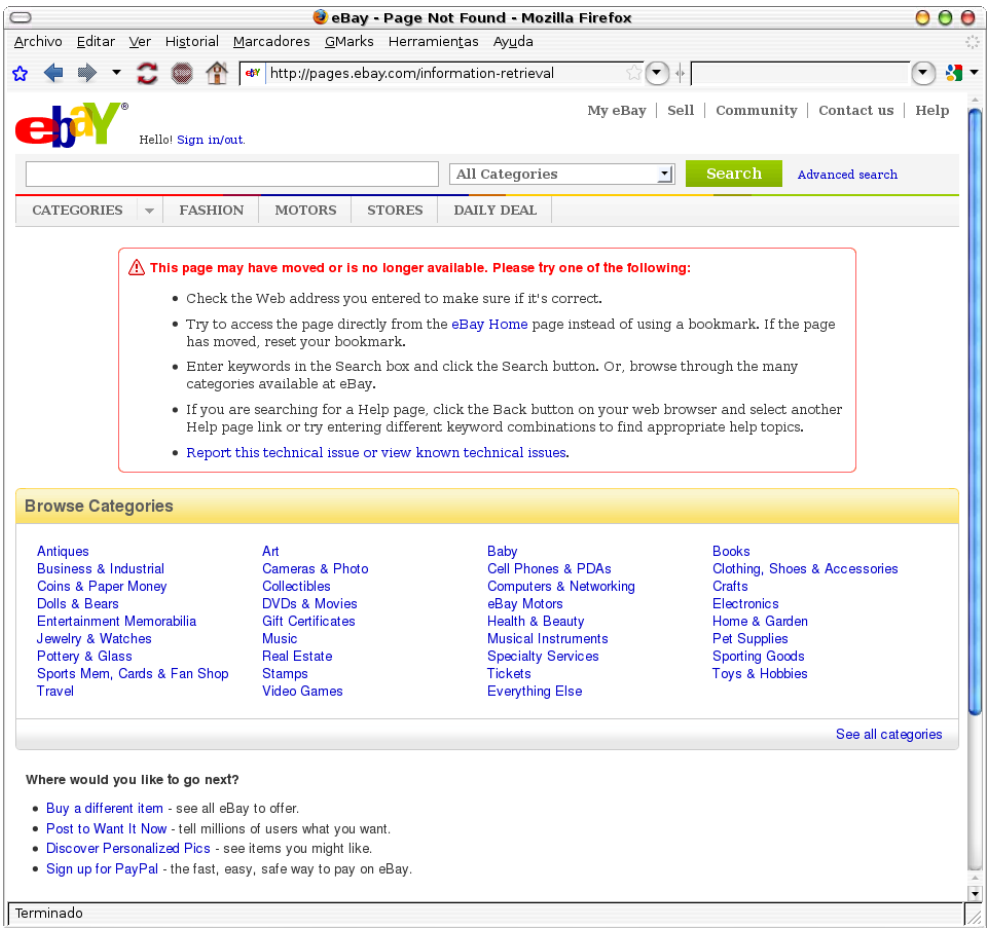

Figura 2.1: *Respuesta del servidor de eBay ante una página no encontrada en la cual no se devuelve un error HTTP 404 sino el código 200 y un texto adaptado.*

el método de detección de errores 404 suaves compara el resultado de las dos peticiones y si el contenido es muy similar, el algoritmo muestra una clara indicación de que un error 404 suave ha sido detectado. Los autores son conscientes de un problema relacionado con el método que proponen, y es el hecho de que dos URLs correctas sean redirigidas a la misma página. Por ejemplo, existen sitios web que después de alguna modificación en la compañía a la que pertenecen pasan a formar parte de un grupo mayor. En este caso, si tienen dos secciones que trataban de la misma temática, una de ellas pasa a ser redirigida a la otra, quedando toda la información centralizada en la segunda. De esta forma, cualquier petición realizada sobre la primera página es redirigida a la segunda. Por lo tanto, si atendemos a la manera de funcionar del método anteriormente descrito, dos peticiones (correcta e incorrecta) sucesivas, podrían ser redirigidas a la segunda página y de esta manera ser considerado como un error 404 suave de manera errónea. Ante este posible fallo del método de detección, sus autores presentan una solución en la que declaran la

raíz de un sitio web como el único caso en el que no se puede presentar un error 404 suave.

Existen por lo tanto diferentes escenarios a la hora de encontrar una página desaparecida y recibir información extra acerca del problema y su posible solución. En un extremo se puede encontrar un error estándar devuelto por un servidor web y en el otro extremo podría encontrarse una aplicación dedicada a este problema, que sugiera una serie de opciones al usuario para intentar resolver el problema. En la Figura 2.2 se muestra un ejemplo con la menor información disponible que puede encontrarse tras intentar acceder a una página desaparecida. Este ejemplo muestra la página por defecto del navegador *Firefox* a la hora de recibir el error HTTP 404, y como se puede comprobar, tan solo se informa del error sin proponer ninguna otra alternativa al usuario excepto comprobaciones triviales como la corrección de la dirección web o el chequeo de la conexión a Internet.

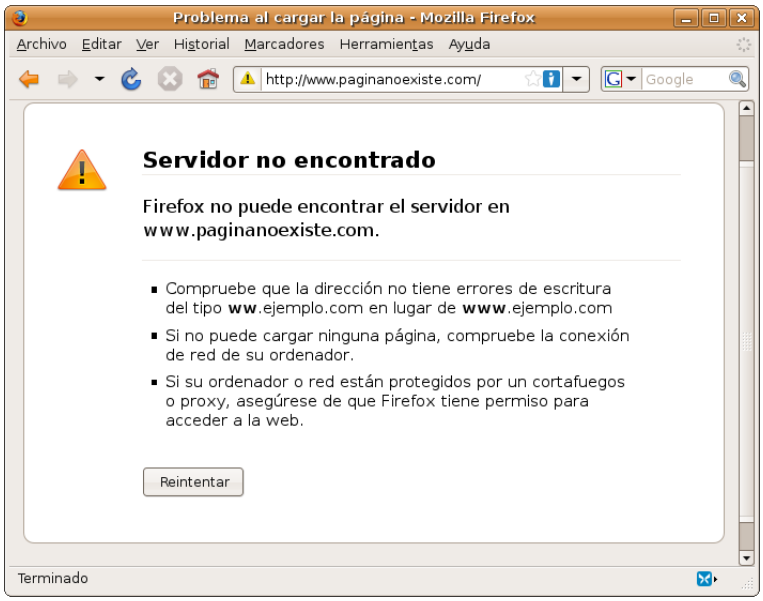

Figura 2.2: *Respuesta de error 404 estándar mostrada por el navegador Firefox.*

En la actualidad, un gran número de sitios web incluyen como parte de sus facilidades de navegación la ayuda al usuario en caso de intentar acceder a una página no existente en su dominio. Un ejemplo de esta ayuda se puede observar en la Figura 2.3, donde tras intentar acceder a una página errónea dentro del dominio del *Ministerio de Educación y Ciencia*, se muestra una página adaptada al usuario en la que se muestran diferentes alternativas para poder encontrar el recurso requerido.

Finalmente el mayor grado de ayuda que se puede encontrar a la hora de tratar el problema de los enlaces rotos, se encuentra en aplicaciones instaladas en el ordenador del usuario. Una de estas aplicaciones es el *plug-in ErrorZilla*[2], que consiste

[2]http://jaybaldwin.com/Projects.ErrorZilla-Mod.aspx

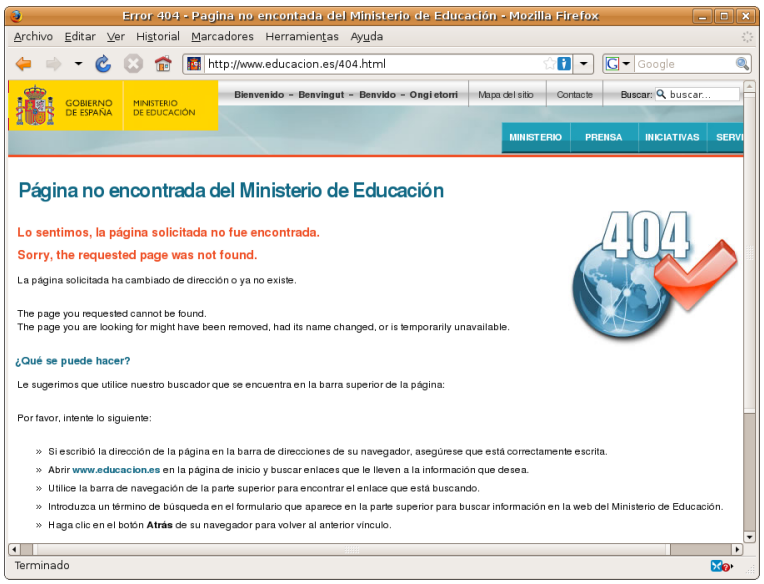

Figura 2.3: *Respuesta de error 404 adaptada por el servidor web del Ministerio de Educación y Ciencia.*

en una aplicación que se instala como un complemento al navegador *Mozilla* y que ofrece al usuario una serie de opciones para encontrar la página que ha originado el error. Este complemento, del cual puede apreciarse un *snapshot* en la Figura 2.4, proporciona al usuario ayuda en diferentes niveles de comunicación de la red. Por un lado dispone de herramientas para comprobar el estado de la conexión física tales como *Ping*, *Trace*, etc. Por otro lado establece mecanismos en el nivel de transporte como la consulta a librerías digitales como la caché de Google, en la que se puede consultar la última versión indexada de una determinada página web, o el *Internet Archive*[3], donde pueden encontrarse múltiples versiones de un sitio web desde la creación de este archivo en 1996. *ErrorZilla* también ofrece la opción de *coralizar*, que consiste en utilizar la aplicación *Coral*[4]. Esta aplicación P2P permite el acceso a una red de distribución de contenidos en la Web, mediante la cual se puede conseguir una versión de la URL solicitada.

Como resultado de los diferentes escenarios que un usuario puede encontrarse a la hora de solicitar una página desaparecida se encuentran las grandes diferencias entre las posibles respuestas del servidor y sobre todo la enorme diferencia entre la página que el usuario esperaba encontrar y el contenido que es finalmente mostrado en su navegador. En este sentido se encuentra el trabajo realizado por Francisco-Revilla et al. [FRSF+01] en el que presentan *Waldens Paths Path Manager*. Esta herramienta permite a los usuarios, principalmente profesores, construir caminos

[3]http://www.archive.org
[4]http://www.coralcdn.org

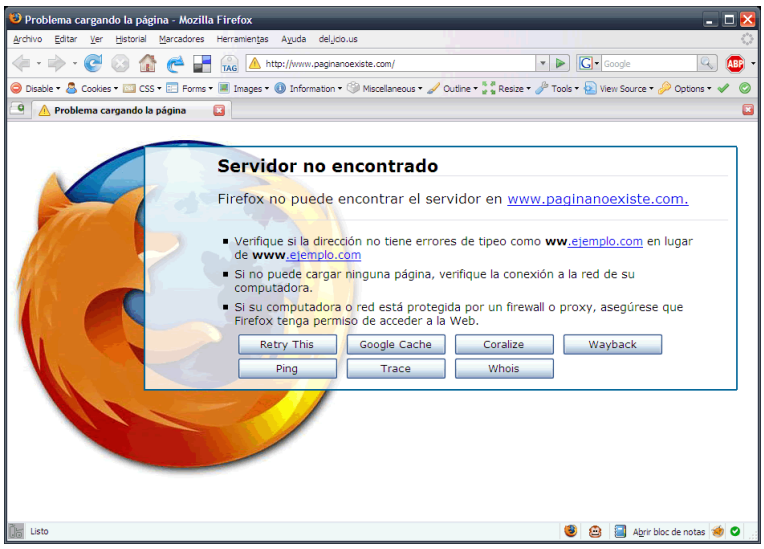

Figura 2.4: *Página de ejemplo mostrada por el plug-in ErrorZilla ante un error 404.*

o rutas empleando para ellos páginas web que normalmente han sido creadas por terceros. Este camino puede verse intuitivamente como un meta-documento que organiza y agrega información contextual de estas páginas. De esta forma, parte de la investigación se centra en descubrir cambios relevantes en sitios web. En algunos casos, comparar la versión actual de un documento con la última copia almacenada en alguna caché no es suficiente ya que la web es altamente dinámica y algunos cambios podrían ser incluso esperados. De esta forma es posible que determinadas páginas (noticias, blogs, etc.) cambien a un ritmo constante, y por tanto una comparación simple de su contenido no sea suficiente. En su trabajo se centran en el descubrimiento de cambios significativos por encima de estas modificaciones circunstanciales. El marco de evaluación que proponen está basado en firmas de documentos en determinadas partes tales como las cabeceras, párrafos, enlaces y palabras clave. La herramienta también dispone de un historial de cambios de estos valores de tal forma que el usuario puede discriminar entre cambios menores y cambios de mayor envergadura.

2.2. Estado del Arte

El problema de la integridad referencial en la Web, a pesar de estar presente desde el inicio de la navegación en Internet, no ha sido un tema ámpliamente tratado por la comunidad científica. En esta sección ofreceremos una visión general de los trabajos que han aparecido en esta línea tratando de mostrar las principales virtudes y defectos de cada uno de ellos. En los últimos años coexisten diferentes confe-

rencias dentro del área de la recuperación de información que de manera periódica inciden en determinados problemas de la Web, incentivando su estudio y análisis. De esta forma, después de repasar los principales trabajos relacionados con los sistemas de recuperación de enlaces rotos, citaremos alguna de estas conferencias que han intentado promover la investigación de este problema.

2.2.1. Métodos para Resolver el Problema de los Enlaces Rotos

A continuación se presenta un resumen de los principales métodos que han sido utilizados para resolver el problema de los enlaces rotos. La principal característica que tienen en común los tres enfoques siguientes es que necesitan algún tipo de recogida de información antes de que se produzca el error, ya sea instalando un software específico o necesitando una configuración específica del servidor web por parte de un administrador.

Protocolo HTTP

El protocolo de transferencia de hipertexto (HTTP, Hy perText Transfer Protocol)[FGM+99] es el protocolo usado en cada transacción de la Web. En el ámbito de los enlaces rotos, este protocolo proporciona la funcionalidad de redirigir una petición a una página que ha sido movida a una localización diferente. Esta funcionalidad se consigue con el código de respuesta 301, representando el aviso al usuario de que el recurso que se está solicitando ha sido asignado a una URL diferente. Además, esta respuesta puede añadir la funcionalidad de redirigir automáticamente al usuario a la nueva dirección. Por otro lado, el código 307 indica la asignación temporal de una nueva URL para el recurso solicitado. Estas respuestas del servidor son de gran ayuda, a la hora de evitar los enlaces rotos, pero requieren de la continua atención del administrador web para avisar y configurar el servidor ante los posibles cambios. En el caso de los usuarios que construyen su própia página web, esta funcionalidad no suele ser de gran ayuda ya que en la mayoría de los casos no tienen acceso a la configuración del servidor.

URLs Persistentes

Para solucionar los problemas de la inestabilidad y la volatilidad en la localización de los documentos en la red, existen una serie de iniciativas que tienen como fin encontrar, de una forma única y normalizada, la localización de cada documento, es decir, identificar el documento mediante una localización inequívoca y que persista a lo largo del tiempo y del espacio en Internet. A continuación se enumeran las iniciativas más importantes llevadas a cabo para estandarizar y normalizar la localización de los recursos electrónicos en Internet:

- *Uniform Resource Locator (URL)*[BLMM94] es el sistema más común de localización de documentos dentro de la web. Una de sus principales características es que este recurso no describe el nombre del documento, sino la forma de acceder a él.

- El *Uniform Resource Name (URN)*[DGIF99] fue una iniciativa de la *Internet Engineering Task Force (IETF)*, la rama de desarrollo de ingeniería y protocolos de Internet, con la premisa de conseguir una forma universal de identificación de recursos, para que cada recurso fuera único y constante.

- *Uniform Resource Identifier (URI)*[BLFM98] también ha sido desarrollado por el *IETF* y pretende crear un sistema mundial para identificar recursos de todo tipo en la Web: documentos, imágenes, programas, servicios, correos electrónicos, etc. Este método combina URNs y URLs, o dicho de otra manera, nombres y direcciones. Se trata de identificar los documentos mediante una secuencia de sintaxis controlada que identifica cada documento de una forma única.

- *Internationalized Resource Identifiers (IRI)*[BLFM98] es un nuevo elemento de protocolo, un complemento para los URIs. Un IRI es una secuencia de caracteres del conjunto de caracteres universales (Universal Character Set) (Unicode/ISO10646). Existe un mapeado de IRIs a URIs, que permite que los IRIs puedan usarse en lugar de URIs cuando esto sea más apropiado para identificar recursos. El uso de IRIs es compatible con los esquemas URI.

- Un *DOI* (Digital Object Identifier)[Pas02] consiste de una única secuencia alfanumérica que contiene dos partes. La primera se conoce como *Publisher ID* e indica el número que le asigna la Agencia *DOI* al editor. La segunda parte, se conoce como *Item ID*, y es un identificador que le asigna el editor concreto y que puede ser una secuencia alfanumérica de caracteres.

- También pretende establecerse una forma universal de identificación de recursos, para que cada recurso sea único y constante. El sistema se denomina *Persistent Uniform Resource Locator* (PURL)[SWJF96] y ha sido desarrollado por *Online Computer Library Center (OCLC)*. Consiste en elaborar una base de datos de URLs a partir del protocolo HTTP. En esta base de datos se almacenan los nombres de los documentos y los servidores donde se alojan. En el caso de que un documento cambie de lugar, tan solo es necesario comunicárselo a la base de datos para que se produzca un redireccionamiento de forma automática. *PURL* hace de intermediario entre la vieja dirección y la nueva. Existen otros muchos proyectos e iniciativas para identificar recursos en la Web. Podemos destacar el *Handle System*[SLB03], un software desarrollado por el *Corporation for National Research Initiatives (CNRI)* que provee un mecanismo para nombrar e identificar objetos digitales.

Existen otras iniciativas de menor relevancia para el problema de los enlaces rotos como son: *Serial Item and Contribution Identifier (SICI)*, *Book Item and Contribution Identifier (BICI)*, *Publisher's Item Identifier (PII)*, *International Standard Work Code (ISWC)* o *Human Friendly Names (HFNs)*.

Una de las primeras propuestas desarrolladas para hacer frente al problema de los enlaces rotos, fue la creación de identificadores persistentes que fueran robustos a la posible reubicación del recurso al que representa. Hugh C. Davis[Dav98, Dav00a] fue uno de los autores que propusieron el uso de identificadores persistentes. En sus artículos estudió las causas que provocan la existencia de los enlaces rotos, y estableció la prevención de estos errores como la solución definitiva. En su trabajo defiende la existencia de un servidor de URNs que de manera centralizada resuelva los nombres de los recursos, muchos de ellos almacenados en bases de datos de hiperenlaces (Hyperbase) capaces de guardar diferentes versiones de los recursos. También propone el desarrollo de herramientas que hagan posible la edición sencilla por parte de los usuarios de los enlaces de sus páginas teniendo en cuenta el modelo propuesto.

Otro de los primeros trabajos que intentaron resolver el problema utilizando servidores e identificadores persistentes fue el de Ingham et al.[ICL96]. Estos autores presentaron un modelo de suministro de recursos web suponiendo la integridad referencial[5]. Para ello introdujeron una novedosa metodología basada en orientación a objetos. En concreto utilizan objetos con herencia para modelar los aspectos de la Web (páginas, enlaces, contenido, etc.) Fue uno de los primeros trabajos que introdujeron los conceptos PURL y LRN (local resources names).

Shimada y Futakata[SF98] 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. En la Figura 2.5, se muestra el complejo esquema del sistema SEDB en el que puede apreciarse como la base de datos de enlaces es el núcleo del sistema, basado en la idea de identificadores persistentes.

El sistema Webvise[GSØ99] integrado con software de *Microsoft*, almacena información acerca de los recursos web en bases de datos hipermedia. Esto permite al sistema un cierto grado de recuperación ante un posible error, ya que esta base de datos sería actualizada cada vez que se crea o modifica un enlace. El sistema a su vez, está basado en el protocolo *Open Hypermedia Protocol (OHP)* que propone que los recursos (enlaces) en la Web estén almacenados en bases de datos (archivos de sonido, html, hojas de calculo, etc.). De esta forma, un *front-end* de esta base de

[5]Propiedad característica en las bases de datos mediante la cual se garantiza que una entidad siempre se relaciona con otras entidades válidas

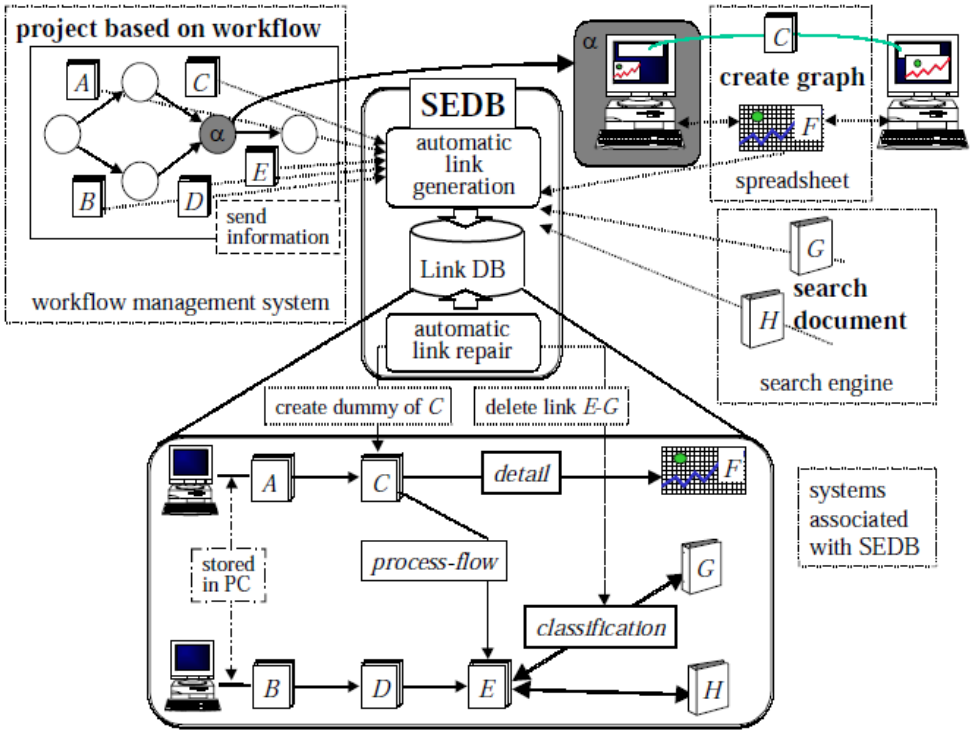

Figura 2.5: *Esquema del complejo sistema SEDB en el que pueden apreciarse como la base de datos de enlaces es el núcleo del sistema, basado en la idea de identificadores persistentes.Fuente:Shimada y Futakata[SF98]*

datos respondería a peticiones de URLs convencionales devolviendo estos recursos almacenados. Otra de las propuestas es la existencia de un servidor proxy que almacene las operaciones del usuario y por la que pasarían todas las peticiones web del usuario mediante un cliente especial (browser). Mediante el cliente, el usuario podría establecer un ancla en una página cualquiera y referenciarla desde otra diferente. Esta relación la almacena el proxy y en sucesivos accesos devolvería al usuario la página enlazada, aunque ésta desaparecería de su sitio original. Al igual que este proxy y el cliente, también proponen extensiones para aplicaciones como Internet Explorer, Word, etc.

Autoridades de enlaces

Intuitivamente, una autoridad en el ámbito de los hiperenlaces, es una página web con un grado alto de mantenimiento y actualizada cada vez que alguno de los recursos que apunta son movidos o eliminados. La Figura 2.6 ilustra el concepto de autoridad de enlaces. u_1 es la página web de un laboratorio en la universidad A, y es enlazada por otras páginas $(v_1, v_2, v_3 \ldots)$. Siendo v_3 una página oficial que contiene

la lista de laboratorios de la universidad. En tal caso, v_3 es una autoridad de u_1, pero cuando u_1 es movido a u_2, el enlace a u_1 en v_3 es siempre actualizado a u_2 en un corto periodo de tiempo.

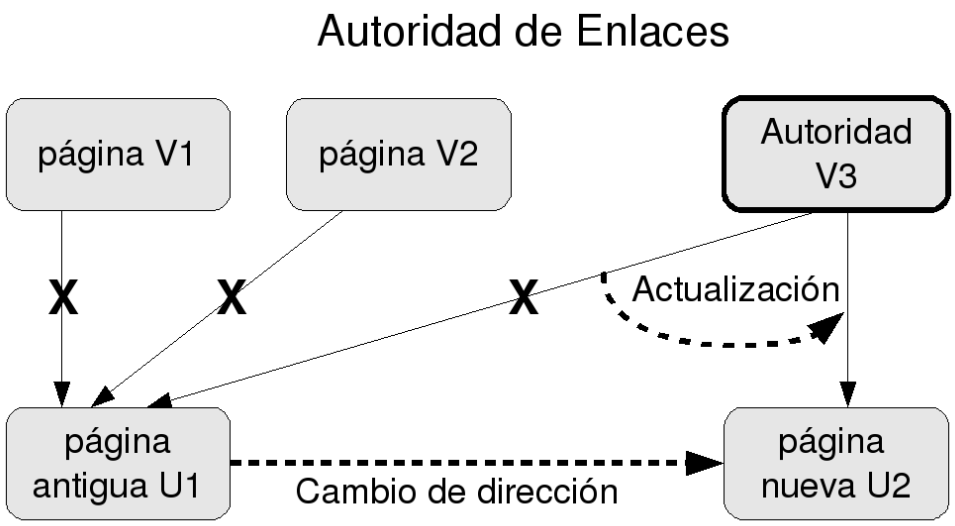

Figura 2.6: *Esquema de una autoridad de enlaces.*

Nakamizo et al.[NIM+05] han desarrollado un sistema de recuperación de enlaces basado en el concepto de "autoridades de enlaces" utilizando para ello servidores especializados. Este sistema descubre la nueva URL de aquel recurso que ha sido movido. Los autores ofrecen un ejemplo en el que una compañía ha cambiado su nombre y por tanto también ha cambiado su sitio web. Sin embargo, ellos son capaces de reencontrar los recursos utilizando para ello las autoridades de enlaces del sitio web antiguo. De esta forma, el nuevo sitio web pudo cambiar su nombre adaptándolo al nuevo nombre de la compañía, actualizando todos sus enlaces de manera fiable.

Morishima et al.[MNI+08] extendieron este trabajo mejorando la herramienta con heurísticas basadas en los cambios de localización más probables que se producen al mover los recursos web. Estas heurísticas corresponden a los típicos movimientos que se producen dentro del mismo sitio web, por tanto no se garantiza el encontrar recursos en otras localizaciones. De esta manera manejan suposiciones como la probabilidad de que una página haya sido movida dentro del mismo dominio, a pesar de haber cambiado la URL. También realizan búsquedas en diferentes profundidades de la URL suponiendo que el recurso se haya movido a otra sección del sitio web, como por ejemplo un empleado que haya cambiado de departamento. Además, utilizan la información que proporcionan las redirecciones del protocolo HTTP para realizar con algunos términos una consulta a un motor de búsqueda. Otro de los aspectos novedosos es que el sistema muestra como salida al usuario

una lista de páginas web ordenadas por la probabilidad de ser una autoridad de enlaces. Esta probabilidad es obtenida en función de una serie de características basadas en la estructura y nombre de los enlaces.

Firma Léxica

Una firma léxica (lexical signature, LS)[KN08a, KN08b] es un pequeño conjunto de términos derivados de un documento que capturan la temática de ese documento. Esta firma léxica podría entenderse como meta-descripción ligera de un documento que representa los términos más significativos de su contenido. La Tabla 2.2 muestra algunos ejemplos de firmas léxicas que fueron generadas a partir de una serie de URLs. Para la generación de esta firma léxica, utilizan un esquema *Tf-Idf* [MRS08] teniendo en cuenta el ranking asignado a esa URL y el numero aproximado de resultados devuelto por el motor de búsqueda (Marzo, 2008). La primera página fue tomada del sitio de Robert Wilensky[6] teniendo en cuenta que se trataba de una descripción de un proyecto sobre procesamiento del lenguaje natural.

URL	Términos de la Firma Léxica
www.cs.berkeley.edu/∼wilensky	texttiling wilensky disambiguation subtopic iago
www.loc.gov	library collections congress thomas american
www.dli2.nsf.gov	nsdl multiagency imls testbeds extramural
www. jcdl2008.org	libraries jcdl digital conference pst

Tabla 2.2: *Términos extraídos mediante una firma léxica a partir de un conjunto de URLs tras varias consultas en Marzo de 2008. Fuente:Klein and Nelson[KN08b]*

Phelps y Wilensky[PW00] introdujeron el uso de firmas léxicas para encontrar contenidos que habían sido movidos a otra URL. El titulo de su primer trabajo "robust hyperlinks cost just 5 words each" y sus primeros experimentos, confirmaron el buen funcionamiento de esta técnica, que se basaba en crear URLs en las que se adosaba al final la firma léxica, separada por el símbolo "?" (http://www.cs.berkeley.edu /∼wilensky?ls=texttiling+wilensky+disambiguation+subtopic+iago).

Park et al.[PPGK04] extendió el trabajo de Phelps y Wilensky analizando el comportamiento de nueve algoritmos diferentes para la creación de la firma léxica. En general concluyeron que los algoritmos pesados por frecuencia de términos (TF) funcionan mejor en la búsqueda de páginas similares. En cambio los algoritmos que utilizan el esquema Tf-Idf tienen un mejor comportamiento a la hora de encontrar la página exacta, con algunas diferencias en el contenido no significativas. En general, estos trabajos no profundizaron más en el tema ni encontraron una explicación al buen funcionamiento de los cinco términos que componen la firma léxica.

[6] http://www.cs.berkeley.edu/∼wilensky

Recientemente Harrison y Nelson[HN06] retomaron esta linea de investigación desarrollando un sistema llamado *Opal*, que hacia uso de la técnica de firmas léxicas para encontrar páginas desaparecidas. Los autores utilizaron el mismo método que Phelps y Wilensky para la extracción de los términos más relevantes, pero tampoco hicieron énfasis en la cantidad idónea de términos. El sistema Opal capturaba los enlaces rotos (error 404) y redirigía al usuario a la misma página que había sido movida o a una nueva página con un contenido relacionado. El principal inconveniente de este método es el esfuerzo necesario por parte del administrador web, ya que se necesita de un servidor específico que sea configurado y mantenido para cada cliente *Opal*.

Wan y Yang[WY06] introdujeron el concepto de *WordRank* basado en la firma léxica, en el cual, mediante la relación semántica entre los términos de una firma léxica, se seleccionaban los términos más representativos. Los autores realizaron los experimentos con una cantidad de cinco términos como los trabajos previos y concluyeron al igual que Park et al.[PPGK04], que los métodos basados en la frecuencia para la selección de la firma léxica favorecían la identificación de páginas relacionadas, mientras que el uso de un esquema Tf-Idf funcionaba mejor a la hora de encontrar una página en concreto.

Nelson et al.[NMSK07] presentaron varios modelos para la preservación de páginas web basadas en los conceptos de *infraestructura web* y *firma léxica*. Los autores argumentaron que los métodos convencionales utilizados en la preservación digital, tales como el almacenamiento digital en archivos específicos y la aplicación de métodos de refresco y migración, no son escalables debido al coste que implica. A su vez, introdujeron un método para la reconstrucción de sitios web mediante el uso de aplicaciones web que proporcionan versiones cache de páginas web, tales como el *Internet Archive* y la caché de motores de búsqueda. En la Figura 2.7 puede apreciarse el esquema del sistema que propusieron basado en las firmas léxicas y la infraestructura web para la recuperación de páginas desaparecidas.

Klein y Nelson[KN08b] extendieron el trabajo de Harrison y Nelson[HN06] analizando en profundidad la generación de las firmas léxicas. Para la generación de la firma léxica, es necesaria la existencia de una colección de referencia y por este motivo los autores tratan de justificar el uso de información inexacta procedente de *Google*, con el objetivo de utilizar el índice de este buscador como colección. En este trabajo analizaron la eficiencia de las firmas léxicas con el paso del tiempo. Para ello utilizaron el *Internet Archive* para extraer diferentes versiones de una página web entre 1996 y 2007. Sobre cada una de estas versiones extrajeron una firma léxica y evaluaron su utilidad analizando el ranking que proporcionaba *Google* a la URL real unos cuantos años después.

A pesar de que todos estos trabajos que emplean métodos como URLs persistentes, autoridades de enlaces y firmas léxicas son la referencia en el área de la recuperación de enlaces rotos, tienen diferentes puntos débiles que nosotros intentaremos corregir. En primer lugar, estos trabajos dan por hecho la existencia de un servidor, que de alguna manera contenga información por adelantado de las páginas que

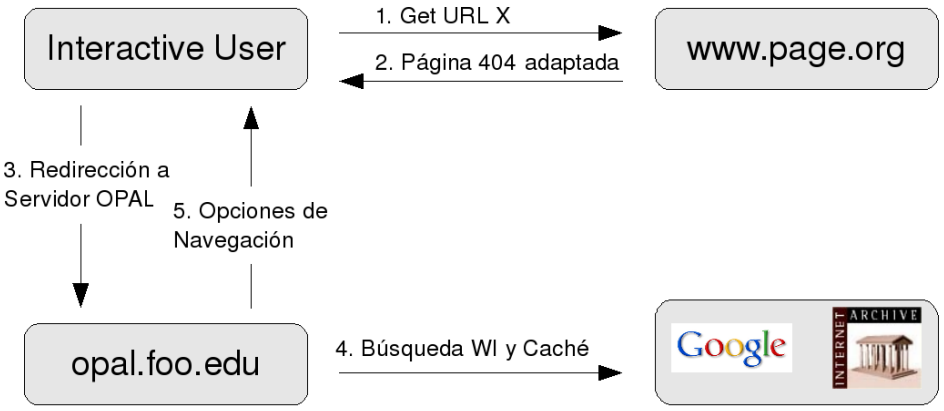

Figura 2.7: *Sistema Opal basado en las firmas léxicas y la infraestructura web para la recuperación de páginas desaparecidas.*

serán recuperadas posteriormente. En segundo lugar, la extracción de terminología que representa a las páginas, no ha sido analizada en profundidad, estableciendo el método más apropiado y examinando la repercusión del número de términos seleccionados. En tercer lugar, algunos trabajos enfocan el problema de los enlaces rotos como un conflicto local, limitando el problema y por tanto su solución. Finalmente, una de las principales carencias es la ausencia de una metodología de evaluación que permita la comparación de resultados.

2.2.2. Búsqueda de Páginas de Inicio de Entidades

En la sección actual vamos a presentar una serie de trabajos enmarcados dentro de la misma competición, que propuso la búsqueda de la página de inicio de un conjunto de entidades. A pesar de que el objetivo de dicha competición no es exactamente el mismo que planteamos en este trabajo, la similitud de algunas técnicas empleadas resultan ser de gran interés.

Las conferencias o foros de evaluación TREC[7] se han convertido en el foro de intercambio científico y de evaluación más prestigioso del campo de la recuperación de información. Tiene un carácter anual desde 1991, reúne a creadores de diferentes sistemas y realizan evaluaciones de los resultados que éstos obtienen en diferentes pruebas, previamente estandarizadas. En la edición del año 2003 llegaron a participar 99 grupos de investigación de 22 países diferentes.

Inicialmente, estas conferencias nacieron con la idea de resolver uno de los mayores problemas de las evaluaciones de los sistemas de recuperación de información. La evaluación solía llevarse a cabo sobre pequeñas colecciones de documentos, y sus resultados resultaban de difícil extrapolación. En 1991, para subsanar

[7] http://trec.nist.gov/

este problema, la agencia *DARPA (Defense Advanced Research Projects Agency)* propuso desarrollar estos foros para propiciar que los investigadores evaluaran sus sistemas sobre una colección de documentos común.

Los cinco objetivos básicos que se propusieron para dar lugar a estas evaluaciones fueron los siguientes:

- Desarrollo de un foro abierto a los entornos académicos, industrial y gubernamentales.
- Frecuencia anual para la presentación de los avances en investigación desarrollados.
- Investigación en recuperación de información sobre grandes colecciones de documentos.
- Incremento de la transferencia tecnológica.
- Avance de las técnicas de evaluación.

A su vez en las conferencias TREC se presentan cada año una serie de talleres (Tracks) que se intentan centrar de manera específica en las diferentes áreas de la recuperación de información. En estos talleres se proponen una serie de tareas orientadas al desarrollo y evaluación de sistemas en un ámbito concreto en el que existe un problema a resolver. Estos talleres crean la infraestructura necesaria para la investigación como son colecciones de prueba, metodologías de evaluación, etc. Algunos de estos talleres son: *Cross-Language Track, Enterprise Track, Filtering Track, Interactive Track, Million Query Track, Question Answering Track, Blog Track, SPAM Track o Web Track.*

Precisamente este último taller (*Web Track*) ha propuesto diferentes tareas relacionadas con la búsqueda web sobre una colección de referencia. En concreto y teniendo en cuenta la relación con este trabajo, destaca la décima edición del TREC (TREC-10)[CH01], en la que se introdujo la tarea de "Homepage Finding". En esta tarea se contaba con una colección de 1.69 millones de páginas web y un conjunto de consultas que consistían en los nombres de una serie de entidades, cuyas páginas de inicio se encontraban en dicha colección. El desafío de esta tarea consistía en generar un ranking de páginas web como respuesta a estas consultas, en las que todas las páginas de inicio que correspondían a una determinada entidad aparecieran en las primeras posiciones.

Uno de los atractivos de estas conferencias es la cantidad de trabajos que se presentan planteando el problema desde diferentes puntos de vista y aplicando técnicas de diversa naturaleza. Entre los trabajos más destacados, destaca el de Westerveld et al.[WKH01], que usó modelos de lenguaje para intentar resolver el problema, teniendo en cuenta además diferentes fuentes de información: el contenido de la página, el número de enlaces entrantes, la profundidad de la URL y los textos de las anclas de los enlaces salientes.

Xi et al.[XFTS02] enfocaron la tarea como un problema de machine learning dividido en tres etapas: en un primer paso se usó el modelo de espacio vectorial para obtener la similitud entre el texto de una consulta y diferentes fuentes de información. Después filtraron las páginas de la colección mediante un árbol de decisión. Finalmente, las páginas resultantes eran ordenadas siguiendo un modelo de regresión logística.

Amati y Carpineto[ACR01] emplearon la técnica de *pseudo-relevance feedback* además del modelo *Divergence from Randomness (DFR)* para la extracción de términos relevantes y la expansión de consultas.

En la edición del TREC-11, el *Web Track* introdujo una variación a la tarea del "Homepage finding" denominada "Named page finding". En esta edición la tarea consistía en encontrar una página en concreto, que no debía ser necesariamente la página de inicio. Entre los trabajos que se presentaron, destaca el de Park et al.[PMRJ02], que desarrollaron un sistema de recuperación de documentos calculando la similitud entre las consultas y todas las frases de un documento. Esta medida de similitud consistía en el número de palabras en común entre los dos ítems aplicando un factor de pesado.

En el año 2003 se fusionaron las tareas de los dos años anteriores, combinando "Homepage finding" y "Named page finding". En esta edición del TREC-12, Ogilvie et al.[OC03] crearon varias representaciones de documentos utilizando para ello texto, título, texto de anclas, etc. Con cada una de estas representaciones, los autores construyeron diferentes modelos de lenguaje. Posteriormente los modelos de lenguaje fueron combinados usando una interpolación lineal para formar un nuevo modelo de lenguaje con el objetivo de estimar la similitud con la consulta.

En la última edición (TREC-13) de este *Web Track*, la tarea fue modificada de nuevo, quedando su finalidad fuera del objetivo de este trabajo.

Aunque estos trabajos en el ámbito de TREC tenían en común la tarea de localizar una o varias páginas web, su naturaleza era diferente a la planteada en este trabajo. El primer lugar, las técnicas empleadas eran específicas de la tarea propuesta en cada edición. En segundo lugar, en la tarea de búsqueda de la página de inicio, el número de candidatos potenciales era mucho más reducido que la realidad que se presenta ante nuestro trabajo. Finalmente, el tamaño de la colección era mucho menor y la información que se disponía de ella (grafo de enlaces completo) mucho mayor que lo que se podría obtener de la Web.

2.3. Sistemas de Recuperación de Información

En esta sección se presentan las principales características de un sistema de recuperación de información (SRI), dado que uno de los objetivos de esta tesis es desarrollar un sistema similar para la recuperación de enlaces rotos.

Es de sobra conocida la cantidad ingente de información que se encuentra en Internet al igual que la necesidad de desarrollar tecnologías de búsqueda que faci-

liten el acceso a esa información. Los SRI aparecieron precisamente para ocupar esa necesidad y ofrecer al usuario final una serie de facilidades a la hora de encontrar los recursos que solicita. Existen muchas definiciones de un SRI, de las cuales consideramos las más relevantes:

- **Salton y McGill[SM86]:** Un sistema de recuperación de información procesa archivos de registros y peticiones de información, e identifica y recupera de los archivos ciertos registros en respuesta a las peticiones de información.
- **Korfhage[Kor97]:** La localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta.
- **Baeza-Yates[BYRN99]:** Parte de la informática que estudia la recuperación de la información (no datos) de una colección de documentos escritos. Los documentos recuperados pueden satisfacer una necesidad de información de un usuario expresada normalmente en lenguaje natural.

Un SRI permite la recuperación de determinada información, previamente almacenada en una colección o base de datos, por medio de la realización de una serie de preguntas en forma de consulta a los documentos contenidos en dicha colección. Esta serie de preguntas o interrogaciones se modelan como expresiones formales que reflejan una necesidad de información, y suelen expresarse por medio de un lenguaje de interrogación. A su vez, a la hora de desarrollar de un sistema de recuperación de información, se deben tener en cuenta las siguientes cuestiones: (1) La recuperación de información es un estudio multidisciplinar en el que intervienen áreas como la lingüística, informática, biblioteconomía, etc. (2) Los documentos podrían estar almacenados en diferentes formatos. (3) La información recuperada debería satisfacer las necesidades de información del usuario.

Los actuales buscadores en Internet utilizan dos formas básicas para almacenar y recuperar información:

- Directorios: Agrupan la información en una estructura temática y jerárquica relacionada. La búsqueda se realiza recorriendo la estructura desde la raíz y descendiendo por los sub-árboles hasta encontrar el nodo requerido.
- Motores de búsqueda: Utilizan robots que visitan páginas automáticamente, analizando sus cambios para enviarlos a un repositorio en donde se indexan para su posterior recuperación.

2.3.1. Clasificación de los SRI

El diseño de un SRI se realiza bajo un modelo en el que se definen las principales características del sistema. Entre estos atributos se encuentra la manera en que se obtienen las representaciones de los documentos y la consulta. También se establece

la estrategia para evaluar la relevancia de un documento respecto a una consulta y los métodos para establecer el orden de los resultados.

En base a estas características, existen diferentes clasificaciones de los SRI. En la Tabla 2.3 se muestra la propuesta de división en cinco grupos según el modelo utilizado:

Modelo	**Características del Modelo**
Modelos Clásicos	Se encuentran los tres modelos más utilizados: booleano, espacio vectorial y probabilístico
Modelos Lógicos	Están basados en la Lógica Formal donde la recuperación de información en un proceso inferencial
Modelos Alternativos	Basados en la Lógica Fuzzy
Modelos Interactivos	Permiten la expansión de la consulta y emplean la retroalimentación mediante la relevancia de los documentos recuperados
Modelos basados en la Inteligencia Artificial	Bases de conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural.

Tabla 2.3: *Clasificación de los SRI elaborada por Dominich[Dom00] según el modelo utilizado.*

Existen otras clasificaciones como la de Baeza-Yates[BYRN99] en la cual los modelos de recuperación de información son divididos según la tarea inicial que el usuario realiza en el sistema: (1) El usuario recupera información mediante una ecuación de búsqueda que escribe en un formulario. (2) El usuario navega por los documentos con el objetivo de buscar referencias incluyendo en el modelo el hipertexto utilizado. Baeza-Yates también divide los modelos de recuperación en dos grupos: (1) Clásicos, que incluyen a los modelos booleano, espacio vectorial y probabilístico. (2) Estructurados (escasamente utilizados), que corresponden a listas de términos sin solapamiento y a nodos próximos.

2.3.2. Modelo de Espacio Vectorial

El modelo de espacio vectorial (MEV) fue planteado y desarrollado por Gerard Salton[SM86] y corresponde a un modelo algebraico utilizado en diversas áreas de la recuperación de información. Está basado en la representación de documentos en lenguaje natural de una manera formal mediante el uso de vectores en un espacio lineal multidimensional. En esta tesis, el MEV se emplea en diferentes tareas de recuperación de información tales como la extracción de terminología, la medida de similitud entre dos páginas web o la relevancia entre una consulta y un documento analizado.

En el MEV, cada documento se representa mediante un vector en el que se refleja la frecuencia de aparición de cada término en dicho documento. Los términos

de dicho vector no pueden ser palabras vacías (*stopwords*), sino que deben tener cierto significado dentro del documento. Por otro lado, los términos del documento son reducidos a un *stem* o tema, utilizando técnicas de *stemming*. Por ejemplo, los términos "bibliotecas" y "bibliotecario" serían reducidas al mismo *stem* que correspondería a "bibliotec". El algoritmo más común de *stemming* es el algoritmo de Porter. Existen además métodos basados en análisis lexicográfico y otros algoritmos similares (kstem, stemming con cuerpo, métodos lingüísticos, etc.).

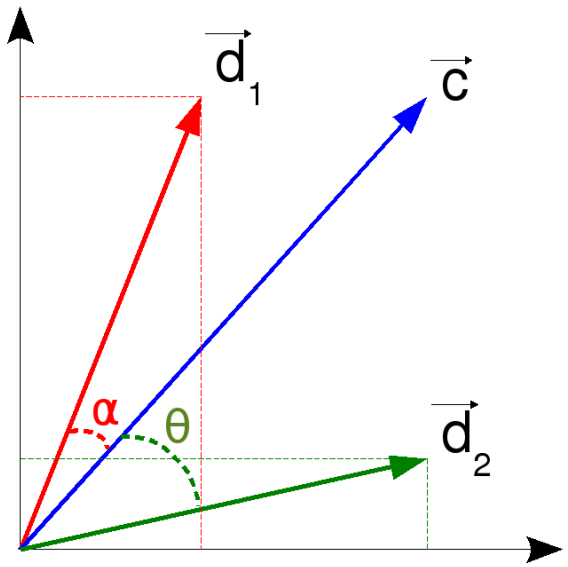

Figura 2.8: *Modelo de Espacio Vectorial en el que se muestra los ángulos entre los vectores formados por la consulta (c) de un usuario y dos documentos (d_1 y d_2).*

En cuanto a la frecuencia de aparición de cada documento, el MEV se comenzó utilizando con la frecuencia de cada término (*tf*) en un documento, pero pronto surgieron argumentos en contra debido a que la importancia de un término en función de su distribución puede llegar a ser desmesurada. Por ejemplo, un valor de frecuencia de 2 es 200 % más importante que un valor de frecuencia de 1, cuando la diferencia aritmética es tan solo de una unidad. A partir de ese momento han surgido diferentes métricas para establecer el peso de un término en el vector que representa a un documento. Ese es el caso de *idf* (frecuencia inversa de documento) que incluye la discriminación de un término frente a otro utilizando para ello una colección de documentos y no solamente un documento individual. Esta medida además incentiva la presencia de aquellos términos que aparecen en menos documentos frente a los que aparecen en muchos, ya que realmente los muy frecuentes discriminan poco o nada a la hora de la representación del contenido de un documento. Finalmente, el peso de un término en un documento se calcula mediante la

combinación de la frecuencia de término (*tf*), y la frecuencia inversa del documento (*idf*). Para calcular el valor de la j-esima entrada del vector que corresponde al documento i, se emplea la ecuación siguiente: $d_{ij} = tf_{ij}\ x\ idf_j$.

Dentro del MEV existen varias medidas de similitud como son el producto escalar, la distancia euclídea y el coseno. A continuación se muestra la función de similitud calculada mediante el producto escalar:

$$SIM(d,c) \ = \ \sum_{i=1}^{n} d_i\ c_i \tag{2.1}$$

siendo d el documento, c la consulta y n el número de términos del vocabulario compuesto por la unión de los términos del documento y de la consulta.

En cuanto a la aplicación práctica del MEV, la teórica básica indica que la relevancia de un documento frente a una consulta puede calcularse usando la diferencia de ángulos (basada en el coseno de esos ángulos) de cada uno de los documentos respecto del vector de la consulta. En la Figura 2.8 se muestran los ángulos entre los vectores formados por la consulta (c) de un usuario y dos documentos (d_1 y d_2). En este caso el coseno de α correspondería a la similitud del documento d_1 con respecto a la consulta y el coseno de θ correspondería a la similitud del documento d_2. Así un valor de coseno de cero significa que la consulta y el documento son ortogonales el uno al otro, y eso significa que no hay coincidencia y por tanto el documento no es relevante.

Para determinar el coseno del ángulo entre dos vectores se usa la siguiente ecuación:

$$cos\ \theta = \frac{V_1 \cdot V_2}{||V_1||\ ||V_2||} \tag{2.2}$$

donde:

- V_1 y V_2 son los vectores.
- θ es el ángulo entre V_1 y V_2.
- $\cdot$ representa el producto escalar.
- $||X||$ representa la magnitud del vector X.

Por tanto para calcular la similitud entre dos unidades de texto, en este caso un documento y una consulta, se utiliza la siguiente fórmula del coseno que se apoya en el producto escalar introducido en la ecuación 2.2.

$$SIM(d,c) \ = \ \frac{\sum_{i=1}^{n} d_i\ c_i}{\sqrt{\sum_{j=1}^{n} (d_j^2\ c_j^2)}} \tag{2.3}$$

siendo d el documento, c la consulta y n el número de términos del vocabulario compuesto por la unión de los términos del documento y de la consulta.

Entre las ventajas de este modelo de recuperación se encuentran:

- Es posible obtener una lista ordenada de documentos que satisfacen la consulta.
- Es posible controlar la respuesta ante una consulta, ya sea limitando el número de resultados o estableciendo un umbral de similitud.

La principal desventaja a tener en cuenta es que este modelo supone que los términos de indexación son independientes.

2.3.3. Modelo Probabilístico

El modelo probabilístico, también conocido como modelo de recuperación de independencia binaria, fue introducido en la década de los setenta por Robertson y Sparck Jones. La base principal del funcionamiento de este modelo es el cálculo de la probabilidad de que un documento sea relevante para una consulta dada.

En este modelo se presupone que existe exactamente un subconjunto de documentos que son relevantes para una consulta dada. Además, para cada documento se intenta evaluar la probabilidad de que el usuario lo considere relevante. En relación a la relevancia de un documento, se considera como el resultado de dividir la probabilidad de que el documento sea relevante para una pregunta entre la probabilidad de que no lo sea. En el hipotético caso de que el usuario supiese los términos que permiten caracterizar tal subconjunto de documentos relevantes (porque aparecen en ellos y no aparecen en el resto de los documentos de la colección), el problema estaría resuelto. En definitiva, se considera exclusivamente de la presencia o ausencia de los términos en los documentos de la colección. Se trata, por lo tanto de un modelo binario, como el modelo booleano. Pero en un caso real el usuario no sabe cuáles son los términos que configurarían la consulta ideal. Tampoco sabe, de hecho, en qué medida los términos empleados en la consulta permiten discernir los documentos relevantes y rechazar simultáneamente los documentos irrelevantes.

El modelo probabilístico no toma en cuenta la frecuencia de aparición de los términos y necesita suponer que todos los términos son independientes unos de otros, siendo esta situación poco realista debido a que como es bien sabido, hay términos cuya presencia suele estar muy vinculada a otros (por ejemplo, “perro” y “gato” suelen estar muy presentes de forma conjunta).

El modelo introduce un refinamiento a los términos que configuran la consulta del usuario, ponderándolos mediante la imposición de un peso, mayor cuanto mejor permita discriminar los documentos relevantes de los irrelevantes, y menor en caso contrario. De esta manera se persigue que el sistema efectúe la recuperación

incidiendo sobre todo en los mejores descriptores de entre los empleados por el usuario en la consulta, minimizando la importancia de aquellos otros términos que, aun figurando en la consulta, son malos descriptores del conjunto respuesta ideal. Basados en estos pesos iniciales, el modelo probabilístico es capaz de calcular el grado de similitud existente entre cada documento de la colección y la consulta ponderada, consiguiendo ordenar los documentos de la colección en orden descendente de probabilidad de relevancia en relación a la consulta. De esta manera el modelo probabilístico supera el gran inconveniente puesto de manifiesto en el modelo booleano que es la equiparación exacta. En efecto, el modelo probabilístico, aun siendo un modelo binario, efectúa equiparación parcial, lo que permite ordenar los documentos de la respuesta conforme a su probabilidad de relevancia.

Si un documento es seleccionado aleatoriamente de la colección hay cierta probabilidad de que sea relevante a la pregunta. Si una base de datos contiene N documentos, y n de ellos son relevantes, entonces la probabilidad se estima en:

$$P(rel) = \frac{n}{N} \tag{2.4}$$

En concordancia con la teoría de la probabilidad, la probabilidad de que un documento no sea relevante para una pregunta dada, viene expresada por la siguiente formula:

$$1 - P(rel) = \frac{N-n}{N} \tag{2.5}$$

2.3.4. Modelos de Lenguaje

En esta sección introduciremos en primer lugar el concepto de modelos de lenguaje que posteriormente será utilizado por el sistema de recuperación de enlaces rotos para tareas tales como la extracción de terminología o el ranking de un conjunto de resultados. Más tarde, se profundizará en la descripción del método de modelado de lenguaje. También se mostrarán algunas comparaciones entre los diferentes métodos de modelado de lenguaje y una serie de posibles extensiones.

En el mundo de la recuperación de información y concretamente en la recuperación de documentos de una colección mediante consultas, se suele aplicar intuitivamente la técnica de componer una consulta con aquellos términos que con una alta probabilidad aparezcan en el documento que queremos recuperar. Los métodos de modelado del lenguaje aplicados a la recuperación de información utilizan precisamente esa idea como su modelo principal: Un documento es un buen resultado para una consulta si el modelo de lenguaje de ese documento generase con una alta probabilidad esa misma consulta, lo que a su vez ocurriría si el documento contuviese los términos de la consulta con una frecuencia alta.

Esta técnica difiere de las funciones de ranking desarrolladas antes de la aparición de los modelos de lenguaje[MRS08]. Previamente se había establecido la

relevancia de un documento d respecto a una consulta q mediante los métodos probabilísticos tradicionales como $P(R = 1|q,d)$. Con la aparición de los modelos de lenguaje se crea un paradigma diferente en el cual se construye un modelo de lenguaje probabilístico M_d por cada documento d, otorgando un ranking a los documentos basándose en la probabilidad de que dicho modelo genere la consulta: $P(q|M_d)$.

La noción de modelo de lenguaje es inherentemente probabilística dado que se trata de una función que otorga una medida de probabilidad a las cadenas extraidas de algún vocabulario. De esta forma, para un modelo de lenguaje M aplicado a un alfabeto $\sum$:

$$\sum_{s \in \sum^*} P(s) = 1 \tag{2.6}$$

Un tipo simple de modelo de lenguaje es equivalente a un autómata finito probabilístico que consiste solamente en un único nodo con una única distribución de probabilidad sobre la producción de los diferentes términos. Después de generar cada palabra, se decide si detener o continuar el bucle y producir una palabra más, por lo que el modelo también requiere una probabilidad para parar en un estado final. Este modelo supone una distribución de probabilidades sobre cualquier secuencia de palabras. De acuerdo con la construcción de los modelos de lenguaje, también son un modelo para la generación de texto en función de su distribución.

Tipos de Modelos de Lenguaje

La forma más simple de modelo de lenguaje simplemente ignora todo el contexto que le condiciona, y estima la probabilidad de cada término independiente. Este modelo se denomina modelo de lenguaje basado en *unigramas*:

$$P_{uni}(t_1\ t_2\ t_3\ t_4) = P(t_1)\ P(t_2)\ P(t_3)\ P(t_4) \tag{2.7}$$

Existen otros tipo de modelos de lenguaje más complejos, como por ejemplo los modelos de lenguaje basados en *bigramas* que establece condiciones sobre el término previo sobre el que se calcula la probabilidad:

$$P_{bi}(t_1\ t_2\ t_3\ t_4) = P(t_1)\ P(t_2|t_1)\ P(t_3|t_2)\ P(t_4|t_3) \tag{2.8}$$

También existen incluso modelos de lenguaje basados en gramáticas tales como las gramáticas libres de contexto probabilísticas. Estos modelos son vitales para tareas como reconocimiento del habla, corrección ortográfica y traducción automática, donde es necesario conocer la probabilidad de un término condicionada por

el contexto. Sin embargo, la mayoría del modelado de lenguajes que son utilizados en recuperación de información suelen usar modelos de lenguaje basados en unigramas. Precisamente la recuperación de información no es un área donde sean necesarios complejos modelos de lenguaje, ya que no depende directamente de la estructura de las frases en la misma medida que lo necesitan otras áreas, por ejemplo en reconocimiento del habla. Los modelos basados en unigramas son a menudo suficientes para extraer o deducir el tema de un texto. Por otra parte, los modelos empleados en recuperación de información son frecuentemente estimados a partir de un documento único, por lo que es dudoso que haya suficientes datos de entrenamiento para conseguir un modelo más informativo. Además, los modelos basados en unigramas son más eficientes a la hora de calcularse y ser aplicados que los modelos más complejos. No obstante, debido a la importancia de la frase y que cada véz más las consultas requieren técnicas de proximidad de términos en la recuperación de información, en general sugiere que el trabajo futuro debería hacer uso de modelos más sofisticados.

Modelo de probabilidad para la generación de consultas

El modelado del lenguaje se ha aplicado en recuperación de información de múltiples formas. Los modelos de lenguaje comenzaron a usarse originalmente mediante el modelo de probabilidad para la generación de consultas. En dicho prototipo, se construye un modelo de lenguaje M_d por cada documento d con el objetivo de ordenar documentos empleando la probabilidad de que un documento sea relevante a la consulta: $P(d|q)$. A partir de dicha fórmula, aplicando la regla de Bayes se obtiene:

$$P(d|q) = \frac{P(q|d)\,P(d)}{P(q)} \tag{2.9}$$

Donde $P(q)$ es la misma para todos los documentos, y por tanto puede ser ignorada. $P(d)$ es a menudo tratada como uniforme para todo d y por tanto también puede ser ignorada, aunque podría ser implementada de manera diferente incluyendo algún criterio acerca de la autoridad, longitud, novedad y el número de visitas que ha obtenido previamente dicho documento. Después de estas modificaciones, la probabilidad queda simplificada a $P(q|d)$, siendo la probabilidad de que la consulta q generada a partir del modelo de lenguaje sea derivada a partir de d. De esta forma, el método a partir del cual se modela el lenguaje tiene el objetivo de modelar el proceso de generación de la consulta. En este proceso los documentos son ordenados mediante la probabilidad de que una consulta sea observada como una muestra aleatoria del modelo de lenguaje correspondiente. La forma más común para hacer esto es usar el modelo de lenguaje multinomial basado en unigramas, que es equivalente a un modelo multinomial de Naive Bayes donde los documentos

son las clases, cada una estimada como un lenguaje "individual". Este modelo se representa mediante la siguiente ecuación:

$$P(q|M_d) \;=\; K_d \prod_{t \in V} P(t|M_d)^{tf_{t,d}} \tag{2.10}$$

Donde $K_d \;=\; L_d! \,/\, (tf_{t_1,d}!\, tf_{t_2,d}! \cdots tf_{t_M,d}!)$ es el coeficiente multinomial para la consulta q, que a partir de ahora será ignorada debido a que es constante para una consulta en concreto.

De cara a la recuperación basada en un modelo de lenguaje, la generación de consultas es tratada como un proceso aleatorio que consiste en tres etapas claramente diferenciadas:

1. Inferir un modelo de lenguaje por cada documento.
2. Estimar la probabilidad de generar la consulta en función de cada uno de estos modelos $P(q|M_{d_i})$.
3. Ordenar los documentos en base a estas probabilidades.

La intuición de este modelo es que el usuario tiene un prototipo de documento en la mente y genera una consulta basada en los términos que aparecen en dicho documento. Ocurre con frecuencia el escenario en que los usuarios tienen una idea razonable de los términos que poseen una alta probabilidad de aparecer en los documentos que le resultan interesantes, y por tanto eligen los términos que forman esa consulta de tal forma que sean capaces de discriminar los documentos que son relevantes de los que no en la colección.

Ventajas frente a otros métodos aplicados en recuperación de información

Los modelos de lenguaje proporcionan una forma novedosa de tratar el problema de la recuperación de información, que además está vinculada a los recientes trabajos que han aparecido en procesamiento del lenguaje y del habla. En general, este método ofrece una metodología diferente a la hora del ordenamiento basado en la relevancia de los documentos a una consulta dada. Además la comunidad científica tiene la esperanza de que la generación del modelado probabilístico mejore las funciones de pesado que se utilizan en la actualidad, y por tanto el rendimiento del modelo aumente sus prestaciones. La cuestión principal es la estimación del modelo del documento, así como la elección del suavizado más efectivo. El método ha alcanzado muy buenos resultados de recuperación. Si se compara con otros métodos probabilísticos, la principal diferencia que se puede apreciar es que el enfoque de los modelos de lenguaje elimina explícitamente el modelado de la relevancia, que además es la variable principal en el enfoque de evaluación de los principales modelos probabilísticos.

Por otro lado, como la mayoría de los modelos de recuperación, es posible argumentar ventajas e inconvenientes. En primer lugar, el supuesto de equivalencia entre el documento y la representación de necesidad de información no es realista. En la actualidad los métodos usados construyen modelos de lenguaje muy simples, generalmente basados en unigramas. Además, sin una noción explícita de relevancia, la retroalimentación es difícil de integrar en el modelo como preferencias de usuario. También parece necesario ir más allá de un modelo de unigramas para dar cabida a las nociones de frase, pasaje e incluso a los operadores booleanos.

El modelo tiene múltiples relaciones con los tradicionales modelos *Tf-Idf*. La frecuencia de términos es directamente representada en *Tf-Idf*, y además varios trabajos recientes han reconocido la importancia de la normalización de la longitud del documento. El efecto de hacer una mezcla de probabilidad de generación de documentos con la probabilidad de generación de la colección es una técnica similar a lo que se consigue con *idf*. Además, los términos poco usuales en la colección, pero frecuentes en algunos documentos tendrán una mayor influencia en el ranking otorgado a los documentos. Si atendemos a otros aspectos como el comportamiento del sistema, existen trabajos recientes que muestran como los modelos de lenguaje son muy efectivos en experimentos de recuperación, logrando mejoras frente a otros modelos como *Tf-Idf* y *BM25*[MRS08].

Además de la directa aplicación en problemas de recuperación de documentos, los modelos de lenguaje han sido extendidos para su aplicación a otras funciones diferentes. En lugar de atender a la probabilidad de que el modelo de lenguaje de un documento M_d genere una consulta, es posible fijarse en la probabilidad de que el modelo de lenguaje de una consulta M_q genere un documento.

La razón principal de hacer estos cambios de dirección y crear un modelo de probabilidad del documento es que hay mucho menos texto disponible para estimar un modelo de lenguaje basado en el texto de la consulta, por lo que el modelo estima que será peor, y dependerá en mayor medida del suavizado que se le aplique mediante un modelo de lenguaje de otra fuente. Por otro lado, es sencillo ver como se podría incorporar relevancia por retroalimentación en tal modelo. En general, se podría expandir la consulta con términos tomados de los documentos relevantes de la manera que venimos describiendo en este capítulo y por tanto actualizar el modelo de lenguaje de la consulta M_q. En lugar de generar el modelo directamente en cualquier dirección, podemos construir un modelo de lenguaje tanto del documento como de la consulta, y después estudiar las diferencias entre ellos. Lafferty y Zhai[LZ01] diseñaron estas tres formas de pensar sobre el problema, que se muestra en la Figura 2.9, y desarrollaron un método de minimización de riesgos generales para la recuperación de documentos.

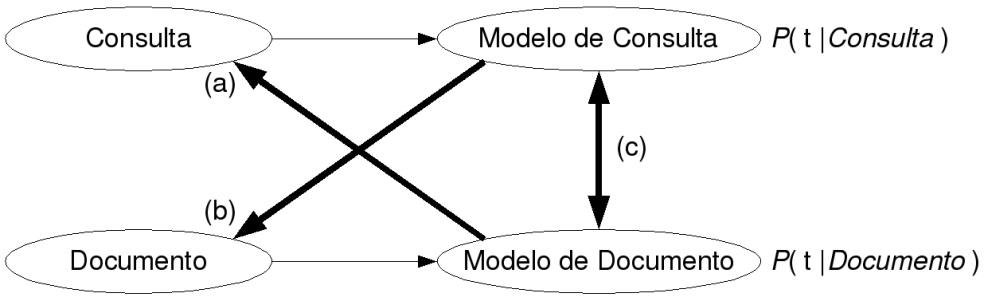

Figura 2.9: *Diferentes alternativas para la aplicación de los modelos de lenguaje.(a) Probabilidad de la consulta. (b) Probabilidad del documento. (c) Comparación de los modelos.*

Por ejemplo, una forma de modelar el riesgo de devolver un documento d como relevante a una consulta c es usar la divergencia de *Kullback-Leibler* (KL) entre los respectivos modelos de lenguaje:

$$R(d;q) = KL(M_d||M_q) = \sum_{t \in V} P(t|M_q)\ log\ \frac{P(t|M_q)}{P(t|M_d)} \tag{2.11}$$

La divergencia de *Kullback-Leibler* es una medida de divergencia asimétrica originaria de la teoría de información que mide cómo de mala es la distribución de probabilidad M_q al modelar M_d. Lafferty y Zhai[LZ01] presentaron una serie de resultados que sugieren que un enfoque de comparación de modelos supera tanto a los enfoques de probabilidad de consultas como a los de probabilidad de documentos. Una desventaja de usar la divergencia KL como una función de ranking es que otorga valores de ranking que no son comparables entre diferentes consultas. Esto no tiene importancia para la recuperación ad-hoc, pero es importante en otras aplicaciones como el seguimiento de temas.

Los modelos de lenguaje básicos no se ocupan de cuestiones de expresión alternativa, como por ejemplo la sinonimia o cualquier desviación en el uso del lenguaje entre consultas y documentos. Berger y Lafferty[BL99] introdujeron modelos de traducción para subsanar los defectos entre las consultas y los documentos. Un modelo de traducción permite traducir las palabras de la consulta mediante la traducción de términos con un significado similar alternativo. Esto también establece la base para el funcionamiento en la recuperación de información en diferentes idiomas. Se supone que el modelo de traducción puede ser representado por una distribución de probabilidad condicional $T(\cdot|\cdot)$ entre los términos del vocabulario. La forma del modelo de generación de traducción de consultas es la siguiente:

$$P(q|M_d) = \prod_{t \in q} \sum_{v \in V} P(v|M_d)\ T(t|v) \tag{2.12}$$

donde el término $P(v|M_d)$ es el modelo de lenguaje básico del documento, y el término $T(t|v)$ realiza la traducción.

Este modelo es claramente más costoso computacionalmente y es necesario construir un modelo de traducción. En general, la construcción de métodos basados en modelos de lenguaje es en estos momentos un área de investigación con mucha actividad. Se ha demostrado que los modelos de traducción, de relevancia con retroalimentación y métodos de comparación mejoran el rendimiento con el uso de modelos de lenguaje basados en la probabilidad de la consulta.

2.3.5. Expansión de Consultas

En la mayoría de las colecciones, el mismo concepto puede ser definido mediante el uso de diferentes palabras. Este fenómeno, conocido como sinonimia, tiene un gran impacto en el funcionamiento de la mayoría de los sistemas de recuperación de información. Por ejemplo, imaginemos que queremos hacer una consulta acerca de *aviones*, pero solo nos interesan los aviones con hélices, no aquellos con reactores. También imaginemos que queremos buscar por *termodinámica* pero tan solo queremos obtener resultados acerca de discusiones acerca del calor. En general, este tipo de problemas son resueltos por el usuario mediante varias consultas en las cuales se consigue un refinamiento progresivo de los resultados. Pues bien, en esta sección veremos cómo un sistema puede ayudar al usuario en este proceso de refinamiento. En concreto, en esta tesis la expansión de consultas será empleada para obtener una mayor precisión en los resultados obtenidos, utilizando diferentes fuentes de información para llevar a cabo esta tarea.

Los métodos que se encargan de resolver este problema se pueden dividir en dos grandes grupos: métodos locales y métodos globales.

Los métodos locales construyen una consulta relativa a los documentos que inicialmente parecen ser relevantes a la consulta original. La principal técnica empleada en estos métodos es la retroalimentación por relevancia, aunque en la mayoría de los casos es necesaria la intervención del usuario final.

Los métodos globales son técnicas para la expansión o reformulación de los términos de una consulta con independencia de la consulta y de sus resultados derivados. La principal técnica empleada en estos métodos es la expansión de consultas. En general estos métodos proponen el cambio de alguna/as palabras de la consulta con el objetivo de componer una nueva consulta que incida en algunos términos semánticamente similares.

A partir de ahora por su directa relación con esta tesis, analizaremos los métodos globales y en concreto la expansión de consultas.

La presencia de varios usuarios colaborando en el proceso de búsqueda de una consulta podría ayudar a un usuario en particular a ver cómo sus búsquedas están funcionando. Esta ayuda podría incluir información acerca de las palabras que el motor de búsqueda ha estado ignorando debido a su presencia en listas de palabras vacías, las palabras que fueron reducidas a una raíz mediante técnicas de *stemming*,

el número de resultados obtenidos por cada término o frase e incluso si las palabras fueron convertidas en frases de forma dinámica. Un sistema de recuperación de información también podría ayudar en gran medida al usuario sugiriendo los términos de búsqueda a través de un tesauro o un vocabulario controlado. Otra alternativa sería que el usuario pudiera navegar en una lista de términos presentes en el índice del buscador, con el objetivo de encontrar los términos que se podrían adaptar mejor a sus necesidades de información.

Una de las principales características de la expansión de consultas es que los usuarios ofrecen determinada información al sistema en forma de palabras o frases mediante la sugerencia o elección de determinados términos. Algunos motores de búsqueda, sugieren al usuario consultas relacionadas en respuesta a una consulta inicial. En ese momento, el usuario es libre de elegir alguna de estas alternativas. La Figura 2.10 muestra un ejemplo de expansión de consulta generada por Google tras la búsqueda del término *termodinámica*. La cuestión principal de esta técnica es cómo se consiguen generar sugerencias útiles para el usuario.

Figura 2.10: *Ejemplo de expansión de consulta propuesta por Google en respuesta a la búsqueda del término termodinamica.*

La forma más común de expansión es mediante un análisis global y usando algún tesauro o colección. Por cada término t en una consulta, dicha consulta puede ser automáticamente expandida con sinónimos y palabras relacionadas con t presentes en el tesauro. El uso de un tesauro puede ser combinado a su vez con ideas

de pesado de términos, por ejemplo asignando unos pesos menores a los términos añadidos que a los términos originales. Existen diversos métodos implicados en la construcción de un tesauro para la expansión de consultas y para cada uno de ellos se deben tener en cuenta algunas cuestiones:

- Los vocabularios controlados deben ser mantenidos por personas. En este caso, hay un término canónico por cada concepto.
- En el caso de un tesauro manual, los editores suelen crear conjuntos de sinónimos para conceptos, sin designar un término canónico.
- Los tesauros derivados automáticamente, tienen en cuenta información acerca de la coocurrencia de términos sobre una colección de documentos de un dominio, usada para generar automáticamente dicho tesauro.
- En relación a la reformulación de consultas basadas en el análisis de logs se suelen explotar las estadísticas de determinadas sugerencias hechas a otros usuarios para realizar dicha expansión. Esto requiere el uso de ficheros de log de un gran tamaño.

Los métodos de expansión basados en la reformulación de consultas tienen la ventaja de no necesitar información interactiva del usuario. En términos generales el uso de expansión de consultas incrementa la cobertura y es ampliamente usado en muchos campos de la ciencia y la ingeniería.

Generación automática de Tesauros

La principal alternativa a la costosa generación manual de tesauros es la generación automática mediante el análisis de una colección de documentos. Existen dos métodos para tal tarea: Por un lado se encuentra la explotación de la coocurrencia de términos, de forma que tal coocurrencia en un documento o párrafo da a los términos una alta probabilidad de ser en algún sentido similares o tener un significado con cierta relación. El otro método consiste en usar un análisis gramatical simple del texto y explotar las relaciones o dependencias gramaticales. Por ejemplo, se podría decir que determinadas entidades como cultivado, cocinado, ingerido y digerido estarán probablemente relacionados con la comida. En términos comparativos, el uso de coocurrencia de términos es más robusto (ante errores del parser) pero las relaciones gramaticales proporcionan una mayor precisión. La manera más simple de calcular un tesauro basado en coocurrencias es tener en cuenta las similaridades de términos. Para empezar, se tendría una matriz de términos y documentos A donde cada posición $A_{t,d}$ sería un valor que tendría en cuenta el peso $w_{t,d}$. Después se calcularía $C = AA^T$, donde $C_{u,v}$ sería un valor de similitud entre los términos u y v, cuyo valor máximo correspondería al grado más alto de similitud.

Palabra	**Términos relacionados**
maquillaje	repelente, loción, brillante, protector solar, piel, gel
absolutamente	absurdo, absoluto, total, exactamente, nada
seductor	brillo, sorprendentemente, magnificamente, valiente, ingenioso
caseta	perro, porche, rastreo, junto, abajo
patógenos	toxinas, bacterias, microorganismos, bacterias, parásitos
litografías	dibujos, Picaso, Dalí, esculturas, Gauguin

Tabla 2.4: *Tesauro generado automáticamente mediante el uso de latent semantic indexing (LSI).*

En resumen, la expansión de consultas es a menudo efectiva en términos de aumento de cobertura. Sin embargo, la expansión de consultas reduce en ocasiones la precisión, particularmente cuando las consultas contienen términos ambiguos.

2.3.6. Evaluación de Sistemas de Recuperación de Información

Los Sistemas de Recuperación de Información[MM04] (SRI), resultan susceptibles, como cualquier otro sistema, de ser sometidos a evaluación, para que sus usuarios puedan valorar su efectividad. La tradición de la evaluación es tan antigua casi como el desarrollo de los primeros SRI, encontrándose estrechamente vinculadas con la investigación y el desarrollo de la recuperación de información.

En esta tesis se propone una nueva metodología de evaluación para los sistemas de recuperación de enlaces rotos. Por este motivo, a continuación se presentan las principales características que se deben tener en cuenta para realizar dicha labor.

La evaluación[SM86, MRS08] de los SRI resulta tener una gran importancia para garantizar su adecuado funcionamiento, es decir, la recuperación de información pertinente y una correcta adaptación a las necesidades de los usuarios: facilidad de uso, respuestas rápidas, coste razonable, etc. Es por ello que el tipo de evaluación elegida depende de los objetivos del SRI. De hecho, a la hora de realizar la evaluación existen principalmente dos criterios que pueden ser medidos: eficiencia y eficacia.

Baeza-Yates y Ribeiro-Neto[BYRN99] manifestaron que "un SRI puede ser evaluado por diversos criterios, incluyendo entre los mismos: la eficacia en la ejecución, el efectivo almacenamiento de los datos, la efectividad en la recuperación de la información y la serie de características que ofrece el sistema al usuario". Estos criterios no deben confundirse, la eficacia en la ejecución es la medida del tiempo para realizar una operación, la eficiencia del almacenamiento es el espacio que se precisa para almacenar los datos y por último está la efectividad de la recuperación "normalmente basada en la relevancia de los documentos recuperados".

Pors[Por00] diferenció entre evaluar el acceso físico y el acceso lógico a los datos, considerando que las evaluaciones han de ser del segundo tipo. El acceso físico es el que concierne a cómo la información es recuperada y representada de forma física al usuario, y está muy vinculado con las técnicas de recuperación y de presentación de la información. El acceso lógico está relacionado con la localización de la información deseada. Pors también distinguió entre "aproximaciones al funcionamiento del sistema y aproximaciones centradas en el usuario", plenamente coincidentes con acceso físico y acceso lógico. De forma parecida, Baeza-Yates[BYRN99] afirmó que existen dos tipos de evaluaciones: la del funcionamiento del sistema y la del funcionamiento de la recuperación, siendo la segunda modalidad la que analiza cómo los documentos recuperados se clasifican de acuerdo a su relevancia con la pregunta efectuada. De esta forma, los SRI son generalmente evaluados en base a una colección de documentos que debería cumplir los siguientes requisitos:

- Colección amplia de documentos que refleje una muestra de la realidad.
- Conjunto de consultas expresando unas necesidades de información.

- Juicios de relevancia sobre los documentos de la colección.

De este modo, el enfoque tradicional para evaluar a los SRI trabaja sobre la noción de documentos relevantes y documentos no relevantes. En términos generales se puede decir que una medida de evaluación cuantifica la similitud entre el conjunto de documentos devueltos por un SRI para un determinado conjunto de consultas y el conjunto de documentos relevantes para dichas consultas.

A grandes rasgos, las medidas de evaluación de los SRI pueden dividirse en dos grupos. Por un lado se encuentran las medidas que tienen en cuenta el orden de las respuestas, asignando mayor relevancia a las respuestas colocadas al principio. Por otro lado se encuentran las medidas que no atienden al orden asignado a las respuestas focalizando su evaluación en la precisión de las respuestas.

Evaluación de conjuntos no ordenados

La capacidad de un sistema de recuperación y organización de la información para proveer al usuario de documentos relevantes se mide con una métrica llamada *Cobertura*, que se define como la proporción de material relevante recuperado, del total de los documentos que son relevantes en la colección de documentos, independientemente de que éstos, se recuperen o no. A continuación se muestra la fórmula de la *Cobertura*.

$$Cobertura = \frac{\#\ documentos\ relevantes\ recuperados}{\#\ documentos\ relevantes\ en\ la\ coleccion}$$

Por otro lado, cualquier sistema de recuperación de información podría conseguir una *Cobertura* del 100 % simplemente devolviendo todos los elementos de la colección. Por ello se utiliza también otra métrica, llamada *Precisión*, que se define como la proporción de material recuperado realmente relevante, del total de los documentos recuperados y cuya fórmula se encuentra a continuación:

$$Precision = \frac{\#\ documentos\ relevantes\ recuperados}{\#\ documentos\ recuperados}$$

En la Figura 2.11, se puede observar de manera gráfica un ejemplo de representación de estas dos métricas.

Otra forma de definir las medidas de Precisión y Cobertura es mediante una tabla de contingencia como la que aparece en la Tabla 2.5, en la que la Precisión y la Cobertura se expresan de la siguiente manera:

$$Precision = \frac{|A \cap B|}{|B|} \qquad Cobertura = \frac{|A \cap B|}{|A|}$$

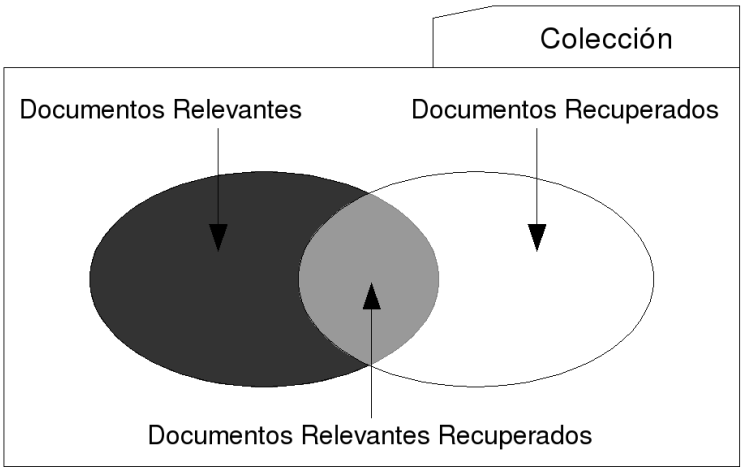

Figura 2.11: *Representación gráfica de precisión y cobertura.*

	RELEVANTES (A)	**NO RELEVANTES** ($\neg A$)
RECUPERADOS (B)	$A \cap B$	$\neg A \cap B$
NO RECUPERADOS ($\neg B$)	$A \cap \neg B$	$\neg A \cap \neg B$

Tabla 2.5: *Tabla de contingencia en el ámbito de la evaluación de recuperación de información.*

Con el fin de facilitar la comparación entre sistemas es deseable poder contar con una medida que combine Precisión y Cobertura. La medida de combinación más utilizada es la media armónica ponderada de ambos valores, conocida como medida F o F_1 debido a que Precisión y Cobertura tienen asignado el mismo peso:

$$F = \frac{2 * Cobertura * Precision}{Cobertura + Precision}$$

En esta medida se puede aplicar un factor β, que consiste en un valor positivo usado para dar más importancia a la Cobertura o a la Precisión. Si $\beta < 1$ se da más importancia a la Precisión mientras que si $\beta > 1$ cobra más importancia la Cobertura como puede apreciarse en la siguiente fórmula:

$$F_\beta = \frac{(\beta^2 + 1) * Cobertura * Precision}{\beta^2 * Cobertura + Precision}$$

Evaluación de conjuntos ordenados

Muchos SRI no hacen afirmaciones explícitas sobre la relevancia o no de un documento, sino que ordenan la colección de mayor a menor relevancia respecto a una consulta. Para un mismo sistema y una misma consulta, Cobertura y Precisión son inversamente proporcionales. En efecto, si se recuperan los n documentos de mayor relevancia, tendremos una alta Precisión y baja Cobertura para valores pequeños de n, así como una baja Precisión y alta Cobertura para valores grandes de n. Para reflejar este hecho se utilizan frecuentemente las curvas Precisión-Cobertura, que muestran el valor de Precisión para distintos niveles de exhaustividad o Cobertura: En la Figura 2.12 se muestra un ejemplo de curva Precisión-Cobertura, en la que además podría intuirse el cálculo del área que encierra la curva con el objetivo de proporcionar un valor único y comparable con los resultados de otros sistemas.

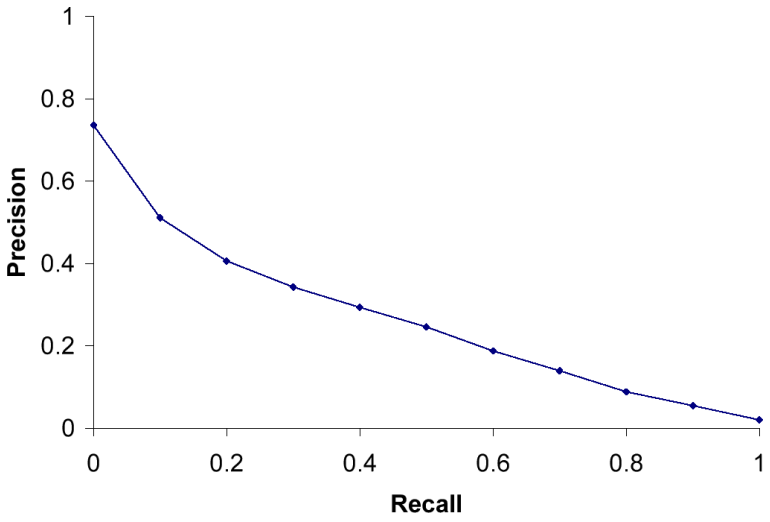

Figura 2.12: *Curva de Precisión y Cobertura.*

En general sería complicado comparar los resultados de dos evaluaciones diferentes tan solo observando la forma de las curvas. Por este motivo existen una serie de medidas de las que se obtiene tan solo un valor numérico y de esta forma es más sencillo comparar el rendimiento de dos SRI. A continuación se muestra un resumen con las principales medidas empleadas en la comparación de SRI:

- **Mean Average Precision (MAP)** es un medida que nos da una idea global del sistema a través de un conjunto de consultas. Para ello se calcula el promedio de las precisiones promedio para ese conjunto de consultas. Esta medida además es ampliamente usada en los foros de evaluación de SRI como *TREC*. Para una determinada consulta se calcula la media de los valores de Precisión obtenidos cada vez que se encuentra un documento relevante. Para el conjun-

to de consultas el valor final es la media de los valores calculados para cada consulta.

$$MAP(Q) = \frac{1}{Q}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk})$$

- **Precision at K** es otra posibilidad, similar a la anterior, de cara a comparar dos conjuntos de resultados ordenados. Consiste en mostrar la precisión obtenida no a determinados valores de Cobertura, sino a un número determinado (K) de documentos devueltos.

- **R-Precision** es la Precisión obtenida a los R documentos devueltos doc_{ret}, donde R es el número de documentos relevantes para esa consulta. Para su uso es necesario tener un conjunto conocido de documentos relevantes R, calculándose la Precisión para los primeros $|R|$ documentos recuperados. Siendo n el número de documentos relevantes para la consulta q_i y $Q = [q_1, q_2, ..., q_m]$ un conjunto de m consultas.

$$R - Precision = \frac{1}{m}\sum\frac{doc_{ret}}{n}$$

- **Mean Reciprocal Rank (MRR)**, es una medida estadística que calcula la media del valor inverso de la posición de la primera respuesta correcta, por cada consulta realizada en una colección. Siendo $rank_i$ el primer documento relevante recuperado para la consulta $q \in Q$.

$$MRR(Q) = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i}$$

2.4. Sistema de Recuperación de Enlaces Rotos

Uno de los objetivos de esta tesis es el desarrollo de un sistema de recuperación de enlaces rotos, y para tal fin, además de tener en cuenta las características generales de un sistema de recuperación de información, es imprescindible la investigación y análisis de determinados aspectos específicos. Además de ciertos métodos que conformarán la base de nuestro trabajo y que fueron descritos en la sección anterior, como el modelo vectorial, los modelos de lenguaje o la expansión de consultas, existen otras técnicas que van a proporcionar las claves del sistema. Por un lado analizaremos algunos trabajos que han demostrado cómo los motores de búsqueda son una herramienta muy útil a la hora de encontrar recursos desaparecidos en la Web. Por otro lado, introduciremos las principales fuentes de información que emplearemos para desarrollar nuestro sistema y cómo es posible discriminar determinada información cuyo valor es superior al del resto del texto.

2.4.1. Redescubrimiento de Recursos Web

A pesar de las limitaciones por el modelo de "bolsa de palabras" que utilizan y de las críticas que reciben en muchos casos, los motores de búsqueda son la principal herramienta para encontrar recursos en la Web. No obstante y aún teniendo en cuenta su indudable utilidad, la manera en la que se realizan las consultas influyen notablemente en el éxito de los resultados obtenidos.

En el año 2000 Jansen et al.[JBJ+00] llevaron a cabo un amplio estudio acerca de las consultas de los usuarios en la Web. En dicho trabajo fueron analizados los ficheros de log del motor de búsqueda *Excite*, encontrando que el 35 % de las consultas fueron *únicas* (primera consulta de un usuario distinto), el 22 % fueron *modificaciones* (consultas modificadas añadiendo o eliminando palabras de una consulta única) y el 43 % fueron *idénticas* (misma consulta que la anterior realizada por el mismo usuario). La gran cantidad de consultas *idénticas* es debido principalmente a dos razones:

- El usuario intenta reencontrar algún resultado obtenido previamente.
- El motor de búsqueda establece como una consulta idéntica los casos en los que el usuario visita las siguientes páginas de resultados.

Otro de los datos significativos que se presentan es que dos tercios de los usuarios realizan tan solo una consulta y la mayoría de usuarios que llevan a cabo más consultas no varían mucho su consulta inicial, sino que lo más frecuente son pequeñas modificaciones tales como añadir o eliminar términos con la misma frecuencia. El número medio de páginas visitadas por cada usuario se sitúa en una cifra de 2.35 mientras que más de la mitad de los usuarios no accedieron a la segunda página de resultados. A su vez, la longitud media de las consultas fue 2.21 términos y los operadores booleanos no fueron usados con demasiada frecuencia (el operados AND solo apareció en el 8 % de las consultas). También fue analizada la distribución de términos de las consultas, encontrando que unos pocos términos eran utilizados con mucha frecuencia y las consultas largas apenas eran empleadas por los usuarios. En relación a la distribución de los términos, se ajusta a una distribución *Zipf*.

En una dirección similar en cuanto al redescubrimiento de recursos en la Web, se encuentran los trabajos de Teevan et al.[TAJP06, TAJP07] que analizan los logs de un buscador con el objetivo de extraer patrones en las búsquedas de los usuarios en un periodo aproximado de 30 minutos. En dicho trabajo analizaron los ficheros de log de Yahoo! durante el periodo de un año, encontrando que las búsquedas repetidas son un suceso muy común. En concreto el 40 % de las consultas tienden a lograr el clic en el mismo resultado por parte del mismo usuario, pero tan solo el 7 % fueron clics en la misma dirección por parte de usuarios diferentes. Este hecho indica que los usuarios tienden con una mayor probabilidad a visitar las direcciones que han visitado previamente. Teevan también propuso un sistema para predecir la

probabilidad de que los usuarios visiten un mismo recurso que tuviera un indudable valor para los motores de búsqueda. De la misma manera fue implementado un complemento para los navegadores con una interfaz que realizaba consultas en los principales motores de búsqueda y que capturaba las consultas del usuario que eran similares a una consulta anterior. Este plug-in recuperaba los resultados de la consulta y los resultados visitados previamente, ofreciendo la posibilidad al usuario de re-encontrar recursos web que ya había visitado en el pasado.

2.4.2. Infraestructura Web

En este trabajo proponemos el uso de toda la información disponible acerca de una página desaparecida, con el objetivo de recuperarla o encontrar otra página con un contenido suficientemente similar como para sustituirla. Esta información está dividida en dos grupos; por un lado la información que se encuentra en la misma página que posee el enlace roto; y por otro lado la información disponible en la infraestructura web.

Infraestructura Web se puede entender como todos los recursos web que están disponibles y que en el ámbito de la recuperación de enlaces podrían ser utilizados para tal fin. En concreto podrían formar parte de esta infraestructura web los actuales motores de búsqueda (Google, Bing, Yahoo!, etc.), las librerías digitales o archivos web (Internet Archive[8], arXive[9], etc.), las redes sociales y sistemas de social tagging (Delicious[10], Webgenio[11], CiteULike[12], Diigo[13],etc.) y diferentes proyectos de investigación (CiteSeer[14], NSDL[15], etc.)

En la Figura 2.13, pueden apreciarse cuatro ejemplos de como la infraestructura web posee información relevante acerca de una determinada URL. En esta figura se han realizado varias consultas sobre la dirección `http://nlp.uned.es`, obteniendo los siguientes recursos: *Google* ofrece la versión en caché como parte de los resultados, pudiendo ser consultada en caso de que la página real esté inactiva. *Wayback Machine* muestra como tiene almacenadas 188 versiones de esta URL desde 2003 hasta 2009. En *Delicious* se pueden observar las etiquetas que los usuarios han asignado a la página (research, academic y university). *Bing* también contiene una versión almacenada de la URL requerida.

Existen varios trabajos que han investigado la utilidad de la infraestructura web para la resolución del problema de los enlaces rotos. Entre estos trabajos destaca el de McCown et al.[MMN09], que encontraron varias razonas por las cuales algunos

[8] http://www.archive.org/
[9] http://arxiv.org/
[10] http://www.delicious.com
[11] http://www.webgenio.com
[12] http://www.citeulike.org
[13] http://www.diigo.com
[14] http://citeseer.ist.psu.edu/
[15] http://nsdl.org/

sitios web desaparecen por completo y como pueden ser potencialmente recuperados. Para esta labor abogan por el uso de estos recursos web, con el objetivo de adquirir la mayor información disponible. A su vez MacCown desarrolló un sistema denominado *Warrick*[MSN06] que realiza un crawling sobre repositorios web tales como las cachés de los motores de búsqueda y el *Internet Archive*. El sistema a su vez, fusiona los resultados y reconstruye el sitio por completo. A pesar de la evidente utilidad de este sistema, el alcance es limitado a pequeños sitios web y además su funcionamiento depende de la presencia de una versión en caché del sitio web requerido, en alguno de los recursos de la infraestructura web. MacCown et al.[MDN07] también investigaron los factores y su impacto en la reconstrucción web basándose en la infraestructura web. En este trabajo se concluye que el valor de *PageRank*[MRS08] ofrecido por *Google*, así como el tiempo desde la desaparición del sitio, afectan al éxito de la tarea. La profundidad del recurso deseado en relación a la raíz del sitio web también influye en su reconstrucción, debido a que muchos *crawlers* tienen limitada la profundidad de sus visitas.

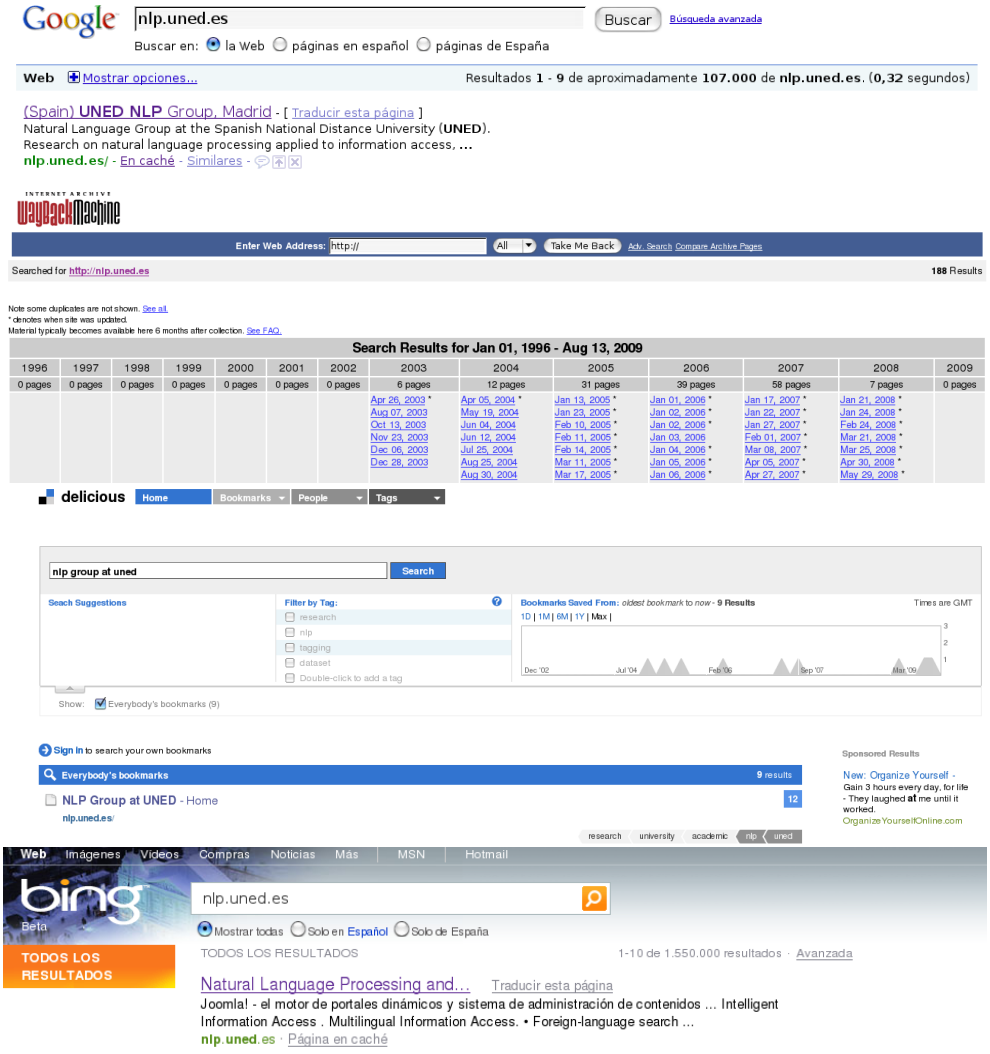

Figura 2.13: *Recursos disponibles en la infraestructura web acerca de una URL: Google ofrece la versión en caché como parte de los resultados; Wayback Machine muestra 188 versiones; En Delicious se pueden observar las etiquetas que se han asignado al sitio; Bing también contiene una versión almacenada de la URL requerida.*

2.4.3. Extracción Automática de Terminología

En la sección anterior expresábamos nuestro objetivo de extraer toda la información disponible acerca de una página desaparecida, con el objetivo de recuperarla o encontrar otra página con un contenido suficientemente similar como para sustituirla. Entre las fuentes de información que proponemos en este trabajo existen fuentes con un tamaño reducido, como puede ser el texto del ancla de un enlace, pero también existen fuentes con un contenido mayor como puede ser el párrafo donde se encuentra el enlace roto. Estos dos ejemplos de fuentes de información pueden tener la misma utilidad siempre y cuando su uso sea el correcto. Dada la cantidad de términos que puede encontrarse en un párrafo o en una página completa, el extraer la información más relevante en unos pocos términos puede convertir un fragmento de texto inmanejable en una fuente de información muy valiosa.

La extracción automática de terminología según[OMC07] *"es el proceso mediante el cual se seleccionan de un texto o conjunto de textos unidades candidatas a constituir términos"*. Esta tarea se usa principalmente para la construcción de tesauros. Un tesauro es una herramienta para control de vocabulario, generalmente diseñado para indexar y buscar información en un área de conocimiento específica. Estos tesauros son muy útiles a los sistemas de recuperación para mejorar la calidad de la información recuperada. De esta forma, los propósitos primarios de un tesauro son la promoción de la consistencia en la indexación de documentos y la ayuda para facilitar las búsquedas.

Existen dos técnicas principales para la extracción automática de terminología:

- La técnica lingüística que se basa en detectar patrones de categorías morfológicas.
- La técnica estadística que se basa en la frecuencia de aparición de una serie de combinaciones de palabras.

Técnica lingüística

Las técnicas lingüísticas de extracción de terminología se basan en la detección de patrones morfológicos. El paso previo, por lo tanto, a la extracción de terminología es el etiquetado morfosintáctico del texto. El etiquetado de textos consiste en añadir información morfológica a cada palabra del texto. Para poder realizar este etiquetado es necesario disponer de un etiquetador o tagger. Una vez etiquetado el texto, la extracción de terminología consistirá básicamente en una búsqueda de patrones que sean típicamente terminológicos.

Consideremos por ejemplo la frase siguiente: *Nuestro sistema de gestión empresarial incluye un programa de facturación y una base de datos de recursos humanos.*

En el caso de utilizar una técnica lingüística se realizaría el siguiente análisis morfosintáctico:
Nuestro {*nuestro* DP1MSP} **sistema** {*sistema* NCMS000} **de** {*de* SPS00} **gestión** {*gestión* NCFS000} **empresarial** {*empresarial* AQ0CS0} **incluye** {*incluir* VMIP3S0} **un** {*uno* DI0MS0} **programa** {*programa* NCMS000} **de** {*de* SPS00} **facturación** {*facturación* NCFS000} **y** {*y* CC} **una** {*uno* DI0FS0} **base** {*base* NCFS000} **de** {*de* SPS00} **datos** {*dato* NCMP000} **de** {*de* SPS00} **recursos** {*recurso* NCMP000} **humanos** {*humano* AQ0MP0}.

Una vez etiquetado el texto, la extracción de terminología consistiría básicamente en una búsqueda de patrones terminológicos. Algunos patrones posibles para el castellano podrían ser:

- NC AQ: substantivo - adjetivo. Detectaría: *gestión empresarial, recursos humanos*.
- NC SP NC: substantivo - preposición - substantivo. Detectaría: *sistema de gestión, programa de facturación, base de datos*.

Debido a las enormes cantidades de páginas web que se manejan en aplicaciones como las que se proponen en esta tesis, este procesamiento lingüístico es dificil de aplicar. Por tanto en la mayoría de los casos se recurre a métodos estadísticos.

Técnica estadística

La información básica que utilizan los sistemas de extracción automática de terminología de tipo estadístico, es la frecuencia de aparición de una serie de combinaciones de palabras. Los sistemas de extracción de terminología estadísticos trabajan con n-gramas de palabras. Los n-gramas de palabras son combinaciones de n palabras consecutivas. Para poder seleccionar las combinaciones con una mayor probabilidad de constituir términos, se puede hacer uso de listas de palabras vacías o *stopwords*. Las palabras vacías, cuando hablamos de extracción de terminología, son aquellas (en su mayoría funcionales) que no pueden estar en ciertas posiciones de la entrada terminológica (normalmente las posiciones extremas, es decir, primera y última).

Teniendo por ejemplo la misma frase que la empleada en la técnica lingüística y utilizando unigramas mediante una técnica estadística, los términos extraídos serían los siguientes: *Nuestro, sistema, gestión, empresarial, incluye, programa, facturación, base, datos, recursos, humanos*.

En la actualidad y con el auge de la web semántica, se ha incrementado la investigación en el área de la extracción de terminología. Este hecho es debido al

crecimiento en el número de comunidades que se han creado y la expansión de la comunicación a través de Internet entre los miembros de las comunidades y entre las propias comunidades. El aumento de la información disponible ha atraído el interés en su investigación con el objetivo de modelar estas comunidades y la información que fluye a través de ellas. La extracción de terminología se divide a su vez en tres pasos:

1. Extracción de términos mediante un análisis morfológico.

2. Pesado de los términos con información estadística, midiendo la relevancia del término en el dominio.

3. Selección, ranking del término y truncado de las listas por umbrales de peso.

En lo relativo a este trabajo, las técnicas más relevantes son aquellas que han utilizado medidas estadísticas para la extracción de terminología. Entre los trabajos más representativos se encuentra el de Justeson y Katz[JK93], que profundizaron en el uso de la frecuencia de términos, mostrando que su uso era la mejor manera de seleccionar aquellos términos que identificaban un dominio. Sin embargo sus estudios también revelaron que la frecuencia funcionaba mejor si era complementada con algunas propiedades lingüísticas. También demostraron que tras analizar los grupos nominales, ciertos patrones en estos grupos nominales tenían una mayor probabilidad de contener términos apropiados que otros patrones.

Cohen[Coh95] mostró que las palabras infrecuentes pueden ser extraídas como términos, si están contenidas en n-gramas con una frecuencia baja. También introdujo otro método estadístico en el que analizaba la frecuencia de los morfemas, concluyendo que buscando estos morfemas que aparecen frecuentemente, se pueden extraer palabras menos frecuentes basadas en la frecuencia de sus morfemas en el dominio.

Sclano y Velardi[SV07] afirmaron que los términos bien distribuidos pueden ser buenos candidatos porque demuestran el reconocimiento universal como términos del dominio. La cohesión es una métrica asociada a las frases y de esta manera los autores midieron la probabilidad de que los términos dentro de un sintagma nominal aparecieran próximos a otros términos de las mismas características. También mostraron que los términos acompañados de otros términos frecuentes podían ser eliminados si la cohesión era baja.

En la Tabla 2.6, se muestra un resumen de las principales métricas estadísticas empleadas en la extracción de terminología, proporcionando además su abreviatura y las razones de su importancia.

Abreviatura	Métrica	Información
TF	Term Frequency	Premia los términos más frecuentes, grandes documentos tienen ventaja
LTF	Logged Term Frequency	Suaviza resultados, efecto similar a normalización
Tf-Idf	Term Frequency - Inverse Document Frequency	Recompensa términos frecuentes en pocos documentos
LTf-Idf	Term Frequency - Logged Inverse Document Frequency	Suaviza la distribución de frecuencia
USN	Document Term Frequency	Recompensa términos muy frecuentes en un solo documento
ED	Evenly Distributed	Todos los documentos aportan el mismo número de documentos
BD	Favor Big Documents	Premia los documentos grandes
NTF	Normalized Term Frequency	Incentiva términos muy frecuentes pero penaliza documentos grandes
DR	Document Relativized	Penaliza documentos muy grandes
CD	Corpus Relativized	Mayor importancia a la colección y menor incentivo a documentos grandes
DRDA	Document Relativized - Document Average Frequency	Frecuencia media sobre documentos
CRDA	Corpus Relativized - Document Average Frequency	Menor recompensa a documentos grandes
DC	Distribution consensus	Premia términos con la misma frecuencia en múltiples documentos
BC	Binary Consensus	Recompensa el consenso, premia la mínima frecuencia

Tabla 2.6: *Resumen de las principales métricas estadísticas empleadas en la extracción de terminología, proporcionando además su abreviatura y las razones de su importancia.*

2.4.4. Reconocimiento de Entidades Nombradas

En esta sección se introducirá el concepto de reconocimiento de entidades nombradas (NER, Named Entity Recognition), definiéndose qué es una entidad nombrada y presentando una descripción lingüística de la tipología de estas entidades. En segundo lugar, se presentarán los diferentes sistemas de reconocimiento de entidades nombradas (NER), clasificando estos sistemas según el enfoque utilizado en tres grupos: basados en reglas, basados en aprendizaje automático e híbridos.

La información que se puede encontrar en una página web acerca de una segunda página enlazada, generalmente tiene un escaso valor descriptivo. En esta tesis, se propone el uso del reconocimiento de entidades nombradas con el objetivo de hacer emerger aquella información más relevante y que puede otorgar un mayor grado de precisión.

El procesamiento del lenguaje natural (PLN) es una rama de la inteligencia artificial que estudia el tratamiento computacional de los lenguajes humanos. Se trata, por una parte, de explotar computacionalmente grandes volúmenes de datos que están representados como lenguaje natural y por otra, de hacer la comunicación menos rígida y más cómoda entre el ser humano y los computadores.

El NER consiste en la identificación y posterior clasificación de las entidades que aparecen en los textos. Una entidad es una palabra o un conjunto de palabras, que pueden ser clasificadas dentro de uno de los siguientes tipos predefinidos: persona, localidad, organización, fecha, tiempo, porcentaje y cantidad. Por ejemplo, si la oración "Juan trabaja para la UNED en Madrid" fuera procesada por un sistema NER quedaría etiquetada así: "<PER>Juan</PER>trabaja para la <ORG>UNED</ORG>en <LOC>Madrid</LOC>". La importancia de esta tarea se vió reflejada en las conferencias MUC (Message Understanding Conferences), que comenzaron a celebrarse en 1987 gracias a la iniciativa de la agencia DARPA (Defense Advance Research Projects Agency). Varios autores propusieron diferentes taxonomías para realizar la clasificación de estas entidades, pero hasta la aparición de estas conferencias no se consiguió establecer un consenso. Entre los avances que se obtuvieron gracias a estas conferencias se encuentran el establecimiento de un conjunto de categorías de entidades que han servido como base para la generación de una taxonomía:

- Entidades con nombre (ENAMEX).
    - Organizaciones: Nombres de compañías, administraciones públicas y otras entidades de tipo organizativo.
    - Personas: Nombres de persona o familia.
    - Localizaciones. Nombres de localizaciones de naturaleza política o geográfica.
- Expresiones temporales (TIMEX)

    - Fechas
    - Horas

- Expresiones numéricas (NUMEX)
    - Expresiones monetarias
    - Porcentajes

Entidades Nombradas

Una entidad nombrada es una palabra o conjunto de palabras que se refieren de forma unívoca a un objeto perteneciente a una entidad por su nombre propio, acrónimo, etc. En general, en un documento suelen encontrarse frases con nombres de persona, animales, objetos, unidades monetarias, organizaciones, localizaciones y fechas, entre otras palabras.

A partir de los diferentes tipos de entidades, por la directa relación con esta tesis y dada la información que ofrecen los principales sistemas NER en la actualidad, nos vamos a centrar en tres tipos de entidades: personas, organizaciones y lugares.

La caracterización lingüística de las entidades nombradas es difícil de realizar y abarca mucho más que encontrar los nombres propios. Muchos de los sistemas NER solamente reconocen nombres propios como entidades nombradas, lo cual es una pérdida importante de información. Por ejemplo, en el caso de la siguiente porción de texto "el presidente del Banco Santander Emilio Botín", cualquier sistema NER obtendría "Banco Santander" como el nombre de una organización y Emilio Botín como el nombre de una persona, sin embargo la entidad nombrada completa es realmente una persona. De esta forma, para delimitar este tipo de entidades se suelen diferenciar dos tipos de entidades nombradas:

- **Entidades nombradas fuertes:** Consisten básicamente en los nombres propios. Por ejemplo, en la oración "El rector de la UNED Juan A. Gimeno" estará compuesto de dos entidades fuertes: "UNED" (organización) y "Juan A. Gimeno" (persona).
- **Entidades nombradas débiles:** Consisten en un disparador (e.g. el alcalde Gallardón) y opcionalmente una entidad fuerte y un determinante (e.g. el alcalde de Madrid). Al mismo tiempo se establecen distintos tipos de entidades débiles en función de su complejidad:
    - Sintáctica
        - ◇ Simples: son aquellas formadas por un único sintagma nominal.
        - ◇ Complejas: son aquellas formadas por sintagmas nominales complejos que incluyen plurales con elipsis, anáfora, sintagmas relativos, etc.

- Semántica
  - Principales: Aquellas que tienen como núcleo una palabra disparadora perteneciente a un conjunto de referencia.
  - Relacionadas: Aquellas que tienen como núcleo un hipónimo, hiperónimo o un sinónimo de una palabra disparadora que viene a partir del conjunto de referencia.
  - Generales: Cualquier sintagma nominal.

Clasificación de sistemas NER

El problema del NER, se ha intentado abordar mediante varias aproximaciones. Una opción es la creación de un sistema que apoye su comprensión del texto en relación con un conjunto de diccionarios creados manualmente y la combinación de estos con una serie de reglas o patrones. Otra alternativa es la creación de un sistema basado en aprendizaje automático, en este caso el sistema aprendería en base a un algoritmo de aprendizaje con la necesidad de disponer de corpus etiquetados para dicha labor. Finalmente, otra posible solución sería construir un sistema NER híbrido que combinara las dos técnicas, consiguiendo combinar la rapidez y precisión de sistemas de conocimiento para dominios cerrados con la eficacia de los sistemas basados en aprendizaje para dominios mucho más generales.

Como ya se ha comentado anteriormente, nos podemos encontrar sistemas NER basados en aprendizaje automático, basados en conocimiento e incluso sistemas que combinen ambas aproximaciones. Cada uno posee sus ventajas y sus inconvenientes. Los sistemas que se apoyan en conocimiento confían su precisión al uso de recursos lingüísticos creados manualmente como son gramáticas libres de contexto, expresiones regulares, listas de disparadores o diccionarios específicos. En determinadas situaciones y siempre bajo dominios muy restringidos, resultaría más ventajosa la utilización de sistemas basados en conocimiento, no olvidando nunca las limitaciones que estos conllevan. La manera de combatir estas limitaciones sería desarrollando sistemas que aprendieran por ellos mismos la tarea NER, que no dependieran del dominio y que aprendieran a reconocer entidades en textos generales. Esto puede ser conseguido, en gran parte, mediante aprendizaje automático.

Un sistema que aprenda la detección y clasificación de entidades mediante algoritmos de aprendizaje automático y sobre textos generales, paliaría los problemas de los sistemas basados en conocimiento, logrando reducir el esfuerzo de crear recursos lingüísticos y mejorando la portabilidad.

A pesar de esto, la precisión de un sistema basado en aprendizaje automático reside en la cantidad y calidad de anotación de los corpus destinados al aprendizaje, así como en proporcionar a los algoritmos las características lingüísticas adecuadas para que realicen un reconocimiento de entidades con precisión. Así, para estos sistemas hay también un gran inconveniente que es el tiempo y el esfuerzo humano

considerable para la creación, etiquetación y corrección de los corpus de aprendizaje que usan este tipo de sistemas.

Como consecuencia de este problema y para solucionarlo, los sistemas basados en aprendizaje automático deben adaptarse para que utilicen la funcionalidad de los métodos semi-supervisados. Este tipo de sistemas usan corpus con un conjunto pequeño de datos etiquetados y la gran cantidad de datos restante no están etiquetados. Así, estos métodos lo que hacen es convertir esa gran cantidad de datos no etiquetados en datos etiquetados para que el sistema los aprenda. El conjunto de características que se han utilizado para convertir los datos no etiquetados en etiquetados no dependen de herramientas específicas del lenguaje.

Información Semántica

Las distintas aproximaciones desarrolladas para NER hasta la fecha se han enfocado en la resolución de las siete categorías de entidades fundamentales que se definieron en las tareas MUC. Sin embargo, los sistemas actuales de procesamiento del lenguaje natural necesitan de sistemas NER mucho más especializados para poder abordar con satisfacción sus tareas. Por ejemplo, para responder a la siguiente cuestión: "¿Quién fue el alcalde de Madrid en 1980?", un sistema de búsqueda de respuestas necesitaría identificar a qué categoría semántica pertenece "alcalde". Del mismo modo, para hacer preguntas a la web del tipo "ganador ACB", el mecanismo de búsqueda debería devolver los documentos relevantes conteniendo los nombres de las organizaciones con la categoría "equipo". Desafortunadamente, los sistemas NER actuales no tienen la suficiente potencia para hacer este tipo de clasificación. Es decir, se necesitan sistemas NER que sean capaces de categorizar no solo que una entidad sea persona, sino que se necesita un mayor nivel de especificidad y detectar que una persona sea o un político, escritor, deportista, etc.

Por otro lado, existen numerosos inconvenientes asociados a la categorización especializada de las entidades. El primero, es que los sistemas necesitan una gran cantidad de textos etiquetados manualmente, y actualmente estos no están disponibles. Además, la creación de los mismos sería muy costosa tanto en tiempo como en esfuerzo de personas especializadas. Esto se une al hecho de que para otras lenguas distintas del inglés, hay una carencia muy grande de disponibilidad de textos o recursos.

Como consecuencia de los inconvenientes anteriores, los sistemas de NER especializados deben enfocarse hacia el desarrollo de aproximaciones independientes del lenguaje. Además este tipo de sistemas deben enfrentarse al problema de la ambigüedad de las entidades, es decir, que diferentes personas, organizaciones o localizaciones tengan el mismo nombre. Por ejemplo, Cambridge es una ciudad del Reino Unido, pero también de los Estados Unidos. ACL se refiere a "The association of Computational linguistics", "The Association of Christian Librarians" o a "Automotive Components Limited" entre otros.

CAPÍTULO 3

Web Spam

Si en el capítulo anterior se ofrecía una visión general del problema de los enlaces rotos, en el primer capítulo de motivación introducíamos su relación directa con otro de los principales problemas que afectan a la Web y a los motores de búsqueda: el web spam. En este capítulo realizaremos un recorrido por las principales técnicas, tanto de generación como de detección de spam, describiremos las colecciones de referencia utilizadas en la investigación de este problema y finalmente analizaremos la utilidad del aprendizaje automático en este área y sus principales medidas de evaluación.

3.1. Definición del Problema

En esta sección introduciremos el concepto de web spam, analizando la clasificación de los tipos de spam existentes atendiendo a las principales técnicas empleadas. Dentro de esta clasificación haremos un repaso de los trabajos más importantes en este área en los últimos años. Finalmente realizaremos un estudio de las áreas de investigación involucradas en esta tesis, analizando algunos temas como la recuperación de información con adversario, que engloba la investigación del spam en buscadores; la clasificación de texto, que será necesaria para automatizar las tareas de detección; o la entropía relativa, que será utilizada para extraer un conjunto rasgos que caractericen a una página web.

3.1.1. Principales Técnicas de Web Spam

Hoy en día, la creciente popularidad de la Web entre los usuarios como fuente de información, ha convertido a los buscadores en un objetivo de la publicidad y el marketing. Los buscadores a su vez, basan su modelo de negocio en la publicidad que añaden a los resultados de una consulta. Pero además de esta publicidad

relevante a las consultas realizadas, una manera muy económica de conseguir publicidad, consiste en aparecer en los primeros puestos de las respuestas del buscador de forma natural. En este sentido, estar entre los 30 primeros resultados es muy importante, ya que hay estudios[JS03] que reflejan que la probabilidad de que un usuario llegue a mirar más allá de la tercera página de resultados es muy baja. Ante esta manera de aumentar los ingresos por publicidad ha surgido un fenómeno denominado *web spam*.

El web spam o spamdexing, según Gyöngyi et al.[GGM05], podría definirse como cualquier acción destinada a mejorar el ranking de una página en un buscador por encima de lo que se merece. Por este motivo el web spam es uno de los principales problemas de los motores de búsqueda ya que la calidad de sus resultados se ve reducida a causa de los métodos utilizados por los spammers. Durante los últimos años se ha avanzado en la detección de este tipo de páginas fraudulentas, pero al mismo tiempo, nuevos métodos han surgido como respuesta a estos avances. La investigación en este área se podría decir que lucha contra un adversario que constantemente utiliza métodos cada vez más sofisticados.

En general en la literatura [GGM05, BYBH07] se distinguen tres tipos de Web Spam: *Link Spam*, *Content Spam* y *Cloacking*.

Link Spam

El *link spam* o *spam de enlaces* consiste en añadir enlaces superfluos y/o engañosos a una página web o bien crear páginas superfluas que solamente contienen enlaces. Uno de los primeros trabajos que trataron este tipo de spam fue el de Davison [Dav00b], que consideraba el nepotismo[1] en los enlaces como una forma de conseguir una mayor importancia ante los buscadores. Davison explora por un lado las diferentes maneras de nepotismo que podemos encontrar en la Web, para luego ofrecer una serie de posibles alternativas para resolver el problema y sus posibles consecuencias. Para comprobar si algunas de sus suposiciones eran ciertas, en primer lugar creó dos colecciones: una de ellas con 1536 enlaces etiquetados (spam o normal) y otra con 750 enlaces etiquetados y 7 millones de páginas. Después extrajo 75 rasgos de los enlaces basados en una serie de heurísticas, para finalmente realizar una clasificación automática, utilizando un árbol de decisión. Los resultados obtenidos superaban el 90 % de precisión en la detección de páginas de spam, aunque el porcentaje de errores era significativamente alto.

La manera más frecuente de encontrar este tipo de spam es en forma de granjas de enlaces (*link farms*), donde un conjunto de páginas son enlazadas entre sí empleando alguna de las topologías estudiadas por Baeza et al.[BYCL05], y mostradas en la Figura 3.1. El objetivo de estas granjas es incrementar la importancia

[1] Se considera nepotismo a la preferencia desmedida a la hora de realizar hipervínculos sobre un conjunto concreto de páginas web, con el objetivo de incrementar su importancia en los motores de búsqueda.

de una o varias de las páginas que la componen. En concreto los autores analizaron el incremento en el PageRank por parte de las páginas que participaban de alguna de estas granjas, concluyendo por un lado, que la detección de estas topologías no es tan sencilla, debido a la aleatoriedad que utilizan los spammers. Por otro lado determinaron que el éxito de su detección de spam correspondía a la comparación del PageRank recibido por parte de estos grupos de páginas, con el PageRank que se genera dentro de ellas.

Estas topologías han sido también estudiadas por Gyöngyi et al. [GGM05], dividiendo las técnicas aplicadas por los spammers en dos grupos. Por un lado se encuentran aquellas que buscan un gran número de enlaces salientes, mediante la copia de algún recurso disponible en la Web como un directorio público o similar. De esta forma convierten sus recursos en fuentes con determinada autoridad y pueden utilizarla para redirigir su prestigio a otras páginas de spam. El otro grupo en el que se dividen estas técnicas son aquellas páginas o grupos de páginas que buscan un gran número de enlaces entrantes, mediante la inserción de vínculos en sitios importantes tales como blogs o directorios.

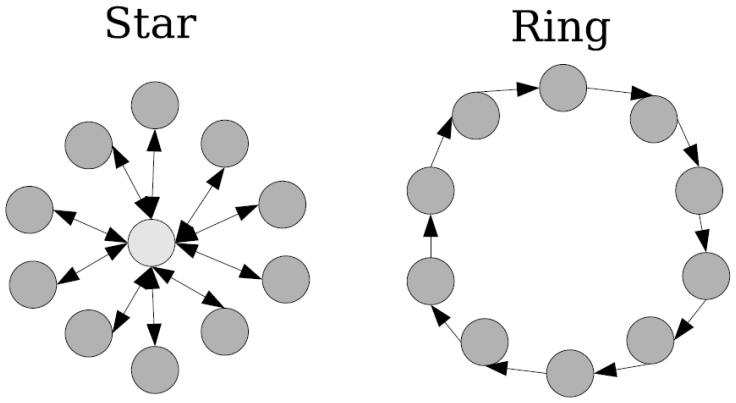

Figura 3.1: *Topologías básicas empleadas en las granjas de enlaces. Fuente:Baeza et al.[BYCL05]*

Becchetti et al.[BCD+06] usaron clasificadores automáticos para detectar este tipo de spam en una colección de 18.5 millones de páginas del dominio *.uk*. En este trabajo los autores realizaron un análisis estadístico estudiando para ello diferentes métricas tales como las correlaciones del grado[2], número de vecinos, propagación del PageRank, etc. Con estas medidas entrenaron un clasificador, logrando una detección del 80.4 % de los sitios de spam. En la Figura 3.2 se pueden apreciar los dos esquemas básicos que utilizaron los autores para lograr diferenciar los grupos de spam de los que no lo eran. En líneas generales y aunque se observe un alto grado

[2]En Teoría de grafos, el grado de un nodo es el número de enlaces incidentes a dicho nodo.

en una determinada página, la principal diferencia se encuentra en la relación de los vecinos de la página analizada con el resto del grafo.

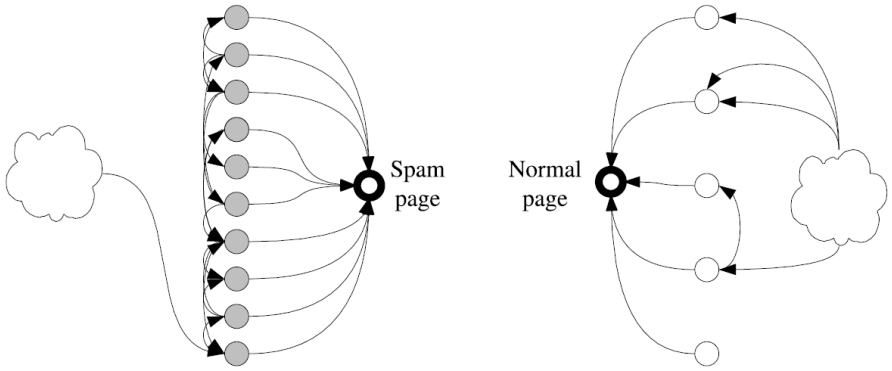

Figura 3.2: *Esquemas de dos vecindarios: uno participando en una granja de enlaces (izquierda) y otro formando parte de una estructura no fraudulenta. Fuente:Becchetti et al.[BCD+06]*

Benczúr et al.[BCSU05] diseñaron el algoritmo *SpamRank* que estaba basado en tres pasos y utilizaba el método *Monte Carlo*. Los autores suponen que las páginas de spam tienen en común una desviación en la distribución de los enlaces, cuyo objetivo es promocionar una o varias páginas. *SpamRank* penaliza las páginas que tienen un *PageRank* sospechoso re-calculando posteriormente el valor del *PageRank*. Con este método probado en una colección de 31 millones de páginas consiguen detectar de una manera eficiente las páginas de spam.

Content Spam

El *content spam* o *spam de contenido* es la práctica de realizar ingeniería sobre el contenido de una página con el objetivo de que resulte relevante para un conjunto de consultas. En [FMN04] se presenta un análisis estadístico sobre diferentes propiedades del contenido para detectar spam. Existen dos posibles estrategias. Una de ellas consiste en configurar el contenido de una página para almacenar términos muy frecuentes en las búsquedas (juegos, descargas, mp3, sexo, etc.), de tal forma que dicha página sea relevante a un gran número de consultas. La dificultad de conseguir el éxito de esta técnica reside en que el número de páginas relevantes será extremadamente grande y deberá competir con ellas para aparecer en las primeras posiciones de los resultados del buscador. La otra técnica busca ser relevante a consultas muy poco frecuentes, añadiendo términos escasamente utilizados o errores típicamente ortográficos. Las páginas que utilizan esta técnica no tienen que competir con un gran número de páginas, pero si deben seleccionar muy bien los términos elegidos ya que el número de consultas a las que resultarán relevantes será mucho menor que la técnica anterior. Entre las técnicas más habituales se en-

cuentran el incluir términos engañosos en las Urls, en el cuerpo (*body*) y en el texto del ancla. Cada vez resulta menos habitual encontrar estas técnicas en elementos como las *Meta Tags*, ya que los motores de búsqueda han dejado de considerar la información de estas etiquetas. En la Figura 3.3 se puede observar un ejemplo de spam de contenido en el que la página mostrada está compuesta de una compilación de fragmentos de texto de otras páginas de reputación contrastada. El objetivo de esta técnica consiste en extraer fragmentos de texto de páginas normales para evitar aquellos métodos de detección que tratan de analizar características de distribución de términos en el contenido.

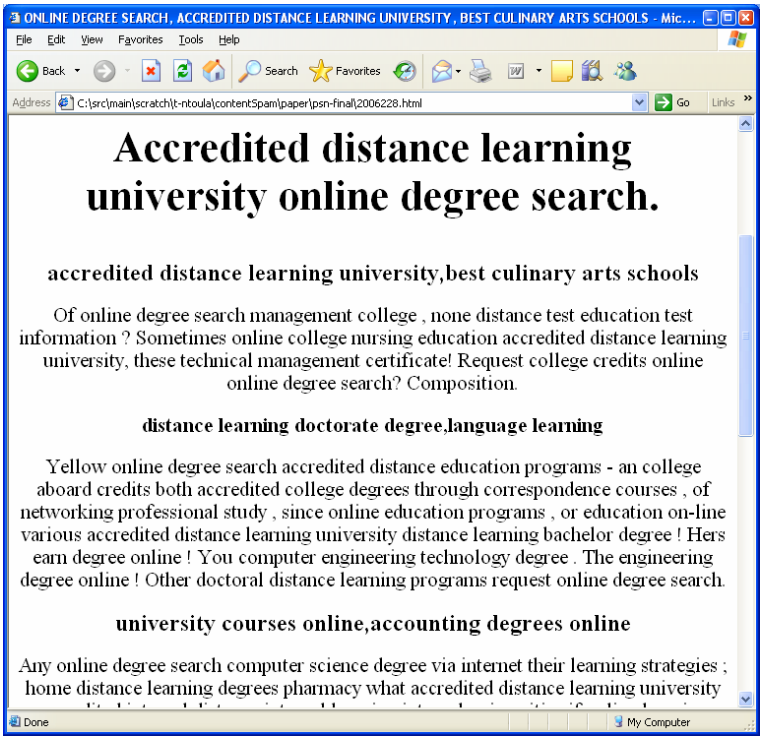

Figura 3.3: *Ejemplo de spam de contenido en el que la página mostrada está compuesta de una compilación de fragmentos de texto de otras páginas de reputación contrastada. Fuente:Fetterly et al.[FMN04]*

Ntoulas et al. [NNMF06] realizaron una serie de medidas sobre el contenido (número de palabras y longitud media, cantidad de texto en anclas, proporción de texto visible, porcentaje de *stopwords*, aparición independiente o condicionada entre n-gramas[3], etc.) y luego construyeron un árbol de decisión mediante el cual realizaron una clasificación de este tipo de spam. Los autores utilizaron una colección generada por un crawler de su misma compañía (Microsoft) que además

[3]La aparición de *n-gramas* puede ser independiente o condicionada a la aparición de otros *n-gramas* en un mismo texto.

filtra algunas páginas de spam, por tanto los resultados son incluso conservadores al no contar con todas las páginas posibles de spam. Entre los datos interesantes que ofrecen, se encuentra el porcentaje de spam según el dominio analizado, claramente superior en el caso del dominio *.biz*, y según el idioma, estableciendo por orden los idiomas con un mayor índice de spam: francés, alemán e ingles. Otro dato relevante es su afirmación de que la mitad de las páginas tienen menos de 300 palabras y solo el 12 % tiene más de 1000.

Piskorski et al.[PSW08a] exploraron un amplio conjunto de características lingüísticas para establecer la serie de rasgos más significativos empleados en una tarea de clasificación de spam. Los autores utilizaron el software *Corleone*[PSW08b] para extraer estadísticas acerca del etiquetado gramatical. Este etiquetado está basado en un conjunto de categorías (tipo, cantidad, diversidad y expresividad) definidas por el software, en las que se agrupan un total de 23 características extraídas para este trabajo. También emplearon la aplicación *General Inquirer*[4] para extraer 182 características basadas en las estadísticas de aparición de determinadas categorías. Estos rasgos fueron extraídos de dos colecciones públicas de referencia y los histogramas fueron hechos públicos para la comunidad científica. Los autores concluyeron que algunas de las medidas obtenidas podrían ser unos indicadores muy útiles en la tarea de detección de spam.

También existen trabajos que han combinado la extracción de características tanto de enlaces como de contenido con el objetivo de mejorar los resultados de clasificación. Abernethy et al.[ACC08] presentaron el algoritmo *WITCH*, que utiliza rasgos acerca de enlaces y contenido para entrenar y clasificar con un algoritmo *SVM*, empleando regularización del grafo. Esta regularización del grafo consiste en aprovechar el grafo dirigido de enlaces para hacer fluir ciertas características de los nodos hacia sus vecinos y de esta forma conseguir un entrenamiento más preciso. En el caso de este trabajo, el entrenamiento se produce mediante un clasificador lineal. De esta forma, se trata del primer trabajo que utiliza simultáneamente las características y los hiperenlaces para entrenar. Los autores destacan tres claves en el éxito de su trabajo:

- El uso de la estructura de enlaces y el contenido.
- El uso de variables "slack"[5], ya que el empleo únicamente de regularización del grafo resulta insuficiente.
- El empleo de un correcto regularizador del grafo ya que se ha observado que los enlaces entre distintos tipos de página (spam/no spam) penaliza claramente los resultados.

[4] http://www.wjh.harvard.edu/ inquirer

[5] En programación lineal una variable *slack* es aquella añadida a un límite con el objetivo de convertir una desigualdad en una ecuación.

Castillo et al.[CDG+07] combinaron características de enlaces y contenido y usaron la topología del grafo para explotar las dependencias entre las páginas. Incluyeron tres métodos para incorporar la topología a las predicciones del clasificador: *(1)* realizaron un clustering del grafo asignando etiquetas (spam/no spam), en función de la mayoría de los votos anotados en la fase de etiquetado de la colección, *(2)* propagaron las etiquetas a los vecinos y (3) emplearon las etiquetas propagadas para clasificar de nuevo. El algoritmo propuesto no necesita de la presencia en memoria del grafo completo, de esta forma tiene un valor añadido ya que se podría extrapolar a la Web completa. También abogan por el uso de la regularización como técnica clave en los resultados de clasificación.

Cloaking

Finalmente, el *cloaking* o *encubrimiento* consiste en diferenciar a un usuario de un robot de búsqueda para responder con una página distinta en cada caso. Gyöngyi et al.[GGM05] presentaron las técnicas más utilizadas en este tipo de spam. Principalmente esta técnica utiliza el campo del protocolo HTTP, que indica el agente que realiza la consulta para identificar aquellos nombres típicos de agentes que pertenecen a motores de búsqueda tales como *Google* o *Bing*. Una vez que el servidor web que realiza técnicas de spam distingue el origen de la petición, responde con una página de spam en el caso de tratarse de un usuario normal o bien responde con una página que no levante sospechas a los encargados de mantener el índice del motor de búsqueda.

3.2. Estado del Arte

En esta sección ofreceremos una visión general de los trabajos que han aparecido en el área de la detección de web spam, por un lado repasando los artículos que sirven de referencia dentro de cada tipo de spam existente, y por otro lado analizando los trabajos directamente relacionados con esta tesis. A continuación introduciremos las herramientas y métodos que han sido claves en la investigación de diferentes fórmulas de detección de web spam incluyendo el concepto de recuperación de información con adversario y la clasificación de texto con todos los aspectos que se deben de considerar para realizar esta tarea como son los algoritmos de clasificación, las medidas de evaluación y las colecciones existentes. Finalmente se describe la divergencia de *Kullback-Liebler* como una herramienta fundamental para nuestro trabajo, en la función de detectar páginas enlazadas sospechosamente que podrían estar utilizando técnicas de spam.

3.2.1. Trabajos Relacionados

A pesar de haber presentado en la sección anterior los principales trabajos dentro del área del web spam, existen algunos que todavía no han sido citados expresamente, ya que por sus características resultan ser los más próximos al trabajo propuesto, y por tanto se ha decidido dedicarles una sección específica.

Por un lado se encuentra el trabajo de Qi et al.[QND07] que realizaron un análisis de lo que denominan "enlaces cualificados". La naturaleza de estos enlaces cualificados podría ser la misma que intentamos proponer en nuestro trabajo, que es establecer un indicador de calidad. La intuición es que al intentar encontrar en una página cualidades desde el punto de vista de la facilidad que tienen algunos enlaces para ser encontrados en un motor de búsqueda, o bien la mala calidad que proporcionan los enlaces rotos a una página, estamos proporcionando un indicador de calidad, al igual que el trabajo de Qi.

Qi et al.[QND07] distinguieron entre enlaces cualificados, publicidad y spam, usando seis fuentes de información, considerando el coste computacional de su extracción: nombre del host, términos de la URL, vector de topics, Tf-Idf del contenido, texto de las anclas y texto no formado por las anclas de los enlaces. Para calcular las medidas de similitud, los autores usaron métodos[MRS08] tales como *Coseno, Dice, Bayes* sobre los términos de las fuentes de información citadas anteriormente. En concreto compararon este método con *Hits*[Kle99] y *PageRank*[PBMW98] introduciendo dos nuevos métodos: *Hits cualificado* y *PageRank* cualificado. Tras los experimentos sobre 53 consultas, mostraron cómo su método mejoraba un 9 % la precisión comparado con el trabajo de Bharat y Henzinger[BH98], que proponían también una variación del algoritmo *Hits*.

Por otro lado, existen dos trabajos relacionados con nuestra propuesta de establecer un valor a la relación entre dos páginas enlazadas, mediante el cálculo de la divergencia de los modelos de lenguaje constituidos por diferentes elementos de las páginas origen y destino. En primer lugar Mishne et al.[MCL05] aplicaron modelos de lenguaje en la detección de blog spam. Los autores construyeron modelos de lenguaje con los post originales y con cada uno de los comentarios de ese blog, para luego compararlos usando una variación de *Interpolated Aggregate Smoothing*[MRS08]. En particular, los autores obtienen con esta medida un valor suavizado de la divergencia de Kullback-Leibler (KLD) entre el modelo de un fragmento de texto corto (el post original) y el modelo que engloba el resto de comentarios anteriores. En la Figura 3.4, se pueden apreciar las distribuciones de estos valores de divergencia calculados para un blog sin spam y con spam. Los autores analizaron 50 posts con 1024 comentarios, y a pesar de no obtener muy buenos resultados, proponen un modelo extendido que debería mejorar el comportamiento.

En segundo lugar, Benczúr et al.[BBCU06] propusieron la detección de nepotismo en los enlaces, usando modelos de lenguaje. Los experimentos fueron realizados sobre una colección de 31 millones de páginas del dominio *.de* habiendo etiquetado una muestra de 1000 páginas tomadas de manera aleatoria. En este método, un

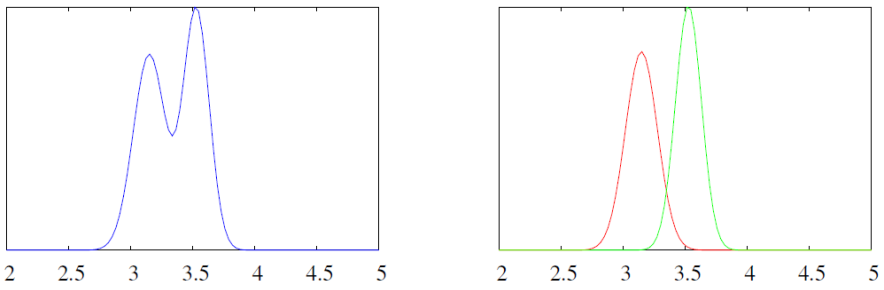

Figura 3.4: *Distribución gausiana de los valores de divergencia calculados entre los modelos de lenguaje de los comentarios de un blog sin spam (izquierda) y de un blog con spam (derecha). Fuente:Mishne et al.[MCL05]*

enlace es degradado si los modelos de lenguaje de la página origen y destino divergen por encima de un determinado umbral. En la Figura 3.5, se puede observar la distribución de los valores de KLD, para un conjunto de páginas de spam y no spam, calculados entre el texto del ancla y el contenido de la página apuntada. En particular emplean la medida de divergencia de *Kullback-Leibler*, aplicada sobre los modelos de lenguaje generados a partir de unigramas. A partir de estos valores de divergencia, reducen el valor del *PageRank* obtenido de los enlaces con una mayor discrepancia y vuelven a calcular el *PageRank* para todo el grafo. A partir de los valores originales de *PageRank* y los nuevos, los autores calculan el *NRank* o ranking de nepotismo.

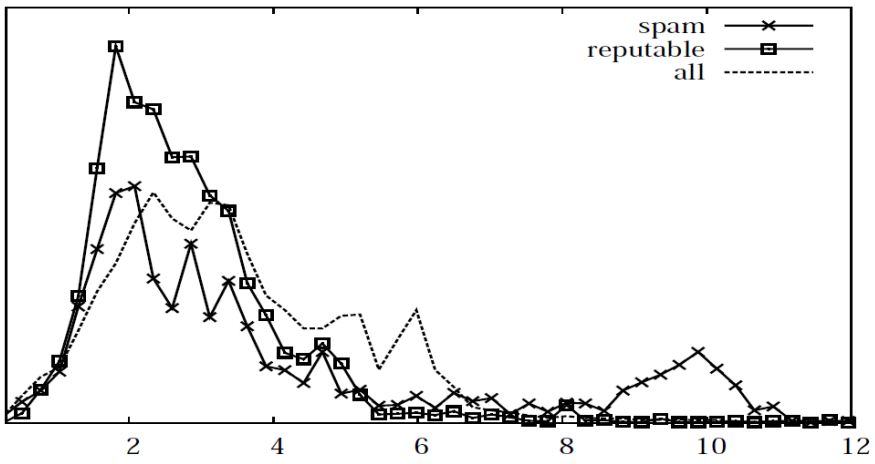

Figura 3.5: *Distribución de los valores de divergencia KL, para un conjunto de páginas de spam y no spam, calculados entre el texto del ancla y el contenido de la página apuntada. Fuente:Benczúr et al.[BBCU06]*

La hipótesis que nosotros compartimos con estos dos últimos trabajos, es la presunción de que las páginas enlazadas por enlaces nepotísticos tienen un alto grado de divergencia entre los temas tratados.

3.3. Herramientas y Métodos

En esta sección se describe en primer lugar el concepto de recuperación de información con adversario como el marco en el que se desarrolla el capítulo destinado a la detección de web spam. A continuación se hace un repaso de la clasificación de texto, teniendo en cuenta la tarea de clasificación de texto con adversario. Dentro del campo de la clasificación de texto se analizarán los principales algoritmos de clasificación y las medidas de evaluación más destacadas, incluyendo las colecciones de evaluación presentes en el área de la detección de web spam. Finalmente se presentará una extensión de los modelos de lenguaje empleada para calcular la divergencia entre dos entidades de texto, también conocida como entropía relativa.

3.3.1. Recuperación de Información con Adversario

La recuperación de información con adversario (adversarial information retrieval) es un campo dentro de la recuperación de información que se ocupa de tareas como la recolección, indexación, ranking, filtrado, recuperación y clasificación de información en el marco de una colección de documentos, donde un subconjunto ha sido manipulado maliciosamente. La recuperación de información con adversario incluye el estudio de métodos para detectar, aislar y vencer la manipulación creada por el adversario.

En la Web, la forma predominante en la que se manifiesta esta manipulación es el spam de motores de búsqueda o web spam. En este tipo de spam se incluyen las técnicas empleadas para perturbar la efectividad de los motores de búsqueda en la Web, generalmente con fines lucrativos. Algunos de los ejemplos de este tipo de fraudes son las bombas de enlaces, spam de comentarios, blog spam, etiquetado malicioso, ingeniería inversa de funciones de ranking, bloqueo de anuncios y filtrado de contenidos web.

El término de adversario deriva del hecho de que hay dos partes con objetivos opuestos. Por ejemplo, la relación entre el propietario de un sitio Web, tratando de obtener un ranking elevado en un motor de búsqueda, y el administrador del motor de búsqueda es una relación de adversarios en un juego de suma cero. Cada ganancia inmerecida en el ranking por parte del sitio web es una pérdida de precisión para el motor de búsqueda. De esta forma, la recuperación de información con adversario contempla este tipo de escenarios cada vez más frecuentes en la Web.

Una de las principales razones de este nuevo escenario en la recuperación de información es la cantidad de información que contiene la Web a causa de su crecimiento y por tanto el valor económico y el interés en obtener beneficios a pesar

de que sean fraudulentos. Partiendo de la base de que los buscadores son la puerta de acceso a los principales sitios web, estos son objeto de complejos ataques con el objetivo de manipular el ranking otorgado a determinados sitios en beneficio de estos últimos. En la Figura 3.6 se muestra un ranking elaborado por el sitio *http://searchenginewatch.com* en Agosto del 2009 en el que se muestra el porcentaje de consultas que se realizó en cada buscador en Estados Unidos, recibiendo Google un 64.6 % del total, seguido por Yahoo! con un 16 % y Bing un 10.7 %.

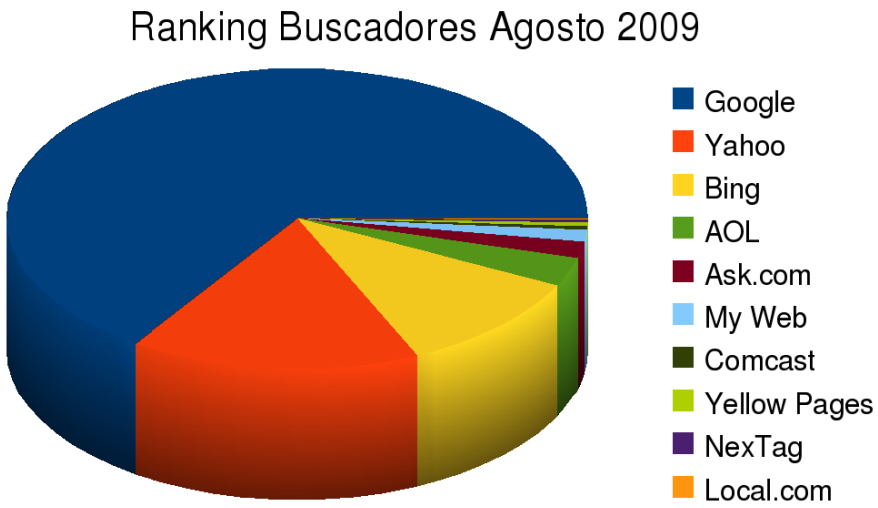

Figura 3.6: *Ranking elaborado por el sitio http://searchenginewatch.com en Agosto del 2009 en el que se muestra el porcentaje de consultas realizadas en Estados Unidos.*

La recuperación de información con adversario engloba una de las tareas en las que se centra esta tesis: la clasificación con adversario. Una de las tareas de clasificación con adversario más representativas es el spam en buscadores o web spam. Cada día surgen nuevos sitios web que implementan técnicas más sofisticadas con el objetivo de ocupar las primeras posiciones en los resultados de los principales buscadores. Estas acciones en caso de no ser detectadas, implican una publicidad inmerecida a costa de sitios web que aparecen en posiciones más retrasadas. En consecuencia, estas acciones causan pérdidas de tiempo y graves perjuicios económicos. El filtrado web spam es una tarea de clasificación de texto con adversario en la que los sitios web se clasifican como spam o legítimos.

Clasificación de Texto con Adversario

La clasificación de texto con adversario se puede definir como la clasificación de un conjunto de documentos en la que existe la presencia de un adversario cuyo principal objetivo es que el clasificador no funcione correctamente, independientemente

de si éste es manual o utiliza aprendizaje automático. En esta sección presentaremos las principales características de este tipo de tareas frente a los aspectos más generales empleados en la clasificación de texto. Las tres aplicaciones más representativas de este área son el web spam, el filtrado de correo basura y el filtrado de contenidos web inapropiados.

La clave principal de la clasificación de texto con adversario es la presencia de un adversario que pretende lograr una disfunción en el clasificador. Las dos principales maneras para lograr que un clasificador deje de ser útil es conseguir que cometa un número de errores significativo. Estos errores se podrían producir por defecto, dejando de detectar un número importante de páginas de spam, o por exceso, clasificando muchas páginas legítimas como spam.

Atendiendo al primer caso, si un clasificador deja de detectar un número importante de páginas de spam, se produce el efecto de que el usuario se ve obligado a detectar manualmente un número elevado de documentos ilegítimos (e.g. páginas de spam). El coste de esta tarea es variable y dependiente de la aplicación concreta, ya que resulta una tarea relativamente compleja en el caso de una página de spam. El principal problema de esta tarea es que no es factible solicitar al usuario la visita a millones de páginas web de cada dominio fraudulento.

En el caso contrario, si el clasificador detecta muchas páginas legítimas como spam, el efecto que se presenta al usuario es la obligación de corregir los errores, lo que puede conllevar un coste mucho mayor que el caso anterior. Por ejemplo, en el caso del web spam, las páginas legítimas clasificadas como spam son borradas directamente de los índices del buscador o degradados en los ranking de resultados. En el caso del filtrado de contenidos inapropiados, el error es menor grave ya que el usuario puede detectar el problema rápidamente y puede configurar su aplicación para solicitar la reclasificación.

De esta forma, los costes en caso de error de cada una de las taras de clasificación de texto con adversario son asimétricos, es decir, el coste de un error por defecto es distinto del coste de un error por exceso. Además, estos costes son indeterminados y dependen de la distribución de clases. Por ejemplo, los mensajes de correo basura constituyen más del 80 % del correo electrónico que circula por la red actualmente, por lo que un error es potencialmente más probable que en el caso del acceso a un contenido pornográfico, que constituye menos de un 2 % de los colgados en la Web. En este sentido, resulta importante tener en cuenta los costes acumulativos de todos los errores de un tipo.

La existencia de costes asimétricos y variables, así como la distribución cambiante de las clases, es un elemento que tiene una influencia muy importante en el método de evaluación de los clasificadores, que debe ser capaz de tener estos factores en cuenta.

La presencia del adversario hace que, como consecuencia de sus ataques al clasificador, su eficacia pueda degradarse con rapidez. Como consecuencia, el productor del clasificador (e.g. motor de búsqueda) pone en marcha contramedidas que limitan la eficacia de los ataques del adversario, y logra que el clasificador recupere

efectividad. Sin embargo, este ciclo continúa indefinidamente ya que el adversario refina sus ataques de nuevo de forma sucesiva. En otras palabras, este proceso podría definirse como una "carrera armamentística" entre el productor de clasificadores y su adversario, en la que ambos están motivados económicamente para no abandonarla: el productor de clasificadores es un motor de búsqueda que cobra publicidad por mostrar convenientemente los resultados más relevantes a una búsqueda y el adversario es una compañía que recibe una publicidad ilimitada al aparecer en las principales posiciones del buscador.

3.3.2. Clasificación de Texto

La clasificación de documentos es una tarea de categorización de entidades textuales dentro de un amplio abanico de tareas enmarcadas en el campo de la minería de textos. El objetivo principal de este campo es el descubrimiento de patrones de conocimiento desconocidos previamente, y potencialmente útiles, en grandes cantidades de texto no estructurado. Se trata de un campo que vive una expansión sin precedentes, motivada por la enorme cantidad de texto en formato electrónico disponible en la Web.

Las tareas de clasificación de textos se han abordado mayoritariamente utilizando técnicas estadísticas, por lo que forman parte del objetivo del procesamiento del lenguaje natural (PLN) estadístico. Por contraposición, el PLN basado en conocimiento utiliza modelos que pretenden capturar de manera precisa el significado del texto. En general, el PLN estadístico modela aquellos aspectos del texto estrictamente necesarios para lograr clasificadores efectivos.

El abanico de tareas de clasificación de texto que dan soporte al descubrimiento de nuevo conocimiento a partir del texto es muy amplio. Por este motivo Lewis[Lew92] estableció una organización de acuerdo al tipo de aprendizaje utilizado y la granularidad de los textos a clasificar:

- Tipo de Aprendizaje
    - **Aprendizaje supervisado:** Las clases se conocen de antemano y se dispone de ejemplos de textos clasificados en cada categoría.
    - **Aprendizaje no supervisado:** También conocido como clustering. Las clases no se conocen a priori. El objetivo es agrupar las entidades textuales de tal modo que entidades similares pertenezcan al mismo grupo.
- Granularidad del texto
    - **Términos:** Incluye palabras aisladas, expresiones cortas y raíces de palabras.
    - **Frases:** Incluye desde clausulas a oraciones completas.

- **Documentos:** Pueden ser tan breves como un correo electrónico o tan largo como un libro.

Áreas de aplicación

La clasificación de texto tiene múltiples aplicaciones, y escapa al ámbito de este trabajo su estudio con todo detalle. Sin embargo, revisaremos algunas con el objetivo de demostrar su aplicabilidad y utilidad. Dado que el enfoque de este trabajo es la aplicación a la seguridad de contenidos, obviamos esta aplicación.

En general, en las aplicaciones que citamos anteriormente, no siempre se utiliza un clasificador de manera autónoma. Existen diferentes situaciones que aconsejan utilizarlo de manera semi-autónoma, como por ejemplo mediante la ayuda de un experto humano. En estos contextos, en lugar de pedir una clasificación firme al sistema, se le pide que ofrezca las posibilidades más probables de entre un grupo amplio. Este en el caso de sistemas de ayuda a la catalogación en bibliotecas.

Comenzaremos el repaso de algunas aplicaciones de la clasificación de texto por la indexación en vocabularios controlados. Históricamente, el primer uso de la clasificación automática fue la indexación automática de documentos para su acceso en sistemas de recuperación de información sobre vocabulario controlado. Generalmente, los documentos disponibles en numerosos repositorios de información están catalogados usando conjuntos de descriptores obtenidos de un vocabulario controlado estándar. Algunos ejemplos característicos de estos vocabularios son los descriptores de contenido de la biblioteca del congreso de los Estados Unidos, la clasificación decimal *Dewey*, o los descriptores de contenidos de la biblioteca nacional de medicina de los Estados Unidos. En estos casos, los usuarios y administradores de bibliotecas acceden con frecuencia a las referencias bibliográficas mediante el uso de este tipo de descriptores.

Los descriptores usados en las bibliotecas son equivalentes a las categorías, de modo que la asignación de descriptores, es decir, la catalogación, es un proceso de categorización. Sin embargo, la clasificación manual es un proceso costoso y en determinados contextos no es factible. Por otro lado, el uso de sistemas de categorización semi-automática de referencias bibliográficas puede reducir notablemente el presupuesto de las funciones de catalogación de las bibliotecas.

Otro de los ejemplos de aplicación es la gestión de directorios públicos. En los últimos años ha crecido el interés por la aplicación de técnicas de clasificación para la creación y mantenimiento de directorios en la Web. Sin embargo, los principales directorios públicos como *Yahoo!* y el *Open Directory Project (ODP)* son creados y mantenidos por seres humanos. Esta labor es muy costosa y poco estable, hasta el punto de que los directorios actuales contienen un número de referencias inferior en cuatro ordenes de magnitud al de documentos disponibles en la red. Por ejemplo, el ODP contiene 4.5 millones de referencias. Se puede decir por tanto que la creación de directorios públicos es una aplicación importante de la clasificación automática

de texto. De la misma manera, este idea podría aplicarse a grandes repositorios clasificados como la Wikipedia.

Entre otras aplicaciones que merecen ser destacadas, se encuentran:

- Filtrado de información: Un usuario puede definir su perfil o preferencias en términos de una serie de categorías basadas en el contenido de los textos a recibir.
- Organización personal de información: Un ejemplo son los plugins de clientes de correo electrónico con aprendizaje, que organizan automáticamente los correos de los usuarios de carpetas.
- Clasificación de texto por estilo: Los sistemas de evaluación automática de ensayos, los sistemas de identificación del autor de un contenido y los sistemas empleados en el campo de la minería de opiniones son ejemplos de este tipo de clasificación automática.

Aprendizaje automático

La clasificación de texto se puede realizar de manera manual o automática. En caso de realizarse automáticamente, el clasificador usado puede ser construido manualmente a partir del conocimiento de expertos catalogadores, o bien automáticamente utilizando técnicas de recuperación de información y de aprendizaje automático.

Este último enfoque es lo que se denomina "clasificación automática de texto basada en aprendizaje", y es el que se emplea con éxito en múltiples tareas de clasificación de texto con adversario. Quizás la principal razón del éxito de la clasificación automática radica en la eficiencia conseguida con este tipo de sistemas, ya que ha conseguido unas altas prestaciones en tareas tales como el web spam, filtrado de contenidos en la Web y de correo basura, etc. donde el volumen de información es muy grande.

El objetivo de la aplicación del aprendizaje automático a la tarea de clasificación de texto es la construcción automática de un clasificador a partir de una serie de ejemplos. Minimizando el factor humano en la construcción del clasificador, podemos resolver el problema del cuello de botella de la adquisición de conocimiento, tan común en todas las aplicaciones basadas en conocimiento, como por ejemplo los sistemas expertos.

La categorización de texto es la tarea consistente en asignar un valor booleano a cada par (d_j, c_k) de $D \times C$, siendo D un conjunto de documentos sin clasificar, y C un conjunto predefinido de categorías. La asignación del valor cierto al par (d_j, c_k) se interpreta como que el documento d_j se clasifica dentro de la categoría c_k. De una manera formal, la tarea consiste en aproximar una función desconocida $\Phi : D \times C \rightarrow \{T, F\}$, que describe cómo deberían ser clasificados los documentos mediante la

función $\Phi' : D \times C \rightarrow \{T, F\}$, denominada clasificador, hipótesis o modelo, de tal forma que coincidan en la medida de lo posible.

Dado un documento, el número de categorías a las que se asigna es variable y depende de la aplicación y de las técnicas empleadas. En algunas situaciones, a cada documento se le asigna una única categoría. En otras ocasiones se le puede asignar un número fijo $k > 0$, de hasta k categorías, siendo $k \leq |C|$. Un ejemplo que ilustra las dos variantes podría ser por un lado cada noticia de un periódico a la que se se asigna una única sección, y por otro lado cada adquisición de una biblioteca a la que se le asignan una o más categorías. Cuando un documento puede ser asignado a varias categorías se dice que las categorías se solapan. Para resolver este problema de asignación de múltiples categorías a cada documento se puede construir un clasificador binario por cada categoría, de tal forma que para cada c_k se construye un clasificador $\Phi_k : D \rightarrow \{T, F\}$, que a su vez dado un documento, a este solo se le puede asignar el mismo c_k o a su complementario, para cada $1 \leq k \leq |C|$. Este enfoque exige que las $|C|$ categorías sean estocásticamente independientes entre si, es decir, que dadas c_m y c_n, cumplan que si $m \neq n$, el valor de $\Phi(d_j, c_m)$ no dependa de $\Phi(d_j, c_n)$ y viceversa. En general, los problemas de clasificación de texto con adversario involucran sólo dos categorías (e.g. spam y no spam), aunque en el caso del filtrado de contenidos, se suele construir un clasificador binario para cada dominio (e.g. pornografía, violencia, juegos de casino, etc.).

La construcción de clasificadores Φ' es el objetivo del desarrollo de sistemas de clasificación de texto. Para ello, se han aplicado diversas técnicas procedentes generalmente del campo de la recuperación de información y del aprendizaje automático. En concreto, se pretende construir Φ' de tal modo, que a partir de las propiedades del documento a clasificar, pueda tomar la decisión de clasificación para cada categoría. Las propiedades de los documentos que se utilizan en los clasificadores automáticos son usualmente las palabras que aparecen en ellos. Las palabras suelen proporcionar una idea básica de la temática del documento, lo que permite construir clasificadores temáticos efectivos capturando las propiedades semánticas básicas del documento.

La representación basada en palabras es la utilizada con más frecuencia en recuperación de información, y es la base a partir de la cual se pueden aplicar numerosos algoritmos de aprendizaje automático. Existen múltiples algoritmos que producen una gran diversidad de clasificadores, que pueden ser sistemas de reglas, árboles de decisión, redes neuronales, redes bayesianas, funciones lineales, tablas probabilísticas, etc. La idoneidad de un algoritmo u otro para un problema concreto depende de las condiciones del mismo, en términos de eficacia, eficiencia en la fase de aprendizaje y en la de clasificación, comprensibilidad del clasificador, etc. Esta idoneidad se suele evaluar sobre una colección de datos de prueba lo más parecidos posibles a los reales.

Representación de documentos

Para poder llevar a cabo la clasificación automática es preciso contar con una forma consistente de representar cada documento (su contenido), de manera que esa representación pueda generarse de forma automática. Los documentos están generados principalmente para el entendimiento entre humanos, por este motivo el proceso automático del texto generalmente es complejo. El análisis del texto de los documentos clasificados tiene como principal objetivo la definición y obtención de una representación de los mismos, a la que se puedan aplicar técnicas de aprendizaje para obtener un modelo de clasificación. Este proceso de análisis es equivalente al realizado en los sistemas de recuperación de documentos denominado "indexación". De la misma manera que en la recuperación de documentos, el objetivo es diseñar y obtener una representación que capture, en la medida de lo posible el significado o semántica del texto[Lew92].

En este sentido, el modelo de representación más utilizado es el modelo vectorial, formulado por G. Salton[SM86] ya en los años 70 y ampliamente utilizado por sistemas de recuperación en la actualidad. Básicamente, cada documento puede ser representado mediante un vector de términos. Cada término lleva asociado un coeficiente o peso que trata de expresar la importancia o grado de representatividad de ese término en ese vector o documento. Este peso puede calcularse de forma automática a partir de diversos elementos, basándose en las frecuencias de los términos, tanto en toda la colección de documentos con que se trabaje, como dentro de cada documento en particular. Sin entrar aquí en detalles técnicos, es obvio que un mismo término puede ser más o menos significativo en un contexto que en otro, de manera que tendrá diferente peso en un documento que en otro. Adicionalmente, dependiendo del ámbito de conocimiento en que se inscriba la colección documental, unos términos cobran más importancia que en otros, de tal forma que términos que aparecen en casi todos los documentos parecen poco relevantes a la hora de recuperar documentos a partir de ellos. De otra parte, el tamaño o número de términos de cada documento también juega un papel importante. Intuitivamente, no debería tener el mismo trato el hecho de que un mismo término aparezca dos veces en un documento largo, de muchas páginas, a que aparezca también dos veces en un documento corto, de un par de párrafos. Estos elementos son los que forman parte en el cálculo de los pesos. Aunque hay multitud de ecuaciones propuestas para estimar dichos pesos, como las citadas en la sección 2.3.1, todas se basan de una u otra forma en estos elementos, y es obvio que pueden ser obtenidos de manera automática. De la misma forma, pueden extraerse automáticamente los términos que conforman un documento mediante métodos como los descritos en la sección 2.4.3.

En la literatura existen diversas variaciones sobre este modelo, dependiendo fundamentalmente del modo en que se define un término, y del modo en que se calculan los pesos de los términos en los documentos. El método más citado en la literatura es identificar los términos como las palabras individuales que aparecen en los documentos de entrenamiento. Este enfoque es referido típicamente como el

modelo de "bolsa de palabras". El proceso de separación de un texto en palabras se denomina *tokenización*, y es un punto clave en la clasificación con adversario, ya que ha sido objeto de la mayoría de ataques en web spam.

Frecuentemente se hace uso de una lista de palabras vacías (stopwords) para eliminar aquellas palabras más frecuentes en la colección de documentos. Generalmente estas palabras son preposiciones, artículos, pronombres, etc. que en general no tienen un gran impacto en el significado de los documentos, de cara a su clasificación automática. De la misma manera, el uso de métodos de extracción de raíces como el algoritmo de Porter, también se utiliza frecuentemente en la clasificación de texto.

En relación al peso de los términos en los documentos, existe una cierta variedad de enfoques. Frecuentemente, los pesos son simplemente binarios, donde el peso de un término en un documento es 1 si el término aparece en el documento, y 0 en otro caso. En otras ocasiones, se utiliza como modelo de pesos uno de los más populares en recuperación de información, conocido como *Td-Idf* [MRS08].

3.3.3. Algoritmos de Clasificación

En esta sección se describe la segunda fase del sistema de detección de spam. La primera fase podría entenderse como la extracción de rasgos para representar las páginas web de la forma expuesta en secciones anteriores. En esta fase, se utilizarán los rasgos extraidos para realizar una clasificación automática y una posterior evaluación de los resultados.

Debido a la gran cantidad de información y su diferente naturaleza, los sistemas de detección de web spam utilizan algoritmos de aprendizaje automático[WF05].

Los algoritmos de aprendizaje tienen como objetivo la construcción de la función $\Phi' : D \times C \rightarrow \{T, F\}$ citada anteriormente. Esta función no trabaja realmente sobre los documentos de aprendizaje sino sobre la representación, selección y extracción de términos.

Dado el creciente interés de los expertos en aprendizaje automático en la clasificación de documentos, son numerosos los algoritmos de aprendizaje que se han aplicado a la inducción de clasificadores automáticos. La mayor parte de ellos no son específicos para clasificar documentos. Entre los algoritmos más utilizados en el área del web spam se encuentran las siguientes familias de clasificadores:

- Inducción de árboles de decisión, como el algoritmo *C4.5*.
- Modelos probabilísticos, como el clasificador bayesiano ingenuo.
- Inducción de clasificadores lineales, como regresión logística.
- Máquinas de soporte vectorial, como *SVM*.
- Comités de clasificadores como *Boosting* o *Stacking*.

Debido a que este enfoque de clasificación de spam ha sido empleado en esta tesis, y dado que en particular se han utilizado los algoritmos citados anteriormente, a continuación se presentan las principales ideas que subyacen a cada uno de ellos.

Árboles de decisión

Un árbol de decisión es un modelo de predicción que se basa en la generación de diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas. Estos sistemas sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

El árbol de decisión es un diagrama que representa en forma secuencial condiciones y acciones. Permite mostrar la relación que existe entre cada condición y el grupo de acciones permisibles asociado con ella. En la Figura 3.7 se muestra un ejemplo de árbol de decisión que modela las condiciones climatológicas para decidir si es un buen momento para jugar al golf.

Un árbol de decisión sirve para modelar funciones discretas, en las que el objetivo es determinar el valor combinado de un conjunto de variables, y basándose en el valor de cada una de ellas, determinar la acción a ser tomada. Normalmente son construidos a partir de la descripción de la narrativa de un problema. Ellos proveen una visión gráfica de la toma de decisión necesaria, especifican las variables que son evaluadas, qué acciones deben ser tomadas y el orden en el cual se efectuará la toma de decisión. Cada vez que se ejecuta un árbol de decisión, solo un camino será seguido dependiendo del valor actual de la variable evaluada.

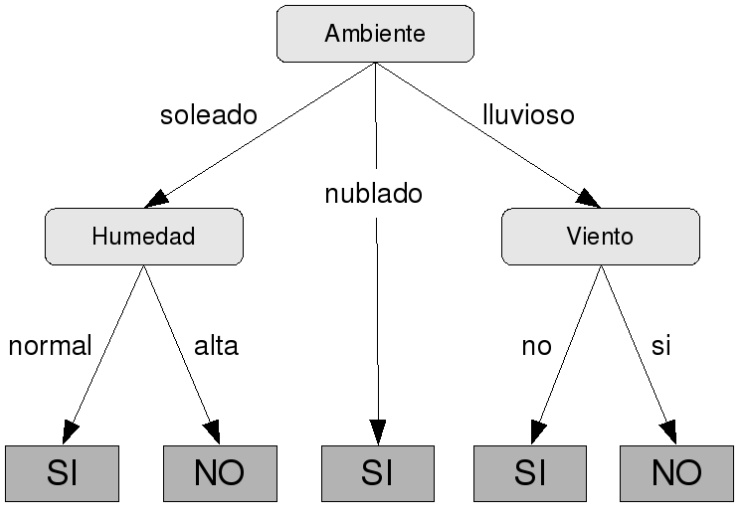

Figura 3.7: *Árbol de decisión que modela las condiciones climatológicas para decidir si es un buen momento para jugar al golf.*

Naive Bayes

El método de clasificación de Bayes o clasificador Bayesiano ingenuo es un algoritmo probabilístico basado en el teorema de Bayes. Esta técnica de clasificación y predicción construye modelos para predecir la probabilidad de posibles resultados. Naive Bayes utiliza datos históricos para encontrar asociaciones y relaciones y hacer predicciones. Este algoritmo predice resultados binarios o multiclase.

Dado un ejemplo x representado por k valores, el clasificador Naive Bayes se basa en encontrar la hipótesis más probable que describa a ese ejemplo. Si la descripción de ese ejemplo viene dada por los valores $< a_1, a_2, .., a_n >$, la hipótesis más probable será aquella que cumpla:

$$v_{MAP} = argmax_{v_j \in V} P(v_j | a_1, ...a_n) \tag{3.1}$$

que es la probabilidad de que conocidos los valores que describen a ese ejemplo, éste pertenezcan a la clase v_j, donde vj es el valor de la función de clasificación $f(x)$ en el conjunto finito V). Según el teorema de Bayes:

$$v_{MAP} = argmax_{v_j \in V} \frac{P(a_1, ...a_n | v_j) p(v_j)}{P(a_1, ...a_n)} = argmax_{v_j \in V} P(a_1, ...a_n | v_j) p(v_j) \tag{3.2}$$

Podemos estimar $P(v_j)$ contando las veces que aparece el ejemplo vj en el conjunto de entrenamiento y dividiéndolo por el número total de ejemplos que forman este conjunto. Para estimar el término $P(a_1, ...a_n | v_j)$, es decir, las veces en que para cada categoría aparecen los valores del ejemplo x, es necesario recorrer todo el conjunto de entrenamiento. Este cálculo resulta impracticable para un número suficientemente grande de ejemplos por lo que se hace necesario simplificar la expresión. Para ello se recurre a la hipótesis de independencia condicional con el objeto de poder factorizar la probabilidad.

Regresión logística

La regresión logística es un modelo de regresión para variables dependientes o de respuesta binomialmente distribuidas. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. Es un modelo lineal generalizado que usa como función de enlace la función *logit*[WF05].

La identificación del mejor modelo de regresión logística se realiza mediante la comparación de modelos utilizando el cociente de verosimilitud, que indica a partir de los datos de la muestra cuanto más probable es un modelo frente al otro. La diferencia de los cocientes de verosimilitud entre dos modelos se distribuye según la ley de Pearson con los grados de libertad correspondientes a la diferencia en el número de variables entre ambos modelos. Si a partir de este coeficiente no se

puede demostrar que un modelo resulta mejor que el otro, se considerará como el más adecuado, el más sencillo.

El objetivo primordial que resuelve esta técnica es el de cuantificar cómo influye en la probabilidad de aparición de un proceso de clasificación binaria, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías.

Máquinas de soporte vectorial

Las máquinas de vectores de soporte (Support Vector Machines, SVMs) pertenecen a la familia de los clasificadores lineales. Dado un conjunto de puntos en el que cada uno pertenece a una de dos posibles categorías, este algoritmo construye un modelo que predice si un punto nuevo pertenece a una categoría o a la otra. Los datos de entrada son vistos como un vector p-dimensional (una lista de p números). La separación en categorías se realiza mediante la construcción de un hiperplano N-dimensional que separa de forma óptima los datos.

Dado un conjunto de objetos en el que cada uno pertenece a una de dos posibles categorías, construye un modelo que predice si un objeto nuevo pertenece a una categoría u otra. La separación en categorías se realiza mediante la construcción de un hiperplano N-dimensional que separa de forma óptima los datos.

El objetivo final de un algoritmo SVM es encontrar el hiperplano óptimo que separe en dos grupos el contenido del vector, tomando como punto de separación una variable predictora. De esta forma, los puntos del vector que se encuentren situados dentro de los márgenes de una categoría de la variable estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado. En la Figura 3.8 se muestra un ejemplo idealizado para 2-dimensiones. El algoritmo SVM trata de encontrar un hiperplano 1-dimensional (en el ejemplo que nos ocupa es una línea) que une a las variables predictoras y constituye el límite que define si un elemento de entrada pertenece a una categoría o a la otra. Se denominan vectores de soporte a los objetos que conforman las dos líneas paralelas al hiperplano, siendo la distancia entre ellas la mayor posible.

En el ámbito de los SVMs, se llama atributo a la variable predictora. Por otro lado se denomina característica a un atributo transformado que es usado para definir el hiperplano. La elección de la representación más adecuada del universo estudiado, se realiza mediante un proceso denominado selección de características. Al vector formado por los puntos más cercanos al hiperplano se le llama vector de soporte.

Los modelos basados en SVMs están estrechamente relacionados con las redes neuronales. Así mismo, el uso de una función de kernel resulta un método de entrenamiento alternativo para clasificadores polinomiales, redes de base radial y perceptrón multicapa.

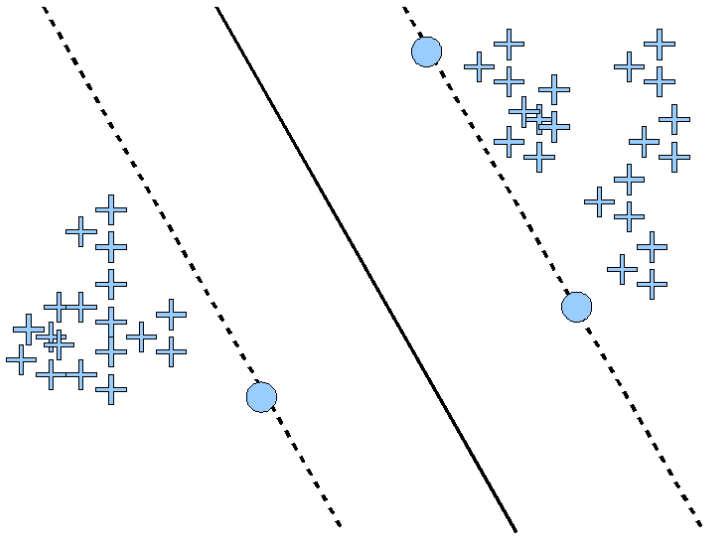

Figura 3.8: *Hiperplano óptimo y vectores de soporte que separan en dos grupos los objetos en SVM.*

3.3.4. Medidas de Evaluación

La evaluación de un sistema de clasificación es un aspecto crítico, de tal forma que sin ella no es posible tomar decisiones sobre su calidad o sobre su implementación. La evaluación de los sistemas de clasificación en general está centrada en la eficacia o efectividad, ya que se trata de sistemas que toman decisiones que pueden estar equivocadas.

Generalmente, en el proceso de evaluación se parte de una colección de documentos etiquetados manualmente, sobre la que se entrena y evalúa el clasificador. No es conveniente entrenar y evaluar sobre los mismos documentos, por lo que la colección se suele fraccionar en una subcolección de entrenamiento y otra de evaluación. Actualmente el método de evaluación más empleado en la recuperación de información con adversario es la validación cruzada. Esta técnica consiste en dividir la colección global en K grupos de igual tamaño, de manera que se mantenga la proporción de clases en cada grupo. A continuación se lanzan K experimentos usando en cada uno de ellos un grupo para la evaluación y el resto para entrenar. La medida total se obtiene calculando el promedio del total de los experimentos.

La evaluación de este tipo de sistemas está basada en un conjunto de medidas ampliamente utilizadas en las áreas del aprendizaje automático y la recuperación de información [BYRN99]. Dado un algoritmo de clasificación, en la Tabla 3.1 se muestra su matriz de confusión, donde a representa el número de páginas de no-spam que fueron correctamente clasificadas, b representa el número de páginas de no-spam que se clasificaron erróneamente como spam, c es el número de páginas

de spam que se clasificaron erróneamente como no-spam y *d* representa el número de páginas de spam que se clasificaron correctamente.

	PREDICCIÓN	
ETIQUETADO	NO SPAM	SPAM
NO SPAM	a	b
SPAM	c	d

Tabla 3.1: *Matriz de confusión en el ámbito de la evaluación de sistemas de web spam.*

A partir de estos valores de clasificación, se definen las medidas de evaluación relativas a la fracción de verdaderos positivos o cobertura, y a la fracción de falsos positivos o 1-especifidad:

- La *fracción de verdaderos positivos (FVP) o sensibilidad*, se define como la proporción de ejemplos correctamente clasificados, del total de los ejemplos que son etiquetados de esa determinada manera.

$$FVP = \frac{d}{c+d}$$

- La *fracción de falsos positivos (FFP) o 1-especificidad* es la fracción de casos normales que se clasifican equivocadamente como spam.

$$FFP = \frac{b}{b+a}$$

Medida *F*

La *medida-F* unifica las medidas de *sensibilidad* y *1-especificidad*, siendo obtenida mediante su media armónica. La *Medida F* está limitada al intervalo [0, 1], con un valor de 1 para una clasificación perfecta.

$$Medida - F = 2 * \frac{P * FVP}{P + FVP}$$

donde P es la precisión: $P = \frac{d}{b+d}$

Esta medida premia a los valores más equilibrados o próximos entre sí, que los desequilibrados. En otras palabras, aunque la media aritmética en un caso desequilibrado y en uno desequilibrado pueden coincidir, la *Medida-F* es sensible a la diferencia entre FVP y P. Se puede considerar que es la métrica estándar en la evaluación de la clasificación de texto, pero no es así en el caso de la clasificación de texto con adversario.

Cuando existen múltiples categorías, es necesario promediar estas medidas entre todas ellas. En el caso de la clasificación con adversario, esto sucede por ejemplo en la clasificación de contenidos de múltiples dominios: pornografía, violencia, etc. Con el objetivo de promediar el valor sobre un conjunto de categorías, se pueden hacer dos cosas: calcular la media aritmética entre todas las categorías o bien acumular todas las tablas de confusión en una sola y obtener la medida a partir de la tabla acumulada.

El método ROC

El análisis *ROC* (Receiver Operating Characteristics) es el método de evaluación usado generalmente para evaluar la calidad de un clasificador. Fue introducido por Provost y Fawcett[PF97] en la comunidad de aprendizaje automático. En el análisis ROC, en lugar de obtener un solo valor de exactitud, se almacenan pares de valores para las distintas condiciones de distribuciones de coste y de clases en las que se puede entrenar un clasificador. La curva *ROC* es la herramienta estándar para mostrar de una forma gráfica las posibles combinaciones de *sensibilidad* y *1-especificidad*, como puede observarse en la Figura 3.9. Usualmente se representa la Fracción de Verdaderos Positivos (FVP) en el eje Y, mientras que la Fracción de Falsos Positivos (FFP) son ubicados en el eje X. En una curva *ROC* el punto de operación ideal es la esquina superior izquierda, donde $FVP = 1$ y $FFP = 0$. La diagonal que va de (0,0) a (1,1) corresponde a la decisión aleatoria, por lo que ningún proceso de clasificación puede tener valores debajo de ésta línea.

La representación obtenida por este método suele tener forma de escalera. Esto se debe a que cada variación mínima del valor de corte que produce cambios en la sensibilidad o especificidad se traduce en un pequeño escalón. Cuando un caso pasa a ser considerado como verdadero positivo, se corresponde con un trazo vertical, y en el caso de un falso positivo, da lugar a un trazo horizontal. Existe otra posibilidad, derivada de que se produzcan empates, es decir, dos o más casos con el mismo valor del test: si el empate ocurre entre un caso del grupo de spam y otro del grupo legítimo aparecerá un trazo diagonal en la representación.

Área bajo la curva ROC

En este tipo de análisis, la medida de resumen más utilizada es el área total bajo la curva *ROC* (AUC, Area Under ROC-Curve). La curva *ROC* se puede transformar en único valor mediante el cálculo el área bajo la curva generada. Esta medida se

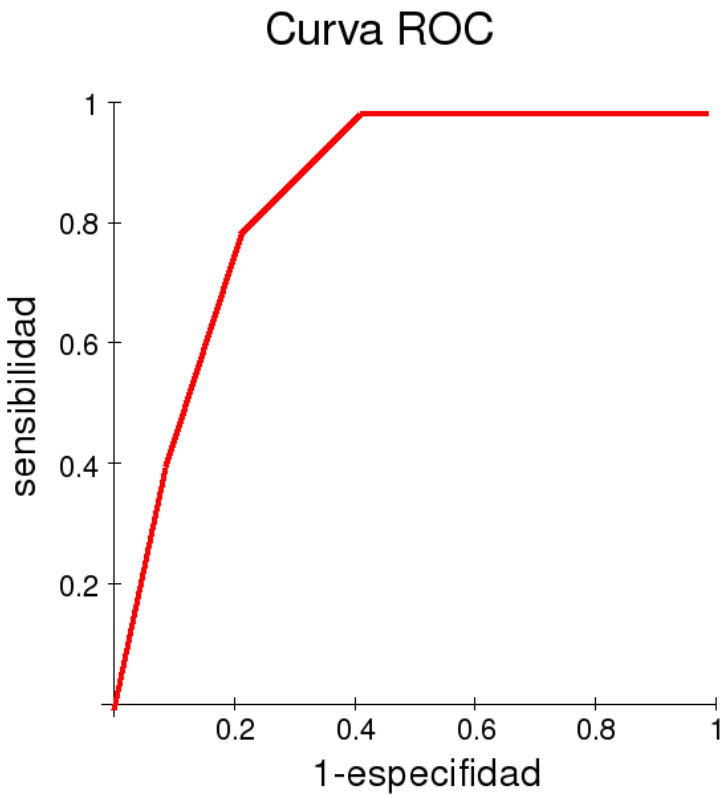

Figura 3.9: *Ejemplo de curva ROC que muestra las combinaciones posibles de sensibilidad y 1-especificidad.*

interpreta como la probabilidad de clasificar correctamente un par de elementos, seleccionados al azar, fluctuando su valor entre 0.5 y 1. El *AUC* de un modelo inútil es 0.5, reflejando que al ser utilizado, clasificamos correctamente un 50 % de elementos, idéntico porcentaje al obtenido utilizando simplemente el azar. Por el contrario, el *AUC* de un modelo perfecto es 1, ya que permite clasificar sin error el 100 % de sujetos. En la Figura 3.10 se muestra un ejemplo de cálculo de esta medida.

Uno de los aspectos positivos de esta medida es que el valor *AUC* puede ser calculado en paralelo con el proceso de aprendizaje y evaluación, por lo que es posible representar una gráfica a medida que estos datos varían. De esta forma, es posible observar la rapidez con la que el sistema aprende, y el tiempo que tarda en llegar a una efectividad estable.

Por otro lado, esta medida se puede interpretar como la probabilidad de que un clasificador ordene o puntúe una instancia positiva elegida aleatoriamente más alta que una negativa. Se puede demostrar que el área bajo la curva ROC es equivalente a la *Prueba de Mann-Whitney*, una prueba no paramétrica aplicada a dos muestras independientes, cuyos datos han sido medidos al menos en una escala de nivel or-

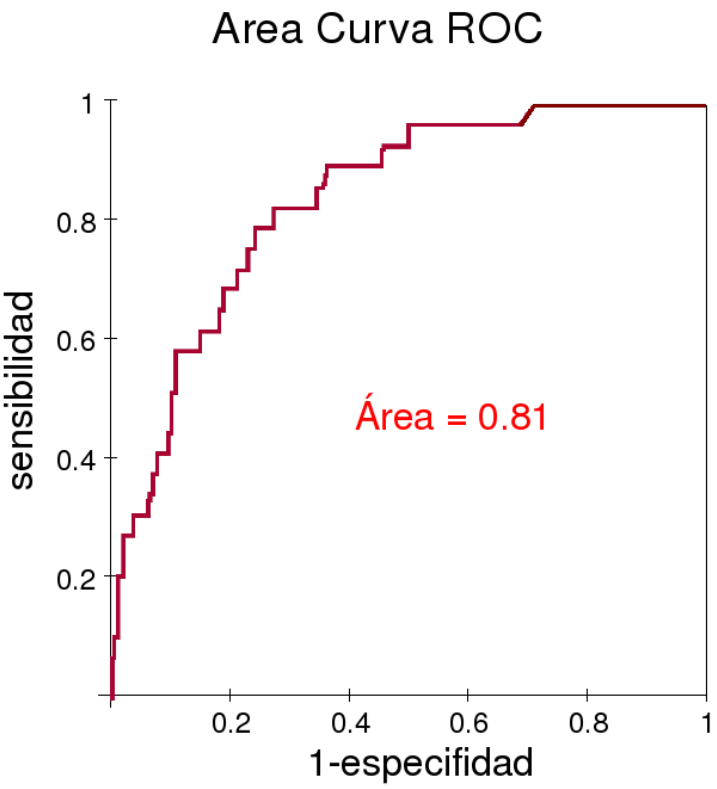

Figura 3.10: *Ejemplo de cálculo de la medida (AUC) que representa el área total bajo la curva ROC.*

dinal. Se trata de una prueba estadística virtualmente idéntica a la realización de una prueba paramétrica ordinaria T de dos muestras en los datos después de haber ordenado las muestras combinadas.

3.3.5. Colecciones de Referencia

En esta sección se describe la principal colección de referencia dentro del área del web spam. De esta forma, la colección que presentamos a continuación será la utilizada en nuestro trabajo para evaluar las cualidades del sistema de detección de spam.

Durante varios años, la investigación en el área del spam ha carecido de una colección de referencia, obligando a los investigadores a construir sus propias colecciones para realizar la experimentación. Desde el año 2000 hasta el 2006, varios grupos de investigación generaron colecciones[CDB+06] de este tipo, variando el tamaño desde los 7 millones de páginas hasta los 1500 millones. También la cantidad de elementos anotados variaron desde los 750 enlaces hasta las 17000 páginas.

En Mayo de 2006 el *Laboratorio de Algoritmos Web*[6] de la *Universidad de Milán*, comenzó a crear la actual colección de referencia[CDB+06] en el área del web spam, realizando un crawling sobre una semilla de 190000 páginas del dominio *.uk*. La colección resultante reunió 78 millones de páginas, 3000 millones de enlaces y 11400 hosts. Esta colección fue almacenada con el formato WARC[7], que comparte con el *Internet Archive*, y alcanzó un tamaño en disco de 8 volúmenes comprimidos de 55GB cada uno.

Uno de los aspectos más significativos fue el etiquetado colaborativo que se propuso para que investigadores del área ayudaran en esta labor. Previamente a esta fase de etiquetado se les había presentado a los voluntarios una serie de ejemplos de los tres tipos de etiquetas (spam, normal y dudoso), que posteriormente debían asignar. Para tal fin se desarrolló una aplicación web en la que al usuario se le presentaban dos paneles. En el panel de la derecha y de manera aleatoria, aparecían un conjunto de al menos 200 sitios web junto con algunas de sus características tales como el número de enlaces entrantes y salientes y alguna otra información que pudiera ayudar a establecer un juicio. En el panel de la izquierda, se proporcionaba una interfaz en la que el usuario debía asignar un juicio acerca de la página que se encontraba cargada en el panel de la derecha. En este panel de votación, el usuario tenía la posibilidad de asignar tres tipos de etiquetas: spam, normal y dudoso. Este proceso se alargó varias semanas involucrando a 33 voluntarios, proporcionando un total de 6552 valoraciones y consiguiendo el etiquetado de 2725 sitios web. Las estadísticas de la colección establecieron que el 62 % de las páginas fueron etiquetadas como normales, un 22 % como spam y un 11 % como dudosas.

La colección `WEBSPAM-UK2006` se hizo posteriormente pública[8], convirtiéndose a partir de entonces en la colección de referencia en este área de investigación y siendo la base de una competición denominada "Web Spam Challenge". Esta competición evaluaba a diferentes sistemas de detección de spam y estaba enmarcada en el workshop AIRWeb (Adversarial Information Retrieval on the Web)[9]. El año siguiente se hizo pública la siguiente versión de esta colección (`WEBSPAM-UK2007`), aumentando su tamaño hasta alcanzar los 106 millones de páginas, 3700 millones de enlaces y 114529 hosts.

3.3.6. Entropía Relativa

La entropía relativa ha sido utilizada en esta tesis como una herramienta fundamental a la hora de detectar páginas enlazadas de manera sospechosa, cuyos contenidos son significativamente diferentes. La divergencia en el contenido de estas páginas ha sido utilizado por tanto como un indicador de spam en el trabajo que se presenta en este capítulo.

[6] http://law.dsi.unimi.it

[7] http://www.niso.org/international/SC4/N595.pdf

[8] http://www.yr-bcn.es/webspam/

[9] http://airweb.cse.lehigh.edu

El concepto de entropía tiene dos ámbitos de uso distintos en la ciencia actual, aunque en lo fundamental se refieran a un mismo concepto. Por un lado, en la termodinámica y en la física en general, designa el grado de evolución (orden) existente en un sistema. Así, la segunda ley de la termodinámica afirma que, para todo sistema cerrado, la entropía siempre tiende a aumentar, es decir, que todo sistema cerrado siempre tiende al desorden o a la incertidumbre estadística.

De otra parte, en la teoría de la información, la entropía designa la información media emitida por una fuente de información, es decir, es una medida de la libertad de elección que muestra dicha fuente. En ambos casos, la definición formal de la entropía es la misma puesto que se basa en el cálculo de probabilidades.

Para nosotros, este concepto tiene interés tanto en su dimensión física, como en su dimensión informativa. Por un lado, la composición de un documento es un fenómeno físico y nuestro objetivo es determinar precisamente en que medida existe orden en dicho comportamiento. Pero tampoco debemos olvidar que se trata de un comportamiento fundamentalmente orientado hacia la transmisión de información, y que, como veremos más adelante, esta dimensión va a ser crucial para la comprensión de los resultados de nuestra investigación.

Dado un sistema con M sucesos posibles, la entropía se define como

$$H = \sum_{j}^{M} p_j \, log(\frac{1}{p_j}) \tag{3.3}$$

donde p_j es la probabilidad del suceso j.

La entropía relativa o divergencia de Kullback-Leibler es un indicador de la similitud entre dos funciones de distribución. Dentro de la teoría de la información también se la conoce como divergencia de la información o ganancia de la información.

Aunque normalmente se utilice el concepto de entropía como sinónimo de desorden, hay que tener presente que se trata de una simplificación en el lenguaje y que realmente es la entropía relativa la que evalúa el desorden de un sistema y la que permite hacer comparaciones entre sistemas distintos. La entropía relativa es independiente de las magnitudes absolutas del sistema que se esté considerando y carece de unidad.

La divergencia de Kullback-Leibler entre dos funciones de distribución P y Q suele representarse así:

$$D_{KL}(P||Q) \tag{3.4}$$

El concepto de divergencia de *Kullback-Liebler*[CT91], introducido también en la sección 2.3.4, modela la distribución de los términos de un documento con el objetivo de calcular la divergencia entre las distribuciones de probabilidad de los términos en una colección y en un documento en particular, o entre dos unidades

de texto. De esta forma la forma general de la divergencia de *Kullback-Liebler* se representa mediante la siguiente ecuación:

$$KLD(T_1||T_2) = \sum_{t \in T_1} P_{T_1}(t) log \frac{P_{T_1}(t)}{P_{T_2}(t)} \tag{3.5}$$

donde $P_{T_1}(t)$ es la probabilidad de que el término t se encuentre en la primera unidad de texto T_1, y $P_{T_2}(t)$ es la probabilidad de que el término t se encuentre en la segunda unidad de texto T_2.

Parte III

Estudios Empíricos

CAPÍTULO **4**

Recuperación Automática de Enlaces Rotos

4.1. Introducción

Como se ha expuesto en el capítulo2, los enlaces en la Web que apuntan a una página que no está disponible, suelen ser la consecuencia de páginas que han desaparecido o han sido movidas a otra dirección. La acojida por parte de las páginas web de este tipo de enlaces tiene varias consecuencias, como la incidencia negativa en el ranking que le otorga el buscador, y por otro lado el efecto disuasorio que lleva a cabo entre los visitantes, que pueden considerar la página que contiene estos errores como obsoletas o poco rigurosas. Existen múltiples sistemas para comprobar la validez de los enlaces de una página, de tal forma que una vez que el propietario de la página ha detectado un enlace roto, ya sea manualmente o utilizando uno de estos sistemas de control, la responsabilidad de actualizar o eliminar el enlace recae sobre él. Sin embargo, es una tarea que requiere cierta dedicación y que además se tiene que realizar con frecuencia. Otro de los escenarios que podemos encontrarnos cuando se navega a través de la web es el hecho de intentar acceder a determinados enlaces que visitamos en el pasado y que nos resultaron interesantes, pero que a la hora de volver a consultar dicho recurso, este ha dejado de funcionar. En esos casos se puede intentar una búsqueda en la Web utilizando información acerca del propio enlace roto y de la página que contiene dicho enlace. Pero esto es un trabajo tedioso que proponemos llevar a cabo de una forma automática, o al menos asistir al usuario en su búsqueda.

En muchas ocasiones habremos comprobado que el enlace roto que al que intentamos acceder no ha desaparecido realmente, sino que ha sido movido dentro de la estructura del mismo sitio web, o incluso a otro sitio web distinto. También hemos podido observar que incluso en los casos en que la página ha desaparecido, es posible encontrar páginas muy relacionadas que realizan la misma función informativa que la página inicial y que por tanto podrían ser utilizadas para reparar el enlace roto.

En la mayoría de los casos, la página que contiene el enlace roto proporciona una gran cantidad de información acerca de la página que apuntaba: el texto del ancla, el texto alrededor del enlace, la propia Url y el texto de la página donde se encuentra el error. Pero también podemos utilizar otros recursos de la infraestructura Web: una versión almacenada en la caché de un motor de búsqueda o en un repositorio Web (*Wayback Machine*), las herramientas públicas proporcionadas por los buscadores e información proporcionada por sitios web de etiquetado social (*Delicious*).

En esta tesis hemos desarrollado un sistema de recomendación de páginas candidatas a sustituir un enlace roto. El sistema comprueba los enlaces de la página dada como entrada. En el caso de los enlaces rotos, el sistema determina si tenemos suficiente información disponible para realizar una recomendación fiable. Si es así, el sistema proporciona al usuario una serie de páginas candidatas para sustituir aquella desaparecida. Este método aplica técnicas de recuperación de información para extraer los términos más relevantes a partir de varias fuentes de información. Las páginas candidatas se obtienen mediante una serie de consultas a un motor de búsqueda compuestas con los términos extraídos de las diferentes fuentes de información y aplicando técnicas de expansión de consultas. Con el objetivo de afinar los resultados, se realiza un proceso de filtrado y ranking con los resultados recuperados de acuerdo a una serie de medidas obtenidas a partir de técnicas de recuperación de información. Finalmente, las páginas candidatas son presentadas al usuario mediante una lista ordenada de resultados.

Resulta evidente que el sistema desconoce el propósito final del autor de una página web a la hora de insertar un enlace, o el interés real de un usuario que busca un determinado recurso. Es por ello que el sistema no trata de reemplazar automáticamente un enlace roto, sino que su función es recomendar una lista ordenada de páginas candidatas que están muy relacionadas con la página desaparecida, y la decisión final de reemplazar el enlace sería siempre del propietario de la página en cuestión.

El primer paso en el desarrollo del sistema ha sido elaborar un análisis acerca de un gran número de páginas web y sus enlaces, con el fin de determinar cuáles son las fuentes de información más útiles y cuáles son las más apropiadas en cada caso. Este estudio nos ha permitido extraer criterios para determinar, en el caso de un enlace roto en particular, si tiene sentido realizar una búsqueda con la información disponible, o si dicha información no es suficiente para lograr una recomendación efectiva. En algunas ocasiones es posible recuperar una página desaparecida simplemente empleando el texto del ancla para realizar una consulta en un buscador. Sin embargo, hay muchos casos en que el texto del ancla no contiene suficiente información para conseguir resultados óptimos. En estos casos, podemos realizar una expansión de la consulta inicial añadiendo términos extraídos de otras fuentes de información.

Antes de entrar en profundidad en los detalles del sistema, debemos establecer un marco de evaluación para poder establecer comparaciones de rendimiento entre

las diferentes alternativas contempladas. De esta forma, hemos desarrollado una metodología que se basa principalmente en la selección aleatoria de páginas y en el uso de sus enlaces activos, lo que permite comprobar en el hipotético caso de que se rompieran en realidad, si nuestras técnicas serían capaces de recuperar la página correcta.

4.2. Diseño del Sistema

El diseño del sistema de recuperación de enlaces rotos (SRER) se compone de cuatro etapas diferenciadas. En la primera etapa se obtiene el texto de las fuentes de información, después se realiza una extracción de terminología para a continuación llevar a cabo una batería de consultas y finalmente realizar un ranking de los documentos obtenidos y presentar los resultados al usuario.

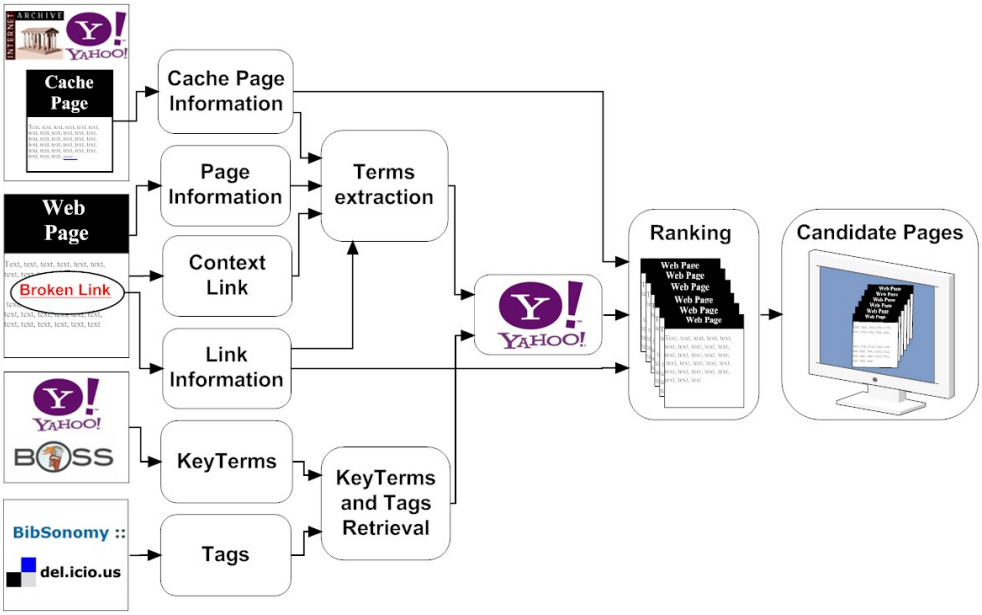

Figura 4.1: *Diseño del sistema de recuperación de enlaces rotos.*

La Figura 4.1 presenta un esquema del diseño propuesto. Cuando se encuentra una página web con algún enlace roto, tanto el texto del ancla como la página analizada permiten extraer términos que pueden estar relacionados con el contenido de la página desaparecida. De esta forma, los términos más relevantes de cada una de estas fuentes de información son utilizados para construir una serie de consultas con el objetivo de ser ejecutadas en un buscador. El sistema SRER adopta algunas técnicas de expansión de consultas[Eft96], ya que resulta ser un método bien conocido para mejorar el rendimiento de los sistemas de recuperación de información.

En este método, los términos de la expansión tienen que ser cuidadosamente seleccionados para evitar un empeoramiento en el rendimiento de las consultas. Muchos enfoques propuestos para la ampliación de consultas utilizan colecciones externas[Voo06, Voo05, Voo03], como documentos de la Web, para extraer los términos candidatos a la expansión. Hay otros métodos para extraer los términos candidatos de la misma colección de la que se extrae la consulta. Algunos de estos métodos se basan en el análisis global, donde la lista de términos candidatos se genera utilizando toda la colección, pero son muy costosos computacionalmente y su eficacia no es mejor que uno de los métodos basados en análisis local[QF93, QF95, JC94, SP97].

En lo que respecta a esta tesis, la consulta original consiste en los términos extraídos del texto del ancla del enlace, y las fuentes de los términos de expansión son los elementos de la página web que contiene el enlace roto (texto, Url, contexto, etc.), y también, si es que existen, una versión de la página desaparecida que se pueda encontrar almacenada en la caché de algún motor de búsqueda y otros recursos extraidos de sitios de etiquetado social y buscadores. Existe un gran número de trabajos que han analizado la importancia del texto del ancla como una fuente de información. Eiron y McCurley[EM03] llevaron a cabo un estudio en el que comparan la utilidad del texto del ancla frente al contenido de la página en una tarea de búsqueda en la Web. Este estudio muestra varios aspectos del texto del ancla, incluyendo la frecuencia de las consultas y los términos que aparecen con una mayor frecuencia en el texto del ancla, el contenido de la página y el título.

El texto del ancla también se ha utilizado para una tarea del TREC en la que se trataba de encontrar sitios web[CHR01], pero esta fuente de información por lo general ha sido usada como una manera de representar una página mientras que en esta tesis, el texto del ancla se utiliza como una fuente de información cuyo objetivo es re-encontrar una página desaparecida. También se ha estudiado el comportamiento de distintas técnicas de extracción de terminología empleadas para la selección de los términos implicados en la expansión de consultas. Algunos de estos enfoques se basan en las frecuencias de términos, mientras que otros se basan en modelos de lenguaje para calcular las diferencias entre la distribución de probabilidad de los términos de una colección y la fuente de información considerada.

Después del proceso de extracción de términos, se realiza una batería de consultas en un motor de búsqueda, y los primeros resultados obtenidos tras cada consulta son recuperados y almacenados con el objetivo de establecer un ranking con todos ellos. También hemos investigado diferentes enfoques para este proceso. Uno de ellos es el modelo de espacio vectorial, según el cual los documentos son ordenados atendiendo a su similitud con un documento de referencia como la página donde se encuentra en enlace roto o una versión en caché de la página desaparecida. La similitud puede ser determinada mediante diferentes medidas normalizadas (Coseno, Dice, Tanimoto) entre los vectores correspondientes. Los modelos de lenguaje son considerados también para esta tarea, construyendo un modelo de len-

guaje probabilístico para cada documento, y realizando la ordenación en función de la probabilidad del modelo que genera la consulta.

4.3. Metodología de Evaluación

La evaluación de la calidad de los resultados mostrados por el sistema, es una tarea difícil ya que en la mayoría de los casos se desconoce el contenido de la página desaparecida por completo. Por otro lado el análisis de las fuentes de información utilizadas también resulta una labor compleja al utilizar directamente enlaces rotos. Por lo tanto, se propone una nueva metodología que emplea hiperenlaces extraidos de la Web de manera aleatoria y que realmente están activos. El objetivo en esta fase de análisis de utilizar enlaces que realmente no están rotos es poder depurar cada una de las fases de desarrollo mediante la comprobación de los enlaces propuestos por el sistema. La idea principal es que al poder contar con la página que estamos intentando recuperar, resulta sencillo evaluar tanto si el sistema ha conseguido recuperar la página que estamos buscando (y que realmente no ha desaparecido) como si el resto de páginas que el sistema recomienda se asemejan a la página requerida.

4.3.1. Selección de Enlaces de Prueba

Para llevar a cabo el análisis del sistema, se toma un conjunto de enlaces de prueba seleccionados al azar mediante una sucesión de peticiones al sitio *www.randomwebsite.com* (un sitio que proporciona páginas web al azar). Inicialmente se imponen una serie de requisitos a las páginas seleccionadas con el objetivo de asegurar unos mínimos de disponibilidad de información. En primer lugar se restringe el uso del idioma al inglés principalmente por la dificultad de evaluar los resultados de otros idiomas. En cuanto al contenido de cada página, es necesario que cuenten con al menos 250 palabras para poder caracterizar el texto. El texto debe contener al menos diez términos que no sean palabras vacías, es decir, palabras que son tan comunes que no se tienen en cuenta en recuperación de información (e.g. artículos, pronombres, etc.). También exigimos que la página tenga al menos cinco enlaces potencialmente analizables, es decir, que tengan un mínimo de información, lo que implica:

- El sistema analiza enlaces externos, por tanto los enlaces que apuntan al mismo sitio son descartados.
- El texto del ancla del enlace no puede ser una secuencia de números, una Url o una cadena vacía.
- Si el texto del ancla está compuesto solamente por un carácter y además es un signo de puntuación, el enlace es descartado.

Algunos experimentos preliminares realizados indican que es frecuente encontrar páginas en las que más del 95 % de los enlaces están activos y otras en las que la mayoría de los enlaces están rotos. En la colección de páginas aleatorias empleada en esta tesis, las páginas tienen un promedio de 17 enlaces. De ellas tan sólo un 9 % tienen más de 100 enlaces y un 57 % no exceden de 10. Dados estos valores, se puede dar el caso de que las páginas que tienen muchos enlaces (e.g. 1000 enlaces), produzca un sesgo de los resultados en un sentido u otro, debido a la gran proporción de los enlaces de estas páginas en el total de los enlaces analizados. Debido a esto, hemos decidido limitar a diez el número de enlaces extraidos de cada página. Este subconjunto de enlaces es seleccionado de forma aleatoria entre los enlaces analizables de la página. Después de esta fase de recuperación y filtrado, disponemos de una colección de 1000 enlaces para estudiar y evaluar nuestro sistema en las siguientes secciones.

4.3.2. Similitud de la Página Recomendada

Una de las claves para decidir si una página está relacionada con otra y sus contenidos son similares es la selección de una medida de similitud y un umbral para descartar casos erróneos. Para la selección de esta medida de similitud se ha realizado una serie de experimentos con el objetivo de poder considerar que un enlace ha sido recuperado y que las página candidata y la desaparecida tienen una semejanza muy alta. En primer lugar, y pensando solamente en la batería de pruebas para la evaluación de las fuentes de información mediante el uso de enlaces activos se verifica si la Url de la página candidata coincide con el enlace analizado (se recuerda que en este primer análisis el enlace está realmente activo). Por otro lado, hemos encontrado algunos casos en los que la página que se ha recuperado tenía exactamente el mismo contenido que el enlace analizado, pero diferente Url. De esta forma, si las direcciones Url no coinciden, se verifica si el contenido de la página web es la misma. Otro detalle relevante para esta tesis es que hemos encontrado varios casos en los que el contenido de las páginas no era idéntico, pero eran muy similares. En estos casos la única variación eran algunos pequeños cambios en el contenido de dichas páginas (analizada y candidata) como los anuncios, fechas, etc.

Por este motivo, hemos decidido utilizar una medida de similitud que no compare simplemente el contenido de las páginas como una cadena de texto, sino que analice término a término el texto con el objetivo de evitar estos casos citados anteriormente en los que existen pequeñas variaciones que no afectan globalmente a la información proporcionada. De esta forma se ha optado por aplicar el modelo de espacio vectorial [MRS08]. En este modelo cada página es representada como un vector de términos y la similitud es calculada mediante el valor de la distancia coseno entre los dos vectores. Antes de decidir el uso del modelo de espacio vectorial, se han analizado otras alternativas como la técnica de *shingling*[BDGM95], que toma un conjunto de términos contiguos o *shingles* de cada documento y compara el número de coincidencias. También podrían utilizarse otros métodos para

esta tarea como *PageSim*[LLK06] o *SimRank*[JW02]. Finalmente, en esta tesis se ha decidido el uso del modelo de espacio vectorial ya que la técnica de *shingling* está orientada al uso en detección de copias y *PageSim* y *SimRank* acarrean un coste computacional muy alto con respecto al uso del modelo de espacio vectorial.

Una vez decidido el uso del modelo de espacio vectorial y el cálculo de la distancia coseno, como la medida de similitud empleada para la comprobación de la similitud con la página desaparecida, en el resto del trabajo, es momento de analizar el umbral de este valor con el objetivo de afinar su efectividad. Con este objetivo hemos realizado una evaluación manual para valorar la efectividad del umbral en cada caso y los resultados se muestran en la Tabla 4.1. Para este estudio se ha empleado el texto del ancla de cada enlace analizado para realizar una consulta en el motor de búsqueda. Por cada enlace se han recuperado una serie de resultados utilizando un umbral diferente para seleccionar las primeras 50 páginas candidatas. En el caso de que la página que se estaba buscando no se encuentre en los 30 primeras posiciones se considera que el enlace no ha sido recuperado.

Umbral de Similitud	1ª pos.	1-10 pos.	Enlace No Recuperado
0.9	253	380	536
0.8	256	384	529
0.7	258	390	521
0.6	262	403	504
0.5	266	425	478

Tabla 4.1: *Resultados al realizar una búsqueda con el texto del ancla en el motor de búsqueda en relación al umbral de similitud empleado. La primera columna indica el umbral de similitud, 1ª pos. representa el número de enlaces recuperados en la primera posición, y 1-10 pos. la cantidad recuperada entre las diez primeras posiciones. Los enlaces no recuperados son aquellos que no aparecen en las 30 primeras posiciones.*

Como se puede comprobar en la Tabla 4.1, si se utiliza un umbral de similitud superior a 0.9, 253 enlaces son recuperados en la primera posición y 380 en las diez primeras posiciones. La reducción del umbral de similitud añade una cantidad no significativa de enlaces recuperados con respecto al valor de 0.9. Además la reducción de este valor aumenta el número de resultados erróneos, incluyendo documentos que no corresponden a la página solicitada entre los resultados. Por estas razones, hemos establecido un umbral de similitud de 0.9 ya que a pesar de ser un valor elevado, al menos se garantiza que la página recuperada es muy próxima a la solicitada. Es cierto que con umbrales más bajos podrían obtenerse mejores resultados ya que en muchos casos algunas páginas recuperadas contienen la información necesaria para reemplazar a una página desaparecida, pero también es cierto que de manera automática no se puede asegurar esta relación. Por tanto, los resultados

que mostraremos a partir de ahora en este trabajo corresponden a un análisis conservador pudiendo obtenerse mejores resultados mediante el ajuste del umbral de similitud.

4.4. El Texto del Ancla de un Enlace

En muchos casos los términos que componen el texto del ancla de un hiperenlace son la principal fuente de información para identificar la página apuntada por dicho hiperenlace. Para comprobar esta teoría, se puede observar el estudio previo sobre el umbral de similitud plasmado en Tabla 4.1. Esta tabla muestra el número de enlaces que se consiguen recuperar entre los diez primeros resultados obtenidos por el motor de búsqueda. Podemos observar que con el uso de un umbral de similitud del 0.9, se consigue recuperar el 41 % de enlaces entre los diez primeros resultados. Además, el 66 % de los enlaces recuperados aparecen en la primera posición. Estos resultados confirman que el texto del ancla es una fuente de información fundamental a la hora de recuperar un enlace roto a partir del rastro de información que deja esa página desaparecida. Por lo tanto, el texto del ancla de un hiperenlace será la principal fuente de información empleada para realizar las consultas en el buscador. Además, estas consultas serán expandidas con términos de otras fuentes de información analizadas en secciones posteriores.

4.4.1. Reconocimiento de Entidades Nombradas

En algunas ocasiones, el texto del ancla tiene un escaso valor descriptivo. Imaginemos un enlace roto cuyo texto de anclaje es el siguiente: “pinche aquí”. En este caso, encontrar la página desaparecida podría ser imposible. Por este motivo, es muy importante analizar estos términos de forma que el sistema sea capaz de decidir qué tareas se deben realizar en función de la cantidad y calidad de dichos términos.

En esta tesis se ha decidido llevar a cabo un reconocimiento de entidades nombradas (personas, organizaciones o lugares) en el texto del ancla a fin de extraer algunos términos cuya importancia sea mayor que el resto. Existen varias soluciones de software para esta tarea como *LingPipe*, *Gate*, *FreeLing*, etc. También existen múltiples recursos, como los *gazetteers*. Pero después de diferentes experimentos, ninguna de estas soluciones han proporcionado resultados precisos a la hora de trabajar con los textos de anclaje. Después de un análisis exhaustivo, hemos llegado a la conclusión de que estos sistemas necesitan un dominio más restringido que el que proponemos en este trabajo. El sistema que hemos desarrollado trata de recuperar páginas web de cualquier ámbito y contenido, por tanto el dominio sobre el que estamos trabajando es muy amplio y eso provoca numerosos problemas a los sistemas de reconocimiento de entidades nombradas disponibles en la actualidad. Además, el tamaño del texto del ancla es muy reducido para este tipo de análisis. En

general, los sistemas de detección de entidades nombradas cuentan con un contexto en el que pueden analizar determinadas características y reglas. Sin embargo, a la hora de recuperar un enlace, es habitual encontrar su texto de anclaje aislado, por ejemplo formando parte de un menú.

Por lo tanto, hemos decidido utilizar la estrategia opuesta. En lugar de encontrar entidades nombradas, hemos optado por compilar un conjunto de diccionarios con el objetivo de descartar las palabras comunes, suponiendo que el resto de las palabras son entidades nombradas. Aunque en las pruebas realizadas hemos encontrado algunos falsos negativos, como por ejemplo la empresa "Apple", en líneas generales hemos obtenido mejores resultados utilizando esta técnica.

Una vez establecido el mecanismo para la detección de entidades, hemos llevado a cabo un análisis con el objetivo de discriminar el valor de los términos del texto de un ancla. La Tabla 4.2 muestra el número de enlaces recuperados en función de la presencia de entidades nombradas y el número términos. Se puede observar que cuando el enlace no contiene ninguna entidad nombrada, el número de enlaces que no son recuperados es mucho mayor que el número de los recuperados, mientras que los valores son similares cuando existen entidades nombradas. Este hecho inicial demuestra que la presencia de cualquier entidad nombrada en el texto del ancla favorece la recuperación del enlace. El resultado más destacado es el número tan reducido de casos en los que el documento correcto se recupera cuando el anclaje se compone de un solo término, y además no es una entidad nombrada[1]. Por otro lado, cuando el ancla contiene entidades nombradas, incluso si sólo hay una, el número de enlaces recuperados es significativo. Otro hecho que también puede ser observado es que a partir de dos términos, el número de términos de anclaje no representan un gran incremento en los resultados.

	Tipo de Ancla			
	Entidades Nombradas		**Sin Entidades Nombradas**	
# Términos	**E. N. R**	**E. R**	**E. N. R**	**E. R**
1	102	67	145	7
2	52	75	91	49
3	29	29	27	45
4+	57	61	33	47
Total	240	232	296	148

Tabla 4.2: *Análisis de los enlaces no recuperados E.N.R y recuperados E.R en función del tipo de ancla — con entidades nombradas y sin entidades — y el número de términos que componen el ancla. En el caso de "4+", se refiere a cuatro términos o más.*

[1]Los casos positivos son generalmente dominios con un nombre común como por ejemplo el ancla "Flock", que ha permitido recuperar www.flock.com

4.5. Principales Fuentes de Información

La principal fuente de información a la hora de recuperar un enlace roto, se ha demostrado que es el texto del ancla. Pero esta información en ocasiones puede ser ambigua y proporcionar consultas diferentes al propósito de la página solicitada. Incluso, como ya hemos visto anteriormente, los términos que componen el ancla pueden no tener ningún valor descriptivo. Por este motivo, es necesario completar la información que proporciona el texto del ancla con otros recursos que permitan por un lado enfocar las consultas ambiguas, y por otro lado otorgar información descriptiva del recurso desaparecido.

Tanto en el contexto de un enlace como en la Web, existen fuentes de información que proporcionan términos de gran interés para la tarea de recuperación de una página desaparecida. Estas fuentes de información tienen características diferentes, tanto en el vocabulario que emplea cada una de ellas, como en la longitud en términos del número de palabras que contiene. La forma de seleccionar estos términos depende de cada tipo de fuente de información, por tanto el propósito de cada una de estas fuentes de información es proporcionar unos pocos términos de gran valor descriptivo, con el objetivo de ser utilizados para expandir las consultas de la manera más eficiente. Por esta razón, en el caso de las fuentes de información con un contenido amplio de texto, la mejor manera de extraer la información más relevante es mediante la selección de un número mínimo de términos para cada una de estas fuentes.

4.5.1. Dirección de Destino de un Hiperenlace

Un hiperenlace tan solo tiene dos unidades de texto que pueden proporcionar alguna información acerca de la página apuntada, el texto del ancla y la Url de destino. Los términos de una dirección Url son muy representativos con respecto del contenido de la página apuntada. De hecho, la mayoría de los motores de búsqueda utilizan los términos Url como uno de los principales parámetros a la hora de elaborar el ranking de relevancia de una página en relación a una consulta.

Además, los términos de la Url pueden ser una fuente de información muy útil en el caso de que una página haya sido movida dentro del mismo sitio, o si por el contrario la página está en un sitio diferente pero al menos mantiene el mismo identificador de usuario, servicio, nombre de la aplicación, recursos, etc.

Una Url se compone principalmente de un protocolo, un dominio, una ruta y un archivo. Estos elementos a su vez se componen de términos que pueden proporcionar información muy valiosa sobre la página de destino. Además, en los últimos años, debido al uso creciente de los buscadores, existen técnicas de optimización de motores de búsqueda (Search Engine Optimization —SEO —) que tratan de realizar ingeniería inversa sobre los algoritmos que utilizan los principales buscadores, con el objetivo de conocer los parámetros que intervienen en el ranking proporcionado y su peso aproximado. Formando parte de estas optimizaciones, se encuentra

la estrategia de colocar determinados términos en la Url de cada página, con el objetivo de conseguir un ranking mayor, aprovechándose de la importancia que los motores de búsqueda otorgan a dichos términos.

Para seleccionar los términos más relevantes de una Url, es importante excluir ciertos términos que son generalmente muy frecuentes en las Urls, como por ejemplo *free*, *mp3*, *download*, etc. Para extraer los términos más relevantes, se ha aplicado un enfoque basado en modelos de lenguaje. En primer lugar, se ha construido un modelo de lenguaje con los términos de las Urls almacenadas en el directorio público `ODP` (Open Directory Project). Para ello han sido indexadas las Urls de este directorio, eliminando las palabras vacías y consiguiendo reunir una colección con los términos de dichas direcciones. El objetivo es utilizar esta colección como referencia para el cálculo de probabilidades que forma parte de la divergencia de *Kullback-Leibler*, que a su vez será la medida utilizada para la extracción de la terminología relevante sobre las direcciones de destino de los hiperenlaces.

4.5.2. Contenido de la Página Analizada

Los términos más relevantes de una página web son una forma de caracterizar el tema principal de dicha página. Pero para aplicar alguna técnica de extracción de terminología con éxito es necesario que el texto de la página sea lo suficientemente largo como para poder discriminar los términos relevantes de los que no lo son. Un claro ejemplo de la utilidad de esta información son los enlaces a las páginas personales. El ancla de un enlace a una página personal suele estar formado por el nombre de la persona a la que corresponde la página. Sin embargo, en muchos casos, el nombre y el apellido no identifica a una persona de una manera única[AGS07], especialmente si son muy comunes. Si realizamos una consulta en un motor de búsqueda solamente con el nombre y un apellido, la página personal de esta persona probablemente no vaya a aparecer entre las primeras páginas recuperadas. Sin embargo, si ampliamos la consulta con algunos términos que caractericen en mayor medida a esa persona mediante la extracción de términos de su página web, entonces su página web ocupará con una alta probabilidad las primeras posiciones de la consulta.

En lineas generales, los términos extraidos de la página donde se encuentra en enlace roto (también denominada *página padre*) no incidirán en la precisión de los resultados de búsqueda. Esto se debe a que la información que puede ser extraida de dicha página no tiene porqué tener una estrecha relación con la página apuntada. A pesar de esto, también es cierto que existe la intuición de que dos páginas enlazadas deberían tener algún tipo de relación. De esta forma, el uso de los términos de la página donde se encuentra en enlace, tienen la función de desambiguar aquellas consultas que por las características del texto del ancla, tienen grupos de diferente naturaleza representados en los resultados obtenidos.

4.5.3. Contexto del Enlace

En ocasiones los textos de anclaje no tienen suficientes términos o estos son muy genéricos, y por tanto su valor descriptivo es practicamente nulo. Imaginemos un enlace cuyo texto del ancla es "nlp group". A la hora de realizar una consulta, a parte de los distintos significados que pueda tener la palabra "nlp" en la Web, aparecerán resultados con distintos grupos de procesamiento de lenguaje natural en el mundo. Por esta razón, el texto alrededor de un enlace puede proporcionar información contextual sobre la página apuntada y ofrecer la información necesaria para completar una búsqueda más específica. Imaginemos ahora que en el enlace se encuentra la siguiente frase "conference organized by nlp group at uned". En este caso, al extraer el término "uned" del contexto, la expansión de la consulta inicial nos devolvería el grupo de procesamiento de lenguaje natural de la Uned como primer resultado.

Por otra parte, Benczúr et al.[BBCU06] midieron la relación entre el contexto de un enlace y la página apuntada, consiguiendo un gran rendimiento cuando el texto del ancla fue ampliado con las palabras vecinas del hiperenlace. En los experimentos realizados en esta tesis, se han utilizado veinte palabras del contexto de cada enlace (10 palabras en cada dirección) para expandir las consultas. A partir de este texto extraído, también se aplicará un proceso para seleccionar los términos más relevantes. Estas palabras han sido extraidas del texto original teniendo en cuenta las etiquetas de bloque de HTML y signos de puntuación tal y como se describe en varios trabajos previos[Pan03, CS07]. Aunque el formato de los documentos Web cambia con el paso del tiempo y originalmente los enlaces se encontraban aislados en un menú, cada vez es posible encontrar más enlaces que están contenidos en un párrafo o en una frase (blogs, Wikipedia, etc.). El principal problema de esta fuente de información es el número limitado de ocasiones en las que puede encontrarse un contexto de calidad para un enlace. Aunque el contexto para un enlace sólo ha podido ser extraido para el 30 % de los enlaces analizados, cuando dicho contexto ha sido recuperado se ha conseguido una gran eficiencia en la expansión de las consultas.

4.5.4. Versión Almacenada en una Librería Digital

En la infraestructura Web es posible encontrar una copia almacenada de una página web relativamente actualizada, al mismo tiempo que se podrían consultar diferentes versiones de dicha página web a lo largo de su vida. La Web es una entidad muy joven, y como tal todavía no existen estudios complejos que analicen su historia y como ha ido evolucionando en este tiempo. Una de las principales herramientas que sin lugar a duda serán fundamentales en el análisis de esta evolución son las librerías digitales que se encargan desde hace años de almacenar versiones de una gran cantidad de sitios web con un objetivo aún sin determinar. En muchos casos la motivación de estas librerías digitales es fotografiar el estado de la Web

cada cierto tiempo, para poder situarse como en una máquina del tiempo en algún momento del pasado, cuando la sociedad requiera información de su historia. De hecho, la principal librería digital que existe se denomina *Wayback Machine* y se trata de un repositorio de versiones de millones de sitios web en el que pueden consultarse múltiples versiones de estos sitios web desde su creación en 1996.

Otro de los recursos que pueden encontrarse en la infraestructura web es lo que se denomina "caché" de los principales buscadores. Por ejemplo, *Yahoo!* y *Google* disponen de este servicio que no es otra cosa que una versión relativamente reciente (según nuestras observaciones alrededor de un mes de antigüedad) de una página web cuyo principal objetivo es poder mostrar al usuario la página requerida cuando esta se encuentra no disponible de manera temporal y a pesar de que dicha página haya podido ser modificada desde el momento de su almacenamiento.

Por los motivos expresados anteriormente, una de las fuentes de información más útiles es una página almacenada en la caché de un motor de búsqueda o en un archivo de Internet. Estos sitios web, con diferentes propósitos, almacenan al menos una copia de un gran número de páginas web con una gran parte de sus recursos (incluso multimedia) disponibles. Sin embargo, esta fuente de información tiene dos inconvenientes: la Web es cada vez más grande y los servidores de estos sitios web no son capaces de almacenar todas las páginas Web disponibles en la red. A pesar de esto, en las pruebas realizadas para este trabajo ha sido posible obtener una versión en caché entre el 50 % y 60 % de los enlaces analizados. Por otra parte, la fecha de la última versión de una página almacenada puede ser vieja en relación a la velocidad con la que avanza la Web, incluso un año en el caso de los archivos de Internet. Aunque en este último caso, la esencia de un sitio no suele cambiar, a pesar de que el contenido haya cambiado. Hay excepciones si tenemos en cuenta sitios de noticias o blogs cuyo contenido varía casi por completo en pocos días. En el caso de la versión en caché de la página desaparecida que queremos recuperar, extraemos los términos más relevantes como en el caso del contenido de la página web en la que se encuentra en enlace roto. De esta forma, los términos más relevantes extraidos de las últimas versiones disponibles en Internet son usados para expandir la consulta formada por los términos del texto del el ancla.

4.6. Extracción de Terminología

Las fuentes de información descritas en la sección anterior, para proporcionar términos de expansión a la consulta construida a partir del texto del ancla, tienen unas características muy diferentes las unas de las otras. En el caso de la dirección de destino de un hiperenlace, el número de términos que podrán ser extraidos será muy reducido y por tanto la selección de estos términos tendrá que hacerse teniendo en cuenta las propiedades del vocabulario utilizado en la creación de las Urls. Algo más extenso será el contexto de un enlace, donde el vocabulario será más parecido al de un texto común aunque con una longitud mucho más reducida. En el caso del

contenido de la página analizada y la versión en caché de la página desaparecida, las características serán muy similares, aunque la selección de los términos en casa caso debería enfocar en aspectos diferentes. Mientras que en el caso de la versión en caché los términos deberían obtener las principales características de la página desaparecida, en el caso de la página analizada el objetivo no es extraer los términos que caractericen dicha página sino captar la información más genérica acerca de la posible relación entre ambas páginas (página que apunta y página apuntada).

En todos estos casos, nuestro principal objetivo es sintetizar esta información y obtener el número mínimo de términos que representen un enlace de la manera más personalizada. Teniendo en cuenta esta idea, se han aplicado algunas de las técnicas clásicas de recuperación de información para extraer los términos más representativos. Después de quitar las palabras vacías, se obtiene una lista ordenada por relevancia de los términos extraidos. En primer lugar los términos de dicha lista se usan para expandir la consulta formada por el texto del ancla, es decir, la consulta se expande con cada uno de esos términos, y los primeros resultados devueltos por el motor de búsqueda son recuperados.

Según la fuente de información que se analiza, la calidad de los términos extraídos será diferente en función del método utilizado. Por lo tanto, hemos utilizado dos familias de métodos de extracción: los enfoques basados en la frecuencia y aquellos basados en modelos de lenguaje.

4.6.1. Métodos Basados en la Frecuencia de Términos

Los métodos basados en la frecuencia son los más sencillos para seleccionar el conjunto de términos de expansión más relevantes. Dentro de este tipo de métodos, hemos considerado los dos criterios más populares para la extracción de terminología. El primero de ellos es la frecuencia de términos (*Tf*) que simplemente atiende al número de veces que cada palabra aparece en la unidad textual. Sin embargo, hay algunos términos cuyo poder discriminativo es muy reducido o incluso nulo como descriptor de la fuente de información, a pesar de que estos términos sean muy frecuentes en el texto. La razón principal es que dichos términos también son frecuentes en muchos otros documentos de una colección o en cualquier otra página web en nuestro caso. Para tener en cuenta estos casos se aplica el conocido esquema de pesado *Tf-Idf* para un término, donde *Idf(t)* es la frecuencia inversa en el documento de ese término:

$$Idf(t) = \log \frac{N}{df(t)} \tag{4.1}$$

siendo N el tamaño de la colección utilizada, y *df(t)* el número de documentos en la colección que contiene el término t. Teniendo en cuenta que para el cálculo del valor de *Tf-Idf* es necesario el uso de una colección, en esta tesis ha sido utilizada

una versión del 2009 de los artículos de la Wikipedia en inglés[2] como colección de referencia.

4.6.2. Modelos de Lenguaje

Uno de los enfoques principales utilizados en la expansión de consultas se basa en el estudio de la diferencia entre la distribución de términos en una colección y en el subconjunto de documentos que pueden ser relevantes para una consulta. Es de esperar que los términos con poco contenido informativo tengan una distribución similar en cualquier documento de la colección. Por el contrario, los términos más representativos de una página o documento se espera que sean más frecuentes en esa página que en otros subconjuntos de la colección de referencia.

Uno de los métodos más eficaces basados en el análisis de distribución de términos utiliza el concepto de *divergencia de Kullback-Liebler*[CT91] para calcular la divergencia entre las distribuciones de probabilidad de los términos en la colección y los documentos considerados. Los términos más factibles a la hora de extender una consulta son aquellos que tienen una alta probabilidad de aparición el documento, y una baja probabilidad en la colección. La divergencia de *Kullback-Liebler* para un término t se representa de la siguiente manera:

$$KLD_{(P_P,P_C)}(t) = P_P(t)\log\frac{P_P(t)}{P_C(t)} \tag{4.2}$$

donde $P_P(t)$ es la probabilidad del término t en la página considerada, y $P_C(t)$ es la probabilidad del término t en toda la colección.

La divergencia de *Kullback-Liebler*, introducida en la sección 2.3.4, es adaptada en esta sección para la extracción de terminología relevante. El cómputo de esta media requiere de una colección de referencia con un conjunto de documentos. La relación entre esta colección de referencia y el documento analizado, es un factor muy importante en los resultados obtenidos mediante este enfoque. Por este motivo, la selección de la colección es un detalle a tener en cuenta, ya que la efectividad de los modelos de lenguaje dependen en parte de las características del conjunto de documentos usados como referencia. Uno de los principales problemas es la gran diferencia entre cada par de páginas tomadas aleatoriamente de la Web, pudiendo cada una de ellas tratar de temas totalmente opuestos, tener una estructura diferente y utilizar vocabularios con muy poca relación. Por este motivo no es fácil encontrar una colección de referencia apropiada ya que, obviamente, no podemos usar toda la Web como colección de referencia. Para estudiar el impacto de este factor en los resultados, se han utilizado tres colecciones diferentes de páginas web, indexadas con Lucene[GH04]:

[2] http://download.wikimedia.org/enwiki/

- *Enwiki*. Esta colección es un conjunto de artículos, plantillas y descripciones de imágenes extraidas de una versión del 2009 de artículos de la Wikipedia en inglés. El tamaño de esta colección es de alrededor de 3.6 millones de documentos.

- *Dmoz*. Esta colección es el resultado de un proceso de crawling sobre el conjunto de Urls presentes en el directorio público ODP (*DMOZ Open Directory Project*). El conjunto de páginas alcanza los 4.5 millones de sitios web, que se corresponde con el número de documentos de la colección ya que tan solo se ha almacenado la página principal de cada sitio.

- *Dmoz Urls*. En esta tesis, se utilizan los términos que componen una Url con el objetivo de extraer información relevante. En el caso de esta colección, los documentos indexados corresponden únicamente a los términos que componen una Url. De esta forma, se han analizado cada una de las Urls correspondientes a los 4.5 millones de sitios web de los que dispone el directorio público ODP. El objetivo de este análisis ha sido extraer los términos que componen una dirección web teniendo en cuenta la estructura de una Url.

4.6.3. Comparación de Métodos de Extracción de Terminología

Con el objetivo de establecer el método más eficiente de extracción de términos relevantes para su posterior uso en la expansión de consultas, se ha realizado una serie de pruebas para analizar el rendimiento de cada uno de los métodos expuestos en la sección anterior.

La Figura 4.2 muestra los resultados obtenidos a partir del uso de los métodos *Tf*, *Tf-Idf* y *KLD* en la extracción de términos de diferentes fuentes de información (Url, página padre, contexto y caché). En el caso del análisis de la Url como fuente de información, debido a que el número de términos que podemos extraer es muy pequeño, hemos sustituido el método *Tf* por el método *KLD* basado en una colección de Urls.

Precisamente si analizamos los métodos empleados para extraer términos relevantes de la dirección de destino de un hiperenlace, de acuerdo con los resultados obtenidos y mostrados en la Figura 4.2(i), el enfoque que tiene el mejor rendimiento es el enfoque *KLD* que ha utilizado la colección de Urls. Parece evidente que un modelo de lenguaje formado por los términos que suelen aparecer en una Url es lo más apropiado en este caso. Tanto la Url como la página que contiene el enlace roto (página padre), son dos fuentes de información que siempre van a estar disponibles, pero en el caso de la página padre, en muchos casos, tiene una gran cantidad de información que no está directamente relacionada con el enlace roto. Por lo tanto, la eficiencia de los términos extraídos de esta fuente es limitada. La Figura 4.2(ii) muestra cómo el sistema consigue recuperar más enlaces mediante uso del método *Tf*. Los motivos de que un método tan simple como *Tf* obtenga los mejores resulta-

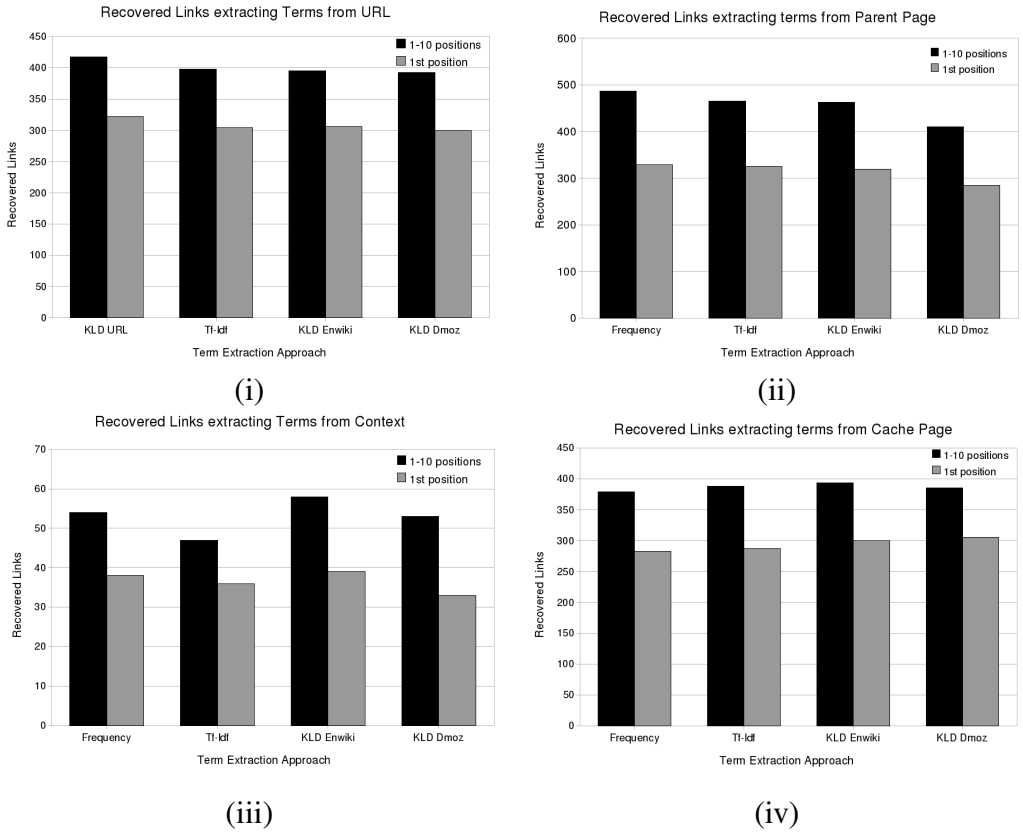

Figura 4.2: *Comparativa de métodos de extracción de terminología en donde se muestran los resultados al realizar la expansión con los términos de (i) la Url del enlace, (ii) la página padre, (iii) el contexto del enlace y (iv) la versión caché de la página desaparecida. Las barras indican el número de casos en los que el motor de búsqueda ha proporcionado la página buscada entre los diez primeros resultados y en la primera posición respectivamente, en función del método empleado. En cuanto al uso de los diferentes métodos de extracción;* `KLD URL` *corresponde al uso de KLD junto con la colección de Urls;* `Tf-Idf` *corresponde al uso del mismo método;* `KLD Enwiki` *corresponde al uso de KLD junto con la colección de artículos de la Wikipedia; y* `KLD Dmoz` *corresponde al uso de KLD junto con la colección de páginas del directorio DMOZ ODP.*

dos, pueden encontrarse en el hecho de que los términos extraídos de dicha página no tienen una gran precisión para esta tarea. De esta forma, un método sencillo basado en la frecuencia, que obtiene los términos de una forma más genérica, funciona mejor que cualquier otro método que consigue una mayor especificidad en los términos que extrae como *KLD*.

Tanto el contexto de un enlace como una versión caché de la página desaparecida son dos fuentes de información muy importantes, ya que el contexto generalmente contiene los términos necesarios para completar la información presente

en el texto del ancla, y la versión caché muestra una foto más o menos actual de lo que realmente estamos buscando. El principal problema en estos dos casos es que la disponibilidad de estas fuentes es más limitada que en el caso de la Url y la página padre. En ambos casos, tal y como se muestra en las Figuras 4.2(iii y iv), el método de extracción que obtiene el mejor rendimiento es *KLD*, utilizando a su vez la Wikipedia en inglés como colección de referencia. El éxito de este método y esta colección se debe principalmente a que los términos que permite extraer son muy precisos, que es justamente lo más adecuado en relación a la síntesis de información que se necesita extraer de estas fuentes de información.

De los resultados obtenidos, se puede concluir que el mejor método de extracción de términos para la recuperación de enlaces rotos es *KLD*. También es posible observar que mediante la aplicación de *KLD*, los resultados obtenidos con el uso de la Wikipedia como colección de referencia son mejores (el número total de páginas recuperadas correctamente). La razón es probablemente que esta colección ofrece una amplia gama de temas, aunque la presencia de páginas de spam en la colección *DMOZ ODP* también podría distorsionar el modelo de lenguaje, y por tanto la extracción de términos.

A pesar de que el método *Tf* basado en la frecuencia es el que obtiene los mejores resultados en la extracción de términos de la página padre, existe una relación limitada entre dichos términos y la página desaparecida en muchos casos. Por lo tanto, esta tarea es una aproximación con un cierto comportamiento aleatorio en el que el objetivo es encontrar los términos más cercanos en relación al contexto del enlace, o los más relacionados con la página requerida.

Además, el contenido de la página que contiene el enlace, a menudo no está estrechamente relacionado con la página apuntada. Por lo tanto resulta inútil el esfuerzo en perfeccionar los métodos para seleccionar los términos más representativos de la página principal, ya que no mejorará los resultados.

Por consiguiente, a partir de ahora los resultados mostrados en esta tesis han utilizado el método KLD apoyándose en la Wikipedia en Inglés como colección de referencia, para la extracción de términos del contexto del enlace y de la versión en caché de la página desaparecida. En el caso de la extracción de los términos de la dirección de destino del hiperenlace se ha utilizado *KLD* junto con la colección de términos extraidos de Urls. Finalmente, se ha utilizado *Tf* como el método de extracción de términos de la página que contiene el enlace roto.

4.6.4. Efecto de la Expansión de Consultas en los Resultados

Como ya se ha descrito en secciones anteriores, el sistema emplea los términos extraidos de diferentes fuentes de información para expandir la consulta formada a partir del texto del ancla del enlace roto. Con el fin de estudiar el efecto de los métodos de expansión de consultas en la relación entre precisión y la cobertura, se ha realizado un análisis de los resultados obtenidos gracias al uso de estos métodos de expansión. En la Tabla 4.3 se muestra dicho estudio en donde aparecen los enlaces

recuperados en la primera y entre las diez primeras posiciones respectivamente en función de la fuente de información empleada para llevar a cabo la expansión. En primer lugar se encuentran los enlaces recuperados sin el empleo de ninguna expansión, de tal forma que esta información sirva de referencia a la hora de analizar los datos obtenidos mediante la expansión. Según los resultados, queda probado que la expansión aumenta considerablemente el número de enlaces que se consiguen recuperar entre las primeras diez posiciones del buscador (cobertura). A pesar de ello, el número de enlaces que se consiguen recuperar en la primera posición se reduce (precisión) si lo comparamos con lo obtenido sin emplear ninguna expansión. Por lo tanto, los resultados se pueden resumir en el hecho de que la expansión de consultas aumenta la cobertura pero en cambio reduce la precisión del sistema. En consecuencia, pensamos que la metodología más apropiada debe aprovechar las ventajas de ambos métodos y combinar los resultados. Posteriormente, dado el conjunto completo de las páginas candidatas se debe aplicar una función de ranking con el objetivo de presentar al usuario una lista con los resultados ordenados, mostrando las páginas candidatas más relevantes en las primeras posiciones.

Análisis	**1ª posición**	**1-10 posiciones**
No Expansión (anchor text)	**380**	462
Expansión (términos Url)	322	417
Expansión (página padre)	329	**486**
Expansión (contexto)	39	58
Expansión (página caché)	300	393

Tabla 4.3: *Análisis del número de documentos recuperados en la primera posición y entre las diez primeras posiciones en función de las diferentes fuentes de información empleadas en la expansión de consultas.*

4.7. Terminología Adicional en la Infraestructura Web

Hasta el momento se ha establecido el texto del ancla como la unidad textual básica para componer una consulta y el entorno de un enlace, como la principal fuente de información a la hora de extraer términos relevantes para expandir la consulta original. Pero además de esta información, es posible encontrar otros recursos en la infraestructura Web. Estos recursos son herramientas disponibles públicamente en la Web cuya funcionalidad no está ideada para el objeto de esta tesis pero que mediante un determinado procesado de esa información puede resultar muy útil en la recuperación de un enlace roto. Entre estas herramientas se encuentran los servicios web que se encuentran disponibles por parte de los motores de búsqueda y los sitios web de etiquetado social.

La Web nos ofrece nuevos recursos cada día que pueden ser utilizados para obtener más información. Nuestro propósito es utilizar los recursos disponibles en la Web para obtener más información acerca de las páginas que ya no existen, pero de las que aún existe información en la Web. Los motores de búsqueda (Google, Yahoo!, Bing, etc.) ofrecen nuevos servicios y aplicaciones cada día más interesantes y útiles (en algunos casos), y que teniendo en cuenta el propósito de esta tesis, pueden resultar de ayuda a la hora de recuperar una página desaparecida. Otro fenómeno cada vez más extendido son los sistemas de etiquetado social, donde una gran comunidad está dispuesta a generar información en un entorno de colaboración. En este caso, las aportaciones de un grupo de personas participando en un sitio web de estas características, pueden ser una fuente muy importante de información para descubrir algo nuevo acerca de una página Web.

4.7.1. Servicios Web disponibles en Buscadores

Los motores de búsqueda trabajan cada día para proporcionarnos un acceso a la información de una manera más sencilla. Además, el hecho de que unas pocas compañías tengan la capacidad para indexar toda la Web (o al menos la gran mayoría) significa que algunos de los servicios que ofrecen tienen una gran utilidad debido a su globalidad.

Recientemente, la plataforma abierta de servicios de búsqueda web de Yahoo! (BOSS [3]), ofrece una nueva característica a través de su API de desarrollo: *Key Terms*. La tecnología utilizada en *Key Terms* es la misma que se utiliza en la búsqueda asistida de Yahoo!, tal y como se muestra en la Figura 4.3. Esta funcionalidad, que está incluida en el buscador, ofrece sugerencias de búsqueda y permite a los usuarios explorar los conceptos relacionados con la consulta. Por su parte, el servicio *Key Terms* utiliza frecuencia de términos y heurísticas posicionales y contextuales para mostrar listas ordenadas de términos que describen una página web. Cada resultado devuelto por una consulta incluye metadatos asociados de hasta 20 términos que describen ese resultado.

La manera en que el sistema presentado en esta tesis utiliza dicho servicio es mediante la inserción de una consulta con la Url del enlace roto. A partir de esa consulta, *Key Terms* devuelve una lista ordenada con los términos que describen esa página, y a continuación son usados los primeros *N* términos para realizar una nueva expansión de consultas. La Tabla 4.4 muestra un caso de uso tras una consulta al servicio *Key Terms* de *Yahoo!*. La consulta realizada es “www.sigir.org”, que corresponde al sitio web del *Grupo de Interés Especial de Recuperación de Información* y también al conocido congreso de recuperación de información. Resulta muy interesante observar cómo el motor de búsqueda proporciona términos tan útiles como *ACM*, *Information Retrieval*, *web search* y *conference*. Además, debido

[3] http://developer.yahoo.com/search/boss/

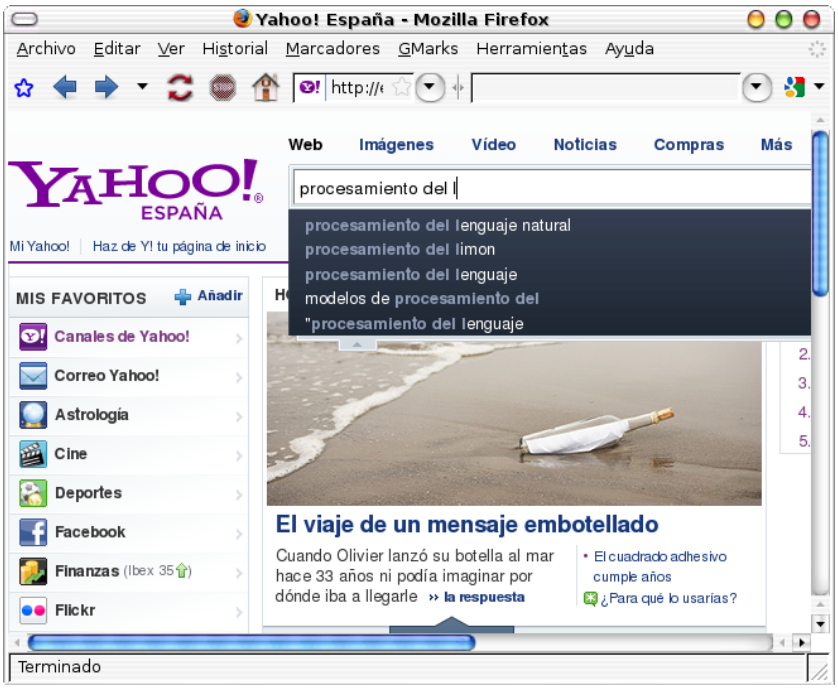

Figura 4.3: *Búsqueda asistida del buscador de Yahoo! que emplea la misma tecnología que el servicio web Key Terms.*

a la conferencia *SIGIR 2009* que se celebró recientemente en *Boston*, el motor de búsqueda proporciona dicha ubicación.

www.sigir.org
SIGIR
ACM SIGIR
Information Retrieval
Special Interest Group on Information Retrieval
conference
retrieval field
SIGIR Forum
web search
text search
Boston

Tabla 4.4: *Primeros diez términos devueltos por el servicio web Key Terms de Yahoo! tras la consulta www.sigir.org.*

4.7.2. Sistemas de Etiquetado Social

Hoy en día, el fenómeno de las redes sociales y en relación con esta tesis del etiquetado social está revolucionando las búsquedas en Internet, ya que detrás de los sitios web como *del.icio.us* o *www.bibsonomy.org* existe una comunidad de usuarios que etiquetan manualmente páginas web y también comparten sus marcadores con otros usuarios. Estos sitios web almacenan una enorme colección de páginas web etiquetadas y se han convertido en una fuente muy importante de información. Además, los motores de búsqueda ya están empleando esta información para mejorar sus búsquedas y están investigando las características de estas etiquetas para intentar sacar un beneficio aún mayor.

En la Figura 4.4 se muestra el resultado de la búsqueda "nlp.uned.es" en el sitio de etiquetado social *del.icio.us* donde se puede apreciar tanto las etiquetas que asigna cada usuario a esta Url (en el centro de la imagen), como una lista de etiquetas ordenada según el número de usuarios que han usado dicha etiqueta (margen derecho).

La información disponible en este tipo de sitios de etiquetado social tiene una gran utilidad a la hora de mejorar las búsquedas. A continuación se presentan varios trabajos que demuestran como las etiquetas de los usuarios pueden ser una fuente de información de una gran utilidad. Bao et al.[BXW+07] han explorado el uso de anotaciones sociales procedentes de *del.icio.us* para mejorar las búsquedas en la Web. Estos investigadores indicaron que las anotaciones son generalmente buenos resúmenes de las páginas web correspondientes y que la cantidad de anotaciones indica la popularidad de las páginas web. Yanbe et al.[YJNT07] propusieron combinar la ampliamente usada métrica de ranking basada en los enlaces (PageRank) con información obtenida a partir de un sitio social de bookmarks. También presentaron el algoritmo *SBRank* que captura la popularidad de una página e im-

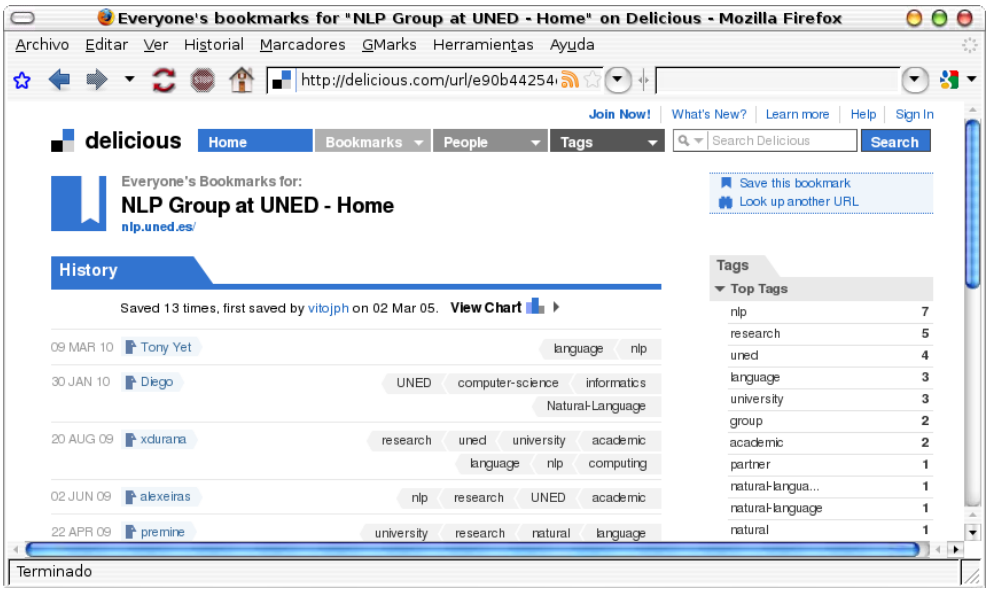

Figura 4.4: *Resultado de la búsqueda "nlp.uned.es" en el sitio de etiquetado social del.icio.us donde se puede apreciar tanto las etiquetas que asigna cada usuario a esta Url (en el centro de la imagen), como una lista de etiquetas ordenada según el número de usuarios que la han usado (margen derecho).*

plementaron una aplicación de búsqueda Web mediante el uso de este etiquetado social. Noll et al.[NM08] presentaron una técnica de personalización que separa los datos de la colección y los perfiles de usuario del sistema de información donde los contenidos y los documentos indexados están siendo buscados. Teniendo en cuenta esta información, utilizan marcadores y etiquetas sociales para realizar un nuevo ranking de los resultados.

Nuestro sistema también considera esta fuente de información y realiza búsquedas en sitios de etiquetado social como *del.icio.us*. La información que trata el sistema se corresponde a una lista de etiquetas acerca de una página, que se encuentra ordenada en función del número de usuarios que han utilizado dicha etiqueta para marcar la página solicitada. Posteriormente, las primeras N etiquetas son utilizadas para extender la consulta formada por el texto del ancla del enlace roto.

4.8. Ranking de Páginas Candidatas

En la sección anterior quedó reflejado el proceso de recolección de la información relacionada con un enlace roto. Las distintas fuentes de información empleadas pueden dividirse en dos tipos, las fuentes de información que pueden emplearse tal cual son extraidas, y aquellas que necesitan un proceso de extracción de terminología para sintetizar la información más relevante. Después de dicho proceso se

realiza una sucesión de expansión de consultas con las fuentes de información citadas anteriormente. En esta fase de exploración, se realizan consultas en los principales motores de búsqueda, recuperando los primeros resultados en cada caso. Una vez que ha finalizado dicha fase, el sistema dispone de un conjunto de páginas candidatas que debe presentar al usuario de manera ordenada según su relevancia. Precisamente esa etapa de ordenación es la que se presenta en esta sección donde analizaremos distintos métodos para establecer un ranking de resultados.

4.8.1. Modelo de Espacio Vectorial y Coeficientes de Coocurrencia

Dentro del estudio comparativo entre los diferentes métodos de ranking, ha sido utilizado el modelo de espacio vectorial[MRS08] para representar los documentos y varios coeficientes de coocurrencia[Rij77] para ordenarlos. Los métodos basados en la coocurrencia de términos se han utilizado frecuentemente para identificar relaciones semánticas entre los documentos. Para las pruebas que mostraremos a continuación hemos utilizado los conocidos coeficientes de coocurrencia *Tanimoto*, *Dice* y el *Coseno* para medir la similitud entre los vectores que representan cada documento de referencia D_1 y el documento candidato D_2. El coeficiente de *Tanimoto* es una extensión el índice de Jaccard, y consiste en una medida estadística que se usa para comparar la similitud y diversidad de dos conjuntos de muestra. El coeficiente de Jaccard mide la similitud entre los conjuntos de muestra, y se define como el tamaño de la intersección dividido por el tamaño de la unión de los conjuntos de muestra. De esta forma, el coeficiente de Tanimoto se representa de la siguiente manera:

$$\text{Tanimoto}(\overrightarrow{D}_1, \overrightarrow{D}_2) = \frac{\overrightarrow{D}_1 \overrightarrow{D}_2}{|\overrightarrow{D}_1|^2 + |\overrightarrow{D}_2|^2 - \overrightarrow{D}_1 \overrightarrow{D}_2} \tag{4.3}$$

El coeficiente de Dice, llamado de esta manera gracias a Lee Raymond Dice, es una medida de similitud relacionada con el índice de Jaccard[Rij77]. Para dos conjuntos X e Y de palabras clave utilizadas en recuperación de información, el coeficiente se puede definir como el doble de la información compartida (intersección) sobre el conjunto combinado (unión):

$$\text{Dice}(\overrightarrow{D}_1, \overrightarrow{D}_2) = \frac{2\overrightarrow{D}_1 \overrightarrow{D}_2}{|\overrightarrow{D}_1|^2 + |\overrightarrow{D}_2|^2} \tag{4.4}$$

El coeficiente de Coseno es una medida de similitud entre dos vectores de n dimensiones, que mide el coseno del ángulo entre ellos. Esta medida se utiliza generalmente para comparar los documentos en el área de la minería de texto. Además,

se utiliza para medir la cohesión dentro los clusters en el ámbito de la minería de datos. El coeficiente de Coseno se representa de la siguiente manera:

$$\text{Cosine}(\overrightarrow{D}_1, \overrightarrow{D}_2) = \frac{\overrightarrow{D}_1 \overrightarrow{D}_2}{|\overrightarrow{D}_1||\overrightarrow{D}_2|} \tag{4.5}$$

4.8.2. Divergencia de Kullback-Liebler

Además de los coeficientes de coocurrencia introducidos en la sección anterior, hemos empleado otro método para realizar un ranking lo más eficiente posible. En este caso se trata de un enfoque basado en modelos de lenguaje donde en lugar de medir la similitud como en el caso anterior, se va a medir la divergencia con el objetivo de posicionar los documentos menos divergentes en las primeras posiciones. En primer lugar se ha decidido construir un modelo de lenguaje para representar los documentos o las unidades textuales a comparar. En segundo lugar, se ha utilizado una media probabilística no simétrica para hallar la divergencia entre la unidad textual de referencia y el documento candidato. En este caso nos fijamos en la diferencia entre las dos distribuciones de probabilidad, empleando para ello la divergencia de *Kullback-Leibler* entre dos documentos:

$$KLD(D_1||D_2) = \sum_{t \in D_1} P_{D_1}(t) log \frac{P_{D_1}(t)}{P_{D_2}(t)} \tag{4.6}$$

donde $P_{D_1}(t)$ es la probabilidad del término t en el documento de referencia, y $P_{D_2}(t)$ es la probabilidad del término t en el documento candidato.

4.8.3. Comparación de Métodos de Ranking

De acuerdo a las medidas descritas en las secciones anteriores, correspondientes a varios coeficientes de coocurrencia y una medida de divergencia, se ha realizado un estudio comparativo para analizar la mejor medida en función de la fuente de información empleada. En cuanto a las fuentes de información, se han seleccionado aquellas más descriptivas teniendo en cuenta las características de la unidad textual frente a la que se quiere comparar. La metodología empleada consiste en seleccionar varias fuentes de información tanto de la página origen como de la página destino (según una relación unidireccional descrita por el hiperenlace), donde cada par de unidades textuales deben cumplir dos características: en primer lugar ser una fuente de información significativamente descriptiva, y en segundo lugar debe tener unas características similares al par correspondiente seleccionado en la comparación.

Las fuentes de información seleccionadas para medir la similitud entre las páginas de origen y destino son el texto del ancla, el contenido de la página, el contenido de la versión en caché y el título. En concreto, realizado el estudio de similitud entre

los siguientes pares de elementos: (i) El texto del ancla de la página origen y el título de la página candidata, (ii) el texto del ancla de la página origen y el contenido de la página candidata, (iii) el contenido de la página origen y el contenido de la página candidata, y (iv) el contenido de la versión en caché de la página desaparecida y el contenido de la página candidata.

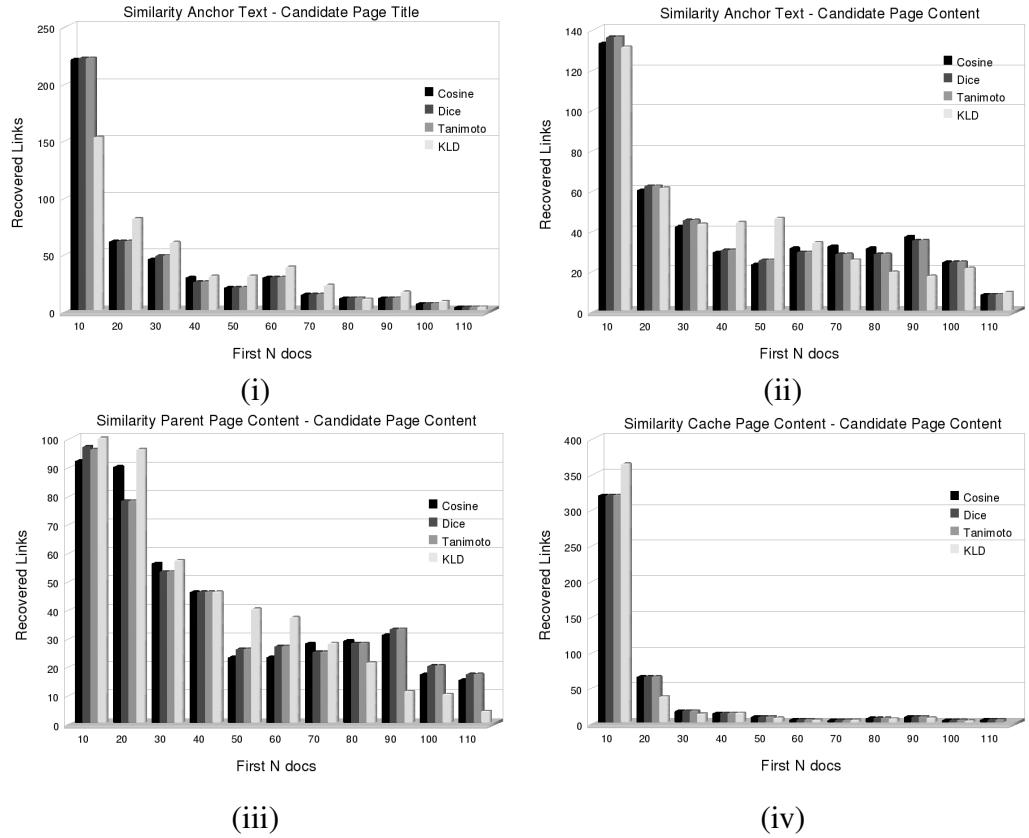

Figura 4.5: *Resultados de diferentes métodos de ranking (Cosine, Dice, Tanimoto y Kullback-Liebler Divergence) empleados para medir la similitud entre (i) El texto del ancla de la página origen y el título de la página candidata, (ii) el texto del ancla de la página origen y el contenido de la página candidata, (iii) el contenido de la página origen y el contenido de la página candidata, y (iv) el contenido de la versión en caché de la página desaparecida y el contenido de la página candidata. Los resultados muestran la posición de la mejor página recuperada por el sistema después del ranking.*

Además de las comparaciones citadas anteriormente, también han sido empleadas la dirección destino del hiperenlace y el *snippet* de la página candidata, pero los resultados no han mejorado lo mostrado en la Figura 4.5. La primera observación en las cuatro Figuras 4.5(i-iv) es que los resultados obtenidos con las tres medidas de coocurrencia, *Coseno*, *Dice* y *Tanimoto*, son muy similares. Se puede observar

en la Figura 4.5(i) y en la Figura 4.5(ii) que los resultados obtenidos con *KLD* son peores que los obtenidos con los coeficientes de coocurrencia, especialmente en la Figura 4.5(i). Casualmente estos peores resultados de *KLD* se reflejan en estas figuras donde se muestra el estudio de la similitud entre un texto muy corto, el texto del ancla, y otros dos donde por un lado el texto también tiene una reducida longitud como es el título de la página de candidata, o con un gran tamaño como en el caso del contenido de la página principal.

Por el contrario, *KLD* se comporta mejor que las medidas de coocurrencia en la Figura 4.5(iii) y en la Figura 4.5(iv), donde estamos midiendo la similitud entre el contenido de dos páginas, la página origen o la versión en caché, y la página candidata. De esta forma podemos concluir *KLD* funciona mejor que los métodos de coocurrencia solamente si se aplica a textos lo suficientemente largos, como el contenido de una página.

En la Figura 4.5(iv) se observa que, como era de esperar, el ranking de resultados obtenido al utilizar la similitud con la página en caché son los mejores. Sin embargo, en muchos casos esta página no está disponible (alrededor de un 40-50 %). En cierto modo estos resultados son los esperados ya que la versión en caché de una página desaparecida es la fuente de información más fiable y más parecida al documento buscado que se puede conseguir. Salvo en el caso de una página web que cambie con mucha frecuencia como ocurre en un sitio de noticias o similar, la versión en caché de una página es una imagen prácticamente idéntica del documento requerido. Además, y a pesar de que dicha página hubiera cambiado desde su almacenamiento, existen ciertos términos característicos en cada documento que hacen de su combinación una especie de firma léxica muy difícilmente equiparable. Si atendemos al resto de figuras que no emplean la versión en caché como fuente de información, se observa que los mejores resultados se obtienen utilizando la técnica de coocurrencia entre el texto del ancla del enlace roto y el título de la página candidata.

De acuerdo con estos resultados, si se puede recuperar la versión en caché de la página desaparecida, esta debería ser la principal fuente de información de la página destino a la hora de medir la similitud, empleando además KLD para realizar el ranking de las páginas candidatas. De lo contrario, utilizaremos la similitud entre el texto del ancla y el título de la página candidata empleando para ello el coeficiente de *Dice*, que funciona un poco mejor que el resto de coeficientes de coocurrencia.

4.9. Análisis de Rendimiento

El sistema de recuperación de enlaces rotos que es descrito en esta tesis emplea diferentes fuentes de información procedentes del contexto del enlace y de los recursos disponibles en la infraestructura web. La extracción y uso de la terminología extraida, así como la recuperación de los resultados obtenidos del motor de búsqueda y su posterior ranking, representan un coste computacional significativo

para el sistema. Por este motivo, en esta sección se presenta un estudio comparativo de los diferentes parámetros que influyen en el rendimiento del sistema para poder configurarlo de la manera más óptima. Este análisis de rendimiento y el posterior ajuste del sistema tiene dos beneficiarios; por un lado el sistema, ya que optimizará sus recursos para obtener el mismo resultado; y por otro lado el usuario que verá reducido el tiempo de respuesta. Para llevar a cabo el siguiente estudio de rendimiento el sistema utilizarán 1000 enlaces para medir los diferentes aspectos que intervienen en el funcionamiento del sistema.

En primer lugar analizaremos la capacidad de cada fuente de información y cómo afecta su presencia en los resultados obtenidos ya que las fuentes de información tienen diferentes disponibilidades y grados de eficacia.

Otra cuestión importante a investigar es el equilibrio entre la cantidad de información recolectada para recuperar los enlaces rotos, y el tiempo requerido para obtener dicha información. En términos generales es de esperar que cuanta más información se tenga a disposición del sistema, mejor sea la recomendación. Sin embargo es importante conocer el coste que acarrea cada incremento en la cantidad de información recuperada.

4.9.1. Efectividad de las Fuentes de Información

Existe una gran diversidad en la disponibilidad de las fuentes de información previamente analizadas. Las fuentes de información tales como el ancla de texto, la Url del enlace roto o la página en la que se encuentra el enlace roto siempre están disponibles ya que forman parte del modelo sobre el que basamos la recuperación. Sin embargo, una versión en caché de la página desaparecida o el contexto del enlace que estamos tratando de recuperar, sólo están disponibles en algunos casos dependiendo de un conjunto de factores que no podemos controlar. Además, si hablamos de los recursos obtenidos de la infraestructura web, tales como los motores de búsqueda, estos no tienen en la actualidad una lista de palabras clave por cada una de las páginas en Internet. Es decir, este tipo de aplicaciones tan solo cuentan en la actualidad con información acerca un subconjunto de páginas de la Web. De la misma forma, y a pesar de que los sitios de etiquetado social tienen un crecimiento practicamente exponencial en el número de páginas etiquetadas, aún distan mucho de tener la Web completa etiquetada, algo que por otro lado no está claro que sea el objetivo principal de las compañías que los mantienen.

Por otra parte y con independencia de la disponibilidad de las fuentes de información, su eficacia varía mucho de un caso a otro. Es decir, existen fuentes de información que cuando es posible contar con ellas, obtienen un índice de recuperación de los enlaces analizados superior al 90 %, mientras que el grado de recuperación con otras fuentes de información es inferior al 60 %. De esta forma los distintos escenarios que se presentan a la hora de recuperar un enlace roto, indican de partida al sistema la probabilidad de recuperación del enlace. Si recordamos, durante este capítulo hemos analizado diferentes datos que influyen en la recuperación de

un enlace roto, como la disponibilidad de una fuente de información, una determinada cantidad de términos o la presencia de entidades nombradas. Pues bien, toda esta información formará parte de la inteligencia del sistema a la hora de estimar al inicio de la recuperación de un enlace, la probabilidad de éxito.

La Figura 4.6 muestra tres parámetros de cada fuente de información. En primer lugar se puede observar la disponibilidad de cada fuente de información en términos del número de ocasiones en las que el sistema ha podido encontrarla. En segundo lugar expone el número de enlaces que han sido recuperados en las primeras diez posiciones gracias a su participación. En tercer lugar se muestra la precisión de cada fuente de información reflejada en el número de enlaces que el sistema ha conseguido recuperar en la primera posición tras emplear dicha fuente de información.

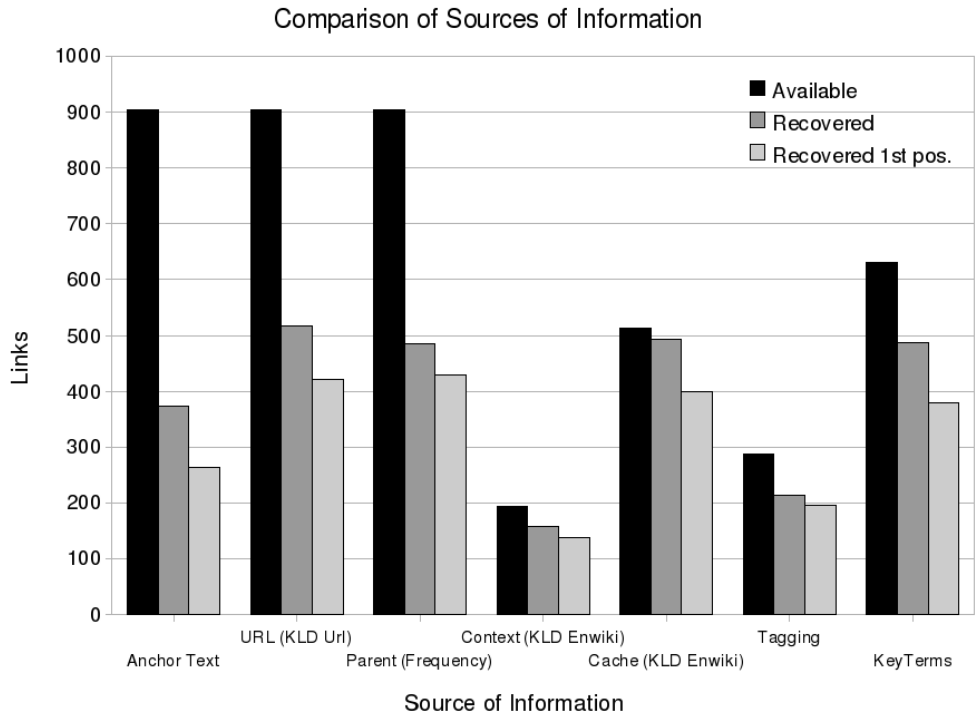

Figura 4.6: *Disponibilidad, número de enlaces recuperados entre las diez primeras posiciones y enlaces recuperados en la primera posición de los resultados proporcionados por el motor de búsqueda, según la fuente de información empleada.*

En la Figura 4.6 se comprueba la gran efectividad de los términos extraidos del contexto del enlace, de la versión en caché y de las etiquetas sociales. A veces el texto del ancla es demasiado genérico y ambiguo, por lo que algunas fuentes pueden encontrar fácilmente un término que es la clave para romper esta ambigüedad, como en el caso del contexto del enlace, de la versión en caché y de las etiquetas sociales. Es de destacar el alto nivel de recuperación obtenido por la versión caché y

los términos clave de *Yahoo!*. Estas fuentes de información son capaces de expandir la consulta con términos muy precisos. Además, como hemos mencionado anteriormente, existen algunas fuentes de información que están disponibles en todos los casos y este hecho queda probado en los resultados. Estas fuentes son el texto del ancla, la página padre y la Url del hiperenlace. Otro factor importante a tener en cuenta es la precisión de cada fuente, de tal manera que cuantos más enlaces son recuperados en la primera posición empleando una determinada fuente de información, este hecho implica una mayor precisión de dicha fuente. Además, cuanto más precisa es una fuente de información, más rápida es la respuesta del sistema. Por lo tanto, el contexto del enlace y las etiquetas sociales se presentan en la Figura 4.6 como las fuentes de información más precisas.

4.9.2. Impacto de la Cantidad de Información Empleada en los Resultados

La cantidad de información con la que cuenta el sistema se basa principalmente en dos parámetros: el número de términos (extraídos de varias fuentes de información) utilizados para extender las consultas formadas por el texto del ancla, y el número de páginas recuperadas de los resultados devueltos por el motor de búsqueda por cada consulta realizada. Con el objetivo de evaluar el efecto de estos parámetros en los resultados del SRER se han realizado una serie de experimentos que se describen a continuación

La primera batería de experimentos tiene como objetivo analizar el número de enlaces recuperados que son conseguidos gracias a la expansión de consultas con los términos de una determinada fuente de información. Cabe mencionar que en muchos casos un enlace se consigue recuperar mediante varias fuentes de información al mismo tiempo y por tanto la suma total de los enlaces recuperados con cada fuente no corresponde al total de los enlaces recuperados por el sistema. Al final de la fase de recolección de resultados del buscador en los que ha intervenido cada fuente de información, las páginas candidatas duplicadas son eliminadas, pero se mantiene la información de qué fuente consiguió recuperar esa página y en qué posición del buscador.

La Figura 4.7 muestra el número de enlaces recuperados según el número de términos utilizados en la expansión. En esta figura se muestran los resultados de las diferentes fuentes de información empleadas por el sistema y los resultados de la combinación de todos ellos. Para el desarrollo de los experimentos se han tomado los 10 primeros resultados devueltos por el buscador después de cada consulta. Obviamente, el método sin expansión que tan solo emplea el texto del ancla no se ve afectado por el número de términos utilizados en la expansión. Sin embargo es interesante apreciar que la expansión con términos de la página padre (aquella donde se encuentra en enlace) obtiene mejores resultados que la consulta formada solamente por el texto del ancla, cuando son utilizados 6 términos o más para expandir

la primera de ellas. La razón principal de este hecho es que mediante la expansión se completa la consulta original con términos que en algunos casos no están relacionados con la página apuntada, pero en otros muchos se refleja la hipótesis de que dos páginas enlazadas tienen una determinada relación de contenido. Además, es destacable la pendiente ascendente que se produce con la expansión de los términos de la página padre, sobre todo empleando los 7 o 8 primeros términos.

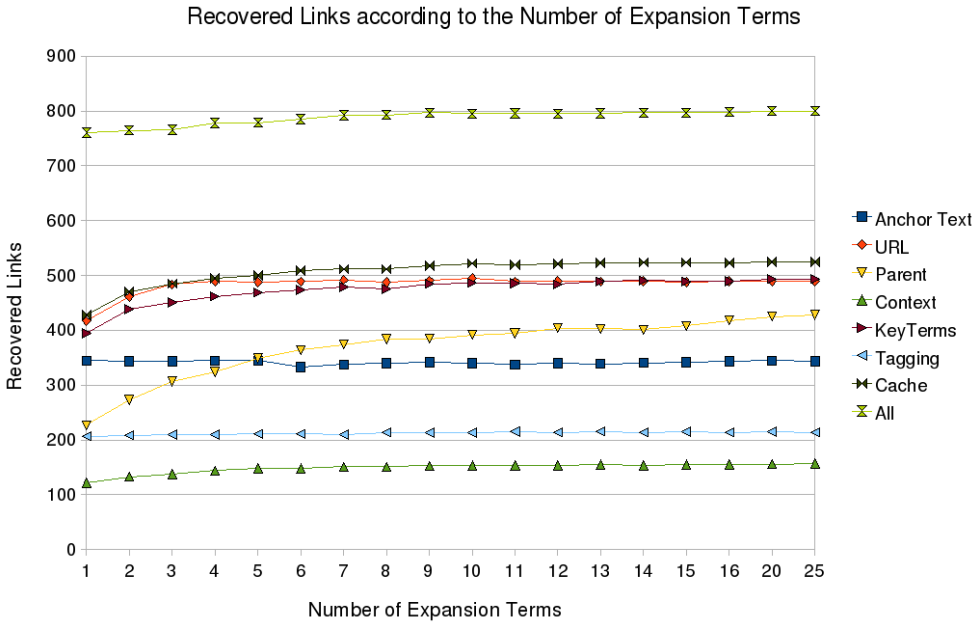

Figura 4.7: *Enlaces recuperados en función del número de términos utilizados para llevar a cabo la expansión de consultas. Esta figura muestra los resultados de utilizar diferentes fuentes de información. El método que fusiona todas las fuentes de información representa la combinación de los resultados, no la suma aritmética.*

El método que utiliza todas las fuentes de información es realmente una combinación de todas ellas, pero este enfoque no es la suma aritmética de los resultados de todas las fuentes ya que algunas obtienen la recuperación de los mismos enlaces. Aparte de esta anotación acerca de la combinación de las fuentes, es posible observar como el número de enlaces recuperados aumenta en todos los casos a medida que aumenta el número de términos implicados en la expansión. Sin embargo, si atendemos a la combinación de todas las fuentes de información, la mejora obtenida es muy pequeña a partir del uso de 10 términos y hasta el máximo. De esta forma y según los datos ofrecidos por los experimentos, podemos concluir que el mejor número de términos usados para extender la consulta es 10. Este valor pude variar en alguna fuente de información en particular, pero como el sistema funciona mediante la combinación de todas ellas, los resultados obtenidos mediante esta combinación son la base del sistema.

La Figura 4.8 muestra el número de enlaces que se han conseguido recuperar en función del número de páginas candidatas recuperadas de los resultados devueltos por el motor de búsqueda. Esta figura presenta los resultados de las diferentes fuentes de información utilizadas por el sistema. Para el desarrollo de los experimentos se han utilizado 10 términos de cada fuente de información para realizar la expansión de las consultas. Entre los detalles más significativos, llama la atención la fuerte pendiente con el uso de 7 documentos recuperados en el caso de la Url, y la baja recuperación cuando se considera tan solo un resultado del motor de búsqueda. También se puede observar cómo el número de enlaces recuperados aumenta con el número de resultados obtenidos, aunque la mejora es mucho menor a partir del segmento 10-15. Además, dado que en la combinación de todas las fuentes de información no hay una clara mejora a partir de la recuperación de 10 documentos, podemos concluir que la cantidad más apropiada para seleccionar los resultados obtenidos por el motor de búsqueda tras cada consulta son 10 páginas.

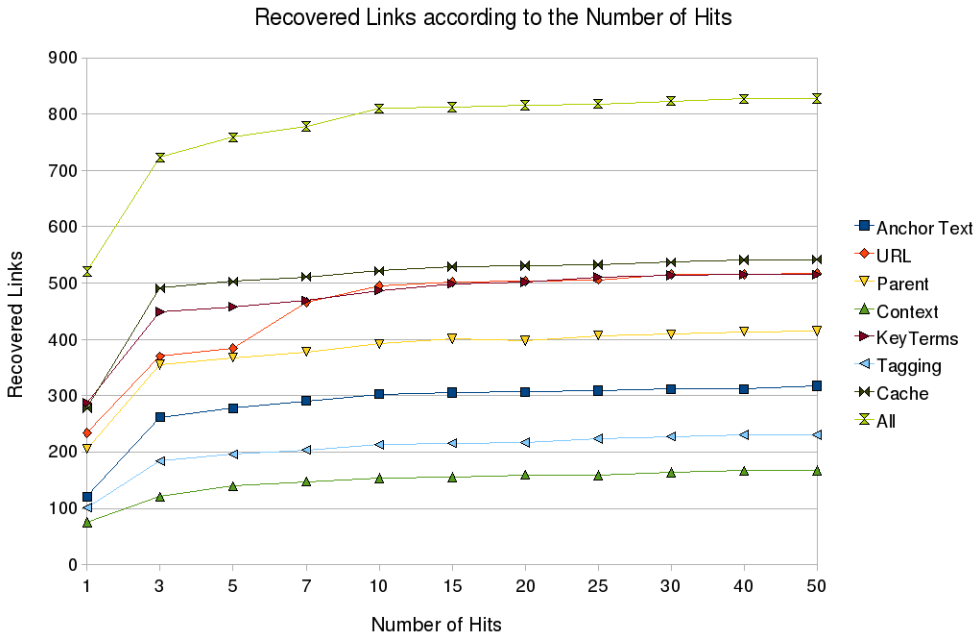

Figura 4.8: *Enlaces recuperados en función del número de páginas candidatas recuperadas de los resultados devueltos por el motor de búsqueda tras cada consulta. Esta figura muestra los resultados de utilizar diferentes fuentes de información. El método que fusiona todas las fuentes de información representa la combinación de los resultados, no la suma aritmética.*

4.9.3. Impacto de la Cantidad de Información Empleada en el Tiempo de Respuesta

En el desarrollo de cualquier aplicación en la que interviene un usuario, es imprescindible realizar un análisis de rendimiento del sistema, de cara a determinar el tiempo de respuesta del mismo. El sistema de recuperación de enlaces rotos que se propone en esta tesis, tiene como entrada por parte del usuario una página donde pueden encontrarse enlaces rotos, y como salida una lista de recomendación de páginas web para sustituir dicho enlace. En el proceso de establecer la lista de recomendación, las tareas a realizar son numerosas y el coste de cada etapa influye determinantemente en el tiempo final de respuesta. En esta sección hemos decidido analizar el tiempo requerido para las dos fases del sistema en donde se invierte la mayor parte del tiempo de ejecución. Estas dos fases se corresponden con los dos parámetros principales del sistema: el número de términos empleados para la expansión de consultas, y el número de resultados recuperados del motor de búsqueda tras cada consulta.

La Figura 4.9 muestra el tiempo medio de ejecución para recuperar todos los enlaces con el uso de las fuentes de información previamente definidas y en función del número de términos empleados en la expansión de consultas y resultados obtenidos del motor de búsqueda tras cada consulta. En los experimentos donde varía el número de resultados obtenidos, el número de términos se ha fijado a 10, y en los experimentos en los que varía el número de términos, el número de resultados obtenidos ha sido fijado a 10. Se puede observar, como era de esperar, que los incrementos de tiempo con la variación de los dos parámetros involucrados son aproximadamente lineales. Si bien, parece que la pendiente es más pronunciada a partir del uso tanto de 10 términos para la expansión como de 10 resultados obtenidos. De esta forma, y teniendo en cuenta las medidas de rendimiento obtenidas en las secciones anteriores, se ha decidido fijar para el resto de los experimentos en 10 tanto el número de términos utilizados para realizar la expansión de consultas como la cantidad de resultados recuperados del motor de búsqueda tras cada consulta.

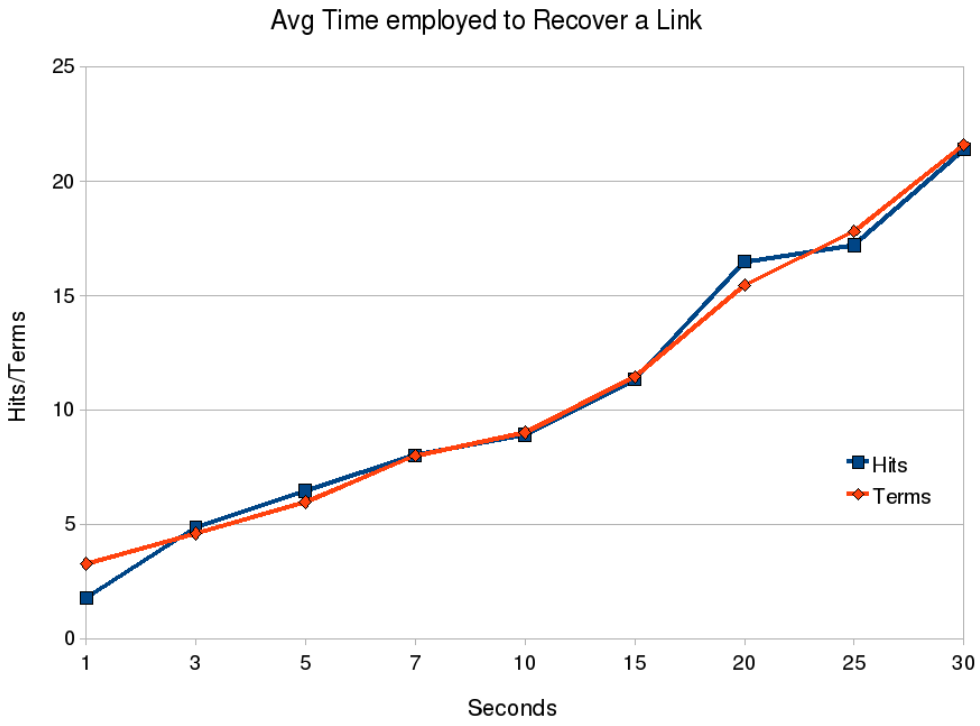

Figura 4.9: *Tiempo medio requerido para recuperar un enlace en función del número de términos empleados en la expansión de consultas y resultados obtenidos del motor de búsqueda tras cada consulta.*

4.10. Algoritmo de Recuperación Automática de Enlaces Rotos

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 4.10. En primer lugar se comprueba si el número de términos del ancla es sólo uno `(length(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 dispone de una versión en caché, entonces se recupera, expandiendo la consulta formada por los términos del ancla con los términos extraidos de la versión en caché, el contexto del enlace y la dirección de destino del hiperenlace, empleando el método *KLD*, y por otro lado con las etiquetas sociales y las palabras clave extraidas de *BOSS*.

```
if length(anchor) = 1 and NoNE(anchor) then
  if InCache(page) then
    docs = web_search(anchor + cache_KLD)
    docs = docs + web_search(anchor + Url_KLD + context_KLD)
    docs = docs + web_search(anchor + tags + keyterms)
    rank(docs, cache_content_KLD)
    if similarity(docs, cache(page) > 0.9) then
      user_recommendation(docs)
    else
      No_recovered
  else
    No_recovered
else
  if InCache(page) then
    docs = docs + web_search(anchor + cache_KLD)
    docs = docs + web_search(anchor + Url_KLD + context_KLD)
    docs = docs + web_search(anchor + tags + keyterms)
    rank(docs, cache_content_KLD)
  else
    docs = docs + web_search(anchor + parent_FREQ)
    docs = docs + web_search(anchor + Url_KLD + context_KLD)
    docs = docs + web_search(anchor + tags + keyterms)
    rank(docs, anchor_title_Dice)
  user_recommendation(docs)
```

Figura 4.10: *Algoritmo de recuperación automática de enlaces rotos.*

Después de este proceso, se ordenan los resultados (mediante el cálculo con *KLD* de la similitud entre la página candidata y la versión en caché) y solamente si hay alguno suficientemente próximo al contenido de la caché se hace la recomendación al usuario. En los casos restantes, es decir en el caso de las anclas con más de un término o que contienen alguna entidad nombrada, la recuperación es llevada a cabo mediante la expansión de la consulta con los términos del contexto y de la Url, aplicando *KLD*, y con las etiquetas sociales y las palabras clave por otro lado. Si la página desaparecida tiene una versión disponible en caché, la consulta es expandida además con los términos extraidos de dicha versión mediante el uso de *KLD*. Si no es así, se utilizan los términos extraidos de la página padre aplicando un método basado en frecuencia (Tf). Después de este proceso, todos los documentos son agrupados y ordenados en función de la relevancia con una versión en caché de la página desaparecida (si está disponible), o en función de la similitud entre el texto del ancla y el título de la página candidata, aplicando el coeficiente de coocurrencia de *Dice* en este último caso.

4.11. Resultados

En esta sección se muestran los resultados obtenidos al aplicar el algoritmo propuesto en la sección anterior a un conjunto de enlaces rotos. Hasta este momento, los experimentos habían sido realizados con enlaces activos, esto es, que la página apuntada se podía consultar en todo caso. En cambio, una vez que han finalizado los experimentos para ajustar los múltiples parámetros que intervienen en el sistema y definir el algoritmo que se encarga de recuperar un enlace roto con la mayor eficacia y velocidad, es el momento de aplicar este conocimiento adquirido a la recuperación de enlaces realmente rotos. La selección de los enlaces para realizar los siguientes experimentos ha seguido la misma metodología expuesta para la recolección de los enlaces usados anteriormente. En total, se ha reunido un conjunto de 1998 enlaces rotos seleccionados aleatoriamente.

Hemos aplicado el algoritmo descrito en la sección anterior a este conjunto de enlaces que están realmente rotos. Además de los requisitos ya conocidos impuestos a los enlaces analizados, para la evaluación de los resultados hemos añadido otro requisito adicional. Sólo hemos seleccionado los enlaces de los que había una versión en caché disponible. La razón de dicho requisito es que sólo en este caso, podemos evaluar los resultados de una manera objetiva. En algunos casos, el sistema es capaz de reunir hasta 700 páginas candidatas (7 fuentes de información x 10 términos x 10 hits) por cada enlace roto. Entonces, de acuerdo con el algoritmo definido anteriormente, el sistema ordena estas páginas candidatas y muestra al usuario las mejores 100 páginas candidatas, como una lista ordenada por relevancia.

Con el fin de analizar los resultados del sistema, se ha realizado una evaluación manual. En dicha evaluación hecha por personas, por cada uno de los enlaces rotos propuestos, al evaluador se le pide que realice un juicio de relevancia sobre la lista ordenada que presenta el sistema. En dicha lista, el juez humano debe indicar si los primeros resultados propuestos por el sistema son realmente relevantes como para poder sustituir en enlace roto. Recordemos que el juez dispone de una versión en caché en todos los casos para comprobar las diferentes características de las páginas propuestas. Además, el evaluador debe dictaminar cual es la primera página relevante que aparece en el ranking de resultados.

Los resultados obtenidos mediante esta evaluación manual se presentan en la Tabla 4.5. De cara a interpretar correctamente los resultados mostrados, *"739 enlaces rotos recuperados"* en *"1-10 Primeros N Documentos"* significa que la evaluación manual ha concluido que la mejor página candidata para reemplazar el enlace roto propuesto, ha sido mostrada por el sistema entre los diez primeros resultados en 739 ocasiones del total de enlaces rotos analizados. Además, se considera que un enlace roto ha sido recuperado cuando una página candidata válida está presente entre los 100 primeros documentos propuestos por el sistema. Una página candidata es válida cuando, según la opinión del evaluador, el contenido de dicha página tiene una gran similitud con la página desaparecida como para poder sustituirla y cumplir una función informativa análoga.

Primeros *N* documentos	Enlaces Rotos Recuperados
1-10	739
10-20	367
20-50	332
50-100	121

Tabla 4.5: *Número de enlaces rotos recuperados (mejor candidato cuyo contenido es muy similar a la página desaparecida) de acuerdo a su similitud con la versión en caché, entre los primeros N documentos utilizando el algoritmo propuesto.*

Según la evaluación manual descrita anteriormente, el sistema ha recuperado 1559 enlaces rotos de un total de 1998 (78 % del total de enlaces). La Tabla 4.5 muestra además la posición del ranking propuesto al usuario, donde se ha ubicado la primera página que, por su similitud con la página desaparecida, podría sustituirla y además podría cumplir una función informativa análoga. Como parte del análisis de los resultados obtenidos se ha probado que, en algunos casos, la página original es encontrada en otra ubicación. Esto es debido generalmente a los frecuentes cambios de páginas web a diferentes dominios o a reestructuraciones de los sitios web. En otros casos, han sido recuperadas páginas con contenidos muy similares (prácticamente el mismo contenido). De esta forma, se puede concluir que el sistema es capaz de proporcionar reemplazos útiles para las páginas web entre las primeras 10 posiciones en el 47 % de los enlaces recuperados, y entre las 20 primeras en un 71 % de los casos.

4.11.1. Influencia de cada Fuente de Información en los Resultados

La Sección 4.9.1 presentó un estudio comparativo donde se mostraba tanto la disponibilidad como la cantidad de enlaces recuperados en los que cada fuente de información había influido. La Figura 4.11 ilustra el mismo estudio comparativo, pero en este caso son empleados los enlaces rotos recuperados en la sección de resultados. Al igual que en la sección anterior 4.11, los enlaces están obligados a tener una versión en caché disponible para poder establecer una evaluación objetiva. Es importante analizar cada fuente de información por separado. Por otra parte, tras analizar los resultados y la influencia de cada fuente de información, hemos demostrado que si eliminamos todas las páginas candidatas recuperadas por los términos de la versión en caché, el sistema recuperaría 1403 enlaces, es decir, un 70 % del total de los enlaces. Este dato refleja que, a pesar de que estamos analizando enlaces rotos con una versión en caché, y la presencia de esta versión en todos los casos podría influir significativamente en los resultados obtenidos, si realmente no con-

táramos con esta fuente de información en ninguna parte del análisis obtendríamos unos resultados muy similares.

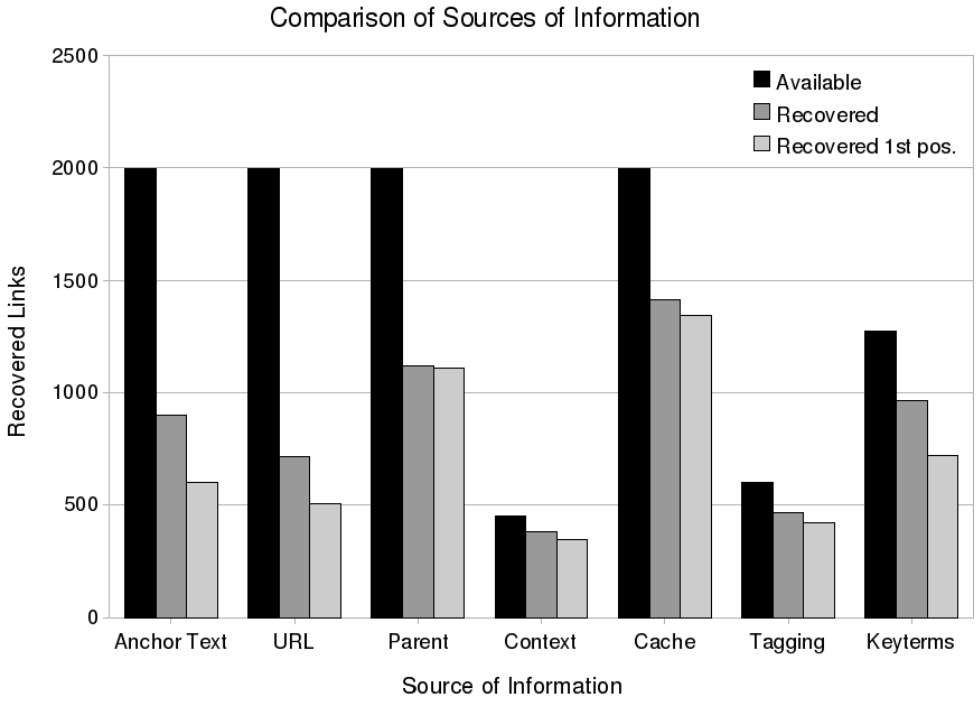

Figura 4.11: *Disponibilidad, número de enlaces rotos recuperados entre las diez primeras posiciones y enlaces rotos recuperados en la primera posición de los resultados proporcionados por el motor de búsqueda, según la fuente de información empleada.*

De hecho, si nos fijamos primero en el análisis de la versión en caché, se puede apreciar que su eficacia (enlaces en los que la fuente está disponible / enlaces recuperados) es más reducida que en el caso de los enlaces activos, mientras que mantiene unos datos similares de precisión (enlaces recuperados / enlaces recuperados en primera posición).

También es posible advertir que la página que contiene el enlace roto (página padre) se ha convertido en la mejor fuente de información después de la versión en caché. Una de las razones principales es que la eficacia de los términos de la Url se ha visto reducida en gran medida. Este hecho es relativamente normal, ya que los enlaces que estamos utilizando en este estudio están realmente rotos y no es suficiente con encontrar la misma Url como ocurría en el caso de los enlaces activos. Por este motivo, este estudio refleja en realidad el comportamiento de la Url como fuente de información, ya que en los estudios previos los términos de la Url provocaban fácilmente la recuperación de la página web que tenía los mismos términos en su dirección web. En cuanto a las palabras clave proporcionadas por *BOSS*, se puede observar que su eficacia también se ha incrementado y se convierten en una

fuente de información realmente relevante. Por otra parte, también se observa un aumento de la eficacia de los términos del ancla. Tal vez en el experimento con los enlaces activos, la función de esta fuente de información estaba oculta por el tipo de experimento, pero ahora se observa que su utilidad es muy alta con el uso de enlaces rotos. Por último, los términos extraídos del contexto y las etiquetas sociales mantienen unos valores similares.

4.12. Conclusiones

En este capítulo de la tesis se han analizado las diferentes fuentes de información que pueden ser utilizadas para llevar a cabo una recuperación automática de enlaces rotos. Los resultados indican que los términos del texto del ancla pueden ser muy útiles, especialmente si hay más de uno y si contienen alguna entidad nombrada. Por otro lado ha sido estudiado el efecto que tiene el uso de diferentes métodos de extracción de terminología aplicados a determinadas fuentes de información, tales como la página que contiene el enlace roto, el texto que rodea al hiperenlace, la dirección de destino del hiperenlace, y una versión en la caché de algún motor de búsqueda o almacenada en alguna biblioteca digital. También se han empleado ciertos recursos de la infraestructura web para obtener más información acerca de las páginas que han desaparecido, pero de las que aún existe información en Internet. En concreto, el sistema ha extraído términos de tecnologías web de reciente aparición tales como sistemas de etiquetado social y herramientas públicas disponibles, proporcionadas por los motores de búsqueda como por ejemplo la API de búsqueda *BOSS* de *Yahoo!*.

Este estudio ha mostrado que los resultados que se obtienen son mejores cuando se utiliza expansión de consultas que cuando se usa el texto del ancla de manera aislada. De esta forma, la expansión de consultas reduce la ambigüedad que supone la cantidad limitada de los términos del ancla. Se ha realizado una comparación de los diferentes métodos de extracción de terminología con el objetivo de ampliar el texto del ancla. Los experimentos han demostrado que los mejores resultados se obtienen utilizando un enfoque basado en modelos de lenguaje a partir de fuentes de información tales como la versión en caché, el contexto y la dirección Url del enlace roto. Por otra parte, la manera más eficiente de obtener los términos de la página que contiene el enlace es mediante un enfoque basado en frecuencia.

El sistema de recuperación de enlaces rotos, combina diferentes métodos de extracción de terminología y el uso de un método de ranking adaptado a esta tarea en particular, a fin de presentar al usuario una lista ordenada con las páginas candidatas a reemplazar el enlace roto. También ha sido llevado a cabo un estudio comparativo entre los diferentes métodos de ranking mediante el uso de varias fuentes de información procedentes de las páginas de origen y destino. El sistema emplea diversos coeficientes de coocurrencia y un enfoque basado en la divergencia de modelos de lenguaje para llevar a cabo el ranking de las páginas candidatas.

Para evaluar el sistema de recomendación, hemos desarrollado una metodología novedosa, sin recurrir a los juicios de usuario, aumentando así la objetividad de los resultados. Para esta evaluación se ha construido una colección de páginas web con enlaces rotos. A través de la metodología de evaluación propuesta, ha sido posible determinar la cantidad óptima de términos utilizados para la expansión de una consulta y los resultados que deben ser recuperados de los motores de búsqueda. Además, esta metodología nos ha permitido realizar una evaluación empírica de la disponibilidad y eficacia de las fuentes de información.

El resultado de este análisis ha permitido diseñar una estrategia que ha sido capaz de recuperar una página que podría reemplazar a otra desaparecida en el 78 % de los casos (1559 de 1998 enlaces rotos). Además, el sistema es capaz de proporcionar el 47 % de estos enlaces recuperados entre los 10 primeros documentos recomendados, y entre los 20 primeros en un 71 % de los casos.

CAPÍTULO **5**

Detección de Web Spam

5.1. Introducción

El spam en buscadores o web spam es uno de los principales problemas de los motores de búsqueda en la actualidad, ya que degrada de una manera considerable la calidad de los resultados. En muchos casos, los usuarios de Internet se sienten frustrados por la aparición constante de sitios de spam en los resultados mostrados por el buscador, cuando realmente están buscando algún contenido legítimo. Además, el web spam tiene un impacto económico, ya que un ranking alto proporciona una gran publicidad gratuita y por lo tanto un aumento en el volumen de tráfico web. Durante los últimos años ha habido muchos avances en la detección de estas páginas fraudulentas, pero como respuesta, han aparecido nuevas técnicas de spam. La investigación en este área se ha convertido en una carrera armamentista para luchar contra un adversario que constantemente utiliza métodos cada vez más sofisticados. Por esta razón, es necesario mejorar las técnicas anti-spam para superar estos ataques. El *spamdexing*, como también se conoce al web spam, incluye todas las técnicas utilizadas con el objetivo de conseguir un ranking alto para una o varias páginas web, sin merecerlo legítimamente. En términos generales, hay tres tipos de web spam: spam de enlaces, spam de contenido, y *cloaking* o encubrimiento, descrita en la sección 3.1.1. Sin embargo, el spam de enlaces y el spam de contenido son los tipos más comunes, y los consideradas en esta tesis.

De acuerdo con Brian Davison[Dav00b] el spam de enlaces, presentado en la sección 3.1.1, se puede definir como "aquellos enlaces entre páginas que están presentes por otras razones diferentes al mérito". Teniendo en cuenta esta descripción hecha por Davison, el spam de enlaces consiste en la creación de una estructura de enlaces cuyo objetivo es aprovechar la información conocida acerca de la influencia de los enlaces en los algoritmos de ranking, tales como *PageRank*. Es bien conocido que este tipo de algoritmos otorgan una ranking mayor a aquellos sitios que

poseen un gran número de enlaces entrantes, y aún mayor si estos sitios están bien enlazados a su vez.

El spam de contenido incluye aquellas técnicas de spam que implican la alteración de la visión lógica que un motor de búsqueda debería tener acerca del contenido de una página[GGM05]. Un ejemplo de este tipo de spam es la inserción de palabras clave, que están más relacionadas con las estadísticas públicas de los términos más buscados en Internet, que con el contenido real de la página.

Una de las técnicas más exitosas en el ámbito de la detección de web spam, como se puede ver en las competiciones establecidas en el AIRWeb[1] (Adversarial Information Retrieval on the Web), es la definición de un conjunto de rasgos que tomen valores diferentes para las páginas de spam y no spam. Estos rasgos son a su vez usados para entrenar un clasificador de aprendizaje automático capaz de detectar páginas de spam.

En este capítulo de la tesis también adoptaremos el esquema descrito anteriormente y propondremos nuevas características para caracterizar las páginas de spam. Sin embargo, aunque la mayoría de los trabajos previos que han tratado la detección del spam de enlaces y contenido han empleado un conjunto de características cuantitativas, en esta tesis se propone el uso de un nuevo conjunto de rasgos cualitativos[QND07] para mejorar la detección de web spam. Estas características están a su vez divididas en dos grupos: *(i)* Un grupo de características basadas en los enlaces que comprueban la fiabilidad de los vínculos establecidos entre diferentes páginas web, y *(ii)* un grupo de características acerca del contenido, extraídas con la ayuda de un enfoque basado en modelos de lenguaje. Por último, se ha construido un clasificador automático que combina ambos tipos de características, alcanzando una precisión que mejora los resultados de cada tipo por separado y los obtenidos por otras propuestas que presentaremos en las siguientes secciones.

En esta tesis, se propone el uso de algunos rasgos que midan de alguna manera la calidad de los enlaces en una página. Típicamente, los enlaces de las páginas que no utilizan técnicas de spam enlazan a otras páginas legítimas y además estos enlaces están debidamente descritos por el correspondiente texto del ancla y su contexto. Es decir, las páginas que no son de spam, generalmente no están involucradas en estrategias conjuntas de spam y por tanto seleccionan sus enlaces de una manera natural, describiendo tales enlaces con una relativa corrección. Sin embargo, las páginas que desarrollan técnicas de spam suelen tener relaciones con otras páginas del mismo tipo, y además suelen tener determinadas características sospechosas en su contenido, como por ejemplo la naturaleza de sus enlaces. De esta forma, se propone la extracción de una serie de características que capturen estas diferencias entre las páginas de spam y el resto. En el capítulo anterior, se describía cómo el texto del ancla de un enlace era una fuente de información muy útil para recuperar la página apuntada. El proceso de recuperación, consistía entonces en una consulta con los términos del ancla en un motor de búsqueda. En este capítulo se intenta modelar

[1] http://airweb.cse.lehigh.edu/

la calidad de los enlaces que apuntan a una determinada página web. El proceso consiste en intentar recuperar dicha página mediante una metodología similar a la empleada en el capítulo anterior. Es natural que la información asociada a un enlace (términos de la Url, texto del ancla y otros términos en el contexto de un enlace) permita recuperar en una posición relativamente alta dicha página. De esta forma, es de esperar que las páginas legítimas puedan ser recuperadas mediante dicha información, mientras que esta afirmación, intuitivamente, no debería cumplirse en las páginas de spam. En términos generales, se espera un comportamiento diferente en la capacidad de recuperación para los enlaces fiables y los engañosos, que pueda ser utilizado como una característica de discriminación para el clasificador.

Además de esta característica que define la capacidad de recuperación de un enlace, se han empleado otros rasgos que tratan de medir la calidad de los enlaces de una página web. Entre estos rasgos se encuentran la cantidad de enlaces rotos en la página analizada, la tipología de los enlaces entrantes y salientes (internos y externos) y algunas características acerca del texto del ancla que describen por si mismos la calidad de un enlace.

Con el objetivo de completar un conjunto de rasgos para la detección de web spam, se ha considerado la extracción de otro conjunto de rasgos de una naturaleza totalmente diferente a los anteriores. En esta ocasión se trata de captar la coherencia entre una página y las páginas que apunta. Es de esperar un cierto grado de relación entre la información asociada a un enlace en la página analizada y el contenido de la página apuntada. Para medir esta coherencia, se ha recurrido a un enfoque basado en el modelado del lenguaje. Los modelos de lenguaje[PC98], descritos en la sección 2.3.4, son métodos probabilísticos que fueron desarrollados para capturar rasgos lingüísticos ocultos en los textos, tales como la probabilidad de las palabras o secuencias de palabras en un lenguaje. Además, algunos trabajos previos han demostrado que las técnicas de discrepancia entre modelos de lenguaje son muy eficientes en tareas tales como la detección de blog spam[MCL05] o la detección de enlaces nepotísticos[BBCU06].

Por lo tanto, en este capítulo se utiliza una extensión del método básico de modelado del lenguaje para analizar varias fuentes de información extraidas de cada sitio web de la colección de referencia. La coherencia entre diferentes combinaciones de las fuentes consideradas, define un conjunto de características para el clasificador de spam.

Con el fin de evaluar el sistema, es necesario recurrir a alguna colección de páginas que hayan sido previamente etiquetadas manualmente como spam o no-spam, y por lo tanto puedan ser utilizadas para entrenar y evaluar el sistema de detección. En esta tesis se ha evaluado la capacidad del sistema para detectar los dos tipos principales de web spam: enlaces y contenido. Para ello se han utilizado dos colecciones de referencia ampliamente utilizadas en el área de web spam[CDB+06].

5.2. Detección de Web Spam mediante el Estudio de la Calidad de los Enlaces

Los hiperenlaces de una página web caracterizan en cierto modo dicha página. Si atendemos a la estructura de un documento *html*, este se compone básicamente de dos elementos; el contenido y los hiperenlaces. En esta sección se analizan una serie de medidas que combinan la información presente tanto en el contenido como en los hiperenlaces de una página, para determinar la calidad de sus enlaces. En esta tesis, el concepto de calidad de los enlaces se describe como la capacidad de una página de ser recuperada en un motor de búsqueda, empleando las técnicas apropiadas y la información relativa a ella que se encuentra en las páginas que la enlazan.

Generalmente, cuando un autor construye una página web y añade un conjunto de enlaces, se sobreentiende una relación de contenido entre la página origen y la enlazada. Precisamente, este supuesto no se cumple en el caso de las páginas de spam, ya que por ejemplo, se dan casos como el de una página que enlaza a otra anunciando en el texto del ancla términos como “gratis”, “vídeos”, “viagra”, “mp3”, etc. En cambio, a la hora de visitar dicho enlace, la página no contiene ninguna información acerca de estos conceptos señalados en la página de origen. Estos hechos se producen frecuentemente en la navegación de un usuario en Internet, y son debidos principalmente a que las páginas de spam utilizan técnicas para hacer creer a los usuarios y al buscador, que una o varias páginas web tienen información sobre una serie de productos ampliamente solicitados, cuando realmente no es así y el único motivo es la visita de usuarios. De esta forma, el concepto de calidad de los enlaces de una página se enfoca en este tipo de sucesos, con el objetivo de encontrar los casos en los que una página no puede ser recuperada según la información que versa sobre ella.

Como parte de esta tesis, se propone un análisis profundo de los hiperenlaces desde un punto de vista de la calidad, continuando con el concepto definido en el trabajo de Qi et al.[QND07]. Este análisis cualitativo no ha sido diseñado para estudiar la topología de la red, ni las características de los enlaces en un grafo. Principalmente, este estudio trata de encontrar enlaces nepostísticos[Dav00b, BBCU06], que están presentes por motivos distintos al mérito. Estos enlaces son generalmente creados teniendo en cuenta una estrategia; el objetivo de crear granjas de enlaces que permitan mediante una determinada estructura, incrementar la importancia de una o varias páginas que forman parte de la organización.

En este capítulo, se propone el estudio de un conjunto de parámetros que determinen la calidad de un sitio web. Algunos de estos parámetros se refieren directamente al tipo y cantidad de los enlaces de una página, tales como los enlaces rotos que se encuentran un dicha página o la proporción entre los enlaces internos y externos o entre los enlaces salientes y entrantes. Otros parámetros se centran en otros aspectos como el contenido del texto de anclaje; si está formado únicamente por

una Url, un número, un signo de puntuación o incluso si dicho texto es simplemente una cadena vacía. A pesar de que dichos detalles puedan parecer insignificantes, por un lado denota en el autor de una página poco interés por añadir un enlace de calidad, y por otro lado resulta sorprendente la cantidad de páginas de spam que apuestan por utilizar este tipo de hiperenlaces sin valor descriptivo. Por último, se presentan otro conjunto de parámetros que están relacionados con diferentes aspectos que definen la coherencia entre un enlace (texto del ancla, contexto del enlace, etc.) y la página apuntada, o entre la página que contiene el enlace y la página destino de cada uno de sus hiperenlaces. La Figura 5.1 muestra un ejemplo en el que una página de spam contiene un conjunto de enlaces con una pésima calidad. En dicho ejemplo, una página que corresponde a una tienda de libros y que aparentemente no es de spam, contiene un enlace a una página dedicada a ganar dinero. Esta última página contiene enlaces cuyo texto de anclaje es limitadamente descriptivo, conteniendo únicamente números y direcciones web. Nótese que en este ejemplo, es practicamente imposible para cualquier motor de búsqueda el poder recuperar la página apuntada, usando únicamente la información que proporciona la página de origen: “Great News” o “Books”.

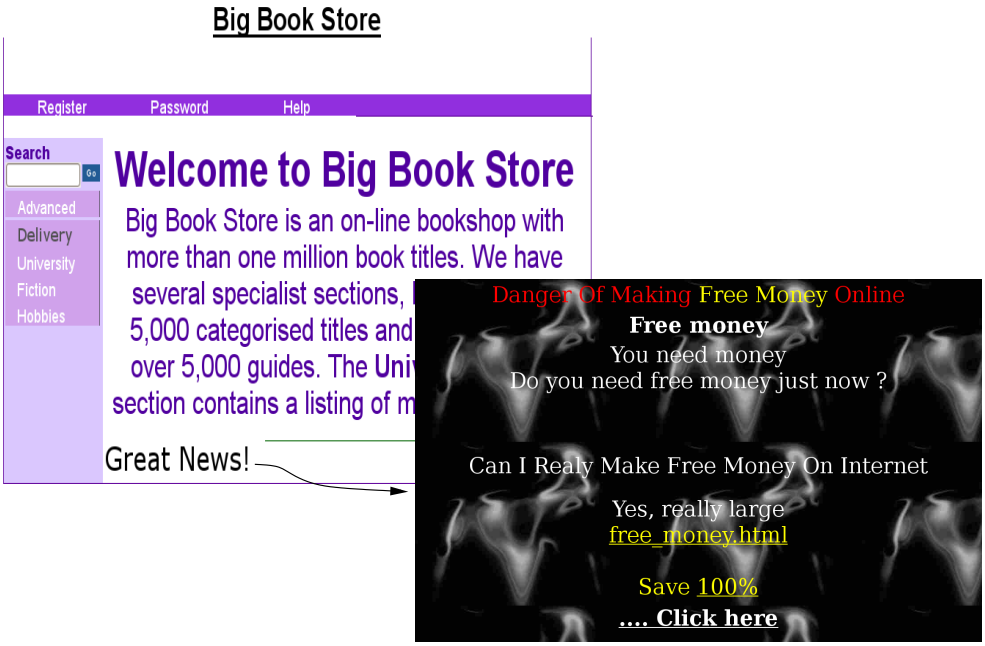

Figura 5.1: *Enlace sospechoso entre dos páginas web cuyas temáticas son muy diferentes: una página de venta online de libros y una página para ganar dinero en Internet.*

Para esta tarea, se ha desarrollado un sistema de recuperación de información que proporciona un factor de calidad por cada página analizada, que a su vez está representado por un conjunto de características acerca de sus enlaces. Esta información es muy útil ya que gracias a ella, el sistema va a ser capaz de detectar un gran número de enlaces cuyo único objetivo es ascender en el ranking de un motor de búsqueda mediante la creación de una red de granjas de enlaces.

Dado que vamos a utilizar una colección de referencia en la que cada sitio web está etiquetado, podría entenderse que existe una notable diferencia de tiempo entre el proceso de etiquetado y la extracción de estas características mediante el sistema de recuperación. Entre estos dos instantes de tiempo, ciertas características de las páginas de la colección analizadas podrían haber cambiado. Sin embargo, este intervalo de tiempo sólo pudo haber empeorado los resultados de la detección de spam, variando las características de las páginas etiquetadas. Por lo tanto, los resultados que se muestran en esta tesis es posible que pudieran ser mejores si la extracción de los rasgos y el proceso de etiquetado se obtuvieran al mismo tiempo.

5.2.1. Análisis de las Características de los Enlaces

En esta sección se describen las principales medidas que se han obtenido en esta tesis para evaluar la calidad de los enlaces de una página web. Para esta tarea, se ha empleado parte del sistema de recuperación de enlaces rotos descrito en el capítulo 4, que a partir de ahora denominaremos sistema de extracción de rasgos de calidad (SERC). A pesar de que este sistema fue diseñado originalmente para la recuperación de enlaces rotos, su funcionalidad ha sido modificada para analizar un conjunto de características acerca de los enlaces activos de una página web. Sobre la base del sistema de recuperación de enlaces, se ha realizado un conjunto de modificaciones. Principalmente el sistema ha sido adaptado con el objetivo de poder analizar todas las páginas de la colección. También la salida del sistema es diferente, ofreciendo un conjunto de medidas acerca de cada página web de tal forma que puedan ser empleadas por un clasificador automático. De esta forma, el SERC no sólo ofrece información acerca del número de enlaces recuperados, sino que muestra mucha más información sobre la calidad de dichos enlaces, representada en forma de un conjunto de valores que pasamos a analizar a continuación. Esta información se refiere a datos sobre la tipología de un enlace, el texto del ancla de cada enlace, el número de sitios que enlazan a la página analizada o la relación entre los enlaces salientes y entrantes.

El SERC analiza los enlaces de cada página y extrae un conjunto de rasgos de calidad acerca de ella. Este sistema se basa en un conjunto de técnicas clásicas de recuperación de información y procesamiento del lenguaje natural para analizar cada uno de los enlaces de cada página. El SERC se compone principalmente de dos fases:

1. **Extracción de Información relevante acerca de un enlace.**
   Existen muchos trabajos que han analizado la importancia del texto del ancla como una fuente de información[CHR01], de tal forma que los términos que lo componen se utiliza como la principal herramienta para recuperar un enlace. Sin embargo, hay muchos casos en los que el ancla no contiene suficiente información. Por esta razón, el sistema realiza una extracción de terminología de otras fuentes de información tales como la dirección de destino del hiperenlace, la página que contiene dicho enlace, el contexto del ancla.

   El sistema utiliza varios métodos de extracción de terminología basados en la frecuencia y en modelos de lenguaje estadísticos, específicamente divergencia *Kullback-Liebler* (KLD), dependiendo de la fuente de información considerada. En concreto, *Tf-Idf* se utiliza para extraer la terminología de la página que contiene el enlace, y *KLD* para extraer la terminología del resto de fuentes de información.

2. **Construcción de consultas y recuperación de los resultados del buscador.**
   La consulta inicial se compone de los términos extraídos del texto del ancla, y esta consulta se expande utilizando los términos extraídos de las otras fuentes de información descritas anteriormente. El conjunto de consultas se ejecutan en un motor de búsqueda, y los diez primeros resultados devueltos por el buscador son recuperados. Para esta tarea de detección de web spam, se considera que un vínculo se ha recuperado si la página apuntada por el enlace analizado se encuentra en el conjunto de páginas recuperadas mediante alguna de las consultas realizadas.

El SERC proporciona una serie de valores acerca de la calidad de los enlaces de una página. Estos valores se basan en un conjunto de medidas que pasamos a analizar.

Grado de Recuperación.

El valor más importante que se extrae gracias al SERC es precisamente el grado de enlaces recuperados. Por cada página, el sistema intenta recuperar todos sus enlaces y como resultado, se obtienen tres valores relacionados con el proceso de recuperación: *(i)* El número de enlaces recuperados, *(ii)* el número de enlaces que no han podido ser recuperados y *(iii)* la diferencia entre los dos valores anteriores, que se representa en la Figura 5.2. En dicha figura se puede observar cómo las páginas de los dos grupos se concentran en un conjunto de valores de la distribución diferentes. También se muestra claramente cómo la tasa de enlaces recuperados en el caso de las páginas de spam es notablemente inferior al de las páginas legítimas. Esta divergencia entre los dos conjuntos de valores tiene un gran valor para el clasificador a la hora de discriminar entre los dos tipos de páginas.

El grado de enlaces recuperados se puede entender como una medida de la coherencia entre la página analizada, uno de sus enlaces y la página apuntada por dicho enlace. La intuición en la interpretación de esta función, es que una página que pertenece a una granja de enlaces se suele conectar a otras páginas desconocidas (que tienen el mismo objetivo) con la única finalidad de aparecer en la parte superior del ranking de los motores de búsqueda. Por lo tanto, estos enlaces son difíciles de recuperar. Así, cuanto más negativa sea la diferencia entre los enlaces recuperados y los que no, mayor será la probabilidad de que este sitio esté haciendo spam.

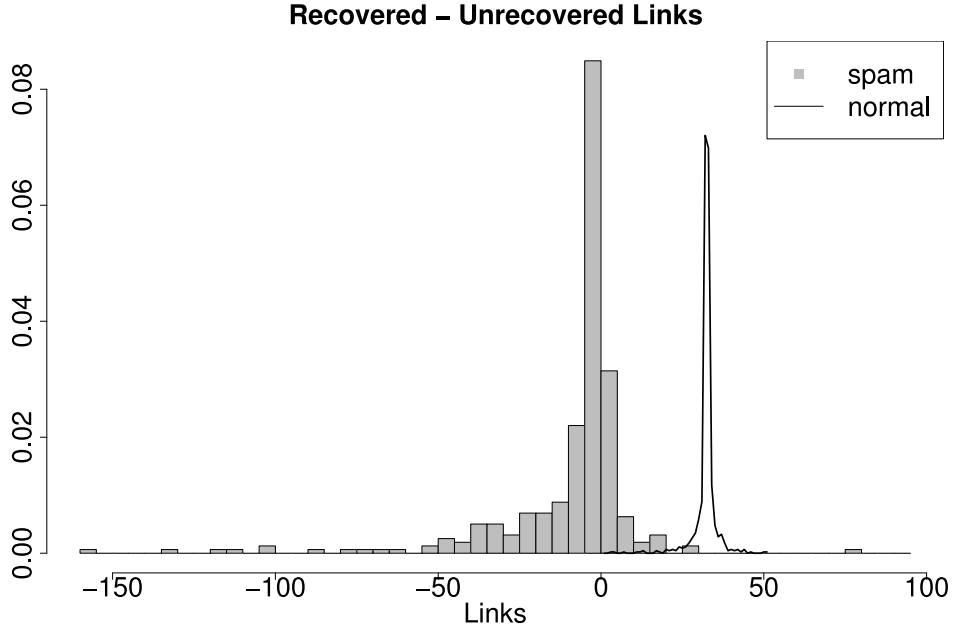

Figura 5.2: *Histograma de la distribución de valores según la diferencia entre los enlaces recuperados y aquellos no recuperados.*

Enlaces entrantes y enlaces salientes.

Es de sobra conocido que las páginas de spam enlazan a páginas legítimas, mientras que las páginas que no hacen spam no incluyen enlaces hacia páginas de spam. Los enlaces entrantes son aquellos que una determinada página web recibe por parte de otras páginas. De esta misma forma, los enlaces que contiene una página cuyo destino es una página diferente, se denominan enlaces salientes. Aprovechando las posibilidades del sistema de extracción de rasgos de calidad para enviar consultas a un motor de búsqueda, se ha incluido una nueva consulta para solicitar el número de sitios que apuntan a la página analizada (enlaces entrantes). La Figura 5.3 representa la diferencia entre la cantidad de enlaces de cada tipo. En esta figura, se puede apreciar como la forma de las distribuciones es diferente, teniendo las páginas de spam una distribución con valores contenidos en un intervalo

menor. Las páginas que no hacen spam, por el contrario distribuye sus valores de una manera más uniforme, no teniendo un pico tan acusado como en el caso de las páginas de spam.

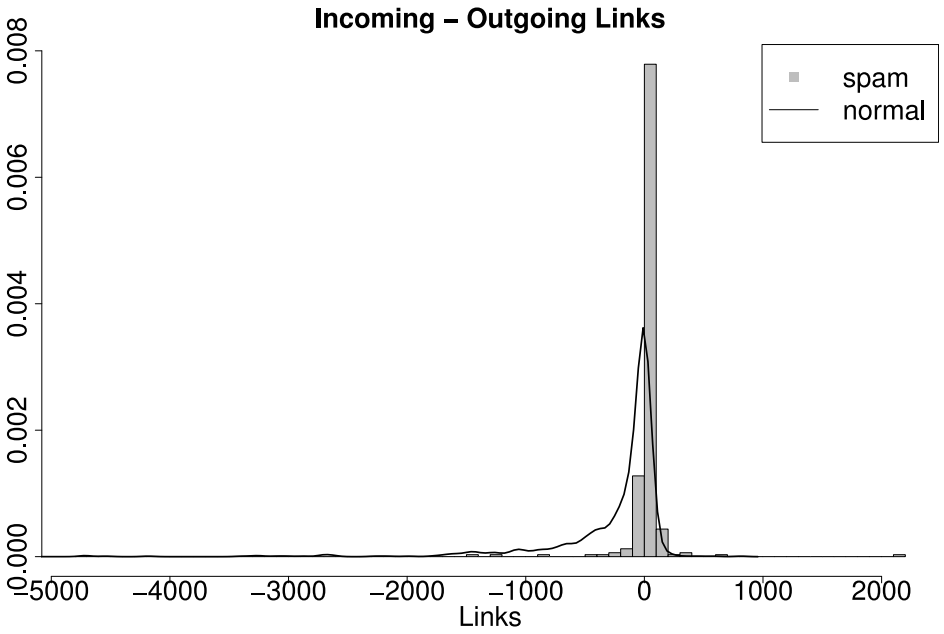

Figura 5.3: *Histograma de la distribución de valores según la diferencia entre los enlaces entrantes y los enlaces salientes.*

Enlaces externos frente a enlaces internos.

Los enlaces externos son aquellos que una determinada página web realiza a páginas de un sitio web diferente al que pertenece. En muchos casos se corresponde también con un dominio diferente. Los enlaces que contiene una página cuyo destino es otra página dentro del mismo sitio web o dominio, se denominan enlaces internos. Existen diversas teorías sobre el impacto de los enlaces internos y los enlaces externos en el *PageRank* asignado a un sitio web. Estas teorías, disponibles en sitios web y foros especializados, se basan en ingeniería inversa sobre los resultados que proporciona un buscador, con el objetivo de extraer los detalles del algoritmo que utiliza. En este tipo de publicaciones, se muestran detalles para incrementar la valoración de una página web. De esta forma, uno de los consejos para mejorar esta valoración es tener en cuenta la proporción entre los enlaces externos e internos.

Aunque no existe una evidencia definitiva para demostrarlo, existe la creencia de que muchos sitios web aplican estas teorías. Por esta razón, se ha tomado el número de enlaces externos e internos para caracterizar a una determinada página web con su proporción entre este tipo de enlaces. La Figura 5.4 representa el ratio

de este tipo de enlaces, en función del grupo (spam y no-spam) en el que se ha etiquetado cada página. En esta figura se observa cómo las páginas de spam emplean una proporción mayor de enlaces externos que las páginas legítimas. El motivo de esta divergencia entre la distribución de valores para un grupo y otro se pude deber por un lado a la existencia de este tipo de teorías descritas anteriormente, o simplemente al hecho de que las páginas de spam no suelen tener un gran interés en enlazar a páginas del mismo sitio web ya que su influencia en la valoración es practicamente nula. Sin embargo, mediante la inserción de enlaces externos a otras páginas que participan en granjas de enlaces, la importancia de los sitios web se ve incrementada significativamente.

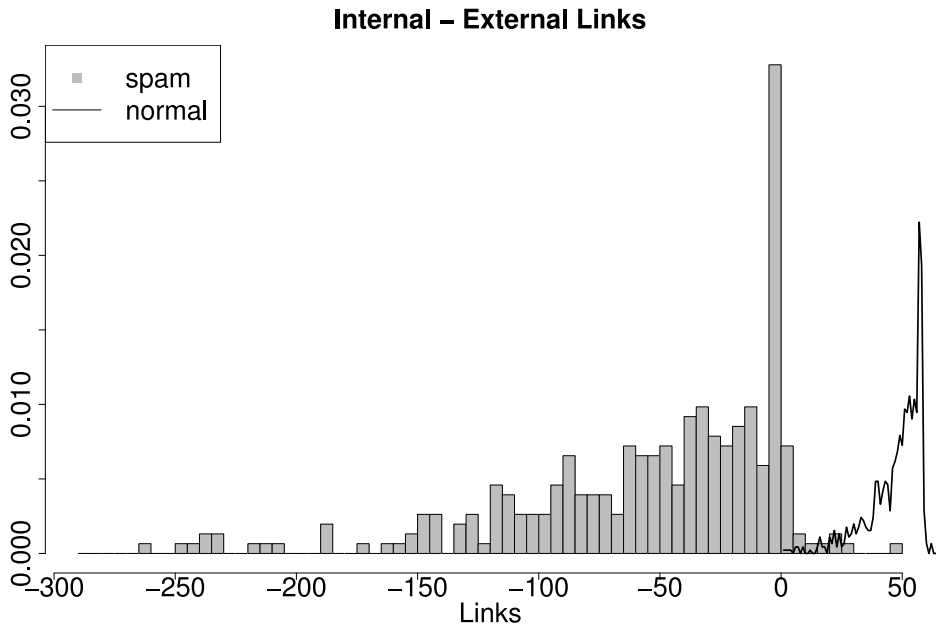

Figura 5.4: *Histograma de la distribución de valores según la diferencia entre los enlaces externos y los enlaces internos.*

Enlaces rotos.

Un enlace se considera roto cuando la respuesta del servidor que aloja dicha página consiste en el código de error 404. Los enlaces rotos son un problema tanto para las páginas de spam como para el resto de las páginas. Además de las consecuencias negativas en el usuario potencial que intenta visitar una página y la desconfianza que supone encontrar en una página varios enlaces rotos, la presencia de este tipo de errores repercute directamente en la valoración que los buscadores otorgan a una página. Es de sobra conocido que los motores de búsqueda penalizan las páginas que tienen fallos, y en concreto este tipo de errores son tenidos en cuenta a la hora de elaborar un ranking de resultados.

Por otro lado, la longevidad de las páginas de spam es menor que el resto de páginas. Las páginas de spam son generadas generalmente para promocionar una o varias páginas, y de esta forma si alguna de las páginas generadas pierde su valor, el spammer pierde el interés por su existencia. Los motores de búsqueda utilizan técnicas para evitar ataques de spam y eliminan de sus índices tanto las páginas que son detectadas como aquellas relacionadas. De esta forma, los spammers crean y destruyen páginas web de una forma muy dinámica dependiendo del tiempo que estas puedan resultarle útiles. Y en consecuencia, el número de enlaces rotos que puede encontrarse en un ámbito de web spam puede resultar superior al del resto de las páginas, dependiendo de si alguna de ellas ha sido detectada.

Por este motivo, se propone en análisis de este tipo de enlaces con el objetivo de encontrar algún patrón común entre las páginas de spam y las páginas legítimas. La Figura 5.5 muestra los valores obtenidos para las páginas de spam y las páginas normales. Se puede observar que, aunque es frecuente encontrar algún enlace roto en cualquier tipo de página, sólamente en las páginas de spam se encuentran grandes cantidades de este tipo de enlaces.

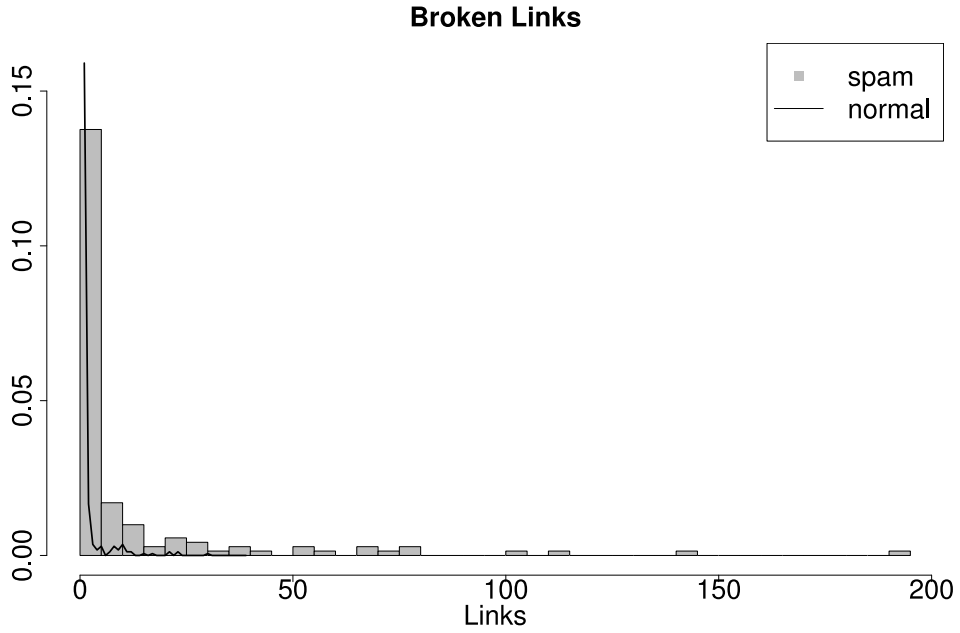

Figura 5.5: *Histograma de la distribución de valores según el número de enlaces rotos.*

Signos de Puntuación en el ancla.

Actualmente es habitual que las páginas de spam contengan texto y enlaces generados automáticamente. Además, el texto del ancla de muchos enlaces suelen generarse pensando en el contexto de los motores de búsqueda en lugar de los usua-

rios. Por lo tanto, se han seleccionado cuatro características con el fin de medir el número de enlaces generados de esta manera que hayan sido formados únicamente por: *(i)* signos de puntuación, *(ii)* dígitos, *(iii)* Url y *(iv)* una cadena de texto vacía. En el caso de los signos de puntuación, es frecuente encontrar páginas con este tipo de símbolos cuyo objetivo en muchos casos es difícil de comprender. En algunos casos encontrar una arroba "@" o un dolar "$" puede tener connotaciones relacionadas con Internet o con el dinero. Pero en otros casos como el uso del punto ".", es difícil encontrar un valor descriptivo en su empleo. La Figura 5.6 muestra la comparación entre el número de enlaces, cuyo texto de anclaje está formado solamente por un signo de puntuación, encontrados en el caso de las páginas de spam y las páginas que no son de spam. Aunque no es uno de los rasgos más discriminantes, si que se puede observar que un número elevado de este tipo de enlaces es más frecuente en las páginas de spam.

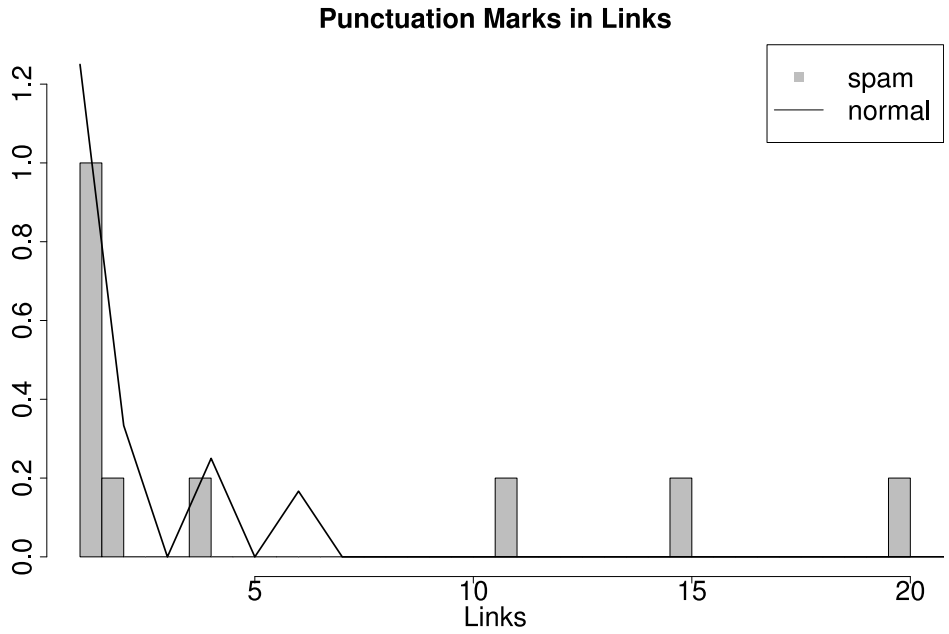

Figura 5.6: *Histograma de la distribución de valores según el número signos de puntuación en el ancla.*

Texto del ancla vacío.

Como se ha mencionado anteriormente, es habitual que las páginas de spam contengan texto y enlaces generados automáticamente. Otra de las consecuencias de este tipo de hábitos son los errores que permiten a un texto de anclaje no disponer de ningún texto que describa la página enlazada. Por otro lado, algunos spammers desean tener un gran número de enlaces en sus páginas web sin que el usuario se percate de la presencia de estos. Si recordamos alguna de las técnicas de spam de

contenido, los spammers solían añadir texto oculto con términos frecuentemente buscados en Internet con el objetivo de ser relevantes para los buscadores pero no llamar la atención entre los usuarios. Para este fin, los spammers utilizaban varias técnicas, y entre ellas se encontraba la de asignar al texto de la fuente de esos términos el mismo color que el fondo de la página web. De esta forma, el usuario ignoraba la presencia de estas listas de términos repetidos. En el caso que nos ocupa, la técnica es muy similar y consiste en eliminar el texto del ancla de los enlaces. De esta forma, el usuario no es capaz de ver los múltiples enlaces que apuntan a una página o varias páginas en concreto, mientras que el motor de búsqueda si detecta la presencia de estos enlaces al parsear el contenido de la página. La Figura 5.7 muestra la comparación entre el número de enlaces, cuyo texto de anclaje está formado solamente por una cadena vacía, encontrados en el caso de las páginas de spam y las páginas que no son de spam. De nuevo se observa que la presencia de este tipo de enlaces tiende a ser más frecuente en las páginas de spam.

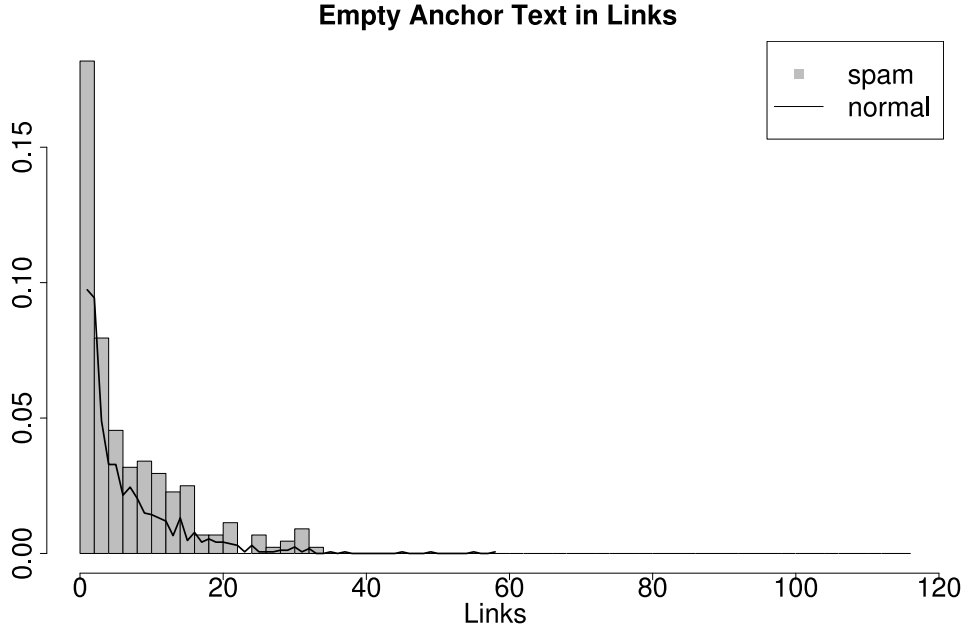

Figura 5.7: *Histograma de la distribución de valores según el número de anclas cuyo texto es una cadena vacía.*

Texto del ancla formado por dígitos.

Otra de las consecuencias de la generación automática de enlaces es la inserción de un conjunto de dígitos para describir el texto del ancla. En la mayoría de los casos, este hecho se produce al insertar como texto del ancla el identificador de dicho recurso en alguna de las bases de datos almacenadas. De esta forma, el enlace queda representado por un conjunto de caracteres, pero su valor descriptivo es nulo de cara

al usuario. Existen otros escenarios donde se selecciona un conjunto de dígitos para formar el texto del ancla. Uno de estos ámbitos es la técnica empleada por algunos spammers de hacer relevante un enlace intentando que el texto del ancla coincida con alguna fecha o número de teléfono. La Figura 5.8 muestra la comparación entre el número de enlaces, cuyo texto de anclaje está formado solamente por dígitos, encontrados en el caso de las páginas de spam y las páginas que no son de spam. De nuevo se observa una ligera tendencia a encontrar más enlaces de este tipo en las páginas de spam.

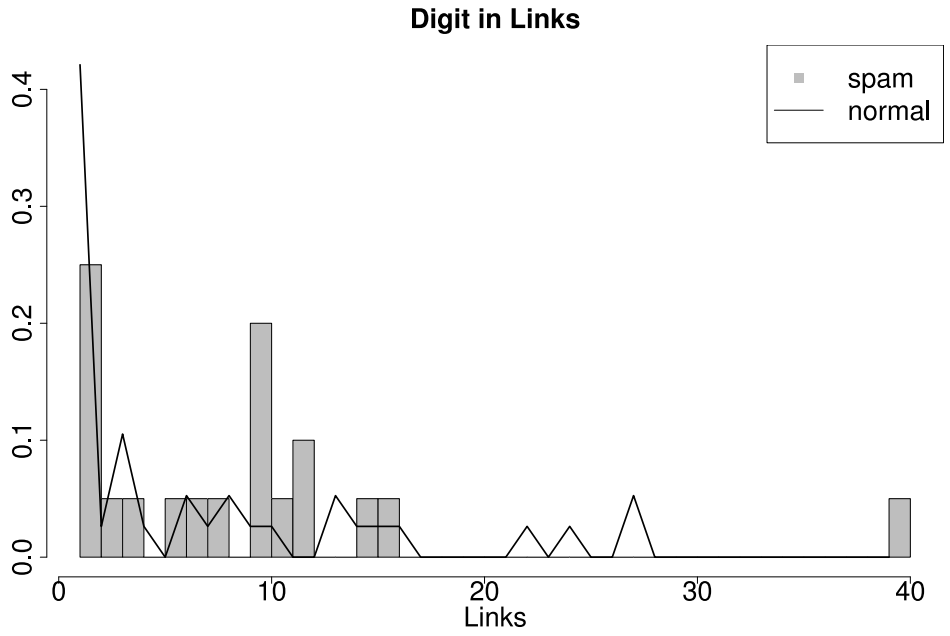

Figura 5.8: *Histograma de la distribución de valores según el número de anclas cuyo texto está formado por dígitos.*

Urls en el texto del ancla.

El texto de ancla, como hemos mencionado en varias ocasiones durante esta tesis, tiene un fuerte poder descriptivo sobre la página que enlaza. Al mismo tiempo, el texto de un ancla tiene una componente de publicidad. Si atendemos al contenido de una página web, los enlaces que contiene son una forma de atraer la atención del usuario hacia otras páginas.

El spam en la web, no solo consiste en ascender en el buscador a toda costa, sino que en muchos casos el mero hecho de cobrar a cambio de la inserción de publicidad, forma parte del negocio del web spam. Uno de los criterios para decidir si una página está haciendo spam, es mediante la inserción excesiva de publicidad. De esta forma, la inserción abusiva de publicidad en forma de enlaces también es considerada como una forma de hacer web spam.

La presencia de una Url en el texto de un ancla puede entenderse como una forma de hacer publicidad, por lo tanto, la presencia masiva de este tipo de enlaces cuyo texto de anclaje está formado únicamente por una Url podría resultar un indicador de spam.

La Figura 5.9 muestra la comparación entre el número de enlaces, cuyo texto de anclaje está formado por una Url, encontrados en el caso de las páginas de spam y las páginas que no son de spam.

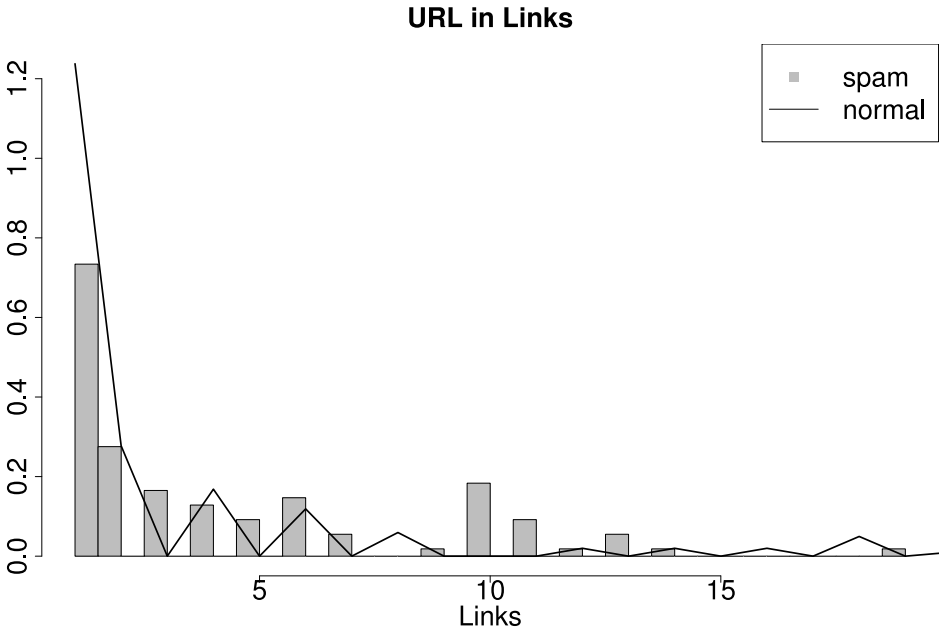

Figura 5.9: *Histograma de la distribución de valores según el número de anclas cuyo texto está formado una Url.*

En resumen, tenemos un total de 12 (alguna medida da lugar a varios rasgos) rasgos para representar cada página web. Analizando los histogramas de cada rasgo, podemos concluir que las características que ofrecen los mejores valores de divergencia entre las páginas de spam y las páginas que no realizan spam, son los siguientes (en orden de importancia): *(i)* diferencia entre enlaces recuperados y no recuperados, *(ii)* número de vínculos con un texto de anclaje vacío, y *(iii)* diferencia entre enlaces externos e internos. Sin embargo, todas las características contribuyen a los resultados del clasificador, ya que cada uno de los valores pueden discriminar casos diferentes.

5.3. Análisis de Divergencia de Contenidos

En esta sección se presenta un estudio de divergencia de contenidos aplicado a la detección de spam. Si en la sección anterior se proponía el análisis de la calidad de los enlaces, en esta sección se pretende caracterizar la información que representa la relación entre dos páginas web mediante el análisis de los contenidos de ambas páginas.

Uno de los métodos con más éxito basados en el análisis de distribución de términos utiliza el concepto de divergencia de Kullback-Liebler [CT91] (KLD) para calcular la divergencia entre las distribuciones de probabilidad de términos de dos documentos. Esta medida que ya se ha aplicado en otros ámbitos de esta tesis, en esta ocasión adopta la función de medida de similitud (valor opuesto a la divergencia). Este método tiene en cuenta las diferencias en la distribución de términos entre dos unidades de texto para calcular el valor de divergencia:

$$KLD(T_1||T_2) = \sum_{t \in T_1} P_{T_1}(t) log \frac{P_{T_1}(t)}{P_{T_2}(t)} \tag{5.1}$$

donde $P_{T_1}(t)$ es la probabilidad del término t en la primera unidad de texto, y $P_{T_2}(t)$ es la probabilidad del término t en la segunda unidad de texto.

En concreto se ha aplicado KLD para medir la divergencia entre dos unidades de texto de las páginas de origen y de destino. En la Figura 5.10 se muestran dos ejemplos de divergencia calculada mediante *KLD*, entre el texto de anclaje de un enlace y el título de la página apuntada por dicho enlace.

Los modelos de lenguaje utilizados en este trabajo estiman la máxima verosimilitud de las probabilidades de ocurrencia basadas en unigramas. En experimentos preliminares, se utilizó la función de suavizado *Jelinkek-Mercer* tal y como se presenta en Mishne et al.[MCL05], que interpola el modelo de lenguaje de dos fuentes de información. Los resultados mostraron que el suavizado mejora los resultados, aunque la diferencia era muy pequeña. Sin embargo el tiempo de cómputo se vio aumentado considerablemente. Por estas dos razones, se ha decidido no usar ningún método de suavizado para los modelos de lenguaje empleados en este trabajo.

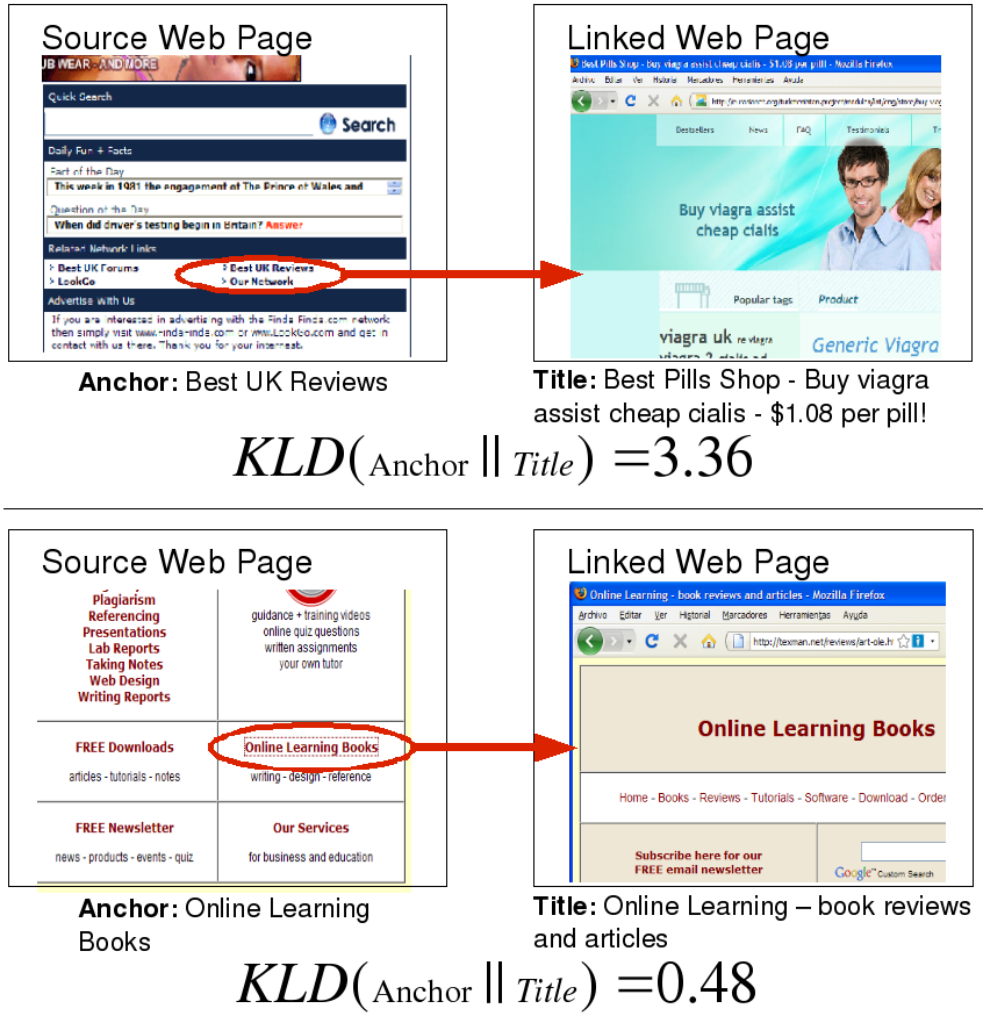

Figura 5.10: *Divergencia calculada aplicando KLD entre el texto del ancla de un enlace y el título de la página apuntada por este enlace.*

5.3.1. Divergencia entre Fuentes de Información

En esta sección se presentan los rasgos extraidos para representar una página web de cara a su posterior uso en un clasificador basado en aprendizaje automático. Los rasgos que se proponen para mejorar la detección de spam, caracterizan la relación entre dos páginas web enlazadas en función de un conjunto de valores de divergencia. Estos valores se obtienen mediante el cálculo de la divergencia de Kullback-Liebler (KLD) entre una o más fuentes de información de cada página, tal y como se mostraba en la Figura 5.10, donde se utilizaban el texto de anclaje de un enlace y el título de la página apuntada por dicho enlace.

En particular, para el cálculo de los valores de divergencia han sido empleadas tres fuentes de información de la página origen: (i) texto del ancla, (ii) contexto del enlace y (iii) términos de la dirección de destino del hiperenlace. También se han utilizado tres fuentes de información de la página destino: (i) Título de la página, (ii) Contenido de la página y (iii) Meta-Etiquetas.

Para medir la divergencia entre dos páginas web se podrían utilizar muchas combinaciones de las fuentes de información de las páginas de origen y destino. Sin embargo, atendiendo al coste computacional de cada combinación y a las características lingüísticas de cada fuente, se han seleccionado un subconjunto de pares cuyo coste computacional está contenido y además son útiles en la detección de web spam.

Por otra parte, para llevar a cabo las operaciones estadísticas con los modelos de lenguaje, se ha utilizado la librería de recuperación de información Lucene[GH04]. Los pares de fuentes de información de los que se han calculado su divergencia, se describen a continuación.

Texto de Anclaje - Contenido de la página destino.

Cuando una pagina enlaza a otra, la única manera que tiene de convencer a un usuario de visitar ese link es mostrando información suficiente y resumida de la pagina destino. Por ese motivo, la divergencia entre este fragmento de texto y la pagina apuntada muestra un claro indicio de Spam. Además, Mishne[MCL05] y Benczúr[BBCU06] probaron en ambos trabajos que la divergencia entre el texto del ancla y la página de destino es una medida muy útil para detectar web spam.

En la Figura 5.11 se muestra la distribución de los valores de divergencia KL entre estas fuentes de información y, al igual que en estudios anteriores de esta distribución, la curva de las páginas que no hacen spam es más compacta que la de las páginas de spam. El texto del ancla por sí solo no es una medida muy discriminante, pero obtiene un mejor rendimiento cuando se combina con el texto que lo rodea y los términos de la Url.

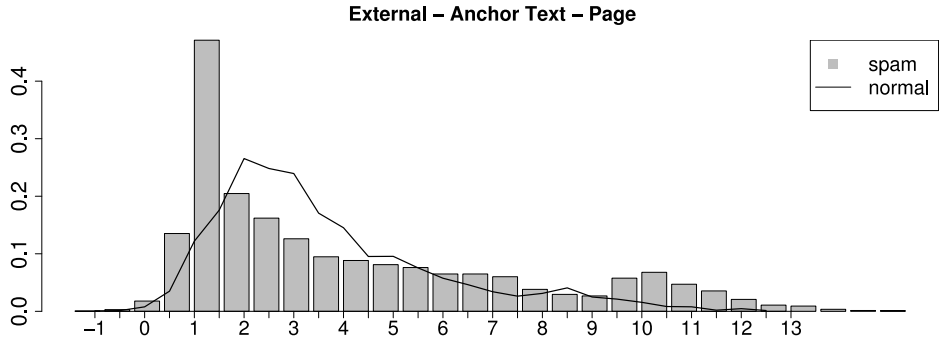

Figura 5.11: *Histograma de la distribución de valores en función de la divergencia entre el texto del ancla y el contenido de la página destino. Se han empleado los enlaces externos de las páginas de la colección de referencia* `WEBSPAM-UK2006`.

Contexto del enlace - Contenido de la página destino.

En ocasiones los términos del ancla de un enlace aportan un escaso valor descriptivo, o incluso nulo. Imaginemos un enlace cuyo texto de anclaje sea e.g. "pinche aquí". En este caso la comparación entre estos términos y los correspondientes a cualquier otra fuente de información van a generar una divergencia que no refleja realmente la diferencia entre el contenido de ambas páginas.

Por esta razón, el texto que rodea un hiperenlace puede proporcionar una información contextual sobre la página enlazada, que permita complementar los casos en los que el texto del ancla no cumple una función descriptiva. Por otra parte, Benczúr et al.[BBCU06] demostraron que el texto del ancla ofrecía un mejor rendimiento cuando era extendido con los términos del texto circundante. De cara a los experimentos realizados como parte de esta tesis, se han utilizado varias palabras alrededor del ancla para extender la información acerca del enlace. En concreto se han extraido 7 palabras en cada sentido: 7 hacia la izquierda y 7 hacia la derecha. En esta selección de términos, se han tenido en cuenta las etiquetas de bloque de HTML y los signos de puntuación tal y como se describe en varios trabajos previos[Pan03, CS07].

El resultado de esta extensión es una fuente de información con una mayor riqueza lingüística y una utilidad superior para detectar web spam. La Figura 5.12 muestra la distribución de los valores de divergencia KL entre estas fuentes de información y se puede observar como la forma de la curva de spam se desplaza hacia valores más altos de divergencia y los valores de las páginas legítimas se concentran en torno a un valor $KL \approx 2{,}5$.

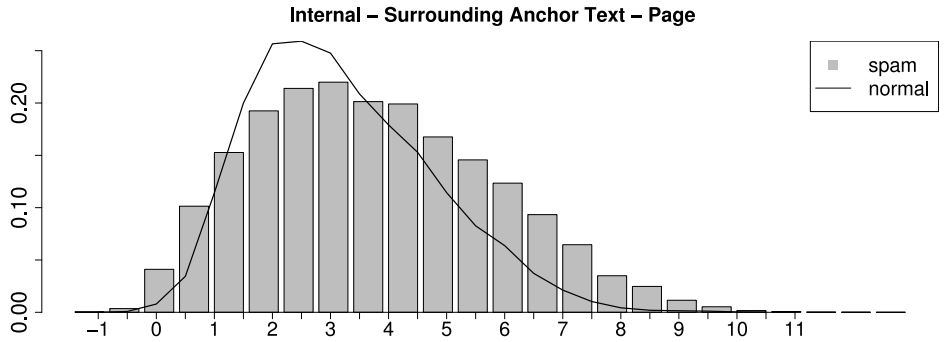

Figura 5.12: *Histograma de la distribución de valores en función de la divergencia entre el contexto del enlace y el contenido de la página destino. Se han empleado los enlaces internos de las páginas de la colección de referencia* `WEBSPAM-UK2006`*.*

Url del hiperenlace - Contenido de la página destino.

Además del texto del ancla, la única información que disponemos de un enlace es su Url. Una Url se compone principalmente de un protocolo, un dominio, un path y un archivo. Estos elementos están compuestos a su vez de términos que pueden ofrecernos información muy rica sobre la página a la que identifica. Además, en los últimos años y debido al auge de los motores de busqueda, existen técnicas de optimización de sitios web en función de los motores de búsqueda, que intentan explotar la importancia de los términos de la Url en una consulta. De esta forma, si tenemos una Url como "www.domain.com/viagra-youtube-free-download-poker-online.html" y al visitar dicha página resulta que se trata de una tienda online de musica, podría decirse que esta pagina está empleando técnicas de spam.

Con el fin de calcular la divergencia con el contenido de la página de destino, se han recuperado los términos más relevantes de cada Url. Para extraer los términos más relevantes, se ha aplicado un enfoque basado en modelos de lenguaje. En primer lugar, se ha construido un modelo de lenguaje con los términos de las Urls almacenadas en el directorio público `ODP` (Open Directory Project). Para ello han sido indexadas las Urls de este directorio, eliminando las palabras vacías y consiguiendo reunir una colección con los términos de dichas direcciones. El objetivo es utilizar esta colección como referencia para el cálculo de probabilidades que forma parte de la divergencia de *Kullback-Leibler*, que a su vez será la medida utilizada para la extracción de la terminología relevante sobre las direcciones de destino de los hiperenlaces.

Por último, y después de compilar una lista ordenada con los términos de la Url, de esta lista se seleccionan tan solo el 60 % de estos términos, con un mayor valor de divergencia, con el objetivo de eliminar términos poco discriminantes. Esta medida se ilustra en la Figura 5.13, en donde se demuestra la gran diferencia entre los histogramas de los dos tipos de páginas.

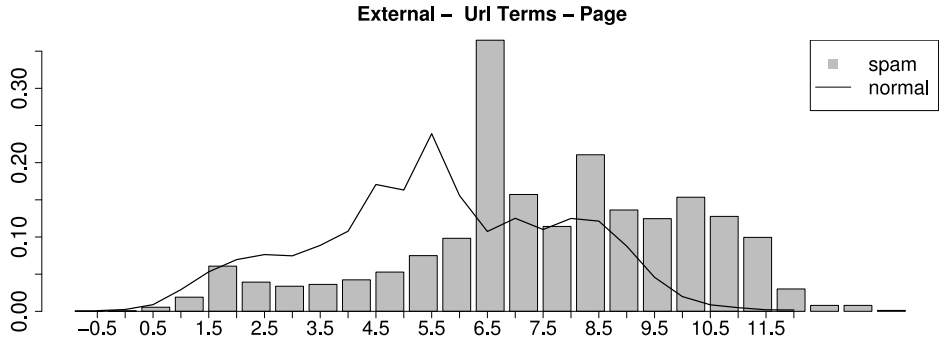

Figura 5.13: *Histograma de la distribución de valores en función de la divergencia entre la dirección de destino del hiperenlace y el contenido de la página destino. Se han empleado los enlaces externos de las páginas de la colección de referencia* `WEBSPAM-UK2007`.

Texto de Anclaje - Título de la página destino.

En la literatura, existen trabajos que expresan la similitud entre el texto de un ancla y el título de una página web. En el caso de estas dos fuentes de información, se trata de una porción de texto de tamaño reducido en el que se resume el contenido de una página. Sin embargo, atendiendo a las diferentes características lingüísticas acerca del origen de cada fuente, en el caso del texto del ancla es una persona ajena a la propia página, y en el caso del título el responsable es el autor. Esta diferencia en cuanto al enfoque entre los responsabes de generar el texto que representa la misma página, puede en algunos casos reducir la similitud entre los términos empleados, y en otros casos hacer más valiosa la comparación. Un posible inconveniente en la comparación de estas dos unidades de texto es la utilización de modelos de lenguaje. En este sentido, se podría pensar que el numero reducido de termino en las dos unidades de texto podría acarrear problemas, pero en los experimentos realizados se ha comprobado que esta medida, en muchos casos, aporta información relevante.

En la Figura 5.14 se muestra la distribución de los valores de divergencia KL entre el texto de anclaje y el título de la página apuntada por el hiperenlace.

Contexto del enlace - Título de la página destino.

De la misma manera que en la medida analizada anteriormente donde intervenían el contexto del enlace y el contenido de la página analizada, en esta ocasión se emplea el contexto del enlace con las mismas características. El texto del ancla es extendido con algunas palabras que circundan dicho hiperenlace con el objetivo de obtener una mayor información descriptiva acerca de la página analizada. Además, en este caso al contar con una fuente de información de similares características al contexto de un enlace, la comparación entre ambas unidades de texto ofrece un rendimiento muy bueno.

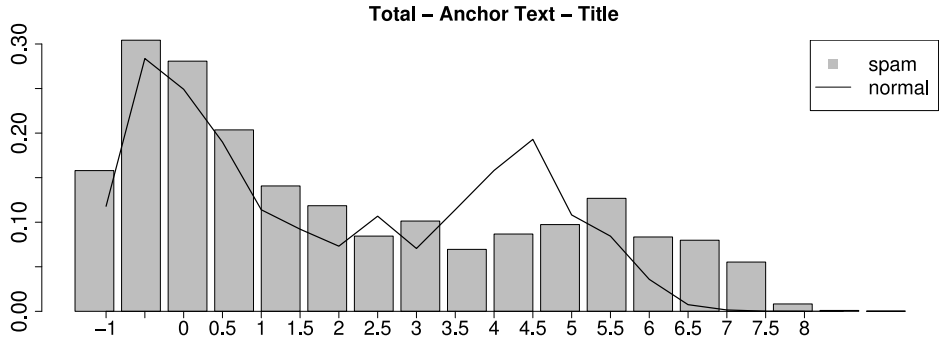

Figura 5.14: *Histograma de la distribución de valores en función de la divergencia entre el texto del ancla y el título de la página destino. Se han empleado los enlaces externos e internos de las páginas de la colección de referencia* `WEBSPAM-UK2007`*.*

En la Figura 5.15 se muestra la distribución de los valores de divergencia KL entre el contexto del enlace y el título de la página apuntada por el hiperenlace. En esta figura, se puede apreciar como el histograma es discriminante y por tanto su utilidad para la detección de web spam es muy relevante. En este caso, la mayoría de los valores de divergencia se encuentran localizados entre $KL \approx 1$ y $KL \approx 3$.

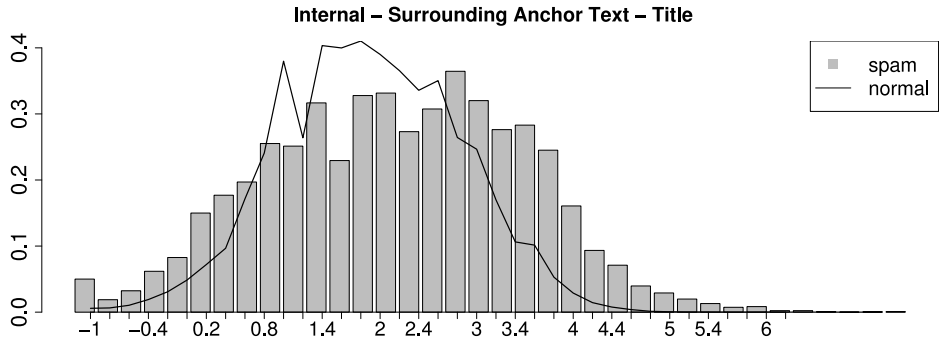

Figura 5.15: *Histograma de la distribución de valores en función de la divergencia entre el contexto del enlace y el título de la página destino. Se han empleado los enlaces internos de las páginas de la colección de referencia* `WEBSPAM-UK2006`*.*

Url del hiperenlace - Título de la página destino.

En los dos casos analizados previamente intervenía el título de la página destino como fuente de información involucrada en la comparación. Sin embargo si atendemos a la fuente de información seleccionada para realizar dicha comparación, se trataba del texto del ancla y del texto de alrededor del hiperenlace. En estos dos casos, la primera unidad de texto estaba generada por una persona ajena a la pagina apuntada. En cambio, la Url es construida por el autor de la página o al menos parte de ella, si atendemos al archivo y al camino de la dirección web. De esta

forma, entre los términos de una Url y el título de una pagina web, debería haber una determinada coherencia ya que por un lado son unidades de texto de reducido tamaño. Por otro lado el autor de ambas fuentes es el mismo y por tanto la síntesis de información que se refleja en el texto que el autor controla, debería tener unas características similares. En otro caso, si la divergencia entre ambas fuentes fuera grande, podríamos estar analizando una pagina de spam.

En la Figura 5.16 se muestra la distribución de los valores de divergencia KL entre la dirección destino del hiperenlace y el título de la página apuntada por el hiperenlace. En ella se puede apreciar como el histograma refleja una cierta discriminancia ya que los valores de las páginas de spam se encuentran más dispersos que en el caso de las páginas legítimas, que se concentran en torno al valor medio.

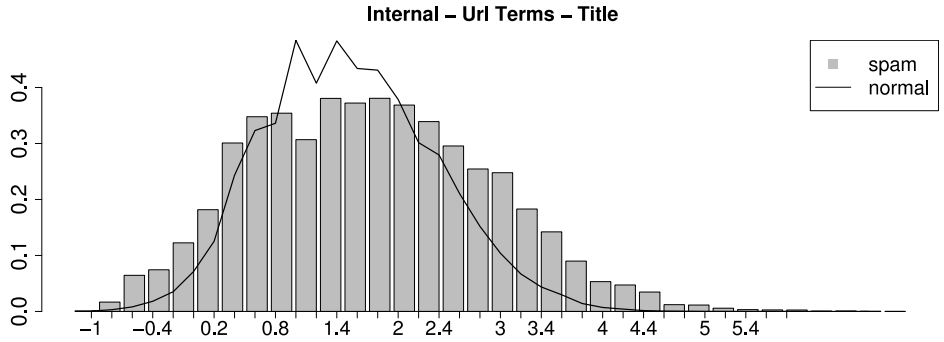

Figura 5.16: *Histograma de la distribución de valores en función de la divergencia entre la dirección de destino del hiperenlace y el título de la página destino. Se han empleado los enlaces internos de las páginas de la colección de referencia* `WEBSPAM-UK2006`.

Título de la página destino - Contenido de la página destino.

Es bien conocido que tanto los términos de la Url como los términos del título de una pagina tienen una gran peso a la hora de decidir si un documento es relevante a una consulta. Además, en los últimos años, debido al uso creciente de los buscadores, existen técnicas de optimización de motores de búsqueda (Search Engine Optimization —SEO —) que tratan de realizar ingeniería inversa sobre los algoritmos que utilizan los principales buscadores, con el objetivo de conocer los parámetros que intervienen en el ranking proporcionado y su peso aproximado. Formando parte de estas optimizaciones, se encuentra la estrategia de colocar determinados términos en la Url y el título de cada página, con el objetivo de conseguir un ranking mayor, aprovechándose de la importancia que los motores de búsqueda otorgan a dichas fuentes de información. Esta información disponible en Internet repercute en que los spammers utilicen frecuentemente estas dos fuentes de información para insertar determinados términos claves en las consultas a un buscador.

El objetivo principal de medir la divergencia entre el título de una página y su propio contenido es tratar de encontrar una divergencia tal que demuestre la

existencia de spam en dicha página. En muchos casos, los buscadores muestran el título de una página como parte del resumen de información que aparece en las listas de resultados. De esta forma, la medida que presentamos en esta sección trata de descubrir la presencia de un conjunto de términos engañosos en el título de una página, que luego no se verán reflejados en el contenido de dicha página.

En la Figura 5.17 se muestra la distribución de los valores de divergencia KL entre el título de la página destino y el propio contenido de dicha página.

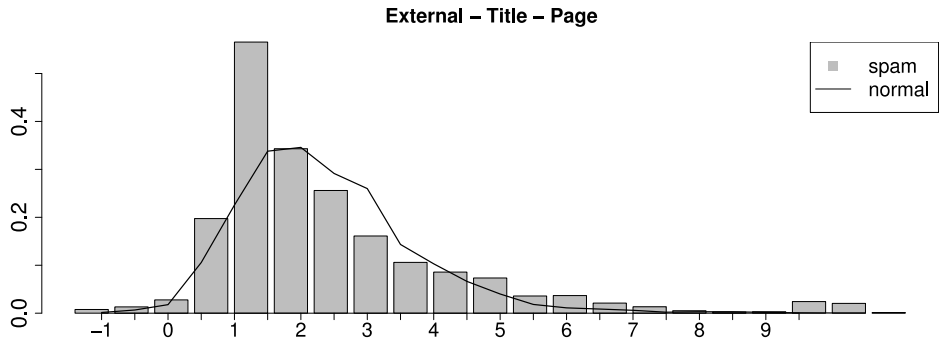

Figura 5.17: *Histograma de la distribución de valores en función de la divergencia entre el título de la página destino y el contenido de dicha página. Se han empleado los enlaces externos de las páginas de la colección de referencia* `WEBSPAM-UK2006`.

Análisis de las *MetaTags*.

Las *MetaTags* son etiquetas html que se incorporan en el encabezado de una página web y que a pesar de que resultan invisibles para un visitante normal, proporcionan una información estructurada que puede ser utilizada por los buscadores. A pesar de que estas etiquetas son el objetivo de spammers desde hace mucho tiempo y que los buscadores cada vez lo tienen menos en cuenta, hay paginas que lo siguen utilizando debido a su clara utilidad. Las MetaTags contienen información estructurada en forma de pares de atributos tales como autor, título, fecha, descripción, etc. En particular vamos a considerar los atributos “description” y “keywords” para construir un documento virtual con estos términos. También se ha decidido utilizar esta información para calcular su divergencia con otras fuentes de información.

En la Figura 5.18 se muestra la distribución de los valores de divergencia KL entre el texto del ancla de la página origen y el documento virtual construido a partir de las MetaTags de la página destino. Como se puede apreciar en esta imagen, los histogramas tienen un rango amplio de valores, en los que la distribución de las páginas de spam y no spam se alternan en tres periodos. A partir de un valor $KLD \approx 8$ se observa que tan solo aparecen páginas de spam, por lo que esa porción de valores de divergencia posee un gran valor discriminante.

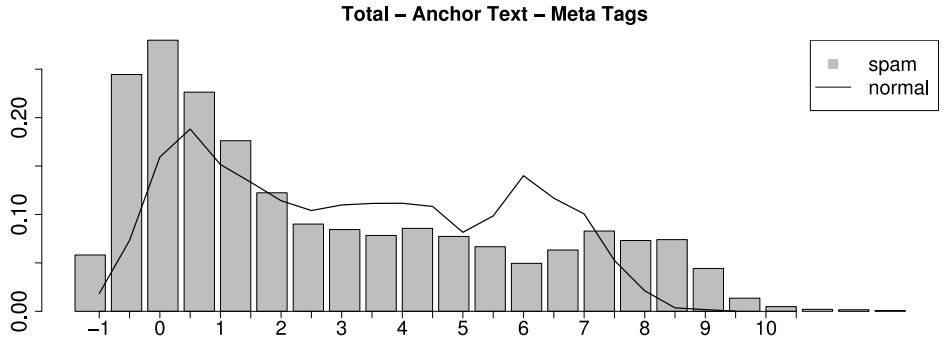

Figura 5.18: *Histograma de la distribución de valores en función de la divergencia entre el texto del ancla y las MetaTags de la página destino. Se han empleado los enlaces externos e internos de las páginas de la colección de referencia* `WEBSPAM-UK2006`.

5.3.2. Combinación de Fuentes de Información

Además de los rasgos presentados en la sección anterior, hemos combinado algunas fuentes de información de la pagina origen con el objetivo de crear documentos virtuales con una mayor información lingüística. Tal y como se presentó en la sección anterior, se han utilizado el texto del ancla (A), el texto circundante del hiperenlace (S) y los términos de la Url (U) como fuentes de información. Además proponemos crear dos nuevas fuentes de información: la combinación del texto del ancla y los términos de la Url (AU), y la combinación del texto circundante del enlace y los términos de la Url (SU). En cuanto a la página destino, el estudio cuenta con tres fuentes de información: el contenido de la página (P), el título de la página (T) y las *MetaTags* (M). La combinación de fuentes de información que se propone en esta sección se ha realizado teniendo en cuenta el coste computacional que supone y la relación lingüística entre dichas unidades de texto. De esta forma, se han descartado ciertas combinaciones que no representaban una aportación significativa. En la Tabla 5.1 se pueden apreciar las 14 comparaciones que se proponen en esta tesis.

Como resultado de esta combinación de fuentes de información se han obtenido modelos de lenguaje mas ricos y descriptivos. En muchos casos podemos encontrar textos de anclas con una pequeña cantidad de texto que nos ocasione resultados engañosos. En cambio, al combinar diferentes fuentes de información como el texto del ancla, el contexto del enlace y los términos de la Url, se obtiene un modelo de lenguaje mucho mas descriptivo. Además, aunque las medidas de divergencia calculada entre fuentes de información individuales, descritas en la sección 5.3.1, ofrecen unos histogramas con una interesante divergencia entre valores de spam y el resto, las mejores medidas propuestas en este trabajo corresponden a aquellas que combinan distintas fuentes de información. Como podemos ver en la Figura 5.19 esta combinación de términos consigue discriminar de manera muy eficiente las paginas de Spam de las que no lo son.

Combinación de diferentes Fuentes de Información
Contenido de la Página Destino (P)
Texto del Ancla $(A \rightarrow P)$
Contexto del Enlace $(S \rightarrow P)$
Términos de la Url $(U \rightarrow P)$
Texto del Ancla $\cup$ Términos de la Url $(AU \rightarrow P)$
Contexto del Enlace $\cup$ Términos de la Url $(SU \rightarrow P)$
Título de la Página Destino vs Contenido de la Página Destino $(T \rightarrow P)$
MetaTags vs Contenido de la Página Destino $(M \rightarrow P)$
Título de la Página Destino (T)
Texto del Ancla $(A \rightarrow T)$
Contexto del Enlace $(S \rightarrow T)$
Términos de la Url $(U \rightarrow T)$
Contexto del Enlace $\cup$ Términos de la Url $(SU \rightarrow T)$
MetaTags de la Página Destino (M)
Texto del Ancla $(A \rightarrow M)$
Contexto del Enlace $(S \rightarrow M)$
Contexto del Enlace $\cup$ Términos de la Url $(SU \rightarrow M)$

Tabla 5.1: *Combinación de diferentes fuentes de información utilizadas para calcular la divergencia de KL. El grupo en la parte superior corresponde a los valores de divergencia calculados entre diferentes fuentes de información de la página origen y el contenido de la página destino (P). El grupo en la parte central corresponde a los valores de divergencia calculados entre diferentes fuentes de información de la página origen y el título de la página destino (T). El último grupo corresponde a los valores de divergencia calculados entre diferentes fuentes de información de la página origen y las MetaTags de la página destino (M).*

Si atendemos a los resultados mostrados en la Figura 5.19, en primer lugar podemos observar las diferentes distribuciones de probabilidad que toman las páginas de spam y las páginas que no son spam. Ambas distribuciones se ajustan a una distribución de Gauss, pero en el caso de los histogramas de las páginas legítimas, éstos son más compactos y sus medias se encuentran cerca de $KL \approx 1{,}2$ y $KL \approx 3{,}5$ respectivamente. Por otro lado los histogramas de las páginas de spam tienen un mayor rango de valores de divergencia, y los valores de su media se encuentran cerca de $KL \approx 4$ y $KL \approx 5$ respectivamente. También se puede observar en el histograma superior (corresponde a la divergencia entre (i) los términos del texto del ancla junto con los del texto de alrededor y junto con los términos de la Url, y (ii) los términos extraidos de las MetaTags asociadas a la página destino) que hay un rango de valores para los que la divergencia obtenida con este rasgo puede discriminar páginas de spam de manera eficaz de las que no lo son (para valores $KL > 2$) con un alto grado de confianza. Los resultados mostrados en la figura inferior (co-

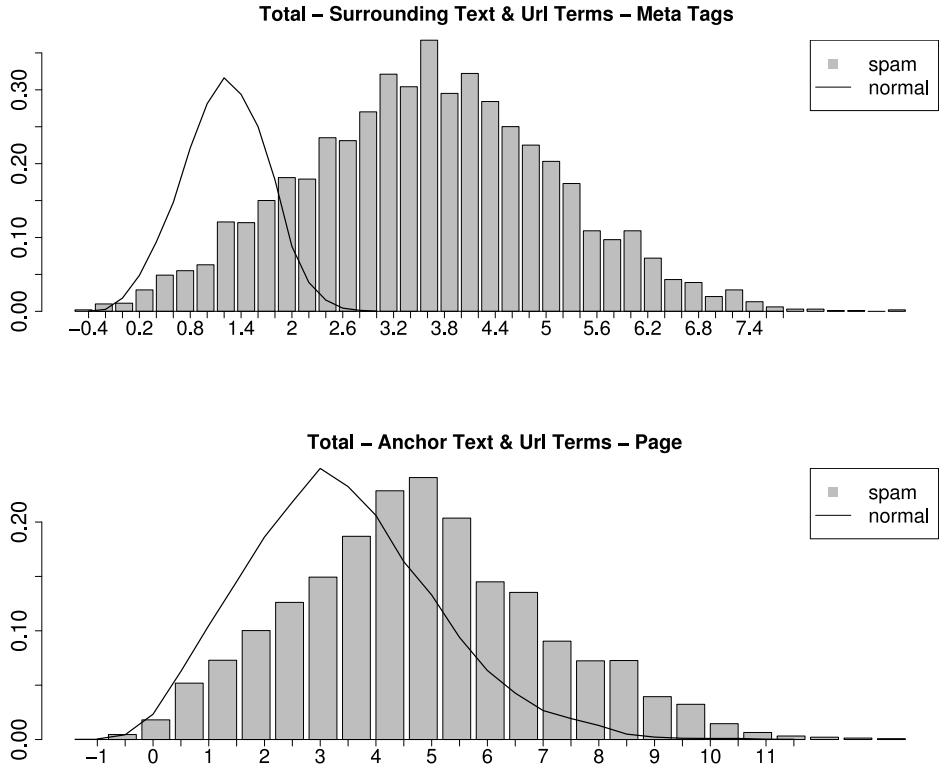

Figura 5.19: *(Arriba) Histograma de la distribución de valores en función de la divergencia entre una combinación del texto del ancla, contexto del enlace y términos de la Url frente al título de la página destino. (Abajo) Histograma de la distribución de valores en función de la divergencia entre una combinación del texto del ancla y términos de la Url frente al contenido de la página destino. La leyenda "Total" se refiere al uso de enlaces externos e internos para el cálculo de los valores de divergencia.*

rresponde a la divergencia entre (i) los términos del texto del ancla junto con los términos de la Url, y (ii) el contenido de la página destino) también muestran un rango de valores (para valores $KL > 6$), aunque es menor en este caso, en el que son un buen discriminante entre las páginas de spam y las que no lo son.

5.3.3. Enlaces Internos y Externos

Los sitios web contienen dos tipos de enlaces: (i) enlaces a páginas del mismo sitio web (enlaces internos) y (ii) enlaces a páginas de sitios web diferentes (enlaces externos). Los enlaces internos, proporcionan profundidad y contexto sobre la información que se presenta en una página web. Sin embargo, los enlaces externos son particularmente útiles para ayudar al lector a recurrir a otras fuentes citadas en el contenido informativo. Por parte de algunos gestores web, existen reticencias a

la hora de insertar enlaces, en muchos casos por una razón de tipo comercial: la de remitir el trafico a otros sitios.

Además, existen técnicas de SEO que consideran la relación entre enlaces internos y externos, es decir, el ratio entre el número de enlaces de cada tipo, como una medida con influencia en la valoración que otorga un buscador a una página. Esto puede hacer pensar que los Spammers puedan estar utilizando algoritmos que tengan en cuenta esta información para promocionar sus paginas.

Por estas razones, en esta tesis se hace una distinción entre enlaces internos y enlaces externos a fin de llevar a cabo el análisis de divergencia. Por lo tanto, para cada página web y por cada rasgo presentado en secciones anteriores, se obtienen valores triples: 14 rasgos para los enlaces internos, 14 rasgos para los enlaces externos y 14 rasgos sin distinción entre el tipo de enlaces. En la Figura 5.3.3 se puede observar la diferencia entre los histogramas de los enlaces internos y externos para una misma combinación de fuentes de información. En dichos histogramas se muestra la distribución de los valores de divergencia entre (i) los términos del texto del ancla junto con los del texto de alrededor, y (ii) el contenido de la página destino. Según los resultados, la curva de spam se desplaza hacia valores más altos de divergencia en el caso de los enlaces internos, mientras que la misma curva se desplaza hacia valores más bajos en el otro caso.

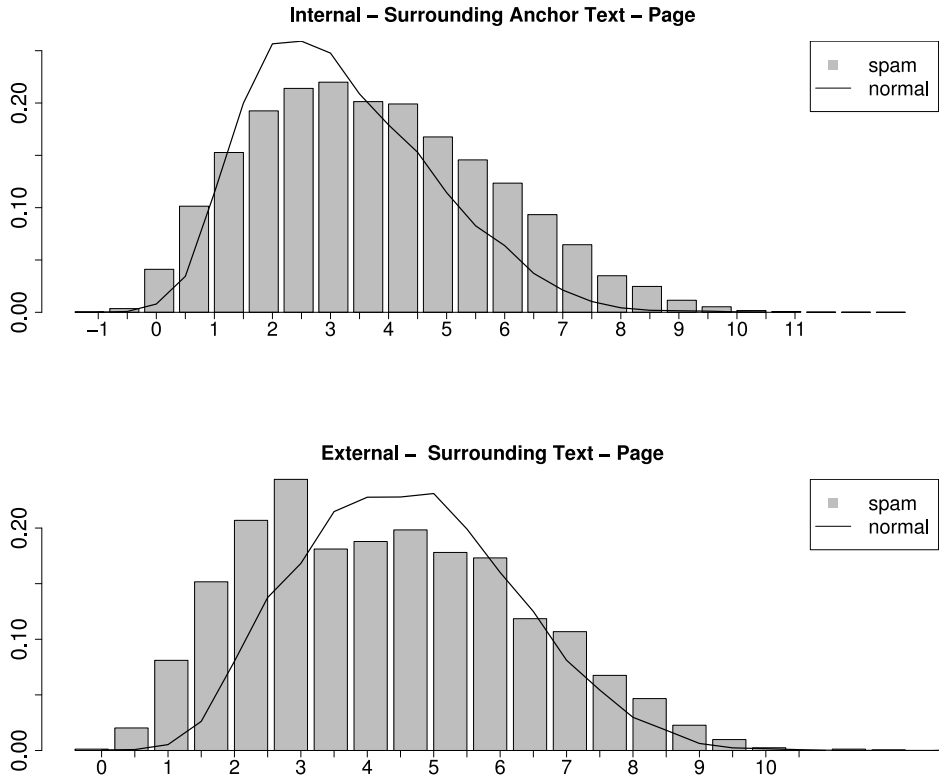

Figura 5.20: *Histogramas de la distribución de valores en función de la divergencia entre el contexto del enlace y el título de la página destino. En la figura de arriba se han empleado solamente enlaces internos, mientras que en la figura de abajo se han empleado solamente enlaces externos. La colección de referencia utilizada ha sido* `WEBSPAM-UK2007`*.*

5.4. Metodología

En esta sección se presenta la metodología seguida para llevar a cabo la tarea de detección de web spam. En primer lugar, se realiza la extracción de los rasgos analizados en las secciones anteriores. Para ello, se utilizan dos colecciones de referencia que describiremos a continuación. Después, es necesaria la selección de un algoritmo de clasificación basado en aprendizaje automático y el ajuste de algunos parámetros para su adaptación al problema que se plantea. Finalmente se presentan los resultados de clasificación obtenidos.

5.4.1. Colecciones de Referencia

En esta tesis se emplean las dos principales colecciones de referencia en el área del web spam[CDB+06] basadas en un *crawling* del dominio *.uk* y realizadas en Mayo del 2006 (`WEBSPAM-UK2006`) y Mayo del 2007 (`WEBSPAM-UK2007`) respectivamente. Aunque estas colecciones fueron presentadas en la sección 3.3.5,

a continuación se describe cómo han sido utilizadas. WEBSPAM-UK2006 incluye 77.9 millones de páginas web, más de 3000 millones de enlaces y alrededor de 11400 dominios. WEBSPAM-UK2007 incluye 105.9 millones de páginas web, más de 3700 millones de enlaces y alrededor de 114529 dominios. Estas colecciones de referencia han sido etiquetadas mediante un proceso manual en el que se solicitaba para cada una de los sitios analizados una clasificación entre tres posibilidades: "spam", "no spam" y "dudoso". En nuestros experimentos, se ha restringido el uso de sitios web a aquellos que contaban con el etiquetado de al menos dos personas y solamente aquellos casos en los que los jueces coincidían en sus apreciaciones. Además, estas colecciones cuentan con una etiqueta adicional extraida del directorio público ODP (Open Directory Project), pero no han sido tenidas en cuenta para este trabajo. La decisión de eliminar este tipo de etiquetas se debe a la presencia de algunos casos que pueden inducir a error debido a su manifiesta ambigüedad. En el siguiente ejemplo, tomado de la colección WEBSPAM-UK2006, se puede apreciar cómo un mismo sitio está marcado como spam por parte de los jueces, y sin embargo la etiqueta ODP indica que la página no es spam:

```
4road.co.uk spam 0.66667 j20:S,j7:S,odp:N
www.4road.co.uk non-spam 0.00000 odp:N
```

Después de la fase de filtrado descrita anteriormente, la colección usada en los experimentos a partir de WEBSPAM-UK2006 cuenta con 3083 dominios, de los cuales 1811 están etiquetados como "no-spam" y 1272 como "spam". Por otra parte, se ha llevado a cabo un análisis de las características de la colección, donde los dominios de "no-spam" tienen una media de enlaces externos e internos de 12.1 y 30.6 respectivamente, y los dominios de "spam" tienen una media de enlaces externos e internos de 7.2 y 15.3. Por su parte, la colección WEBSPAM-UK2007 tiene un tamaño de 4166 dominios, 4012 de ellos etiquetados como "normales" y 154 como "spam". En esta colección los dominios "normales" tienen una media de enlaces externos e internos de 3.7 y 13.4 respectivamente, y los dominios de "spam" tienen una media de enlaces externos e internos de 9.3 y 12.06.

Las colecciones son etiquetadas a nivel de dominio[CDB+06], es decir, todas las páginas dentro de un mismo sitio web tienen la misma etiqueta. Por lo tanto, es necesario agregar todos los rasgos obtenidos de la divergencia de contenido y del análisis de enlaces cualificados a este nivel, ya que la información se encuentra disponible por cada página. Con el objetivo de llevar a cabo la extracción de rasgos sobre la divergencia de contenido y el análisis de enlaces cualificados, y teniendo en cuenta el coste computacional que supone el análisis del sitio web completo, se han analizado exclusivamente una página por dominio. En concreto, se han utilizado la página principal del sitio de origen y cada página apuntada por un enlace en dicha página de origen. Por otra parte, los dominios que no tienen enlaces salientes son descartados, por lo que el tamaño final del conjunto de datos se reduce ligeramente.

El análisis de la divergencia de contenido requiere el estudio de todos los enlaces para cada dominio etiquetado. De esta forma, tan solo se han analizado los enlaces que tienen términos en el texto del ancla. Por lo tanto, han sido filtradas las imágenes, enlaces a la misma página (marcadores HTML), números, Urls y cadenas vacías. También se han descartado enlaces cuyo protocolo no sea `HTTP` y los documentos que no sean `HTML`. Después de todo este proceso, se obtienen 42 medidas de divergencia para representar cada enlace de una página web, y de cara a agregar esta información a nivel de sitio web, se calcula la media de los valores de todos los enlaces. Por tanto, se puede decir que cada sitio web está representado por 42 medidas que corresponden a un valor medio.

En cuanto al análisis de rendimiento del sistema, el tiempo medio para procesar un sitio web es alrededor de 4 segundos. Aunque este valor tiene una gran variabilidad (una desviación estándar de 2 segundos) ya que depende directamente del número de enlaces de cada página y el número de páginas analizadas por cada enlace.

5.4.2. Algoritmos de Clasificación

En esta sección se describe la tarea de clasificación, así como algunas medidas de evaluación empleadas para estimar el rendimiento del sistema.

Para realizar las tareas de clasificación, hemos utilizado el software de Weka-[WF05], debido a su amplia colección de algoritmos de aprendizaje automático.

El primer paso para obtener los mejores resultados en la tarea de clasificación es seleccionar el clasificador más adecuado. Se han seleccionado diferentes algoritmos de clasificación para evaluar los rasgos introducidos. En particular, hemos elegido los siguientes algoritmos de clasificación:

- **Metacost**. Proporciona un recubrimiento para otro clasificador teniendo en cuenta el coste. El clasificador base utilizado es un árbol de decisión con "bagging" (Bootstrap Aggregating) C4.5[WF05]. Este árbol de decisión puede manejar atributos numéricos, y por lo tanto no es necesario discretizar cualquiera de los atributos. Además produce modelos que son fáciles de interpretar y que son muy robustos al efecto de los valores extremos.

- **Naive Bayes**. Es un clasificador estadístico basado en el teorema de Bayes que utiliza las probabilidades conjuntas de observaciones de muestra para estimar las probabilidades condicionales de las clases dada una observación.

- **Regresión Logística**. Es un modelo lineal generalizado para aplicar la regresión a un conjunto de variables categóricas.

- **Máquinas de Soporte Vectorial (SVM)**. Tiene por objeto la búsqueda de un hiperplano que separa dos clases de datos con el mayor margen posible.

Se ha realizado una evaluación exhaustiva de estos clasificadores que forman parte del conjunto de herramientas de Weka. En cuanto al árbol de decisión, se ha empleado la implementación específica de Weka, denominada *J48*. Para los algoritmos de Naive Bayes y Regresión Logística se ha utilizado aquellos determinados por la herramienta. Finalmente en el caso de *SVM*, se ha empleado la implementación *Sequential Minimal Optimization* (SMO) de un kernel polinómico[WF05].

En relación al ajuste de parámetros, se han utilizado las opciones por defecto de estos algoritmos, excepto en el caso del árbol de decisión para el que se fijó un método reducido de poda en caso de error. Los detalles algorítmicos de estos clasificadores están fuera del alcance de esta tesis, aunque se ha dado una paqueña descripción en la sección 3.3.3. Además, en la mayoría de textos de aprendizaje automático[WF05] se encuentra disponible información adicional acerca de dichos clasificadores. Los resultados mostrados en esta tesis podrían ser mejorados mediante la optimización del proceso de aprendizaje y clasificación, por tanto se puede considerar que mediante nuestro enfoque se obtiene una cota inferior de los resultados.

La evaluación de los algoritmos de aprendizaje utilizados en todas las predicciones de este trabajo, han sido realizadas mediante una validación cruzada de 10 subconjuntos. Para cada evaluación, el conjunto de datos se divide en 10 subconjuntos iguales elegidos al azar y se realiza la clasificación 10 veces. En cada iteración, el clasificador entrena con 9 porciones y emplea la décima parte para clasificar.

Por otra parte, se ha adoptado un conjunto de medidas de evaluación ampliamente usadas en la investigación de web spam: tasa de verdaderos positivos (TP o cobertura), tasa de falsos positivos (FP) y la *Medida-F*. Esta última medida, mostrada en la Ecuación 5.2, combina precisión (P) y cobertura (R). De esta forma, para la evaluación de los algoritmos de clasificación, nos centraremos en la *Medida-F* ya que es una medida estándar para resumir tanto la precisión (P) como la cobertura (R).

$$Medida - F = 2\frac{PR}{P+R} \tag{5.2}$$

La Tabla 5.2 muestra la *Medida-F* para cada uno de los algoritmos propuestos y los indicadores de spam que fueron presentados en las secciones anteriores. El mejor clasificador en la mayoría de los conjuntos de rasgos es el árbol de decisión, seguido por el clasificador *SVM*. A pesar de que el árbol de decisión proporciona los mejores resultados, más de la mitad de los spammers están clasificados como no spammers. Por este motivo, se ha empleado el algoritmo de recubrimiento *Metacost*, que por un lado mantiene las cualidades del árbol de decisión, y por otro lado introduce penalizaciones en el caso de clasificaciones erróneas.

	Medida-F			
	UK-2006		UK-2007	
Algoritmo	**Divergencia**	**Calidad Enlaces**	**Divergencia**	**Calidad Enlaces**
C4.5	**0.55**	**0.67**	**0.24**	**0.32**
NB	**0.55**	0.57	0.20	0.18
SVM	0.54	0.65	0.22	**0.32**
LR	0.53	0.58	0.18	0.26

Tabla 5.2: *Medida-F obtenida por los algoritmos: árbol de decisión, Naive Bayes, SVM y regresión logística, basados en los rasgos obtenidos a partir de la divergencia de contenido y los enlaces cualificados según la colección de referencia utilizada.*

5.4.3. Penalización en el Algoritmo de Aprendizaje

Si atendemos a los resultados de clasificación de un algoritmo de aprendizaje, los errores producidos como consecuencia de clasificar erróneamente páginas legítimas como spam, no tienen el mismo impacto que clasificar páginas de spam como legítimas. El motivo principal de esta diferencia reside en que las colecciones no están bien balanceadas y contienen un número mucho mayor de páginas legítimas que de spam. Por este motivo, si no se penaliza la clasificación errónea de las páginas de spam, podría darse el caso paradójico de que un sistema que no detecta ninguna página de spam pudiera obtener mejores resultados que cualquier otro.

Por este motivo se ha empleado el algoritmo *Metacost* de Weka ya que proporciona un encubrimiento para un árbol de decisión, implementando además una matriz de costes. Este algoritmo permite establecer diferentes costes de penalización por clasificación errónea. De esta forma, antes de aprender un modelo en la fase de entrenamiento, los datos son re-pesados para incrementar la sensibilidad de los casos de spam.

En relación al ajuste de la matriz de costes del algoritmo, se ha seguido una metodología similar a la descrita en el artículo de Castillo et al.[CDB+06]. Se ha impuesto un coste cero a las predicciones correctas, mientras que las páginas de spam clasificadas erróneamente penalizan al sistema mediante un parámetro R, que implica un coste R veces mayor que si una página legítima es clasificada como spam. Además, como el objetivo de este trabajo es maximizar la *Medida-F*, se ha realizado un análisis para optimizar este parámetro R. En la Figura 5.21 se ilustra la evolución de la *Medida-F* obtenida mediante el ajuste de los diferentes costes de R. De acuerdo con estos resultados, se ha fijado un valor $R = 4$ en el caso de la colección `WEBSPAM-UK2006` y $R = 14$ en el caso de la colección `WEBSPAM-UK2007`.

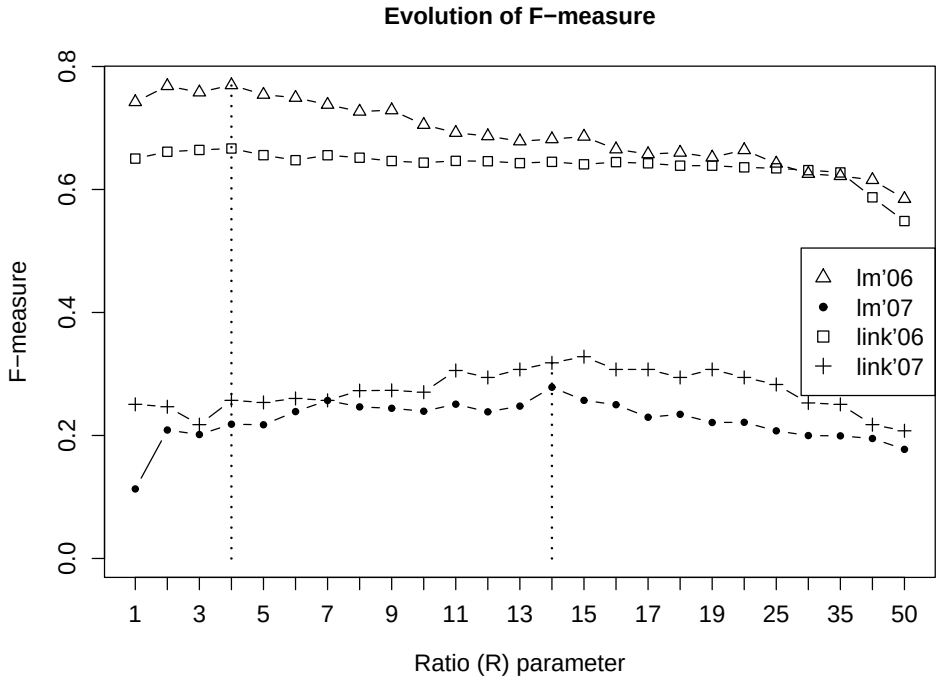

Figura 5.21: *Evolución de la Medida-F obtenida mediante la aplicación de diferentes costes de penalización (R) en la matriz de confusión. Han sido utilizadas las características basadas en divergencia de contenido (lm) y enlaces cualificados (link) sobre las colecciones de referencia* `WEBSPAM-UK2006` *(06) y* `WEBSPAM-UK2007` *(07).*

5.5. Resultados

Con el fin de comprobar si los rasgos propuestos mejoran la precisión del sistema de detección de web spam, se han utilizado un conjunto de características disponibles junto de las colecciones de referencia. Estas características son un conjunto de indicadores acerca de cada sitio web divididas en dos conjuntos: un conjunto de características acerca del contenido y detalles lingüísticos, y otro conjunto de características basadas en información de los enlaces en cuanto a su estructura. Estas características son el resultado de una compilación de rasgos publicados en diferentes artículos[BCD+06, NNMF06], que demostraban su utilidad para detectar spam. De esta forma, estos conjuntos de características han sido empleados como el baseline para la comparación de los resultados obtenidos.

También se han combinado diferentes conjuntos de características con el objetivo de desarrollar un clasificador capaz de detectar tanto páginas de spam que utilizan técnicas de contenido como aquellas que utilizan técnicas de spam de enlaces. Por último, se han combinado todas las características: de contenido, enlaces, diver-

gencia y enlaces cualificados, logrando un clasificador más preciso. Como *baseline* para los experimentos realizados, se han combinado además las características precomputadas para detectar diferentes tipos de páginas spam.

Los resultados de los experimentos de clasificación presentados en esta tesis sobre las colecciones WEBSPAM-UK2006 y WEBSPAM-UK2007, se muestran en las Tabla 5.3 y 5.4 respectivamente.

En primer lugar puede apreciarse que en el caso de usar tan solo las características precomputadas a partir de la colección analizada, los mejores resultados se obtienen mediante la combinación de las características de contenido y de enlaces (C ∪ L). Por esta razón, se ha establecido esta combinación de características como la base para los experimentos que se presentan en esta sección.

Resultados de la colección WEBSPAM-UK2006.

La Tabla 5.3 muestra los resultados sobre la colección WEBSPAM-UK2006. La primera observación es que en dicha tabla se observa que los rasgos de enlaces cualificados obtienen una *Medida-F* mayor (0.67) que el los de contenido (0.63), enlaces (0.66) o divergencia de contenido (0.55). Este resultado es digno de reseñar ya que el número de características utilizadas por este enfoque es mucho más pequeña (12) que en el caso del contenido (98) o de los enlaces (139). Aun así, las características de los enlaces cualificados no son tan eficientes por sí mismos como la combinación de contenido y enlaces (0.75). Cuando se combinan el baseline con las características de divergencia de contenido y enlaces cualificados, se obtienen mejoras significativas. En concreto, cuando el clasificador emplea la combinación de baseline y contenido (C ∪ L ∪ DC) se obtiene una mejora del 6 % en la *Medida-F*. Por otra parte, la combinación del baseline y los enlaces cualificados (C ∪ L ∪ QL) mejora un 8 %, pasando de 0.75 a 0.83. La observación más importante acerca de estos resultados, es que si se tiene en cuenta el baseline, el clasificador mejora un 11 % la *Medida-F*, pasando de 0.75 a 0.86. Estos resultados se producen al combinar el baseline, la divergencia de contenido y los enlaces cualificados (C ∪ L ∪ DC ∪ QL).

Resultados de la colección WEBSPAM-UK2007.

La Tabla 5.4 muestra los resultados sobre la colección WEBSPAM-UK2007. En el caso de esta colección, la tasa de detección es menor que para el caso de la colección anterior. La causa principal de esta reducción en la tasa de aciertos es que la colección está mucho peor equilibrada que la anterior, tendiendo 4012 dominios etiquetados como no-spam mientras que en el caso de los dominios de spam tan solo cuenta con 154. Si lo comparamos con la colección anterior, esta disponía de 1811 dominios de no-spam, frente a 1272 dominios de spam. De esta forma, tanto si atendemos al número de dominios de spam etiquetados de manera individual, como si comparamos la proporción entre los tipos de etiquetas, la colección

WEBSPAM-UK2006					
Conjunto de Rasgos	**Número de Rasgos**	***TP***	***FP***	***F***	***AUC***
Contenido (C)	98	0.61	0.08	0.63	0.82
Enlaces (L)	139	0.67	0.09	0.66	0.83
Divergencia de Contenido (DC)	42	0.43	**0.05**	0.55	0.76
Enlaces Cualificados (QL)	12	0.87	0.27	0.67	0.83
C ∪ L (baseline)	237	0.84	0.14	0.75	0.85
C ∪ L ∪ DC	279	0.87	0.11	0.81	0.86
C ∪ L ∪ QL	249	**0.92**	0.14	0.83	0.86
C ∪ L ∪ DC ∪ QL	291	0.89	0.10	**0.86**	**0.88**

Tabla 5.3: *Número de características, tasa de verdaderos positivos (TP), tasa de falsos positivos (FP), Medida-F (F) y área bajo la curva ROC (AUC) en función de las fuentes de información empleadas. Los mejores resultados son mostrados en negrita.*

`WEBSPAM-UK2007` tiene un número de dominios de spam significativamente pequeño.

A pesar de este problema, los experimentos muestran resultados consistentes en comparación con las mejoras obtenidas en el conjunto de datos anterior. En cualquier caso, la Tabla 5.4 muestra que las características basadas en enlaces cualificados obtienen una *Medida-F* superior (0.32) a las de contenido (0.30), enlaces (0.20) y divergencia de contenido (0.24). Al igual que en los resultados obtenidos con la colección anterior, cuando se combina el baseline con las características de divergencia de contenido y de enlaces cualificados, se obtienen mejoras significativas. En concreto, cuando el clasificador utiliza la combinación del baseline con la divergencia de contenido (C ∪ L ∪ DC) se consigue una mejora del 2 % en la *Medida-F*. Por otra parte, con la combinación del baseline y enlaces cualificados (C ∪ L ∪ QL) el sistema obtiene una mejora del 7 %, pasando de 0.31 a 0.38. Los mejores resultados que el sistema es capaz de lograr, consisten en una mejora del 9 % en la *Medida-F* con respecto al baseline, pasando de 0.31 a 0.40. Esta mejora es obtenida mediante la combinación del baseline, divergencia de contenido y enlaces cualificados (C ∪ L ∪ DC ∪ QL).

Teniendo en cuenta los datos ofrecidos sobre las dos colecciones de web spam, se puede concluir que a partir de los valores mostrados en la Tabla 5.3 y en la Tabla 5.4, las mejoras más significativas se obtienen mediante la combinación de características de divergencia de contenido y enlaces cualificados. La combinación de los cuatro conjuntos de características producen los mejores resultados, ya que cada conjunto de rasgos se centra en un tipo diferente de spam y tienen características complementarias. Por lo tanto, esta combinación logra detectar spam de contenido, spam de enlaces, enlaces nepotísticos y enlaces cualificados. Por otra parte, si tenemos en cuenta cada conjunto por separado, cada uno uno de ellos tiene

Colección WEBSPAM-UK2007					
Conjunto de Rasgos	**Número de Rasgos**	***TP***	***FP***	***F***	***AUC***
Contenido (C)	98	0.33	0.04	0.30	0.72
Enlaces (L)	139	0.39	0.12	0.20	0.68
Divergencia de Contenido (DC)	42	0.24	**0.03**	0.24	0.72
Enlaces Cualificados (QL)	12	0.40	0.06	0.32	0.72
C ∪ L (baseline)	237	0.31	0.05	0.31	0.73
C ∪ L ∪ DC	279	0.33	**0.03**	0.33	0.75
C ∪ L ∪ QL	249	0.48	0.06	0.38	0.75
C ∪ L ∪ DC ∪ QL	291	**0.50**	0.06	**0.40**	**0.76**

Tabla 5.4: *Número de características, tasa de verdaderos positivos (TP), tasa de falsos positivos (FP), Medida-F (F) y área bajo la curva ROC (AUC) en función de las fuentes de información empleadas. Los mejores resultados son mostrados en negrita.*

un impacto diferente en los parámetros que intervienen en la *Medida-F*. Mientras QL consigue la mejor precisión, también obtiene la peor cobertura. En el caso de DC obtiene los peores valores de precisión, sin embargo su cobertura es la mejor. Por último, la combinación de los cuatro grupos de rasgos obtiene una precisión muy alta, sin afectar a la cobertura.

5.6. Conclusiones

En este capítulo, se propone una nueva metodología para detectar web spam, basada en un análisis de enlaces cualificados y un estudio de la divergencia entre el contenido de las páginas enlazadas.

En primer lugar, se introduce un conjunto novedoso de rasgos que se centran en la capacidad de los motores de búsqueda para recuperar los enlaces de una página. Este análisis proporciona un factor de calidad para cada pagina, que resulta representada por un conjunto de 12 rasgos de gran efectividad. A pesar de la pequeña cantidad de características extradas por el sistema, se obtiene una mejora notable, especialmente cuando se combina con características basadas en el contenido y en la estructura de los enlaces.

En segundo lugar, se presenta un método que aprovecha la potencia de los modelos estadísticos y el procesamiento del lenguaje natural. En concreto se emplean modelos de lenguaje para representar un documento web y calcular la divergencia entre diferentes fuentes de información tanto de la página de origen como de la página de destino. Otra de las aportaciones de este trabajo es la combinación de

diferentes fuentes de información con el objetivo de reunir en una pequeña unidad textual la síntesis de información que representa a una página.

En este capítulo, también se propone el uso de un clasificador sensible al coste de los errores. Para la evaluación de este clasificador se han utilizado las dos colecciones de referencia dentro del área del web spam.

En cuanto a los resultados obtenidos, mediante la combinación de las características basadas en contenido, enlaces, enlaces cualificados y divergencia de contenido, se consigue detectar un 89.4 % en el caso de la colección `WEBSPAM-UK2006` y un 54.2 % para la colección `WEBSPAM-UK2007`. Para las mismas colecciones, se obtiene una *Medida-F* de 0.86 y 0.40, lo que significa una mejora del 11 % y 9 % respectivamente.

CAPÍTULO 6

Conclusiones, Perspectivas de Futuro y Contribuciones

6.1. Conclusiones

En esta tesis, se han abordado dos de los problemas más importantes que afectan a la Web en la actualidad: los enlaces rotos y el web spam. Para la resolución de dichos problemas se ha empleado un sistema de recuperación web que aporta información útil tanto para la recuperación automática de un enlace roto como para la detección de web spam.

En el caso de la desaparición de una página web, el sistema de recuperación adopta la forma de un sistema de recomendación. Este sistema de recomendación, toma como entrada una página web, analiza sus enlaces, y recomienda una lista de páginas candidatas para reemplazar cada enlace roto.

Este Sistema de Recuperación de Enlaces Rotos (SRER) está gestionado por un algoritmo que analiza toda la información disponible acerca del enlace analizado y decide en cada caso los pasos a seguir, ya sea descartando su recuperación, o mostrando al usuario un conjunto de resultados ordenados por relevancia.

El algoritmo de SRER se compone de cuatro etapas para llevar a cabo la recuperación de un enlace. La primera etapa consiste en un proceso de selección de información, en el cual se ha obtenido, en primer lugar, el texto del ancla del hiperenlace que ha dejado de funcionar

En segundo lugar se han obtenido las fuentes de información en el entorno del enlace, como la propia página que contiene el enlace roto, el texto que rodea al hiperenlace y la dirección de destino del hiperenlace. También han sido empleados ciertos recursos de la infraestructura web para obtener más información acerca de las páginas que han desaparecido, pero de las que aún existe información en la red. Tras el análisis de estas fuentes de información, se puede concluir que tanto la utilidad de cada una como la disponibilidad varían en cada caso. El texto del ancla, la dirección de destino del hiperenlace y la página donde se encuentra el enlace

roto son fuentes de información con las que el sistema puede contar en la mayoría de los casos. Sin embargo la utilidad de los términos del ancla son superiores al resto de fuentes. Por otro lado, la versión en caché de la página desaparecida tan solo puede encontrarse entre un 50 % y un 60 % de los casos, sin embargo cuando el sistema puede contar con dicha fuente de información, los resultados mejoran notablemente.

En la segunda etapa, se realiza un proceso de extracción de información teniendo en cuenta las fuentes de información disponibles para cada enlace. Atendiendo a los resultados obtenidos, se puede concluir que el texto del ancla es la principal fuente de información, resultando especialmente útil en el caso de estar formado por más de un término y aún más si contiene alguna entidad nombrada. A partir de algunas de estas fuentes de información, se ha llevado a cabo un proceso de extracción de terminología a fin de sintetizar la mayor información posible en el menor número de términos. Tras el análisis de diferentes técnicas de extracción de términos relevantes, se puede concluir que el uso de modelos de lenguaje junto con una colección de referencia próxima a la naturaleza de cada fuente de información, proporciona los mejores resultados de selección de términos.

En la tercera etapa del sistema y contando con las fuentes de información recuperadas en la etapa de selección, el algoritmo puede tomar la decisión de informar al usuario de la imposibilidad de recuperar en enlace roto, debido a la ausencia de información, o bien continuar un proceso de expansión de consultas. De acuerdo a los estudios realizados en esta tesis, los resultados obtenidos mediante la expansión de consultas son superiores a los casos en los que el texto del ancla fue utilizado de manera aislada. De esta forma, la expansión de consultas reduce la ambigüedad que supone la cantidad limitada de los términos del ancla. El resultado de esta etapa es la recuperación de los primeros diez resultados devueltos por el buscador por cada consulta realizada.

En la etapa final, el algoritmo dispone por un lado de la información relativa a la página desaparecida, y por otro lado del conjunto de resultados obtenidos del buscador. En esta cuarta etapa el algoritmo emplea diferentes métodos de ranking en función de las fuentes de información disponibles. De esta forma, SRER combina diferentes métodos de ranking, a fin de presentar al usuario una lista ordenada con las páginas candidatas a reemplazar el enlace roto. También ha sido llevado a cabo un estudio comparativo entre los diferentes métodos de ranking mediante el uso de varias fuentes de información procedentes de las páginas de origen y destino. El sistema emplea diversos coeficientes de coocurrencia y un enfoque basado en la divergencia de modelos de lenguaje para llevar a cabo el ranking de las páginas candidatas, en cada caso.

Como parte de esta tesis, también se ha desarrollado una metodología de evaluación novedosa, que permite aumentar la objetividad de los resultados sin recurrir a los juicios de usuario. Para esta evaluación, se ha construido una colección de páginas web con la particularidad de que todas ellas contienen enlaces rotos. A través de la metodología de evaluación propuesta, ha sido posible determinar la cantidad

óptima de términos utilizados para la expansión de una consulta y los resultados que deben ser recuperados de los motores de búsqueda. Además, esta metodología ha permitido realizar una evaluación empírica de la disponibilidad y eficacia de las fuentes de información.

El resultado de este análisis ha dado lugar al diseño de un algoritmo, que ha sido capaz de recuperar una página que podría reemplazar a otra desaparecida en el 78 % de los casos. Además, el sistema es capaz de proporcionar el 47 % de estos enlaces recuperados entre los 10 primeros documentos recomendados, y entre los 20 primeros en un 71 % de los casos.

Al inicio de las conclusiones se indicaba el uso común de un sistema de recuperación web tanto para el problema de los enlaces rotos como para la detección de web spam. En cuanto al segundo problema, el sistema de recuperación ha sido utilizado para extraer un conjunto de indicadores que pueden ser útiles en esta tarea de detección.

La recuperación de un enlace está basada en la hipótesis de que la información relativa a una página apuntada por parte de la página que apunta es coherente. No obstante, existen casos es los que los autores de páginas web manipulan la información relativa a una determinada página con el objetivo de obtener algún beneficio, en concreto promocionarla en un buscador. De esta forma, el sistema de recuperación web adopta la forma de un analizador de la calidad de los enlaces que se encuentran en una página.

Gracias a este sistema, se ha propuesto un conjunto novedoso de rasgos basados en la capacidad de los motores de búsqueda para recuperar los enlaces de una página. Este análisis proporciona un factor de calidad para cada pagina, que resulta representada por un conjunto de 12 rasgos de gran efectividad.

En una segunda fase de la detección de web spam, se ha decidido completar esta tesis con un análisis de la divergencia entre el contenido de dos páginas enlazadas. El objetivo de esta segunda fase, ha sido complementar los tipos de web spam que el sistema de recuperación web es capaz de detectar. De esta forma, se ha propuesto la utilización de un nuevo conjunto de características, que enfocan en el contenido de las páginas enlazadas, con el objetivo de encontrar signos de spam en la divergencia entre diferentes fuentes de información de ambas páginas.

La principal aportación en esta segunda parte de la tesis, dedicada a la detección de web spam, ha sido la propuesta de utilización de dos nuevos conjuntos de indicadores. Entre los resultados obtenidos destaca el buen rendimiento de los dos conjuntos de indicadores por separado, logrando el conjunto de rasgos acerca de la calidad de los enlaces, una *Medida-F* superior a los otros dos conjuntos utilizados como baseline, basados en contenido y enlaces. Además, la combinación de ambas características da lugar a un sistema ortogonal que mejora los resultados de detección de ambos conjuntos por separado. De esta forma, al combinar los cuatro conjuntos de características sobre las colecciones de referencia `WEBSPAM-UK2006` y `WEBSPAM-UK2007`, se consigue una detección del 89.4 % y 54.2 % de los dominios de spam, obteniendo el sistema una *Medida-F* de 0.86 y 0.40, respectivamente. Es-

tos datos implican una mejora en la Medida-F del 11 % en el caso de la colección del 2006 y un 9 % al emplear la colección de spam del 2007.

6.2. Futuros Trabajos

En esta sección se describen las posibles mejoras que podrían hacerse para cada uno de los trabajos de investigación descritos en esta tesis.

El sistema de recomendación de enlaces rotos, como ya se ha mencionado en capítulos anteriores, consta de cuatro etapas. En cuanto a la primera etapa de recuperación de fuentes de información, se podrían incluir nuevos recursos de la infraestructura web, no contemplados en esta tesis. La etapa de extracción de terminología podría ser extendida, incluyendo algunas técnicas como "Latent Semantic Analysis" (LSA) o "Hyperspace Analogue to Language" (HAL). De cara a la expansión de consultas, se podrían establecer otras estrategias, incluyendo pesos para cada uno de los términos. Recientemente los motores de búsqueda han presentado modificaciones en las opciones de búsqueda, ofreciendo al usuario la posibilidad de realizar consultas más específicas tanto por el tipo de sitios donde buscar (noticias, blogs, etc.) como por el periodo de tiempo en el que se deberían buscar los documentos relevantes. Estas nuevas opciones podrían ser estudiadas e integradas en caso de aportar nueva información al sistema. En el caso de la etapa de ordenamiento de las páginas candidatas, se podría profundizar en el estudio de funciones de ranking típicas como BM25, BM25F, etc. Aunque algunos estudios preliminares, no ofrecieron resultados positivos. La evaluación de sistemas de recuperación de enlaces rotos es un campo prácticamente inexistente en la actualidad, por este motivo, el estudio de nuevas medidas de evaluación que se ajusten a las características de la tarea, podrían tener una gran aceptación en la comunidad científica.

En cuanto a la detección de web spam, una de las partes fundamentales del sistema, es la clasificación automática. En esta tesis, no se ha realizado un trabajo en profundidad en cuanto a las labores de optimización del algoritmo de clasificación, ya que el objetivo principal era la proposición de nuevas características que pudieran mejorar la detección de spam. De esta forma, es de suponer que podrían obtenerse aún mejores resultados a partir de las características propuestas mediante una optimización del algoritmo de clasificación empleado. Si atendemos a las características basadas en la divergencia de contenido, podría resultar interesante analizar la relación entre la página analizada y aquellas páginas que la apuntan, es decir analizando los enlaces entrantes. También se podrían añadir nuevas fuentes de información al estudio de divergencia, con el objetivo de completar la caracterización de la relación entre dos páginas. Por otro lado, el sistema que se presenta no es una aplicación en tiempo real, aunque en el futuro sería interesante reducir el tiempo de ejecución sin influir negativamente en el rendimiento. Este coste computacional podría verse afectado en el caso de reducir tanto el número de enlaces analizados por cada página como el número de páginas recuperadas por cada enlace. También

podría resultar interesante extender la investigación realizada para la detección de web spam a otras áreas tales como el spam en e-mail.

6.3. Contribuciones

El trabajo realizado durante la realización de esta tesis ha dado lugar a distintas publicaciones que se citan a continuación:

- Lourdes Araujo y Juan Martinez-Romo "Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models" IEEE Transactions on Information Forensics & Security, ISSN 1556-6013. (Aceptado, pendiente de publicación).
- Juan Martinez-Romo y Lourdes Araujo "Analyzing Information Retrieval Methods to Recover Broken Web Links" in Proceedigs of the 32nd European Conference on Information Retrieval, ECIR 2010. Milton Keynes, UK. March 28th - 31st March, 2010.
- Juan Martinez-Romo y Lourdes Araujo "Retrieving Broken Web Links using an Approach based on Contextual Information" in Proceedings of the 20th ACM conference on Hypertext, HT'09. Torino, Italy. June 29th - July 1th, 2009.
- Juan Martinez-Romo y Lourdes Araujo "Web Spam Identification Through Language Model Analysis" in Proceedings of the Fifth International Workshop on Adversarial Information Retrieval on the Web, AIRWeb'09 (Co-located with the WWW2009 conference). Madrid, Spain. April 20-24, 2009.
- Juan Martinez-Romo y Lourdes Araujo "Web People Search Disambiguation using Language Model Techniques" in Proceedings of the Second Web People Search Evaluation Workshop, WePS'09 (Co-located with the WWW2009 conference). Madrid, Spain. April 20-24, 2009.
- Lourdes Araujo y Juan Martinez-Romo "Detección de Web Spam basada en la Recuperación Automática de Enlaces" Revista de Procesamiento del lenguaje natural. vol. 42 (marzo 2009). ISSN 1135-5948, pp. 39-46
- Enrique Amigó, Juan Martinez-Romo, Lourdes Araujo y Victor Peinado "UNED at WebCLEF 2008: Applying High Restrictive Summarization, Low Restrictive Information Retrieval y Multilingual Techniques" Lecture Notes in Computer Science. 2008.
- Juan Martinez-Romo y Lourdes Araujo "Recommendation System for Automatic Recovery of Broken Web Links" in Proceedings of the 11th Ibero-American Conference on Artificial Intelligence, IBERAMIA 2008. Lisbon, Portugal, October 14 -17, 2008.

- Juan Martinez-Romo y Lourdes Araujo “Sistema de Recomendación para la Recuperación Automática de Enlaces Web Rotos” Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN’08. Leganés (Madrid), Spain, 10-12 Sept. 2008.

BIBLIOGRAFÍA

[ACC08] Jacob Abernethy, Olivier Chapelle, and Carlos Castillo. Webspam identification through content and hyperlinks. In *Proceedings of the fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[ACR01] Gianni Amati, Claudio Carpineto, and Giovanni Romano. Fub at trec-10 web track: A probabilistic framework for topic relevance term weighting. In *In Proceedings of the 10th Text Retrieval Conference TREC 2001*, pages 182–191, 2001.

[ADWC98] Helen Ashman, Hugh Davis, Jim Whitehead, and Steve Caughey. Missing the 404: link integrity on the world wide web. *Comput. Netw. ISDN Syst.*, 30(1-7):761–762, 1998.

[AGS07] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64–69, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[AMM08] Eyhab Al-Masri and Qusay H. Mahmoud. Investigating web services on the world wide web. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 795–804, New York, NY, USA, 2008. ACM.

[Ash00] Helen Ashman. Electronic document addressing: dealing with change. *ACM Comput. Surv.*, 32(3):201–212, 2000.

[BBCU06] András A. Benczúr, István Bíró, Károly Csalogány, and Máté Uher. Detecting nepotistic links by language model disagreement. In *WWW*

'06: Proceedings of the 15th international conference on World Wide Web, pages 939–940, New York, NY, USA, 2006. ACM.

[BCD+06] Luca Becchetti, Carlos Castillo, Debora Donato, Stefano Leonardi, and Ricardo Baeza-Yates. Link-based characterization and detection of web spam. In *AIRWeb'06: Proceedings of the 2th international workshop on Adversarial information retrieval on the web*, 2006.

[BCSU05] András A. Benczúr, Károly Csalogány, Tamás Sarlós, and Máté Uher. Spamrank - fully automatic link spam detection. In *In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb*, 2005.

[BDGM95] Sergey Brin, James Davis, and Héctor García-Molina. Copy detection mechanisms for digital documents. In *SIGMOD '95: Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 398–409, New York, NY, USA, 1995. ACM.

[BH98] Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 104–111, New York, NY, USA, 1998. ACM.

[BL98] Tim Berners-Lee. Cool uris dont't change. `http://www.w3.org/Provider/Style/URI`, 1998.

[BL99] Adam Berger and John Lafferty. Information retrieval as statistical translation. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 222–229, New York, NY, USA, 1999. ACM.

[BLFM98] T. Berners-Lee, R. Fielding, and L. Masinter. Uniform resource identifiers (uri): Generic syntax, 1998.

[BLMM94] T. Berners-Lee, L. Masinter, and M. McCahill. Uniform resource locators (url), 1994.

[BXW+07] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, New York, NY, USA, 2007. ACM.

[BYBH07] Ricardo Baeza-Yates, Paolo Boldi, and José María Gómez Hidalgo. Recuperación de información con adversario en la web. *Novática: Revista de la Asociación de Técnicos de Informática*, 185:29–35, 2007.

[BYBKT04] Ziv Bar-Yossef, Andrei Z. Broder, Ravi Kumar, and Andrew Tomkins. Sic transit gloria telae: towards an understanding of the web's decay. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 328–337, New York, NY, USA, 2004. ACM.

[BYCL05] Ricardo A. Baeza-Yates, Carlos Castillo, and Vicente López. Pagerank increase under different collusion topologies. In *AIRWeb*, pages 17–24, 2005.

[BYRN99] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[CDB+06] Carlos Castillo, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi, Massimo Santini, and Sebastiano Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, 2006.

[CDG+07] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. Know your neighbors: web spam detection using the web topology. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, New York, NY, USA, 2007. ACM.

[CH01] Nick Craswell and David Hawking. Overview of the trec-2001 web track. In *In Proceedings of TREC-2001*, 2001.

[CHR01] Nick Craswell, David Hawking, and Stephen Robertson. Effective site finding using link anchor information. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 250–257, New York, NY, USA, 2001. ACM.

[Coh95] Jonathan D. Cohen. Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science*, 42(3):162–174, 1995.

[CS07] Naresh Chauhan and A. K. Sharma. Analyzing anchor-links to extract semantic inferences of a web page. *Information Technology, International Conference on*, 0:277–282, 2007.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[Dav98] Hugh C. Davis. Referential integrity of links in open hypermedia systems. In *HYPERTEXT '98: Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems*, pages 207–216, New York, NY, USA, 1998. ACM.

[Dav00a] H.C Davis. Hypertext link integrity. *ACM Computing Surveys Electronic Symposium on Hypertext and Hypermedia*, 31(4), 2000.

[Dav00b] B. Davison. Recognizing nepotistic links on the web, 2000.

[DGIF99] L. Daigle, D. van Gulik, R. Iannella, and P. Falstrom. Urn namespace definition mechanisms, 1999.

[Dom00] Sándor Dominich. A unified mathematical definition of classical information retrieval. *J. Am. Soc. Inf. Sci.*, 51(7):614–624, 2000.

[DRP03] Heilig Lauren F. Drake Amanda L. Kuntzman Jeff W. Graber Marla Schilling Lisa M. Dellavalle Robert P., Hester Eric J. Going, going, gone: Lost internet references. *Information Science*, 302(5646):787–788, 2003.

[Eft96] E. N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology*, 31:121–187, 1996.

[EM03] Nadav Eiron and Kevin S. McCurley. Analysis of anchor text for web search. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 459–460, New York, NY, USA, 2003. ACM.

[EMT04] Nadav Eiron, Kevin S. McCurley, and John A. Tomlin. Ranking the web frontier. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 309–318, New York, NY, USA, 2004. ACM.

[FGM+99] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext transfer protocol – http/1.1, 1999.

[FMN04] Dennis Fetterly, Mark Manasse, and Marc Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In *WebDB '04: Proceedings of the 7th International Workshop on the Web and Databases*, pages 1–6, New York, NY, USA, 2004. ACM.

[FRSF+01] Luis Francisco-Revilla, Frank Shipman, Richard Furuta, Unmil Karadkar, and Avital Arora. Managing change on the web. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 67–76, New York, NY, USA, 2001. ACM.

[GGM05] Zoltán Gyöngyi and Hector Garcia-Molina. Web spam taxonomy. In *Proceedings of the first International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[GH04] Otis Gospodnetic and Erik Hatcher. *Lucene in Action*. Manning, 2004.

[GSØ99] Kaj Grønbæk, Lennert Sloth, and Peter Ørbæk. Webvise: Browser and proxy support for open hypermedia structuring mechanisms on the world wide web. *Computer Networks*, 31(11-16):1331–1345, 1999.

[HN06] Terry L. Harrison and Michael L. Nelson. Just-in-time recovery of missing web pages. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 145–156, New York, NY, USA, 2006. ACM.

[ICL96] David Ingham, Steve Caughey, and Mark Little. Fixing the "broken-link" problem: the w3objects approach. *Comput. Netw. ISDN Syst.*, 28(7-11):1255–1268, 1996.

[II08] Terry Dunford II. *Advanced Search Engine Optimization: A Logical Approach*. American Creations, 2008.

[JBJ+00] Bernard Jansen, Major Bernard, J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.

[JC94] Y. Jing and W. Bruce Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 146–160, New York, US, 1994.

[JK93] J. Justeson and S. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, pages 9–27, 1993.

[JS03] Bernard J. Jansen and Amanda Spink. An analysis of web documents retrieved and viewed. In *International Conference on Internet Computing*, pages 65–69, 2003.

[JW02] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, New York, NY, USA, 2002. ACM.

[Kah97] Brewster Kahle. Preserving the internet. *Scientific American*, 276(3):82–83, March 1997.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[KN08a] Martin Klein and Michael L. Nelson. A comparison of techniques for estimating idf values to generate lexical signatures for the web. In *WIDM*, pages 39–46, 2008.

[KN08b] Martin Klein and Michael L. Nelson. Revisiting lexical signatures to (re-)discover web pages. In *ECDL '08: Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries*, pages 371–382, Berlin, Heidelberg, 2008. Springer-Verlag.

[Koe99] Wallace C Koehler. Digital libraries and world wide web sites and page persistence. *Information Research*, 4(4), 1999.

[Koe02] Wallace Koehler. Web page change and persistence—a four-year longitudinal study. *J. Am. Soc. Inf. Sci. Technol.*, 53(2):162–171, 2002.

[Koe04] Wallace Koehler. A longitudinal study of web pages continued: a consideration of document persistence. *Information Research*, 9(2), 2004.

[Kor97] R. R. Korfhage. Information storage and retrieval. *New York: Wiley Computer Publisher*, 1997.

[Lew92] David Dolan Lewis. *Representation and learning in information retrieval*. PhD thesis, Amherst, MA, USA, 1992.

[LLK06] Zhenjiang Lin, Michael R. Lyu, and Irwin King. Pagesim: a novel link-based measure of web page aimilarity. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 1019–1020, New York, NY, USA, 2006. ACM.

[LPF+01] Steve Lawrence, David M. Pennock, Gary William Flake, Robert Krovetz, Frans M. Coetzee, Eric Glover, Finn Arup Nielsen, Andries Kruger, and C. Lee Giles. Persistence of web references in scientific research. *Computer*, 34(2):26–31, 2001.

[LZ01] John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–119, New York, NY, USA, 2001. ACM.

[MB02] John Markwell and David W. Brooks. Broken links: The ephemeral nature of educational www hyperlinks. *Journal of Science Education and Technology*, 11(2):105–108, June 2002.

[MCL05] Gilad Mishne, David Carmel, and Ronny Lempel. Blocking blog spam with language model disagreement. In *In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[MCNB05] Frank McCown, Sheffan Chan, Michael L. Nelson, and Johan Bollen. The availability and persistence of web references in d-lib magazine. *CoRR*, abs/cs/0511077, 2005.

[MDN07] Frank McCown, Norou Diawara, and Michael L. Nelson. Factors affecting website reconstruction from the web infrastructure. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 39–48, New York, NY, USA, 2007. ACM.

[MM04] Francisco Javier Martínez Méndez and José Vicente Rodríguez Muñoz. Reflexiones sobre la evaluación de los sistemas de recuperación de información: necesidad, utilidad y viabilidad. *Anales de Documentación*, 7:153–170, 2004.

[MMN09] Frank McCown, Catherine C. Marshall, and Michael L. Nelson. Why web sites are lost (and how they're sometimes found). *Commun. ACM*, 52(11):141–145, 2009.

[MNI+08] Atsuyuki Morishima, Akiyoshi Nakamizo, Toshinari Iida, Shigeo Sugimoto, and Hiroyuki Kitagawa. Pagechaser: A tool for the automatic correction of broken web links. In *ICDE*, pages 1486–1488, 2008.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[MSN06] Frank McCown, Joan A. Smith, and Michael L. Nelson. Lazy preservation: reconstructing websites by crawling the crawlers. In *WIDM '06: Proceedings of the 8th annual ACM international workshop on Web information and data management*, pages 67–74, New York, NY, USA, 2006. ACM.

[NA02] Michael L. Nelson and B. Danette Allen. Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1), 2002.

[NIM+05] A. Nakamizo, T. Iida, A. Morishima, S. Sugimoto, , and H. Kitagawa. A tool to compute reliable web links and its applications. In *SWOD*

'05: Proc. International Special Workshop on Databases for Next Generation Researchers, pages 146–149. IEEE Computer Society, 2005.

[NM08] Michael G. Noll and Christoph Meinel. Web search personalization via social bookmarking and tagging. In *Proceedings of 6th International Semantic Web Conference (ISWC)*, pages 367–380, 2008.

[NMSK07] Michael L. Nelson, Frank McCown, Joan A. Smith, and Martin Klein. Using the web infrastructure to preserve web pages. *Int. J. Digit. Libr.*, 6(4):327–349, 2007.

[NNMF06] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 83–92, New York, NY, USA, 2006. ACM.

[OC03] Paul Ogilvie and Jamie Callan. Combining structural information and the use of priors in mixed named-page and homepage finding. In *In Proceedings of the Twelfth Text Retrieval Conference TREC-12*, pages 177–184, 2003.

[OMC07] Antoni Oliver, Joaquím Moré, and Salvador Climent. *Traducción y tecnologías*. Editorial UOC, 2007.

[Pan03] Gautam Pant. Deriving link-context from html tag tree. In *DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 49–55, New York, NY, USA, 2003. ACM.

[Pas02] Norman Paskin. Digital object identifiers. *Inf. Serv. Use*, 22(2-3):97–112, 2002.

[PBMW98] Larry Page, Sergey Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1998.

[PC98] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM.

[PF97] Foster Provost and Tom Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. 1997.

[PMRJ02] Eui Kyu Park, Seong In Moon, Dong Yul Ra, and Myung Gil Jang. Web document retrieval using sentence-query similarity. In *In Proceedings of the 11 th Text REtrieval Conference (TREC-11), notebook version*, 2002.

[Por00] Niels Ole Pors. Information retrieval, experimental models and statistical analysis. *Journal of Documentation*, 56:55–70(16), 1 January 2000.

[PPGK04] Seung-Taek Park, David M. Pennock, C. Lee Giles, and Robert Krovetz. Analysis of lexical signatures for improving information persistence on the world wide web. *ACM Trans. Inf. Syst.*, 22(4):540–572, 2004.

[PSW08a] Jakub Piskorski, Marcin Sydow, and Dawid Weiss. Exploring linguistic features for web spam detection: a preliminary study. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 25–28, New York, NY, USA, 2008. ACM.

[PSW08b] Jakub Piskorski, Marcin Sydow, and Dawid Weiss. Exploring linguistic features for web spam detection: a preliminary study. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 25–28, New York, NY, USA, 2008. ACM.

[PW00] Thomas A. Phelps and Robert Wilensky. Robust hyperlinks cost just five words each. Technical report, Berkeley, CA, USA, 2000.

[QF93] Yonggang Qiu and Hans-Peter Frei. Concept based query expansion. In *SIGIR*, pages 160–169, 1993.

[QF95] Yonggang Qiu and Hans-Peter Frei. Improving the retrieval effectiveness by a similarity thesaurus. Technical Report 225, Zürich, Switzerland, 1995.

[QND07] Xiaoguang Qi, Lan Nie, and Brian D. Davison. Measuring similarity to detect qualified links. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 49–56, New York, NY, USA, 2007. ACM.

[Rij77] C. J. Van Rijsbergen. A theoretical basis for the use of cooccurrence data in information retrieval. *Journal of Documentation*, (33):106–119, 1977.

[SF98] Takehiro Shimada and Atsushi Futakata. Automatic link generation and repair mechanism for document management. In *HICSS '98: Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences-Volume 2*, page 226, Washington, DC, USA, 1998. IEEE Computer Society.

[SLB03] S. Sun, L. Lannom, and B. Boesch. Handle system overview, 2003.

[SM86] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[SP97] Hinrich Schütze and Jan O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. *Inf. Process. Manage.*, 33(3):307–318, 1997.

[Spi03] Diomidis Spinellis. The decay and failures of web references. *Commun. ACM*, 46(1):71–77, 2003.

[SV07] Francesco Sclano and Paola Velardi. Termextractor: a web application to learn the shared terminology of emergent web communities, 2007.

[SWJF96] K. Shafer, S. Weibel, E. Jul, and J. Fausey. Introduction to persistent uniform resource locators. `http://purl.oclc.org/OCLC/PURL/INET96`, 1996.

[TAJP06] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael Potts. History repeats itself: repeat queries in yahoo's logs. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 703–704, New York, NY, USA, 2006. ACM.

[TAJP07] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 151–158, New York, NY, USA, 2007. ACM.

[Voo03] Ellen M. Voorhees. Overview of the trec 2003 robust retrieval track. In *TREC*, pages 69–77, 2003.

[Voo05] Ellen M. Voorhees. The trec robust retrieval track. *SIGIR Forum*, 39(1):11–20, 2005.

[Voo06] Ellen M. Voorhees. The trec 2005 robust track. *SIGIR Forum*, 40(1):41–48, 2006.

[WF05] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2 edition, 2005.

[WKH01] Thijs Westerveld, Wessel Kraaij, and Djoerd Hiemstra. Retrieving web pages using content, links, urls and anchors. In *Proceedings of TREC10*, pages 663–672, Gaithersburg, MD, NIST, 2001.

[WY06] Xiaojun Wan and Jianwu Yang. Wordrank-based lexical signatures for finding lost or related web pages. In *Frontiers of WWW Research and Development - APWeb 2006*, pages 843–849, 2006.

[XFTS02] Wensi Xi, Edward A. Fox, Roy P. Tan, and Jiang Shu. Machine learning approach for homepage finding task. In *SPIRE 2002: Proceedings of the 9th International Symposium on String Processing and Information Retrieval*, pages 145–159, London, UK, 2002. Springer-Verlag.

[YJNT07] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Can social bookmarking enhance search in the web? In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 107–116, New York, NY, USA, 2007. ACM.

APÉNDICE **A**

Entorno de Recuperación de Enlaces Rotos

Este apéndice contiene la información relativa a la aplicación web "Detective Brooklynk". Esta aplicación contiene el motor del sistema de recuperación de enlaces rotos presentado en el capítulo 4.

Esta aplicación consiste en un sistema de recomendación de páginas candidatas a sustituir un enlace roto. La aplicación comprueba los enlaces de la página dada como entrada. En el caso de detectar algún enlace roto, determina si tenemos suficiente información disponible para realizar una recomendación fiable. En caso negativo, se avisa al usuario de dicha situación informándole de los motivos de la imposibilidad de recuperar el enlace. Si por el contrario la aplicación dispone de suficiente información, proporciona al usuario una serie de páginas candidatas para sustituir aquella desaparecida.

A continuación, se realizará un recorrido por los estados principales de la aplicación "Detective Brooklynk", repasando las principales decisiones tomadas por el algoritmo de recuperación de enlaces rotos.

En primer lugar, tal y como se muestra en la Figura A.1, el sistema cuenta con una interfaz de sencillo uso en donde el usuario puede insertar una Url para que sea analizada. Para dicha inserción, el usuario dispone de un cuadro de texto y un botón de búsqueda.

En la Figura A.2, se observa un ejemplo de inserción de una Url. Después de escribir la dirección de la página que el usuario solicita ser analizada, se encuentra a la derecha un botón de búsqueda que dará lugar al inicio del sistema de recuperación.

En esta primera etapa del sistema de recuperación de enlaces, se realiza un análisis de todos y cada uno de los enlaces diferenciando por un lado los enlaces activos de los que están rotos. En cuanto a los enlaces rotos, recordamos que el sistema tan solo analiza los enlaces externos, por tanto los enlaces que apuntan al mismo sitio son descartados. También es un requisito imprescindible que el texto del ancla del enlace no sea una secuencia de números, una Url o una cadena vacía. Finalmente, si el texto del ancla está compuesto solamente por un carácter y además es un signo de puntuación, el enlace es descartado. Este proceso de análisis se realiza mientras

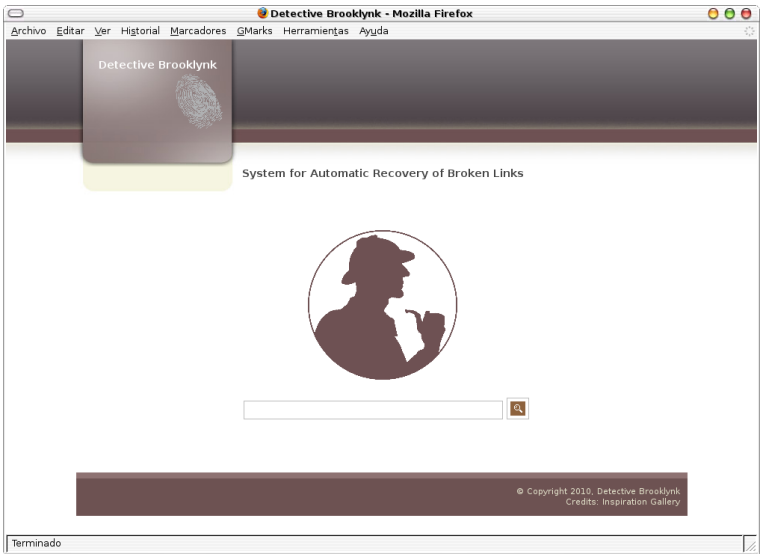

Figura A.1: *Página de inicio de SRER.*

la aplicación muestra una página de transición, tal y como la que se muestra en la Figura A.3.

En la Figura A.4 se pueden observar los enlaces analizados por el sistema. En el caso de los enlaces rotos la aplicación los muestra al principio, indícandolos mediante dos elementos diferentes. Por un lado se muestra un símbolo rojo con un aspa en el interior, que indica que la página a la que apunta ha desaparecido. El otro elemento que caracteriza cada enlace roto, es un botón que inicia el proceso de recuperación del enlace.

En el caso de que existan enlaces activos, la aplicación “Detective Brooklynk”, tal y como aparece en la Figura A.5, muestra un símbolo verde que representa la corrección de dicho enlace.

El proceso de recuperación de un enlace roto, se inicia cuando el usuario hace click en el botón “Investigate”, tal y como se refleja en la Figura A.4. A partir de ese momento, siguiendo el esquema definido en esta tesis, se desarrolla parte de la primera etapa. En esta etapa, el sistema recolecta un conjunto de fuentes de información tanto en el contexto del enlace roto como en la infraestructura web. Desde el inicio de este proceso, el algoritmo de recuperación de enlaces rotos, gestiona cada paso del sistema. De esta forma, tras esta fase de búsqueda de fuentes de información, el algoritmo estima la probabilidad de recuperación del enlace analizado, teniendo en cuenta la información que el sistema ha sido capaz de recuperar. El resultado de esta estimación se puede observar en la Figura A.6, donde el sistema informa al usuario mediante un valor numérico de dicha probabilidad de éxito. El sistema también pregunta al usuario si desea seguir adelante con el proceso de recuperación o bien lo descarta.

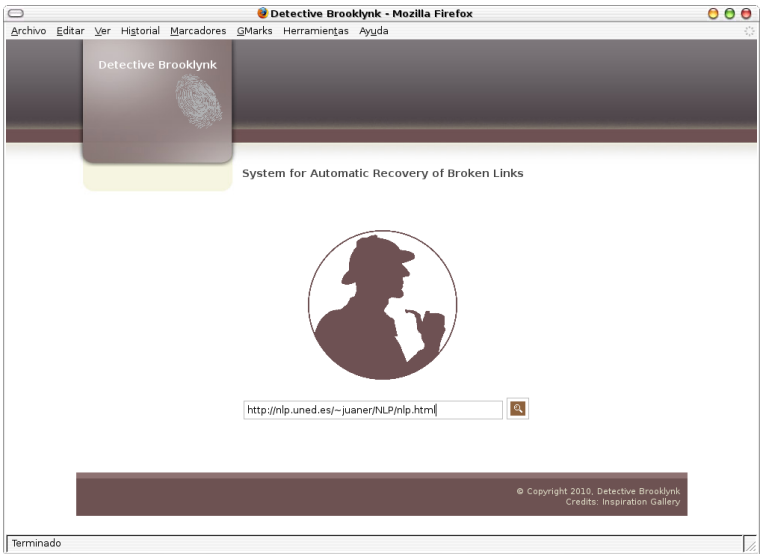

Figura A.2: *Proceso de inserción de la Url en el cuadro de texto para su posterior análisis en SRER.*

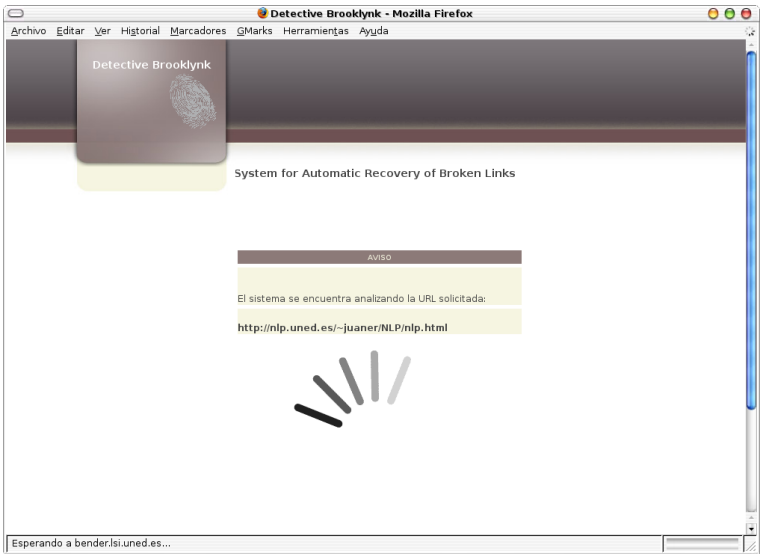

Figura A.3: *Página mostrada al usuario mientras se realiza el proceso de análisis de los enlaces de una página web.*

Finalmente, si el usuario decide seguir adelante, el sistema completa el resto del proceso de recuperación del enlace dirigido por el algoritmo de recuperación. Al final de este proceso, el sistema muestra al usuario una lista de las páginas candidatas a reemplazar el enlace roto. Esta lista se encuentra además ordenada según la rele-

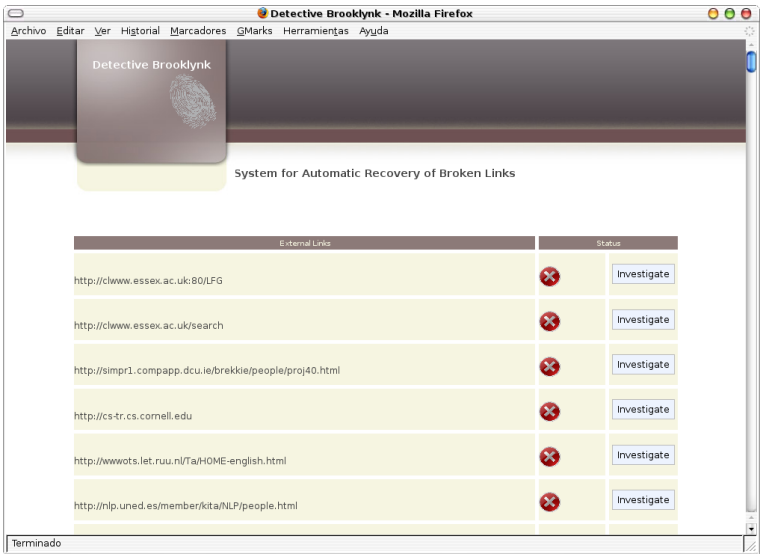

Figura A.4: *Página de resultados donde se muestran los enlaces analizados por SRER. En el caso de los enlaces rotos se muestra un símbolo rojo y un botón que permite la recuperación de dicho enlace.*

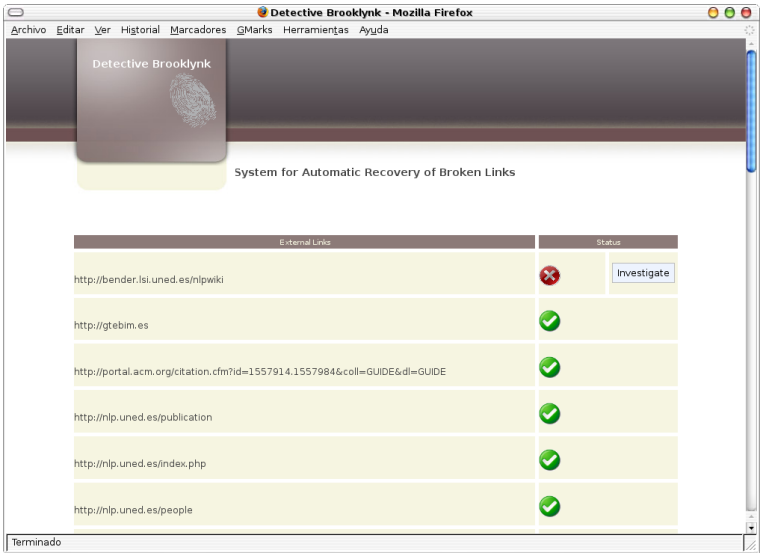

Figura A.5: *Página de resultados donde se muestran los enlaces analizados por SRER. En el caso de los enlaces activos se muestra un símbolo verde que indica la corrección del enlace.*

vancia con la información que el sistema ha conseguido recolectar en la fase inicial. En la Figura A.7 se pueden ver los primeros resultados que “Detective Brooklynk”

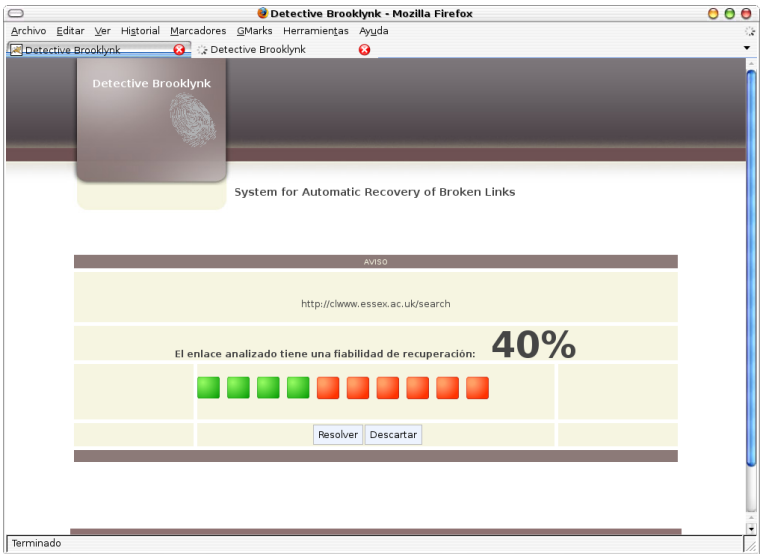

Figura A.6: *Página informativa en la que se muestra al usuario una estimación de la probabilidad de recuperación del enlace roto. El usuario debe optar por intentar recuperar el enlace o descartar la operación.*

ha recomendado al usuario. En esta figura también se puede apreciar la síntesis de información que se muestra de cada página candidata, que consta del título, la Url y un breve resumen del contenido.

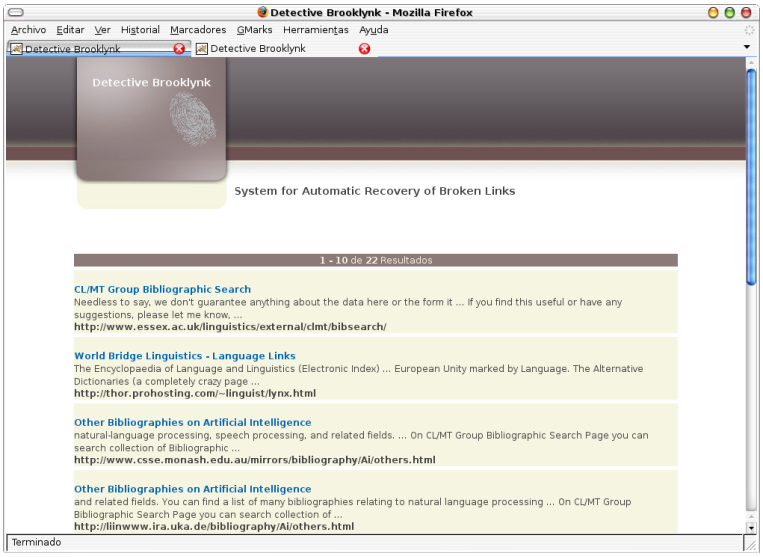

Figura A.7: *Página de resultados de SRER donde se muestra una síntesis de información de cada página candidata.*