Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación

Caracterización de la Web Chilena y Extensiones a un Buscador Web

Carlos Alberto Castillo Ocaranza

Comisión Examinadora:

Profesor Guía: Sr. Ricardo Baeza-Yates
Profesor Co-Guía: Sr. Gonzalo Navarro
Profesor Integrante: Sr. José Pino

Memoria para optar al título de
Ingeniero Civil, Mención Computación

Santiago, Julio 2000

Resumen

Este trabajo consta de dos partes: una caracterización de la Web Chilena y el diseño e implementación de extensiones a una máquina de búsqueda (TodoCL).

Respecto a la caracterización de la Web Chilena, durante los meses de mayo y junio del año 2000, se llevó a cabo un estudio basado en datos obtenidos con el recolector de páginas del buscador TodoCL. Dicho estudio contempla tanto características individuales de las páginas, como del conjunto de las páginas a nivel de sitio y de dominio. Se presentan numerosos datos estadísticos y modelos que configuran los aspectos fundamentales de la Web Chilena.

Mediante una clasificación de los dominios se muestra una representación concisa de la conectividad entre ellos, así como características de las estructuras que esta representación revela.

Las extensiones al buscador TodoCL consistieron en mejoras de la interfaz y de las capacidades de búsqueda. En la interfaz, se incluye un localizador de búsqueda y un sistema de expansión de consultas. Respecto a las capacidades de búsqueda, se desarrollaron rutinas para reconocer textos en español e inglés y para mantener datos sobre los multimedios insertados en cada página, así como para descartar potenciales archivos binarios que se incorporaran a la consulta.

Finalmente, se agregó a TodoCL un directorio de más de 3200 páginas Web Chilenas clasificadas por tema, provenientes del Open Directory Project.

Índice General

Índice de Figuras

Capítulo 1

Introducción

1.1 El World Wide Web

La capacidad de realizar publicación global de documentos a costos muy bajos, es parte de una línea de evolución tecnológica que tiene por antecesores a la escritura, la imprenta, y la radio/televisión. Cada uno de estos avances está inspirado en la voluntad de comunicar, y se observa en cada etapa una variación en los siguientes factores:

- Aumento de la capacidad multiplicadora del medio, o equivalentemente, reducción del costo por copia.
- Aumento del rango de alcance del medio, o equivalentemente, reducción del costo por transporte de la información.
- Aumento de la capacidad de información del medio, o equivalentemente, aumento del ancho de banda disponible para la información.

En las dos primeras variables, el World Wide Web se acerca a un óptimo teórico, en términos de que el costo por copia y el costo por transporte son muy cercanos a cero.

En la tercera variable, el límite teórico es el ancho de banda de los sentidos del ser humano contemplados en su totalidad; incluyendo la vista, el olfato, el oído, el gusto y el tacto, agregando las sensaciones térmicas, de presión y roce; es en esta área donde podemos esperar nuevos avances que impliquen cambios cualitativos en el medio.

Mientras todas ellas tienen una tendencia constante, hay un aspecto del cambio tecnológico que no presenta un comportamiento monótono al revisar la historia de las tecnologías de comunicación: la cantidad de fuentes de información, o dicho en términos económicos, los costos o barreras de entrada a la hora de establecer una fuente de información. En términos comparativos, estos son los de la tabla 1.1.

Esta reducción en las barreras de entrada es lo que ha generado un fenómeno que se ha llamado desde "explosión de la información" hasta "anarquía informática" Los desafíos más relevantes en la Web se relacionan con la cantidad, calidad, falta de estructuración, redundancia y heterogeneidad de los datos[BYRN1999] y todas salvo la primera son producto de este giro inesperado de la historia tecnológica que baja las barreras de entrada repentinamente.

Medio	Costo de Habilitación
Escritura	Muy Bajo
Imprenta	Medio
Radio	Alto
Televisión	Muy Alto
Web	**Bajo**

Tabla 1.1: Costos de habilitación de diversos medios de información.

1.2 Buscando Información

A fines de la segunda guerra mundial, en el año 1945, el oficial federal Vannevar Bush advertía sobre los peligros del hecho de que el hombre creara información a un ritmo más rápido del que podía digerir. Este visionario científico imaginaba avances en las técnicas fotográficas o de creación de fotofacsímiles que permitirían, algún día, "gran cantidad de duplicados de un documento que probablemente serían distribuidos con costos muy inferiores a un centavo" y que "La suma de la experiencia humana se expande a una tasa prodigiosa, y los medios que usamos para (recorrer) el laberinto que se produce para llegar a los ítems que son importantes son los mismos que eran usados en los días de las naves con velas cuadradas"[VBUSH1945].

Hoy en día, la tasa de aumento de la información ha crecido, pero estamos algo más preparados para lidiar con ella que en los tiempos de V. Bush. Por ejemplo, sabemos que la búsqueda de información se facilita al permitir a quien la necesita consultar en un universo más pequeño y mejor estructurado que los propios documentos que constituyen su objetivo.

Esto es "información sobre la información" o de una manera más suscinta, *meta-información*. Una característica mínima requerida para que ella presente utilidad a quien tiene una necesidad de información es que tenga una relación estrecha e inequívoca con el objeto representado.

Es propicio citar un famoso cuadro de René Magritte, que se reproduce en la figura 1.1. El sentido original del pintor surrealista es delatar la incongruencia de un universo de azar[V1979] y la fascinación del hombre con las representaciones que se hacen de los objetos más que sobre los objetos mismos[1], pero una segunda lectura lleva a lo chocante que resultan las incongruencias entre el objeto (la información provista por la pintura de la pipa) y la representación del objeto (la meta-información del texto que la acompaña).

Casos como éste ocurren con mucha frecuencia en recuperación de la información, particularmente en la Web. La presencia de un porcentaje alto de respuestas que no tienen ninguna relación con el campo en el cual se encuentra la necesidad del usuario es una de las principales quejas de éstos; los métodos para permitir a los usuarios encontrar documentos que cubran sus necesidades de información en forma satisfactoria, principalmente que sean actuales, válidos y de calidad, requiere contar con meta-información de calidad sobre los documentos.

El mejoramiento de la meta-información disponible sobre la Web corresponde a dos áreas de trabajo: con los proveedores de contenido y con los proveedores de meta-contenido (directorios o buscadores de páginas).

[1] Como en otro cuadro del mismo autor: "La condición humana".

Figura 1.1: "Esto no es una pipa" dice René Magritte.

El trabajo con los proveedores de contenido consiste en entregarles herramientas y estándares para etiquetar los documentos (palabras claves, descripciones o etiquetas PICS[W3C]). El éxito de tales iniciativas es parcial y está limitado a un número pequeño de documentos, pero aún así representa una línea en la cual se deben seguir realizando esfuerzos.

El trabajo con los proveedores de meta-contenido se relaciona con el desarrollo de algoritmos y software para clasificación y búsqueda de documentos. Al respecto existe investigación realizada desde antes del nacimiento de la Web en el área de recuperación de la información, y con anterioridad, en el área de catalogos bibliográficos.

1.3 Objetivos

Este trabajo se divide en dos áreas:

- Caracterización de la Web Chilena: estudio estadístico y cualitativo sobre las características de la Web Chilena, para compararlo con caracterizaciones de la Web mundial y otros estudios locales.
- Investigar extensiones a la máquina de búsqueda existente e implementar algunas de esas extensiones.

Ambas apuntan a poder desarrollar a futuro mejores sistemas verticales (es decir, en un contexto definido, como un país o región) de búsqueda.

En el transcurso de la memoria se definieron objetivos más específicos respecto a cómo extender la máquina de búsqueda:

- Extensiones de la interfaz: Diseño e implementación de consejos de búsqueda, localizador de búsqueda, expansión de consultas, historial de consultas y opciones de presentación de resultados.
- Extensiones del recolector: Diseño e implementación de rutinas para reconocer textos en español y en inglés, para almacenar datos sobre los multimedios insertados en la

página (audio, video u otros) con el fin de buscar sobre ellos y para descartar archivos binarios que eventualmente pasan los primeros filtros del recolector y se incorporan erróneamente a la colección.

- Extensión del buscador: Incorporación de un directorio de páginas.

1.4 Principales Resultados

Se realizó el estudio más completo existente a la fecha sobre la Web Chilena, a nivel de página, sitio y dominio; en particular, se extendieron los resultados de [B2000] al estudiar características de las componentes que se extraen al estudiar la conectividad a nivel de dominio.

Se diseñaron e implementaron varias extensiones a la máquina de búsqueda de TodoCL que amplian las capacidades de búsqueda y visualización de los resultados a una consulta, probándose ideas en un entorno real de consultas y manteniéndose la complejidad en tiempo y espacio suficientemente baja como para permitir que estas extensiones sigan siendo aplicables ante cambios de escala.

1.5 Organización de esta Memoria

En el capítulo 2 (Preliminares) se discuten conceptos y estado del arte en el tema de buscadores en Internet y estudios sobre la Web. Así mismo, se explica el entorno y la metodología empleada en este trabajo.

En el capítulo 3 (Caracterización) se presenta el estudio sobre la Web Chilena, dividido a nivel de colección, página, sitio y dominio[BYC2000].

En el capítulo 4 (Extensiones) se discuten las extensiones realizadas a la máquina de búsqueda, en términos conceptuales, de diseño e implementación.

Finalmente, en las Conclusiones se explican los resultados obtenidos en los temas que componen esta memoria.

Capítulo 2

Preliminares

2.1 Conceptos sobre Máquinas de Búsqueda

Una máquina de búsqueda es un proveedor de meta-contenido que selecciona un conjunto de páginas Web, tomando como base el requerimiento de un usuario, y despliega meta-información respecto a ellas.

El requerimiento del usuario puede ser formulado en lenguaje natural (ej: "¿dónde puedo encontrar información sobre automóviles deportivos?"), en términos de estructura (ej: documentos en que el título sea "automóviles deportivos"), basado en patrones (ej. : "(auto*) (deport*)") o como palabras clave (ej. : "automóvil deportivo").

La selección del conjunto de páginas Web que será desplegado depende fuertemente de la forma de almacenar los documentos en la base de datos de la máquina de búsqueda. Esta forma puede consistir en guardar el texto completo, una porción del texto, la estructura del documento, combinaciones de las anteriores, o formas reducidas (representaciones o índices).

Adicionalmente y como se trata de hipertextos, puede guardarse información sobre la estructura de los enlaces o links entre los documentos, así como los demás metadatos que se hayan incorporado a la página Web mediante el uso de marcas especiales de HTML existentes a tal efecto.

Las máquinas de búsqueda están compuestas usualmente de varios módulos, como se indica en la figura 2.1.

2.2 Buscadores Locales

2.2.1 TodoBR

TodoBR [TODOBR] es una máquina de búsqueda desarrollada en el Departamento de Ciencias de la Computación de la Universidad Federal de Minas Geráis, en Belo Horizonte, Brasil.

TodoBR recibe requerimientos del usuario expresados como palabras clave y modificadores contextuales restringidos o *método de búsqueda* el cual puede ser "todas las palabras", "alguna de las palabras" o "frase exacta". Se almacena un índice invertido (que es una

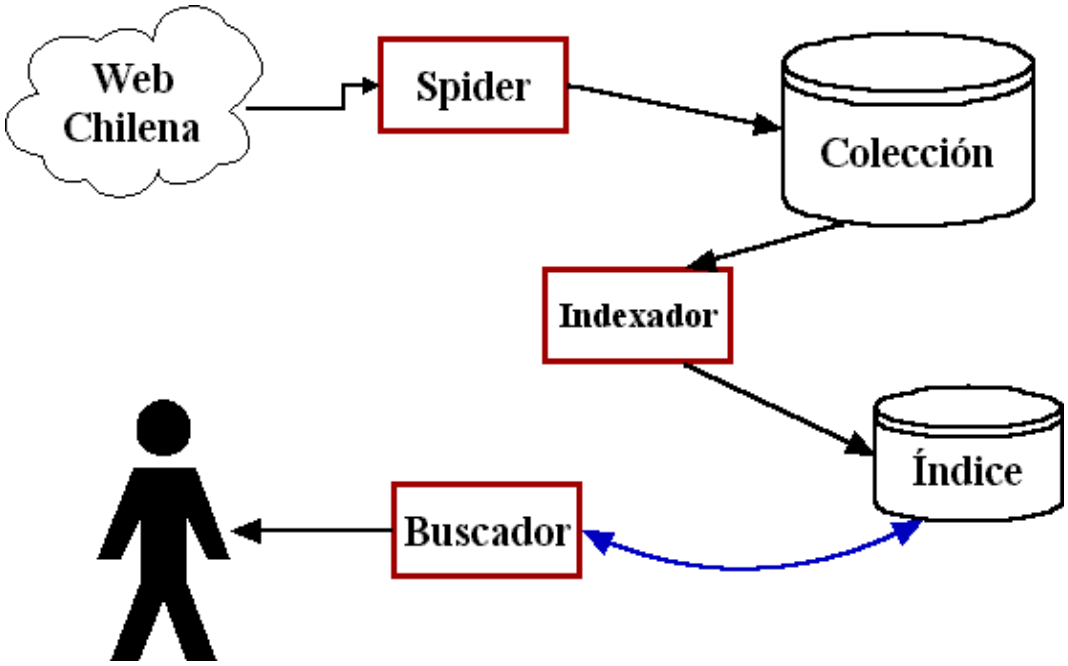

Figura 2.1: Esquema de una máquina de búsqueda.

lista de las palabras existentes en la colección de documentos y punteros a los documentos en que cada palabra aparece) creado sobre el texto completo de las páginas.

En las primeras etapas de TodoBR, se almacenaba también el texto completo de las páginas extraídas, sin embargo, para el tamaño de la Web brasileña, esto representa una gran cantidad de espacio, del orden de los 100Gb.

2.2.2 TodoCL

TodoCL [TODOCL] es la versión para Chile de TodoBR. Consiste en los mismos programas que TodoBR más extensiones desarrolladas localmente. Los programas son actualizados cuando ocurre una actualización mayor del software utilizado en TodoBR. Se incluyen en este sitio las estadísticas de la Web Chilena que se presentan en el capítulo 3.

Para efectos de este trabajo, se creó una copia de todo el sitio que se abrió al público bajo el nombre de "Laboratorio de TodoCL"[TODOCL2]. Este fue el ambiente de experimentación donde se llevaron a cabo las extensiones al buscador que se abordan en el capítulo 4.

2.2.3 Ventajas de un Buscador Local

TodoBR y TodoCL poseen algunas características destacadas en el ámbito de los buscadores, relacionadas con la construcción del índice y la forma de presentar los resultados y con la cobertura y actualización del buscador.

El índice es creado sobre el texto completo de las páginas Web, lo que es muy difícil

y costoso en buscadores internacionales por el gran volumen de datos involucrado; en ellos es común indexar sólo una pequeña porción de los documentos, usualmente los primeros 200 bytes.

El hecho de almacenar localmente el texto completo da la posibilidad de presentar en los resultados el párrafo en el cual ocurre el calce con los términos de búsqueda provistos por el usuario. En el ámbito de los buscadores de mayor prestigio, sólo [GOOGLE] realiza tal tarea.

La cobertura también es un problema importante, puesto que al existir una gran cantidad de páginas en el mundo, cada buscador indexa sólo una porción de ella (alrededor de un 30% en los mejores casos, como [GOOGLE]) y para indexarla se requiere de mucho tiempo de procesador. En el caso de TodoBR, indexar sobre el 95% de las páginas brasileñas toma del orden de 15 días[VMG1999] y en el caso de TodoCL se puede lograr la misma tasa de cobertura en 4-5 días.

Estas cifras están muy por debajo de las de un buscador Web global que puede tomar del orden de 2 años para rotar por completo su índice [ATW], siendo esta una estimación, por cuanto ningún buscador internacional tiene tasas de cobertura tan altas.

Una ventaja adicional de un buscador local corresponde a las características propias de cada lenguaje, en particular de aquellos que utilizan alfabetos latinos como `ISO-8859-1`. En estos lenguajes, es muy usual omitir tildes, umlauts y cedillas al escribir, existiendo una tabla de conversión entre los caracteres del lenguaje y las letras del alfabeto convencional `US-ASCII`. Este hecho normalmente no es contemplado por los buscadores internacionales.

2.3 Estado del Arte

2.3.1 Estudios sobre la Web

A nivel global

En términos de número de servidores y de dominios registrados, el referente más destacado es [NETCRAFT], que mide el número de sitios en Internet (estimado en cerca de 15 millones); este sitio está principalmente orientado a determinar la participación de mercado de cada software utilizado como servidor Web.

Respecto al número de dominios inscritos por país, o `ccTLD` [1] [DSTATS] mantiene un registro que es actualizado periódicamente. Los países con más dominios registrados son Alemania (`.de`) e Inglaterra (`.uk`), con 2/5 de los dominios totales registrados bajo algún país.

Estudios relativos al tamaño de la Web en términos de páginas son llevados a cabo periódicamente por las máquinas de búsqueda más importantes, en particular por [AV] y [GOOGLE].

Para países de latinoamérica, el resumen de más de 40 estudios, focalizados preferentemente en el tema de penetración de comercio electrónico en el continente, está disponible en [NUA].

[1] `ccTLD`: *Country Code Top Level Domain* - Dominio de máximo nivel por código de país

Sobre la Web Chilena

NIC Chile[NIC-CL] mantiene el registro del número de dominios inscritos bajo `.cl`. Además de esto, no hay otros estudios sobre las características de la Web Chilena hasta la fecha.

Respecto al comportamiento y preferencias de los usuarios, una encuesta realizada por Ekhos e Interaccess [NAP99] sobre uso de Internet fue realizada en 1998 y 1999. Adicionalmente las empresas Certifica.COM e Interating.COM realizan medición de audiencia de algunos sitios Web, pero ninguna ha emitido un reporte al respecto.

Adicionalmente, existe un estudio de mercado sobre empresas y usuarios en Internet realizado por la Facultad de Ciencias Económicas y Administrativas de la Universidad de Chile [FACEA99].

2.3.2 Buscadores

A nivel global

Varios informes sobre distintas características (audiencia, número de sitios, ratings) de las mayores máquinas de búsqueda son mantenidos por Search Engine Watch [SEWATCH]. Algunas máquinas de búsqueda que utilizan índices creados automáticamente aparecen en la tabla 2.1; en la tabla se indica cuales tienen localización de búsqueda o expansión de consultas. Mediante el uso de técnicas de muestreo apropiadas, se puede estimar que en el mejor de los casos la tasa de cobertura de estos sistemas es del orden del 20-30%[HHMN2000].

Buscador	Millones de Páginas Indexadas	Extensiones
Google	560	Localización de búsqueda
Altavista	350	Expansión de consultas
AllTheWeb	340	
NorthernLight	260	Expansión de consultas
Excite	214	Expansión de consultas
Inktomi	110	

Tabla 2.1: Características relevantes de las mayores máquinas de búsqueda.

La mayoría de los buscadores provee de verificación ortográfica y algún tipo de proceso para consultas en lenguaje natural.

Sobre la Web Chilena

A la fecha no existen otros índices automáticos específicamente sobre la Web Chilena, sino más bien directorios de tipo editorial.

Sin embargo, todos los buscadores globales en la tabla 2.1, salvo Google, tienen en su búsqueda avanzada opciones que permiten buscar en la Web Chilena (indicando como parámetro el dominio `.cl`). Por ejemplo, si se busca "universidad" en Chile, el siguiente es el procedimiento en los mayores buscadores:

- En Altavista, consultar por "universidad host:.cl"
- En Northern Light, consultar por "universidad URL:.cl"

- En Excite, en la página de búsqueda avanzada, ingresar término "universidad", y bajo *choose country/domain*, seleccionar "chile".
- En All The Web, en la página de búsqueda avanzada, ingresar término "universidad", y bajo *domain filter, only include*, ingresar "cl".
- En Inktomi, en la página de búsqueda avanzada, seleccionar en una de las cajas de búsqueda "must contain" y "in URL" e ingresar ".cl". En otra de las cajas ingresar "universidad".

Google ha instalado buscadores con filtros por idioma, pero aún no provee de operadores para filtro por dominio.

2.4 Metodología Empleada

La metodología adoptada para ordenar el trabajo tiene una relación estrecha con la forma en que se dieron los objetivos, y consistió en ir en paralelo desarrollando una extensión y explorando las siguientes. Las extensiones que iban surgiendo (motivadas por el alumno o el profesor guía, o en conversaciones entre ambos) se fueron poniendo en una lista de prioridades. Cuando se terminaba de realizar una extensión, se tomaba la primera en la lista de las pendientes y se implementaba.

Este ciclo se repitió varias veces, pero se reservó tiempo para documentación y para la caracterización de la Web Chilena. Cómo se trataba de experimentar con ideas e implementaciones, no existía claridad sobre los tiempos de desarrollo; de hecho, si sólo una extensión hubiera ocupado todo el tiempo de esta memoria, también nos encontraríamos frente a un escenario aceptable, de acuerdo a lo definido al comienzo entre alumno y profesor.

2.5 Ambiente de Trabajo

El trabajo se desarrolló en dependencias del Departamento de Ciencias de la Computación.

2.5.1 Grupo Humano

Además del profesor guía de esta memoria, Ricardo Baeza-Yates, se trabajó permanentemente con el profesor Gonzalo Navarro respecto a algoritmos y heurísticas sobre texto y con Edleno de Moura y Eveline Veloso de la Universidad Federal de Minas Gerais en Brasil respecto a la implementación de la máquina de búsqueda (consultas por e-mail). Además Eveline viajó a Chile en Enero del 2000 a instalar el recolector, explicando el código de la máquina de búsqueda y los procedimientos de mantenimiento.

2.5.2 Hardware

Las máquinas utilizadas fueron 2 PC con procesadores Intel de 500Mhz, 256Mb de memoria y 30-40Gb de disco duro. Por las características del sistema de búsqueda se requería capacidad masiva de almacenamiento, pero no particularmente mucha velocidad de procesamiento, de modo que estos computadores cubrían razonablemente bien las demandas del motor de búsquedas.

Un PC ejecuta el indexador y recolector y el otro el servidor Web y la aplicación de consultas.

2.5.3 Software

Los PC operaban bajo sistema operativo Linux, distribución RedHat/6.1 y servidor Web Apache con plug-in de PHP.

El código base es el de TodoBR, versión Marzo 2000.

Como lenguaje de programación se usó C/C++ para el buscador y Perl para el recolector. PHP fue utilizado sólo tangencialmente para ciertos aspectos de presentación que no son relevantes.

Se utilizaron varios filtros y subrutinas de Gonzalo Navarro que se mencionan en el resto del documento.

Para graficar se utilizó [GRAPHVIZ] de la AT&T y [RDFPARSE] para convertir el directorio de páginas del *Open Directory Project* en páginas HTML. Ambos programas son distribuídos bajo licencia GPL.

No se utilizó ningún otro software (ej: aplicación de base de datos) en esta memoria.

Capítulo 3

Caracterización de la Web Chilena

Nota: Este trabajo fue realizado con Ricardo Baeza-Yates y presentado a las Jornadas Chilenas de Ciencias de la Computación 2000

3.1 Introducción

El interés de realizar una descripción de la Web es proveer de información para aplicaciones técnicas, comerciales y sociológicas, en particular para minería de datos y de comportamiento de los usuarios. La Web es altamente dinámica y su caracterización también permite entender como evoluciona. También es importante estudiar subconjuntos de la Web, en particular por contextos culturales o geográficos, que en nuestro caso es la Web Chilena. Para este estudio, usamos el buscador TodoCL [TODOCL], desarrollado en el Departamento de Ciencias de la Computación de la Univ. de Chile.

En el año 1993 se instaló el primer servidor Web Chileno (y uno de los primeros en Iberoamérica) en el DCC de la Universidad de Chile, y conforme fueron apareciendo más servidores, se decidió implementar un mapa sensible de Chile[SSCHILE] dividido por regiones, en el cual aparecería cada sitio basado en su localización geográfica. Este mapa permitía visualizar el estado de la Web Chilena de una forma clara e inequívoca, pero 6 meses más tarde era imposible seguir manteniéndolo, pues diaramente aparecían nuevos sitios. Hoy es necesaria una nueva forma de visualización de los sitios, basada en una estructura virtual, con criterios más complejos que el geográfico.

Con el boom de la Web, en los últimos años se ha comenzado a entender cómo es la Web, desde el punto de vista de su estructura y de cómo se usa. En particular, un estudio reciente muestra el hecho de que pocos sitios Web concentran la mayoría de las visitas de los usuarios [AH1999]. Con respecto a la macroestructura, en [B2000] se analiza una posible caracterización de la Web global basada en la conectividad y se propone la clasificación de sitios que adoptamos en este trabajo. En [PPR1999] se estudia una caracterización de las páginas que utiliza información tanto de conectividad como del comportamiento de los usuarios que utilizan un sitio.

Dado el tamaño y crecimiento de la Web, estos estudios son difíciles de realizar periódicamente. Sin embargo, esto es más sencillo en subconjuntos de ella y permite verificar

si las características globales se replican a estructuras locales y su nivel de desarrollo. Por ejemplo, una descripción al nivel de página y sitio fue realizada durante 1999 para la Web Brasileña [VMG1999] usando **TodoBR**. TodoBR es un buscador de páginas desarrollado en el Departamento de Ciencia de la Computación de la Universidad Federal de Minas Gerais, en Belo Horizonte, Brasil [1]. Del mismo modo, el estudio de la Web Chilena es interesante por si mismo, además de permitir su comparación con estudios similares.

Una motivación comercial importante, es que muchos negocios en Internet están mantenidos por medio de publicidad, en donde el beneficio de cada sitio comercial depende directamente del número de visitas que recibirá el sitio. Este número de visitas está fuertemente correlacionado con la cantidad de referencias que tiene un sitio Web en el resto de la colección. Además las visitas siguen una ley del "ganador se lo lleva todo" pues sólo unos pocos sitios atraen prácticamente toda la atención de los usuarios.

En este capítulo realizamos un estudio similar a los anteriores, pero el análisis de conectividad lo realizamos en base a dominios y no a páginas, ya que creemos que la conectividad a nivel macro es más interesante que a nivel micro. Además, realizamos el primer estudio de la correlación que existe entre la macroestructura de un subconjunto de la Web con páginas clasificadas manualmente y el comportamiento de los usuarios cuando buscan. En la siguiente sección explicamos brevemente la metodología del estudio. En las siguientes secciones caracterizamos la Web Chilena a nivel de colección, páginas, sitios y dominios, respectivamente. Para finalizar, ofrecemos las conclusiones principales de nuestro estudio y extensiones al mismo.

3.2 Conceptos Básicos

Un buscador Web que utilice indexación automática, como TodoCL y al igual que Altavista, AlltheWeb, Inktomi o NorthernLight, usualmente incluye un recolector de páginas o *spider* que comienza recorriendo e indexando un conjunto de sitios predeterminado (puntos de partida), para luego seguir indexando las páginas que son apuntadas desde estos sitios mediante un procedimiento recursivo. Este proceso es realizado simultáneamente por varios spiders a la vez, los que se comunican con un planificador o *scheduler* de indexación. Dicho proceso puede ser optimizado si se conocen a priori datos sobre las páginas que integran la colección, principalmente el tamaño de estas, qué tan frecuentemente se actualizan y cuántos links posee en promedio cada una.

Para obtener estos datos, se utilizó entonces el recolector y el *scheduler* de visita a sitios Web de TodoCL, que son adaptaciones del recolector[CoBWeb99] desarrollado para TodoBR. Se consideraron fundamentalmente páginas bajo el dominio `.CL` más algunas páginas en el dominio `.NET` pertenecientes a empresas Chilenas, principalmente proveedores de Internet (ISPs).

Este recolector realiza un proceso de filtrado de las páginas, en el cual ellas son convertidas a formato de texto plano. Se usan filtros para varios formatos comunes de documento, incluyendo `HTML`, `PDF`, `PostScript` y `Word`.

Los archivos binarios (gráficos, archivos comprimidos u otros) no se incorporan en la colección. Se utilizó una heurística para eliminar bloques de archivo que hubieran pasado a través de los filtros sin ser documentos de texto, con lo que se descartó cerca del 4% de la colección, en su mayoría archivos binarios con encabezados mal formados.

[1] Está localizado en `http://www.todobr.com.br`

La descripción que presentamos se divide en 4 niveles:

- Colección: cifras globales y estudio del vocabulario.
- Página: tamaño, tipo de documento e idioma.
- Sitio: profundidad de las páginas, número de páginas por sitio y contenido de texto total por sitio.
- Dominio: número de referencias hacia y desde un dominio, representación de la estructura global de hipervínculos entre dominios y preferencias de los usuarios.

3.3 Nivel Colección

3.3.1 Cifras Globales

En la tabla 3.1 se muestra el tamaño de la base para el estudio, obtenida en 15 días de recolección y que estimamos corresponde a más del 95% de la Web Chilena.

Puntos de partida	19.390
Páginas	730.673
Sitios	10,352
Dominios	9,102
Tamaño de la Colección	2.3 Gb
Tamaño del Vocabulario	1.9 Millones

Tabla 3.1: Tamaño de la colección.

Los puntos de partida corresponden a 19.200 nombres de dominio y alrededor de 200 direcciones ingresadas por los usuarios de TodoCL. El tamaño de la colección considera *sólo el texto* de los archivos que fue posible convertir.

La mayoría de los dominios (90%) se encuentra en Santiago, lo siguen las regiones V (780 dominios), VIII (268 dominios) y X (211 dominios).

3.3.2 Vocabulario

El conjunto de palabras distintas en la colección se denomina *vocabulario*. El modelo más utilizado para el tamaño del vocabulario (V) en función del tamaño de la colección (n) es la *Ley de Heaps* [HEAPS1978], que establece que:

$$V = Kn^{\beta}$$

donde los parámetros K y β dependen de la colección.

El problema con estimar el tamaño del vocabulario en la Web radica en que, a pesar de utilizar varias heurísticas para incorporar sólo documentos con texto, siempre una porción pequeña de archivos binarios logra traspasar los filtros, incorporándose a la colección; esto distorsiona los resultado, porque un archivo binario, por pequeño que sea, contiene muchas palabras distintas.

Para resolver este problema, se incorporó una heurística de descarte de bloques basada en la frecuencia de ocurrencia de cada carácter. Un documento normal presenta sólo unos pocos caracteres muy frecuentes (vocales, así como consonantes de uso común), mientras que un documento binario presenta una distribución de caracteres más equilibrada. Esto permite discriminar con bastante precisión cuando un archivo no debe incorporarse a la colección. Existen más datos acerca de cómo se realiza este proceso en la sección 4.2.1.

Se estimó $K \approx 2.22$ y $\beta \approx 0.63$. Para la colección del texto encontrado en páginas de la Web Chilena, el modelo sublineal se cumple con bastante precisión, y es más alto que colecciones de texto editado en inglés donde $\beta \approx 0.5$. Esto se debe en parte a la cantidad de palabras con errores y a la diversidad de lenguajes.

3.3.3 Palabras más Frecuentes

Descartando los números, así como artículos, preposiciones, y otras palabras funcionales, las palabras que aparecen en más documentos en la Web Chilena son las de la tabla 3.2. Se consideran las palabras escritas con y sin acento (por ejemplo, en 7000 documentos de la Web Chilena, "informacion" aparece sin acento, que corresponde a casi el 5% del total de ocurrencias de esta palabra).

chile	31%
cl	20%
home, información	19%
copyright, santiago, mail	17%
internet	15%
www	14%
software, page	11%

Tabla 3.2: Palabras más frecuentes.

3.4 Nivel Página

3.4.1 Tamaño

Se estudió la distribución del tamaño de cada página y de la porción de este tamaño que corresponde a texto, descartando ilustraciones y comandos de formato en los tipos de datos analizados.

Un gráfico con los tamaños se muestra en la figura 3.1. La mayoría de las páginas tienen poco texto (incluso un 50% aproximadamente sólo tiene imágenes o *tags*), y el promedio de texto es de 3.4 Kb, mientras que el promedio de la página en su totalidad es de 15.3Kb.

Alrededor del 75% del tamaño de un archivo HTML es usado por las marcas de formato (*tags*), siendo sólo el 25% restante texto. Los archivos mayores de 40kb tienen algo más de texto, alrededor de un 30% del tamaño total del archivo.

Sólo un porcentaje muy pequeño (3% de las páginas) tiene más de 40Kb.[2]

[2] Por motivos de eficiencia, el buscador trunca las páginas de más de 1Mb, por lo que no se puede

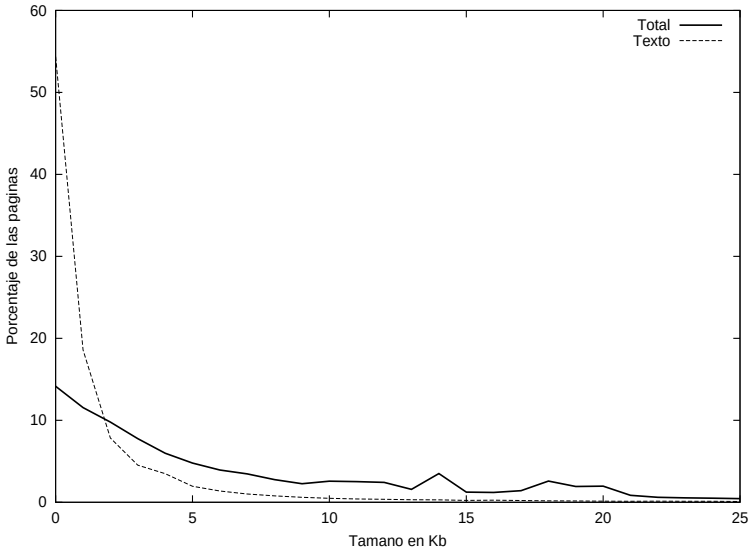

Figura 3.1: Tamaños de Texto en Páginas.

3.4.2 Tipo

El tipo de documento más común es HTML, con más del 95% de las páginas, los demás formatos de documento lo siguen bastante más atrás, como se aprecia en la tabla 3.3.

HTML	96.18%
TXT (texto plano)	2.14%
PDF (Adobe Portable Document Format)	0.85%
DOC (Microsoft Word)	0.75%
PS (Adobe Postscript)	0.01%

Tabla 3.3: Tipos de archivo.

Cabe destacar que estos datos no se obtuvieron basándose en la extensión del archivo, sino en el resultado de los filtros y de los primeros bytes de cada documento (*magic numbers*).

3.4.3 Idioma

Observaciones preliminares sobre muestras de 200-300 documentos indicaron que además de español, un porcentaje importante de las páginas Chilenas en inglés. Las mismas observaciones hacen suponer que un 1-2% de las páginas están en ambos idiomas a la vez. Otros idiomas como el francés, portugués y alemán fueron observados, pero combinados no alcanzan el 1% del total de documentos; estos datos orientaron el estudio a establecer cuál es el porcentaje de documentos en inglés sobre el total de documentos.

determinar el tamaño máximo de texto en una página. Los archivos con más texto usualmente se distribuyen en formato `PDF`.

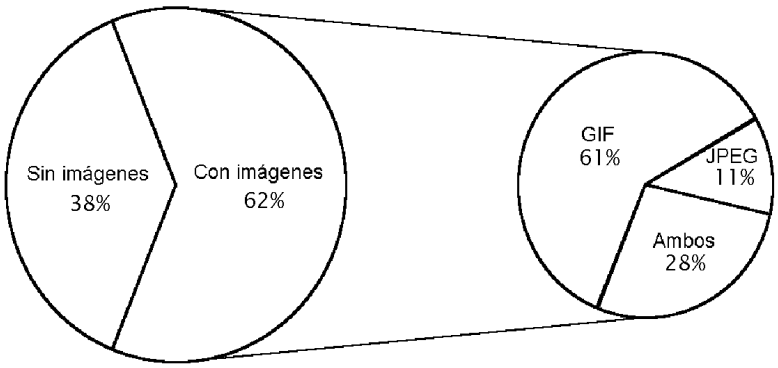

Figura 3.2: Documentos con y sin imágenes y formatos más comunes de imágen.

Es un hecho conocido que aproximadamente un 40%-50% de las palabras en un texto son *stopwords* [BYRN1999] o palabras funcionales, y esto es común a ambos idiomas. Esto permite utilizar una heurística de discriminación de lenguaje basada en el hecho de que un texto normal [3] en inglés contiene muchas stopwords de inglés y un texto normal en español contiene muchas stopwords en español.

Utilizando esa heurística, se obtiene que alrededor de un 7-8% de las páginas de la Web Chilena están en inglés. Estas páginas pueden ser destacadas en la página de resultados de una búsqueda, tal como se muestra en la sección 4.2.3.

3.4.4 Multimedios y otros formatos

Se estudió la presencia o ausencia de enlaces a formatos multimediales y de otros contenidos, en total unos 30 de los más conocidos y usados. Se incluyeron archivos de imagen, video y animaciones, programas y archivos comprimidos.

Respecto a los formatos de imagen, prácticamente sólo `GIF` (*Compuserve Graphics Interchange Format*) y `JPEG` (*Joint Photographics Expert Group*) son usados, seguidos de `PNG` (*Portable Network Graphics*) pero sólo a nivel muy incipiente (menos del 1%). En el gráfico 3.2 se aprecia la proporción entre páginas con y sin imágenes y entre los dos formatos mayores de imagen.

Las páginas que contienen programas, audio o archivos comprimidos [4] son interesantes desde el punto de vista de las necesidades de usuario, puesto que en el último tiempo se ha masificado el uso de formatos de audio y video digital. El porcentaje de páginas que contienen archivos de tipo multimedial o comprimido es pequeño, como se observa en la figura 3.3.

Nota: Los tipos de dato `RPM` (*Redhat package*) y `DEB` (*Debian package*) se consideraron comprimidos pues no sólo son usados para distribuír programas.

En la sección 4.2.2 se explica cómo se recolectan y utilizan estos datos en un entorno real de búsquedas.

[3] Quedan fuera casos anómalos como por listas de nombres propios, que no tienen stopwords.

[4] Además de los formatos de compresión estándar, los archivos con extensión `RPM` (*Redhat package*) y `DEB` (*Debian package*) se consideraron comprimidos pues no sólo son usados para distribuir programas.

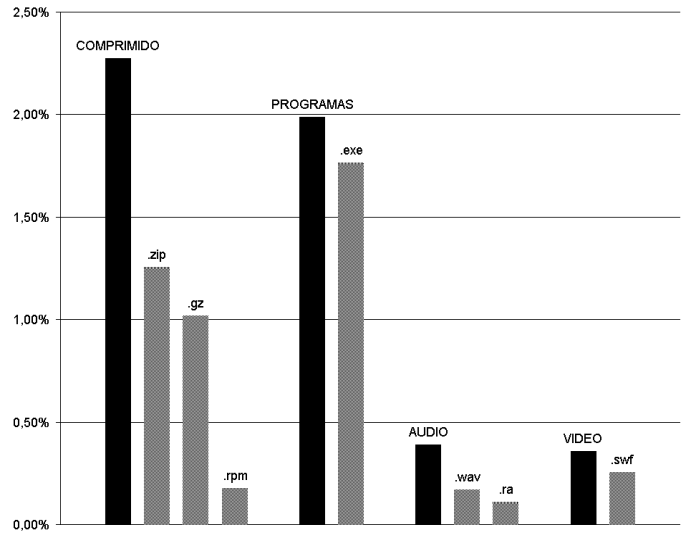

Figura 3.3: Porcentaje de páginas que incluyen otros tipos de archivo.

3.5 Nivel Sitio

3.5.1 Número de Páginas

El 52% de los sitios en la colección tienen sólo una página, y prácticamente todos los sitios tienen menos de 100 páginas, como se puede ver en la figura 3.4. La observación de que de los 20.000 dominios registrados sólo se utilicen 10.000 para poner sitios Web, y de estos sitios sólo 5.000 vayan más allá de una simple portada, dice bastante sobre la tendencia a "estar en Internet" de las empresas y organizaciones más que a "hacer cosas en Internet"

Si la mayoría de los sitios tiene tan pocas páginas, ¿cómo se llega a un total sobre 700.000? La respuesta es que las páginas están muy concentradas en unos pocos sitios, por ejemplo, la mitad de las páginas de la colección se encuentran en los 1300 sitios más grandes, cada uno con 70 o más páginas. Otro indicio de este fenómeno es que los 100 sitios más grandes (1000 o más páginas) contienen un tercio de las páginas de la Web Chilena.

3.5.2 Profundidad de las Páginas

Si bien entre el conjunto de los sitios las relaciones por hipervínculo comumnente no obedecen a clasificaciones pre-establecidas, la estructura de las relaciones al interior de un sitio resulta más bien jerárquica, incluyendo una portada de la cual se cuelgan varias secciones y subsecciones. Una forma de estudiar este árbol jerárquico de relaciones es observar la *profundidad* de las páginas dentro del árbol.

Una aproximación razonable a la profundidad de una página dentro del árbol jerárquico de relaciones es su profundidad física dentro del árbol de directorios. Así, por ejemplo, una página cuya url es: `www.uchile.cl/aa.html` tiene profundidad 1, `www.uchile.cl/bb/aa.html` tiene profundidad dos y así sucesivamente. Se observa en términos gruesos que la mitad de las páginas están a profundidad 2 o 3, y el resto sigue la distribución de la figura 3.5.

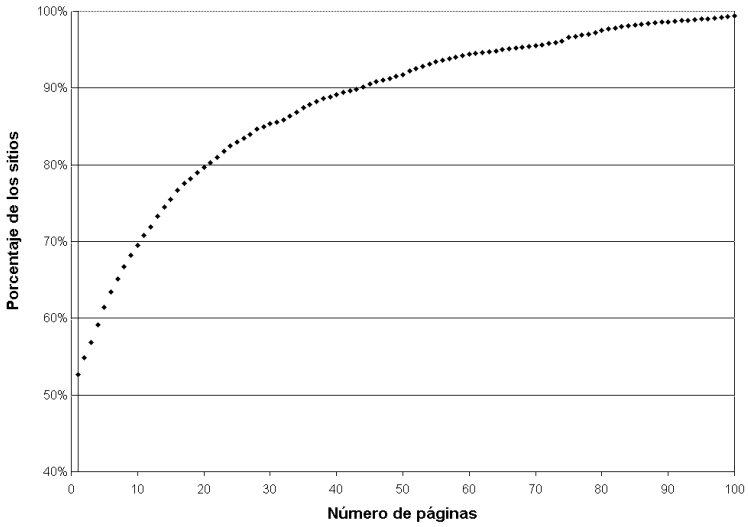

Figura 3.4: Número de páginas por sitio.

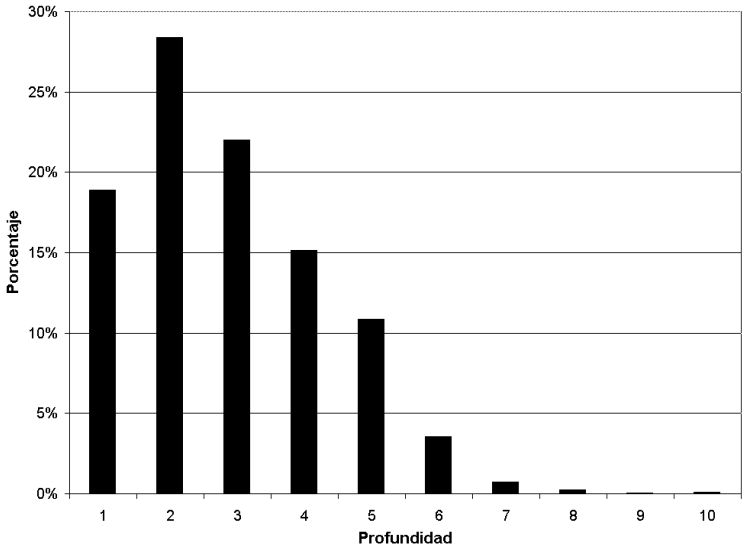

Figura 3.5: Profundidad de las páginas.

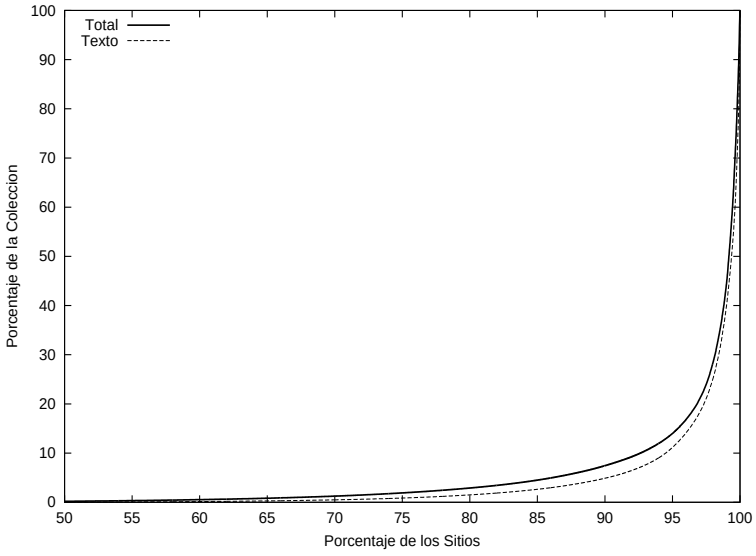

Figura 3.6: Tamaño del texto y tamaño total de las páginas, agrupadas por sitio.

3.5.3 Tamaño Total

El tamaño del texto en cada sitio (la suma del tamaño del texto de las páginas que lo componen), sigue una distribución similar al número de páginas por sitio. El 1% de los sitios más grandes en contenido aportan el 60% del texto total en la colección. En el otro extremo, los sitios de una sóla página (como se mostró más arriba, cerca del 50% de los sitios) prácticamente no aportan texto. Esto se muestra en la figura 3.6, en la que también se incluye la distribución del tamaño total de las páginas Web, considerando *tags*.

3.6 Nivel Dominio

Se estima que las páginas bajo un mismo dominio tienen relación entre sí [5]. Las relaciones entre dominios pueden representarse como un grafo dirigido, en que cada vértice representa un dominio D_i y un arco va de D_i a D_j si existe un *link* (enlace) desde una página en el primer dominio hacia una página en el segundo (es decir, varios enlaces se colapsan en uno sólo). En adelante llamaremos a este grafo, el grafo de links (enlaces), y mostraremos un análisis del mismo sobre una muestra de aproximadamente 6200 dominios.

3.6.1 Grado Interno y Externo

El número de enlaces externos[6] hacia un dominio (grado interno en el grafo de links) y el número de enlaces externos desde un dominio (grado externo), siguen una distribución

[5] Aunque es usual que se utilicen varios subdominios para usos distintos y que un mismo sitio tenga páginas no relacionadas, en general páginas bajo el mismo dominio están relacionadas.

[6] Es decir, descontando los enlaces dentro del mismo dominio.

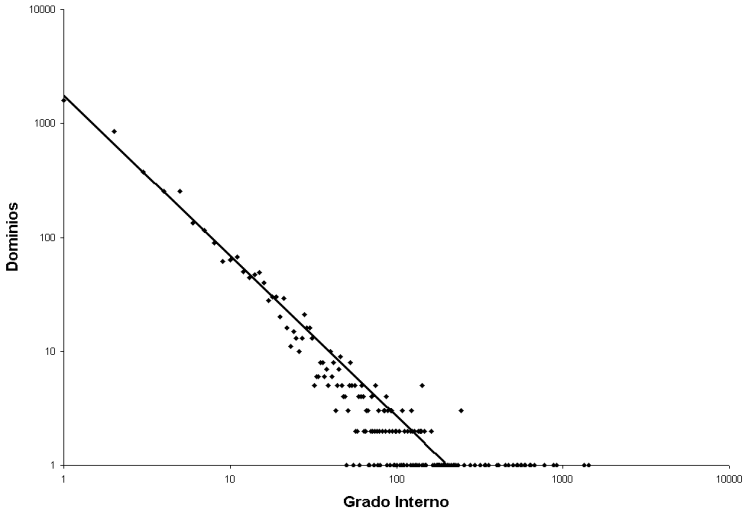

Figura 3.7: Grado Interno.

potencial de exponente negativo, de acuerdo a la ecuación

$$frecuencia = \alpha(Grado)^{-\theta} \ .$$

Usando esta ecuación, estimamos para el grado interno, $\alpha = 1781,9$ y $\theta = 1,4092$ y para el grado externo, $\alpha = 1265,3$ y $\theta = 0,9744$. Se presentan gráficos al respecto en las figuras 3.7 y 3.8.

Los 10 dominios hacia los que llegan más enlaces son los que se muestran en la tabla 3.4; esta lista representa los sitios más "populares" entre los administradores de los demás sitios Web.

`uchile.cl`	406
`chilnet.cl`	267
`elmercurio.cl`	210
`brujula.cl`	182
`puc.cl`	169
`meteochile.cl`	158
`tercera.cl`	157
`bcentral.cl`	147
`udec.cl`	128
`sii.cl`	127

Tabla 3.4: Dominios que reciben más referencias.

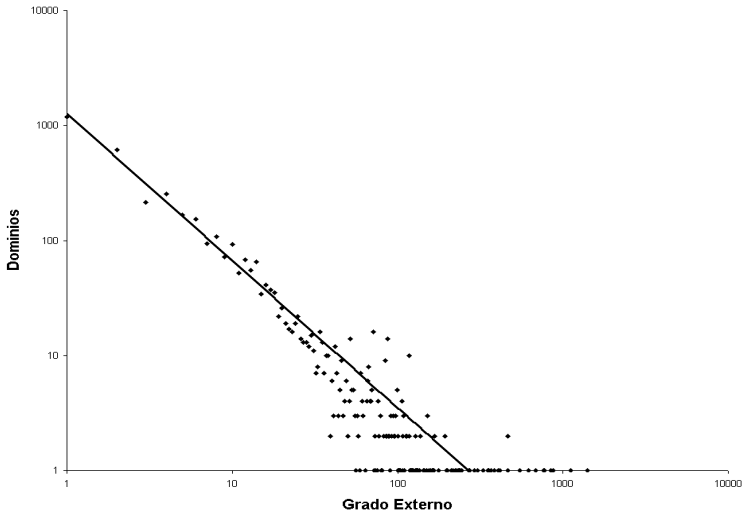

Figura 3.8: Grado Externo.

3.6.2 Largo de Caminos

Nos preguntamos si dados dos nodos D_1 y D_2, escogidos al azar en el grafo de links, existe un camino dirigido de D_1 a D_2, y si es así, cual es el número máximo y promedio de dominios que habría que visitar. Si no consideramos la dirección de los arcos, prácticamente siempre existe un camino, pues la componente conexa más grande ocupa el 94% de los dominios.

Considerando la dirección de los links, sólo un 25% de los nodos pertenece a la componente *fuertemente* conexa principal. Sólo dentro de esta componente es posible encontrar un camino entre dos nodos cualesquiera. Estudios sobre esta porción del grafo muestran que el camino promedio pasa por 3 dominios y tiene un largo máximo de 13 dominios (este es el diámetro de la componente conexa). Considerando página a página (pues por lo general las páginas en un dominio están fuertemente conectadas) y la profundidad promedio de las páginas mencionada anterioremente, el promedio y el máximo probablemente sean 2 o 3 veces mayores.

3.6.3 Macroestructura

En [B2000] se propone una forma de clasificar las páginas Web en base a su conectividad que comienza observando la presencia de una gran componente fuertemente conexa [7] en el grafo de links, que incluye aproximadamente un 25% de las páginas.

Adoptamos la siguiente nomenclatura para clasificar los dominios de acuerdo a su relación con esta componente fuertemente conexa principal:

- `MAIN`: Componente fuertemente conexa principal.
- `IN`: Dominios desde los cuales `MAIN` es alcanzable (tienen enlaces hacia `MAIN`, o tienen enlaces hacia sitios que apuntan a `MAIN`, y así sucesivamente).

[7] En el caso Chileno, la segunda componente fuertemente conexa tiene sólo 10 sitios.

- OUT: Dominios que son alcanzables *desde* MAIN
- TENTÁCULOS: Dominios que tienen relación con IN o OUT, pero no con MAIN
- ISLAS: Dominios que no tienen relación alguna con ninguna de las anteriores

Un dominio pertenece a una y sólo una de las anteriores. Se estudiaron también subconjuntos de las componentes anteriores, denominados de la siguiente forma:

- MAIN-IN: Dominios en MAIN que tienen enlaces directos desde IN
- MAIN-OUT: Dominios en MAIN que tienen enlaces directos hacia OUT
- MAIN-MAIN: Intersección de las dos anteriores
- MAIN-NORM: Dominios en MAIN que no están en MAIN-IN ni en MAIN-OUT, es decir, que no tienen enlaces directos desde o hacia sitios fuera de MAIN.
- TENTÁCULOS-IN: Sitios relacionados con IN, pero no con MAIN.
- TENTÁCULOS-OUT: Sitios relacionados con OUT, pero no con MAIN.
- TÚNEL: Dominios que permiten conectar a IN y OUT sin pasar por la componente principal, corresponden a una clase especial de TENTÁCULOS.

Las relaciones de conectividad, así como el grado interno y externo promedio de cada componente se observan en la figura 3.9 que corresponde a una representación esquemática de la Web Chilena.

Los sitios en la componente IN se identifican con sitios nuevos que poseen referencias hacia la componente principal, pero que no poseen una referencia recíproca desde aquella componente (por ejemplo: no pertenecen a ningún directorio todavía), mientras que los de la componente OUT son en su mayoría sitios corporativos que proveen de información sobre alguna organización sin poner enlaces hacia otro dominio. También pueden representar páginas más viejas, creadas antes de la mayoría de las páginas en MAIN y que no pasaron a formar parte del núcleo de páginas más conocidas.

El número de dominios de cada componente se puede observar en el gráfico 3.10 y corresponde a los datos de la tabla 3.5.

Para tener una idea de cómo se ven estas estructuras en el grafo de links, se utilizó un software de visualización de grafos llamado Graphviz, desarrollado por AT&T [8], y una muestra al azar del 10% de los dominios recolectados. En la figura 3.11 se observa abajo a la izquierda el grupo IN, al centro el grupo MAIN y arriba a la derecha el grupo OUT.

Lo más relevante que se obtiene de la figura 3.11 es la observación de que los nodos que no están en la componente principal prácticamente no se enlazan entre sí, lo que concuerda con las medidas de número de links hacia y desde los dominios en cada componente: mientras que a un dominio en OUT llegan en promedio enlaces desde 4 dominios, a sólo 1 de cada 10 de los dominios en IN llega un link. En el otro extremo los dominios que, perteneciendo a MAIN apuntan hacia afuera, tienen unos 30 enlaces de salida (a dominios distintos). Lo mismo pasa en los sitios que corresponden a los tentáculos de salida y entrada.

[8] Disponible en http://www.research.att.com/sw/tools/graphviz/

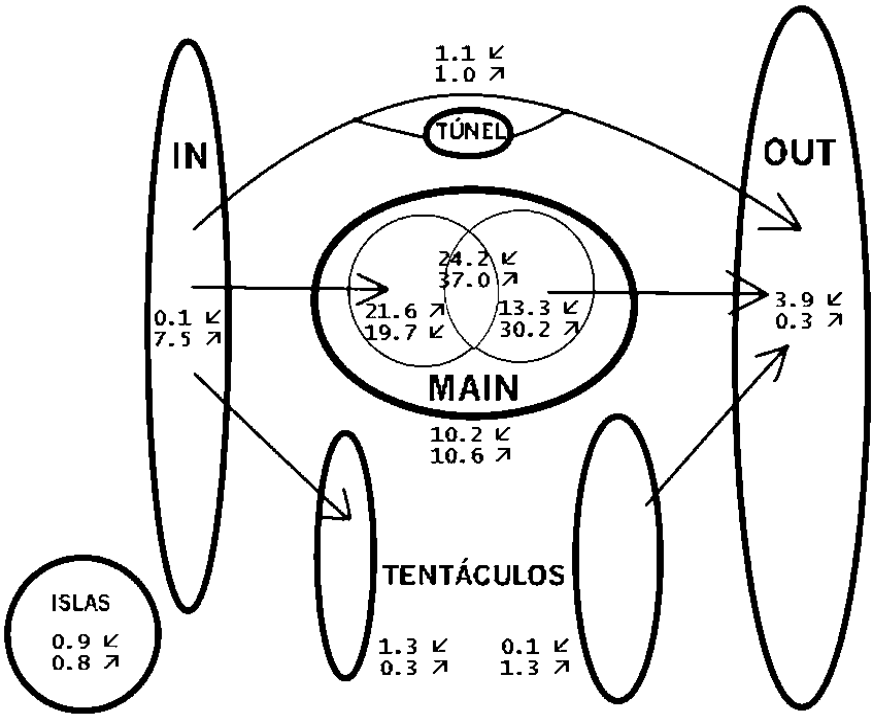

Figura 3.9: Macroestructura de hipervínculos, con grados interno y externo promedio.

3.6.4 Preferencias de los Usuarios

El objetivo de esta sección es estudiar a qué componentes pertenecen mayoritariamente los sitios escogidos por usuarios, y tener indicios de si existen o no diferencias sustanciales entre los sitios a los que un usuario accede si utiliza una máquina de búsqueda automatizada o un directorio con sitios clasificados a mano.

Se utilizaron dos muestras independientes: Editores ODP y Usuarios TodoCL. La muestra ODP contempla 3.100 sitios clasificados por editores del Open Directory Project [9] en la categoría *World/Español/Regional/Chile* que corresponden a 1.000 dominios dinstintos bajo `.cl`.

La muestra Usuarios TodoCL corresponde a la observación de 18.000 enlaces seguidos por los usuarios de entre los contenidos en las páginas de respuesta, que pertenecen a 2.500 dominios distintos. TodoCL cuenta con un sistema de redireccionamiento que permite tener un registro de los enlaces escogidos para cada consulta.

Ambas muestras pueden interpretarse como: "Los sitios de la Web Chilena que cumplen ciertos criterios de calidad" (ODP) y "Los sitios de la Web Chilena que parecen relevantes a los usuarios al ser entregados por la máquina de búsqueda" (Usuarios).

La ubicación de los sitios escogidos por editores ODP y usuarios TodoCL en las componentes antes descritas da origen a la figura 3.12. La primera observación es el hecho

[9] Disponible en `http://odp.org`. Este es un proyecto que entrega su base de datos bajo licencia *Netscape Public License*, una variante de GPL, que es utilizado como directorio de páginas en TodoCL.

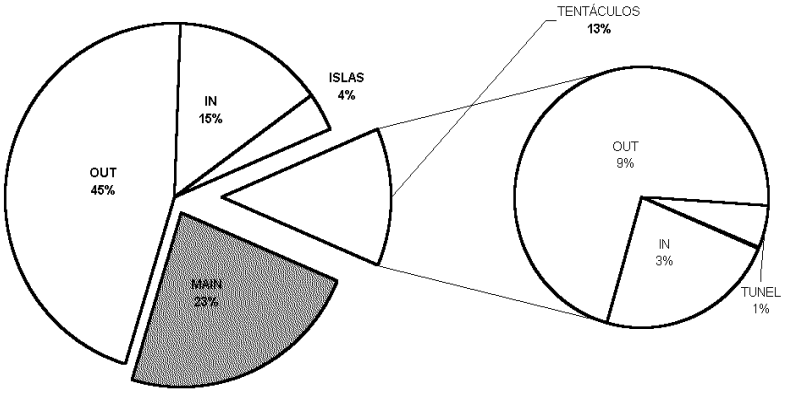

Figura 3.10: Tamaño de las componentes.

MAIN	23%
MAIN-IN	5%
MAIN-OUT	8%
MAIN-MAIN: MAIN-IN ∩ MAIN-OUT	2%
MAIN-NORM: MAIN - (MAIN-IN ∪ MAIN-OUT)	11%
OUT	45%
IN	15%
TENTÁCULOS	14%
TENTÁCULOS-IN	3%
TENTÁCULOS-OUT	9%
TÚNEL: TENTÁCULOS-IN ∩ TENTÁCULOS-OUT	1%

Tabla 3.5: Tamaño de las componentes.

de que la mayoría de los sitios escogidos por editores ODP se encuentran en `MAIN-NORM`[10], mientras que la máquina de búsqueda usada tiende a llevar a los usuarios hacia páginas que son usualmente directorios de otras páginas (`MAIN-OUT`), por el hecho de que éstas incluyen a menudo muchas palabras distintas y eso las lleva a aparecer como respuesta en varias consultas distintas.

La segunda observación es que el número de sitios en la componente `OUT` ofrecidos por TodoCL es más bien bajo, probablemente aumente con el uso de algún algoritmo de análisis de enlaces como PageRank [PBMW1998].

[10] Esto es, dominios que están en `MAIN`, pero que no tienen enlaces hacia o desde dominios fuera de `MAIN`

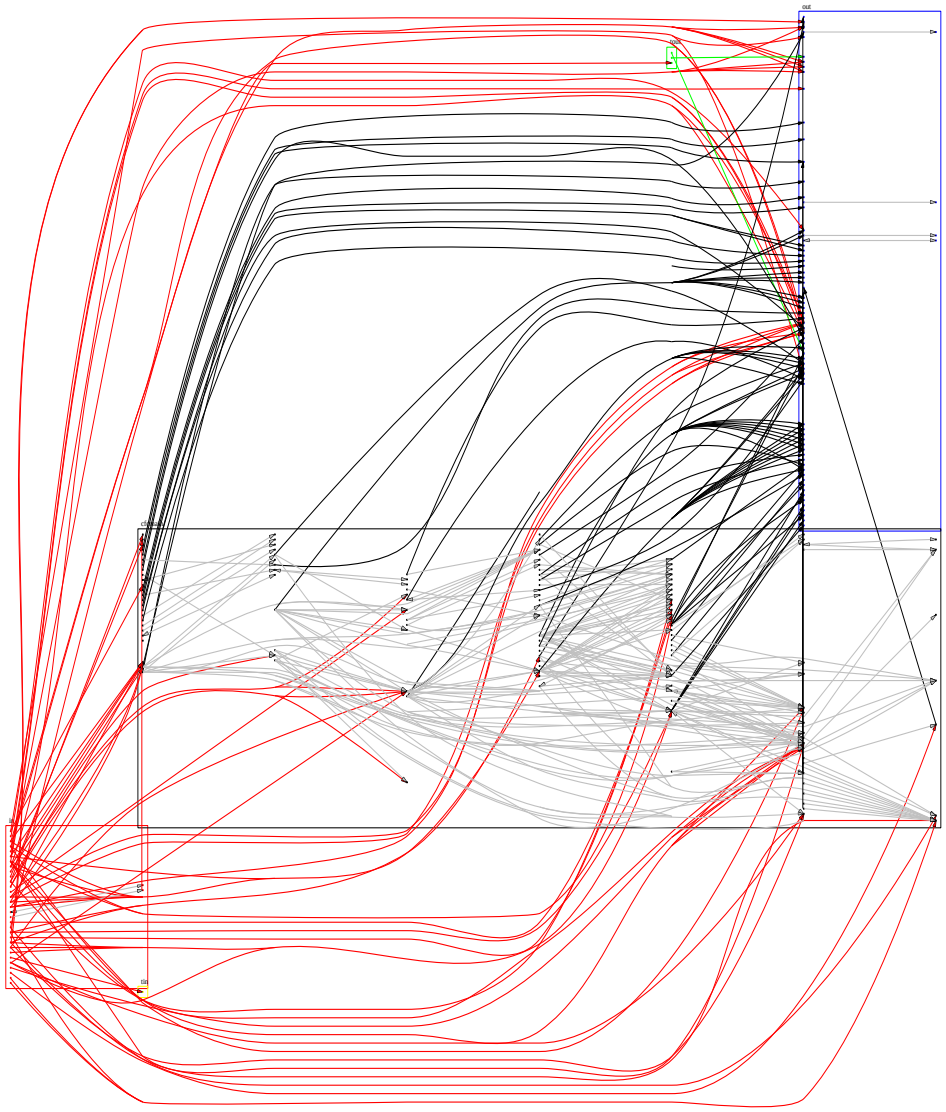

Figura 3.11: Conectividad del 10% de los dominios bajo `.cl`. Cada punto representa un dominio. La componente `IN` está abajo a la izquierda, al centro `MAIN` y arriba a la derecha `OUT`.

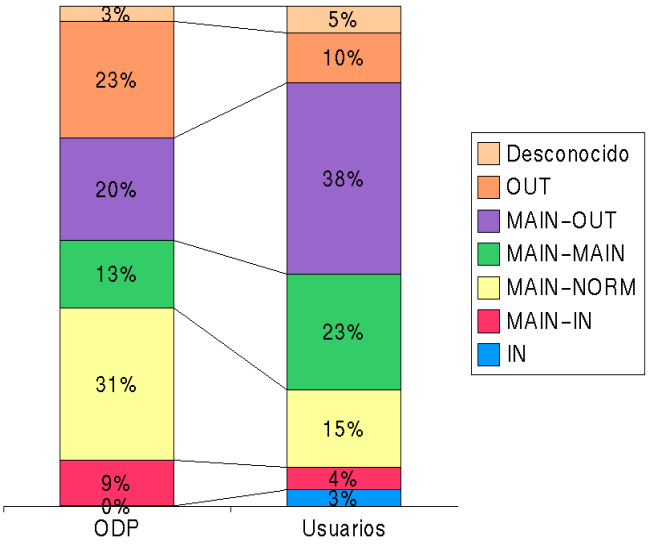

Figura 3.12: Ubicación de los sitios escogidos.

Capítulo 4

Extensiones al Buscador Web

En este capítulo se describen las extensiones más importantes realizadas. Para entenderlas es necesario conocer las distintas componentes que conforman la máquina de búsqueda, el flujo de información entre ellas y los protocolos utilizados; esto se muestra en la figura 4.1.

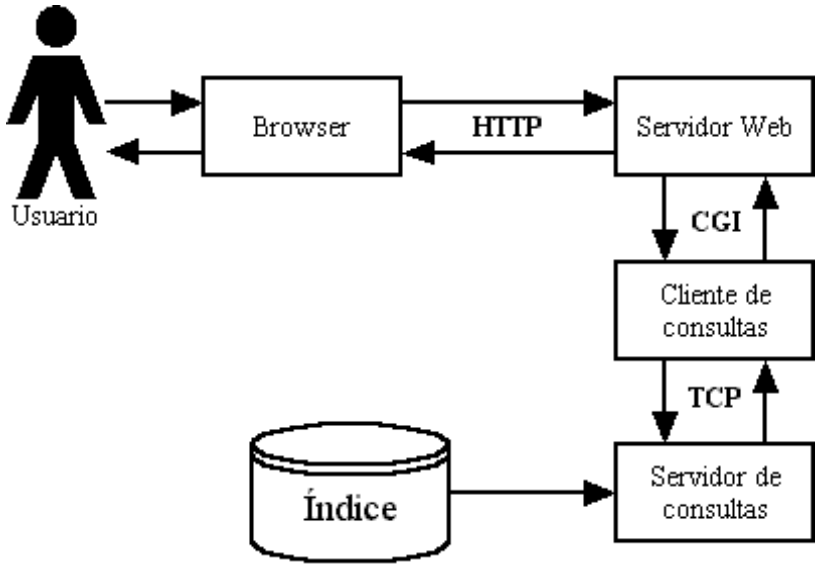

Figura 4.1: Flujo de la información en una consulta y protocolos utilizados.

Existe un servidor y un cliente de consultas; la necesidad de un servidor viene dada por que se realiza un proceso de inicialización que carga en memoria algunas estructuras relevantes para la búsqueda, y este proceso no puede ser realizado cada vez que se realiza una consulta.

Las extensiones se presentan clasificadas en tres secciones de acuerdo a la componente a la que afectan principalmente:

- Interfaz e interactividad: afectan al cliente y servidor de consultas.

- Recolector: afectan la creación del índice, correspondiendo a formas de almacenar información para realizar consultas más específicas.
- Buscador: afecta al cliente de búsqueda y a las páginas del sitio.

A su vez, para cada extensión se indica:

- Conceptos y/o Motivación: por qué es importante la extensión propuesta y cuáles son los principales algoritmos involucrados.
- Diseño: cómo se llevará a cabo, cuales son las restricciones del problema.
- Implementación: aspectos de la implementación e implantación de la solución propuesta en el diseño.

La mayoría de los cambios se realizaron en el cliente de consultas, para evitar modificaciones en el protocolo y así hacer más fácil la implantación de la extensión en las aplicaciones que componen TodoBR.

4.1 Extensiones de Interfaz

El proceso habitual de un usuario de un buscador Web es:

1. Realizar una consulta
   ... si ésta consulta retorna páginas ...
2. Examinar los resultados
   ... si alguno de los resultados, tal como es presentado por la máquina de búsqueda, parece relevante...
3. Seguir un link hasta la página Web correspondiente
   ... si la página encontrada es relevante para el usuario[1], marcarla o leerla.

Este es un proceso iterativo con mucho *backtracking*, es decir, se producen varios pasos hacia atrás, porque el usuario puede volver a examinar la lista de resultados, refinar la consulta realizada o hacer una nueva consulta. Para un usuario promedio, aproximadamente un 30% de las veces se elije un documento de la lista de resultados; una vez revisado ese documento, un 50% pasará a navegar por los links de esa página y un 30% volverá a la máquina de búsqueda para formular una nueva consulta. El 20% de las veces buscará otra máquina de búsqueda o un nuevo punto de partida.

Este comportamiento es estudiado en detalle en [HS2000] y depende principalmente de la experiencia del usuario tanto en el uso de Internet como en el área en la cual reside su necesidad de información. En la figura 4.2 se reproduce un diagrama del comportamiento de los usuarios de acuerdo al estudio citado.

El tiempo requerido para examinar los resultados y llegar a páginas Web relevantes tiene dos componentes:

[1] Esto es, que su contenido cubra la necesidad de información del usuario, o contenga links a páginas que sí lo hagan

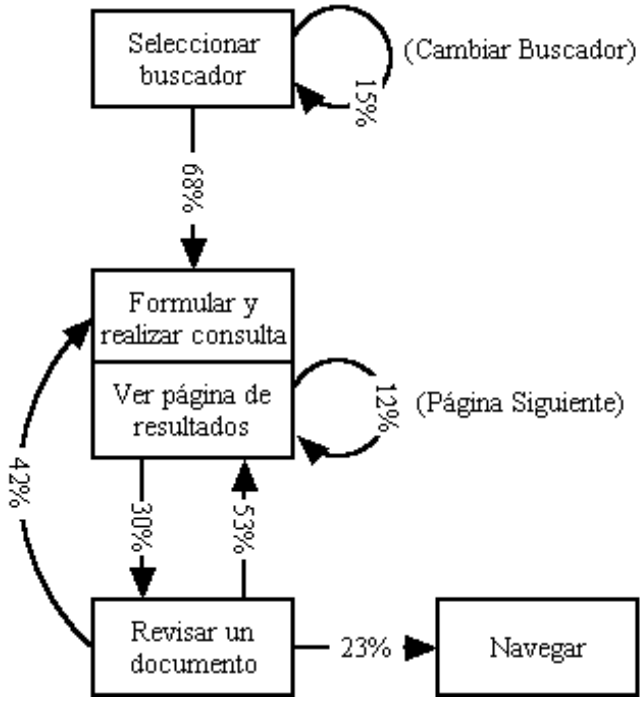

Figura 4.2: Comportamiento de los usuarios.

- **Intra-Sistema**: el tiempo ocupado en formular la consulta, esperar los resultados y examinarlos.
- **Extra-Sistema**: el tiempo ocupado en revisar algunos documentos hasta encontrar alguno relevante.

Estos tiempos deben balancearse adecuadamente, en términos de cuánta información se provee sobre cada candidato retornado por la máquina de búsqueda, la cual no puede ser ni muy extensa (porque aumenta demasiado el tiempo Intra-Sistema utilizado en leer las descripciones) ni demasiado suscinta (porque el usuario perderá tiempo Extra-Sistema revisando documentos irrelevantes, por no poder juzgar a priori si satisfacen o no su necesidad de información).

Las extensiones de interfaz propuestas tienen relación con: disminuir la sobrecarga de memoria en el usuario para realizar este backtracking (historial de consultas), orientarlo para obtener resultados más relevantes (consejos de búsqueda y revisión de ortografía) y disminuir el tiempo requerido para examinar los resultados (opciones de presentación, y localización de búsqueda).

4.1.1 Opciones de Presentación

Conceptos y Motivación

Existe un cierto nivel de consenso [IA1998, capítulo 6] (también en [UW]) respecto a que una buena interfaz debe proveer de dos formas de comunicarse con la aplicación: una simple y una avanzada. La forma simple debe ser la opción por omisión, y contener sólo un conjunto mínimo de controles sobre la búsqueda, mientras que la forma avanzada debe contener opciones para usuarios que tienen un mayor manejo del sistema. Esta mayor flexibilidad no tiene por qué estar sólo en la forma de realizar la consulta, sino también en la forma de desplegar la lista de resultados; por ejemplo, un usuario avanzado podría querer ver la fecha de última actualización de los documentos, mientras que para uno novicio esta información resultaría inútil e incluso afectaría negativamente la claridad de los resultados.

Además, dentro de los planes a futuro de TodoCL, se considera la posibilidad de que un un administrador de sitio Web pueda realizar *outsourcing* del servicio de búsquedas sobre su sitio, siendo estas manejadas por TodoCL por medio de una consulta restringida a su dominio.

Estos antecedentes indican que hay variedad en la forma en que deben presentarse los resultados; se elige una parametrización completa de los resultados frente a proveer de dos o tres formas preestablecidas por ser el primer mecanismo más flexible y extensible.

Diseño

Cada ítem desplegado en la pantalla de resultados será susceptible de ser encendido o apagado, para conseguirlo lo más apropiado es una máscara de bits, en que cada bit indica uno de tales aspectos, ej.: el Bit 0 contendrá el título, el Bit 1 la URL, etc. Se consideran 4 bytes (32 bits) como máscara completa, para dar cabida a los siguientes ítems:

```
// RO: Result Options

#define RO_NONE          0
#define RO_TITLE         1
#define RO_URL           2
#define RO_TIMESTAMP     4
#define RO_SCORE         8
#define RO_SIZE         16
#define RO_SUMMARY      32
#define RO_LOCATE_MATCH 64
#define RO_FREQUENCIES 128
#define RO_TIP         256
#define RO_FEEDBACK_LINK 512
#define RO_CACHE_LINK   1024
#define RO_DIRECTORY    2048
#define RO_DOMAIN_LINK  4096
#define RO_MEDIA        8192

#define RO_DEFAULTVAL (RO_TITLE|RO_URL|RO_FEEDBACK_LINK|
 RO_SUMMARY|RO_LOCATE_MATCH|RO_TIP|RO_DIRECTORY|RO_MEDIA)
```

Un perfil de resultados será un valor de esta máscara de bits; por ejemplo, lo más simple es el valor 35 (título + url + resumen).

Implementación

Se debe establecer coordinación entre las interfaces de búsqueda (en HTML) y el cliente de búsquedas. En la interfaz de búsqueda se incluyen los valores de la máscara considerados estándar y en el cliente opciones de despliegue o no despliegue condicionadas a que un bit esté presente o no.

La implementación requiere de bastante cuidado puesto que el cliente de búsquedas genera su salida en HTML y por lo tanto incluir o no incluir un elemento tiene efectos colaterales (por ejemplo, en ocasiones es necesario iniciar una tabla, o una lista descriptiva, y más adelante cerrarla dependiendo de combinaciones de bits más que de bits individuales); aparte de eso la implementación no presentó mayores problemas.

Esta fue una de las primeras extensiones en realizarse y simplificó el trabajo de implementar el resto, puesto que permitió un marco apropiado para agregar información sobre las páginas conforme se creaban más extensiones.

4.1.2 Consejos de Búsqueda

Conceptos

Una sesión típica de búsqueda consiste en que el usuario ingresa inicialmente sólo unos pocos términos (muchas veces sólo uno) para determinar qué tipo de respuestas entrega el motor de búsqueda y luego busca entre los resultados (por ejemplo: si está buscando información turística sobre Osorno, comenzará buscando sólo "osorno" e incluso sólo "turismo"). Esto puede deberse a que la mayoría de quienes utilizan estos sistemas no dimensionan la vastedad de la red y la gran cantidad de páginas existentes. Al utilizar un solo término de búsqueda el número de resultados suele ser inmanejable, del orden de los cientos o miles de páginas Web.

Otro escenario se produce por dificultades en la comprensión del significado de las opciones de búsqueda: si alguien busca información sobre "Como agua para chocolate" con método: alguna de las palabras, probablemente no encuentre información sobre la novela que lleva este nombre, sino principalmente páginas que tengan que ver con "chocolate" que es la palabra menos frecuente y que por lo tanto tiene más peso como término de búsqueda. Además sabemos que cuando se busca con "alguna de las palabras" o "todas las palabras" es poco recomendable utilizar conectivos, mientras que estos son necesarios en la mayoría de las búsquedas de "frase exacta". En este caso, el sistema debería comunicarle al usuario que debe eliminar los conectivos de su búsqueda o buscar por la frase exacta.

Aparte de los dos errores comunes antes descritos, existen otras circunstancias en que se puede orientar al usuario en términos de cambios en su estrategia de búsqueda. El sistema debería ser capaz de comunicarle correctamente al usuario cómo restringir su búsqueda y obtener finalmente sólo unas pocas páginas muy relevantes para su necesidad.

Diseño

Esta extensión implementa el subsistema que sugiere al usuario cambios en su estrategia de búsqueda con el fin de permitirle mejores resultados.

En un esquema ideal, sería más necesario o menos necesario hacer una sugerencia al usuario dependiendo de si su consulta fue bien o mal formulada. En el entendido de que tal cosa es difícil de evaluar por la máquina de búsqueda (aunque podría ser en el futuro tema de análisis), la heurística utilizada responde entre otras cosas a que si una consulta retornó demasiados resultados, o ninguno, entonces habría que plantearse una nueva consulta.

Otras variables consideradas además del tamaño de la lista de resultados son:

- Presencia de términos demasiado frecuentes, excepto cuando se busca una frase completa.
- Presencia de términos que no se encuentran en el vocabulario del índice (se presume en este caso que hay un error ortográfico).
- Método de búsqueda elegido.
- Cantidad de términos de búsqueda empleados (demasiados términos y pocos resultados puede significar una búsqueda mal enfocada, y que sería preferente una aproximación incremental).

Respecto a la forma de desplegar los consejos de búsqueda, la interactividad de un sitio en la red puede entenderse por la capacidad de establecer un diálogo entre el sitio y el usuario [JF1992]. Una de las formas de hacer sentir al usuario más cómodo es aproximar este diálogo en la medida de lo posible a un diálogo entre humanos; una de las dificultades de esta aproximación es que las respuestas de un computador por su predicictibilidad tienden a ser ignoradas e incluso a cansar al usuario; es por esto que las respuestas se escogen aleatoriamente de entre un conjunto de oraciones que tienen el mismo significado.

Implementación

Dada la cantidad de casos posibles, una forma natural de describir el diseño será un diagrama de flujo. Una versión *simplificada* del procedimiento utilizado se muestra en la figura 4.3; se destacan en ese diagrama la ubicación de los módulos de **Expansión de Consultas** y **Verificación de Ortografía** que son invocados selectivamente sólo en los casos indicados.

En la figura 4.3 para simplificar el diagrama, entre otras cosas, se omitió el hecho de que varios consejos pueden ser generados para una misma consulta. La clave de los métodos de búsqueda es: `AND`: Todas las palabras, `OR`: Alguna de las palabras y `FRASE`: Frase exacta.

Se ilustra el funcionamiento de esta extensión en la figura 4.4.

4.1.3 Localización de Búsqueda

Conceptos

El principal motivo por el cual el espacio requerido para almacenar todo el texto de los cerca de 25000 sitios de la Web Chilena es relativamente pequeño, de alrededor de 3Gb, es porque una página típica contiene en promedio sólo 12Kb de texto, tal como se estableció en la sección 3.4.1.

Teniendo estas copias locales de los documentos, parece lógico mostrar al usuario que realiza una consulta un extracto del documento retornado por la máquina de búsqueda, entendido en principio como el párrafo donde aparecen sus términos de búsqueda.

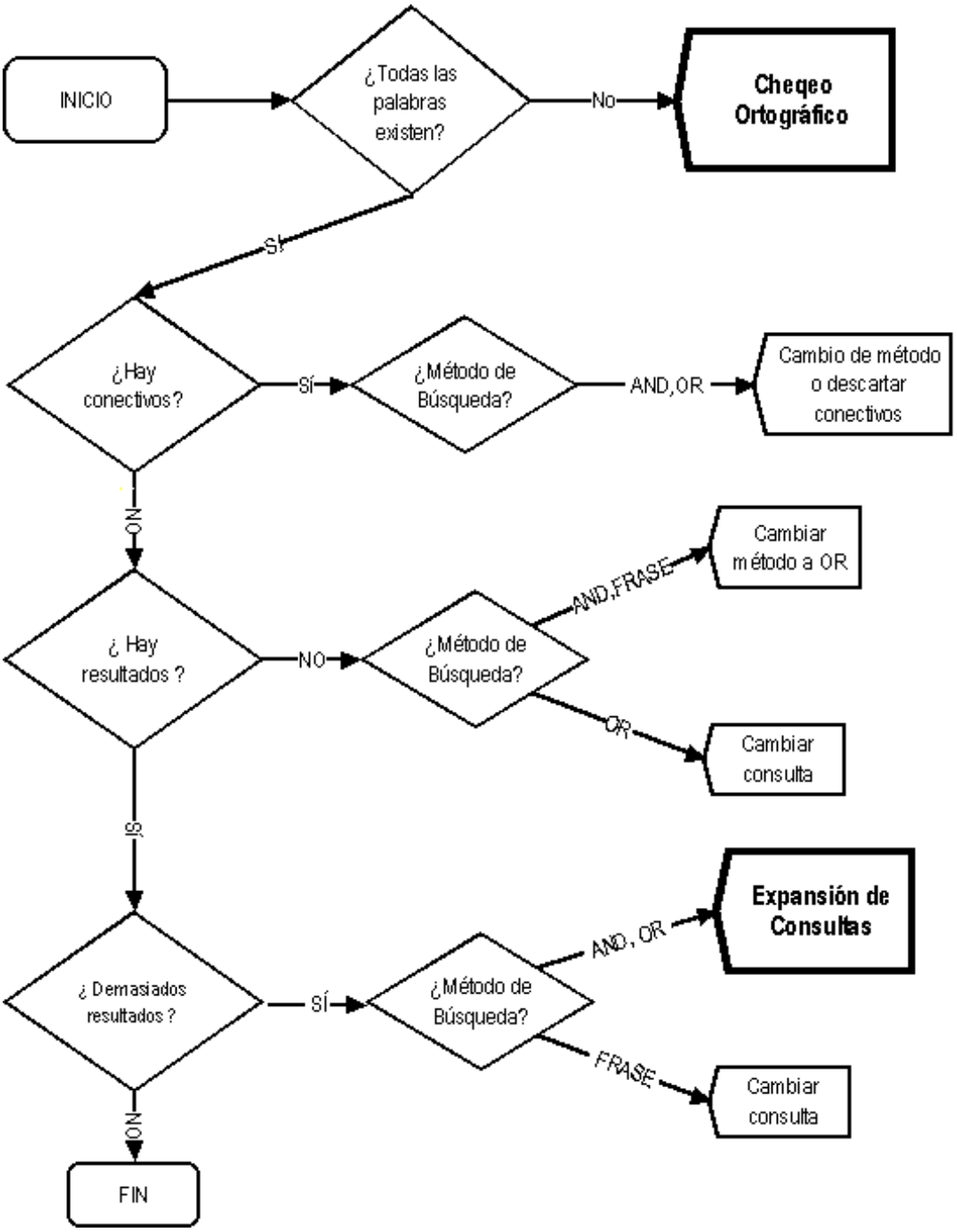

Figura 4.3: Esquema de consejos al usuario.

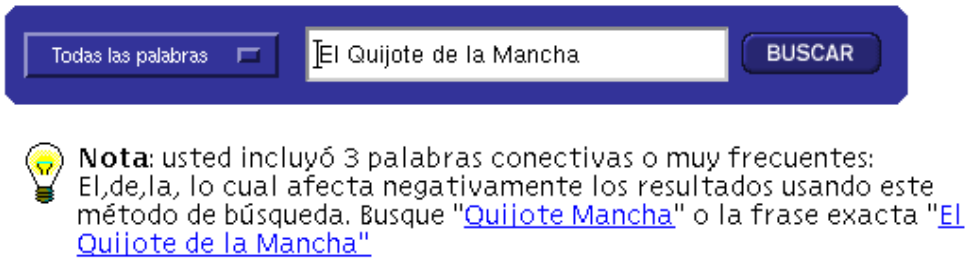

Figura 4.4: Ejemplo de consejo de búsqueda.

Diseño

Se consideró que la porción de la página que aparecía en la lista de documentos debía ser de alrededor de cinco líneas, que a 80 caracteres por línea representa unos 400 bytes de texto. Nótese que para que los fragmentos o pasajes de texto retornados provean de información que tenga sentido, debe proveerse de un contexto apropiado para juzgar el rol de los términos de búsqueda dentro del documento; particularmente si la página Web se refiere en extenso a ellos o sólo los toca tangencialmente. Entender el significado de la palabra sólo requiere de la oración en que está inserta (en la mayoría de las ocasiones), pero visualizar si su rol es accesorio o principal requiere de mucho más texto, incluso varios párrafos.

Un dato a tener en cuenta es que TodoCL almacena en un archivo separado los primeros 200 bytes de cada archivo, considerándolo un "resumen" del texto. En el futuro pueden implementarse formas más elaboradas de generar este resumen que consideren más elementos, por ejemplo, la estructura del documento.

Se resolvió de la siguiente manera: devolver siempre el primer pasaje del texto o resumen (primeros 200 bytes), junto con el primer pasaje del texto en que aparezca algún término de búsqueda (200 bytes, centrados en torno a ese término); en ambos pasajes se destaca cada aparición de alguno de los términos de búsqueda.

Como el proceso de ir al archivo donde están los textos es más bien lento, se optó por no buscar un párrafo en el texto si algún término de búsqueda aparecía en el resumen.

Implementación

Para la implementación, se utilizó un algoritmo de búsqueda multipatró n, en el que un trie con los patrones se construye y se desliza sobre el texto. El diseño e implementación de este algoritmo de búsqueda corresponden a Gonzalo Navarro.

El localizador de búsqueda debe accesar a disco, por lo que se tomaron varias provisiones para evitar demorar excesivamente la consulta; por ejemplo, no se busca nada más allá de los primeros 16Kb de texto del archivo, y si por algún motivo el acceso a disco es muy lento en ese momento (i.e.: toma más de 2 segundos) el proceso es abortado, retornándose en ese caso sólo los primeros 200 bytes del archivo, contengan o no los términos de búsqueda.

Se muestra un ejemplo de la búsqueda "hielo" y cómo se localiza y marca el pasaje del texto en que ocurre el calce.

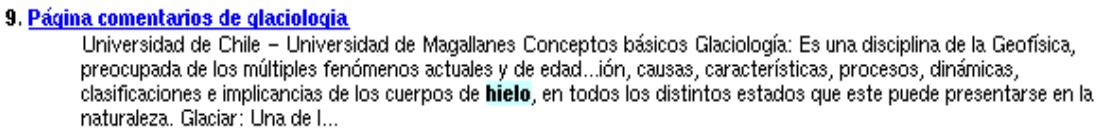

Figura 4.5: Ejemplo de localización de búsqueda.

4.1.4 Expansión de Consultas

Conceptos

En directa relación con la extensión de "Consejos de Búsqueda" se intenta orientar al usuario a utilizar la máquina de búsqueda de manera de obtener sólo unos pocos resultados muy relevantes, dado que un porcentaje menor de los usuarios va más allá de las primeras 2 páginas de resultados. Si consideramos que la lista de documentos retornada por el buscador es de más de 100 documentos, es altamente probable que lo mejor sea restringir la búsqueda más que llevar al usuario a revisar esta lista en detalle.

La idea es buscar la manera de obtener una sub-lista de resultados. Para ello, es necesario realizar una consulta más específica, utilizando un mayor número de términos de búsqueda.

Desde el punto de vista del usuario, la simple indicación "utilice más palabras" no parece suficiente y sería mejor poder recibir sugerencias sobre cuáles podrían ser esas palabras. Esto se conoce como "expansión de consultas".

Los mejores términos para expansión de consultas son aquellos que permiten particionar el espacio de resultados de la manera más natural posible. Por ejemplo, si el usuario busca "estrella" sería ideal poder desplegar algo como:

```
Usted busca ''estrella'' relacionado con:

1. Planetas, Galaxias, Universo, Satélites ...
2. Cine, Televisión, Música, Espectáculos ...
3. Mar, Zoología, Equinodermos, ...
```

Existen dos grandes tipos de método para conseguir una expansión de consultas: análisis de contexto global y análisis de contexto local[BYRN1999]. El análisis de contexto global implica la construcción de una red estática de términos relacionados (en que los nodos son las palabras y las aristas las relaciones) en base a la colección completa, mientras que el análisis de contexto local consiste en la construcción de una red dinámica basada en los documentos retornados por la consulta del usuario.

En TodoCL se utilizó una estrategia de análisis de contexto local. Dentro de esta clase existen varias heurísticas para encontrar los nodos y aristas de la red de relaciones: los nodos pueden ser todas las palabras o un subconjunto más pequeño y las aristas pueden ser obtenidas por proximidad dentro de los textos estudiados o en base a factores estadísticos de co-ocurrencia dentro de los documentos.

Diseño

Las palabras que se considerarán útiles para expansión de consulta serán todas las que no sean funcionales (preposiciones, artículos, pronombres), no estén dentro de los sustantivos más frecuentes en la Web (Internet, Web, página, sitio, información, HTML) ni en la consulta.

Se considerará que una palabra esta más estrechamente relacionada con otra si aparece cerca de ella en varios documentos.

Esta cercanía se define como su distancia en palabras del término de consulta (por ejemplo, en la frase "El caballo corría por la pradera" las palabras "pradera" y "caballo" están a distancia 4), considerando que esta distancia es infinita pasado un límite de puntuación. Esto último es una heurística para emular en algún sentido la idea de frase como unidad de información sin la necesidad de entrar en un análisis sintáctico más complejo.

Se construye dinámicamente una red de términos relacionados con el siguiente algoritmo (que es una variación de la técnica de grupos métricos - *Metric Clusters*).

1. Construir una lista con los términos de la consulta Q y con los r documentos de mayor ranking D.
2. Para cada término en la lista Q_i y cada documento D_j:
3. Localizar las ocurrencias de Q_i el término en el documento D_j.
4. Construir una lista de términos en la misma "frase" de la ocurrencia de Q_i en D_j, descartar las palabras definidas como no útiles más arriba.
5. Insertar las palabras que quedan en una lista de candidatos, sumándoles a su puntaje $(1/d)^k$ al puntaje que tuvieran antes en la lista de candidatos, en que k es un parámetro de la heurística y d la "cercanía" entre dicha palabra y Q_i
6. Retornar las p palabras de mayor puntaje en la lista de candidatos.

Los parámetros p, k y r escogidos fueron: (p) 5 términos para expansión, (k,d) el puntaje es la inversa de la distancia al cuadrado y (r) se estudian los primeros 60 documentos.

Existe una distancia máxima en la cual se detiene el análisis incluso si no se han encontrado símbolos de puntuación. Esto define un puntaje máximo o inicial que se va dividiendo por la constante de caída exponencial K del puntaje al encontrarse una palabra que no es stopword. Las fuentes de las listas de stopwords son [ALLIENDE] y [OVID].

Se asigna un puntaje cuadráticamente decreciente en la distancia con el término original a cada palabra que aparece en el fragmento. Una stopword o palabra demasiado frecuente no se considera para el cálculo de distancia. Por ejemplo, si el puntaje máximo es 1024, en la frase "el oxígeno es un elemento vital" si el término de consulta es "oxígeno" los puntajes son asignados como se indica en la figura 4.6.

Implementación

La implementación hace uso del sistema ya desarrollado de localización de búsquedas, del sistema de consejos de búsqueda y de un módulo de identificación léxica simple.

El módulo de identificación léxica simple contiene un diccionario de palabras conocidas y una etiqueta léxica; esto permite descartar de la expansión de consultas términos demasiado frecuentes y stopwords.

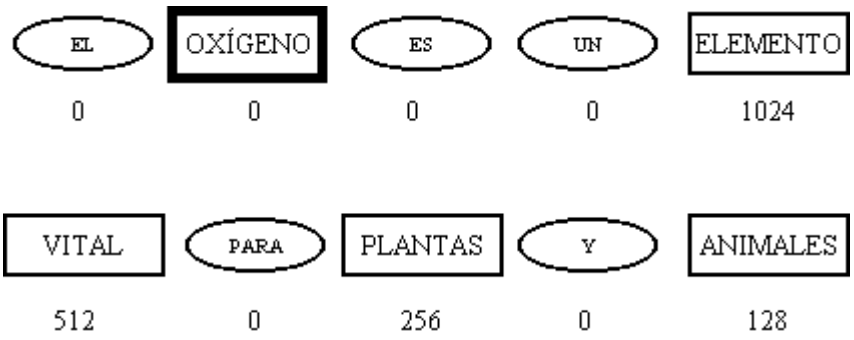

Figura 4.6: Expansión de consultas.

La expansión de consultas se relaciona con la localización de búsqueda porque este módulo, antes de retornar el texto con las marcas, pasa el texto original con la posición de los términos de búsqueda a otra función. Esta última función para cada término de consulta revisa hacia atrás y hacia delante en el texto las palabras cercanas antes de un símbolo de puntuación.

Un ejemplo de cómo se ve el resultado frente a la consulta "banco" se observa en la figura 4.7.

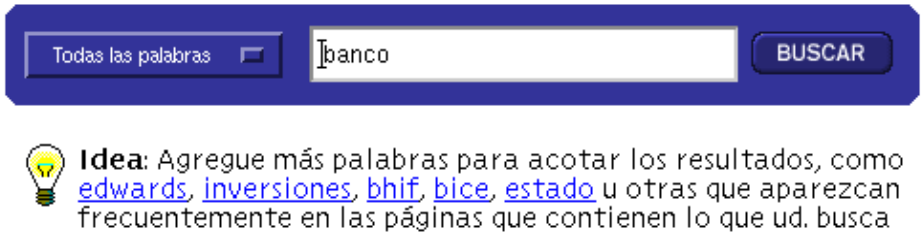

Figura 4.7: Ejemplo de expansión de consulta.

Para ilustrar el funcionamiento en general de la expansión de consultas, a continuación se muestran más ejemplos de las sugerencias que entrega el sistema frente a distintas consultas:

- **puerto**: puertos, arica, iquique, antofagasta, naviera
- **medicina**: facultad, escuela, enfermería
- **hubble**: ley, secuencia, video, telescopio, space
- **universidad**: frontera, católica, prat, ibáñez, austral
- **compras**: podrás, informes, detalle, cliente, pedidos

Se observa que por lo general los términos sugeridos guardan relación con la consulta inicial. Aparece manifiesto en el ejemplo de la palabra "compras" que el sistema no elimina

los verbos que no aparecen en el diccionario (que son conjugaciones de verbos auxiliares como ser y haber, principalmente).

4.1.5 Historial de Consultas

Conceptos

Una característica deseable de una interfaz es mantener a un mínimo posible la sobrecarga de memoria de corto plazo de quien utiliza el sistema. Esto ha sido evaluado, en particular para las máquinas de búsqueda más destacadas, en [G1999]. En TodoCL, el objetivo es mantener al usuario informado de:

- Los términos de consulta utilizados.
- Las principales opciones de búsqueda elegidas.
- La página de resultados en la cual está y cuántos resultados quedan por revisar.
- Las últimas consultas realizadas en esta sesión [2]. Estas consultas constituyen lo que llamaremos *historial de consultas.*

Diseño

Para mantener la presentación del historial lo más simple posible, sólo se incluyen los términos de búsqueda en él, esto en principio no excluye la posibilidad de que se almacenen también las opciones de búsqueda, pero sólo se desplieguen los términos. Se descartó esta posibilidad porque parece confuso que botones o links que se ven similares tengan acciones distintas, y se estimo que tendería a aumentar la necesidad de memoria de corto plazo más que a acortarla (pues el usuario tendría que recordar que tal o cual frase la buscó con tal o cual método). No se consideró desplegar el número de resultados que se obtuvo en cada consulta por un argumento similar[3].

El historial se limita a lo que el usuario puede examinar en un golpe de vista sin necesidad de buscar demasiado: las 5 últimas consultas. Estas son persistentes entre sesiones para ayudar al usuario a recordar el tipo de consultas que se pueden realizar y que él lo asocie con el tipo de resultados que se obtienen.

Por motivos de privacidad se ofrece al usuario la posibilidad de que borre el historial de búsquedas o lo desactive en forma muy simple; esto porque para algunos usuarios podía parecer una intromisión en su privacidad el hecho de estar almacenándose sus consultas; por no entender que tal almacenamiento ocurre sólo en su computador.

Al ser poco usual en máquinas de búsqueda y por tanto revestir un cierto carácter de novedad se incluyó en la portada del "Laboratorio de TodoCL" para incentivar a más usuarios a utilizarla.

[2] Una sesión se define (en diseño Web) como el conjunto de acciones que un usuario realiza en el programa navegador de su preferencia desde que lo inicia en el computador hasta que lo finaliza; usualmente en aquellos que tienen conexión telefónica corresponde al intervalo de tiempo que permanecen conectados a internet sin desconectarse.

[3] Sería interesante experimentar a futuro con mostrar el número de resultados obtenidos, siempre que las posibilidades de formato físico de HTML o D-HTML permitieran hacerlo de una manera atractiva y clara para todos los usuarios

Implementación

La implementación del historial está basada en *cookies*, que son espacios para almacenamiento de información del servidor en el cliente, de 1Kb por servidor. Éstas cookies son administradas en TodoCL por funciones en Javascript. El servidor provee de algunos scripts al browser para que sea éste quien mantenga el historial.

Las últimas consultas se almacenan en un sólo string. Esto impone un límite de 1KB (largo máximo de una cookie) sobre el total de los strings almacenados, pero al considerar que sólo se almacenan 5 consultas y que el formulario permite el ingreso de hasta 120 caracteres por consulta, el espacio es suficiente. Este string expira en un mes (el historial se borra del PC del usuario al cabo de ese tiempo).

Al desplegarse el formulario, los programas en javascript generan el texto y los links apropiados para desplegar el historial. Al hacer click sobre alguna de las consultas anteriores, esta se copia en la caja para ingresar los términos de búsqueda, pero el formulario no es enviado, para permitir al usuario reutilizar una consulta ya realizada, con nuevas opciones.

En el momento de realizarse una consulta, se invocan las funciones que almacenan la consulta en la cookie y se borra la consulta más antigua en caso de excederse el límite de cinco consultas.

Comentarios

Si bien esta extensión resultaba interesante en principio, los browsers están comenzando a incorporar sistemas que recuerdan y despliegan lo que se ha ingresado en los formularios que eventualmente aparecen en páginas Web. Este sistema está pensado para ahorrar al usuario tiempo de tipeo de su nombre, dirección, datos personales, números de tarjeta de crédito, etc.

Tales opciones están consideradas en Netscape 6 y Explorer 5, y por lo tanto alcanzarán a una gran mayoría de los usuarios; es por esto que mantener el historial como estaba planteado en esta sección se implementó pero no se recomienda utilizarlo.

4.1.6 Verificación Ortográfica

Conceptos

Cuando el usuario ingresa una palabra que no existe en el vocabulario de la colección, existen dos posibilidades: a) la palabra existe, pero ningún documento la contiene b) la palabra no existe y fue mal ingresada por un error de ortografía (probablemente causado por un problema de tipeo).

Un corrector ortográfico automático funciona buscando las palabras *más parecidas* a la que ingresó el usuario que estén presentes en un vocabulario, en este caso, el vocabulario de la colección[4]. Esta noción de similaridad se puede cuantificar en términos de distancia, y para esto existen dos clases de medida: las medidas de tipo Hamming y de tipo Levenshtein.

La distancia de tipo Hamming entre dos palabras será el número de caracteres distintos entre ambas, y es una extensión natural de la distancia de Hamming entre vectores

[4] Se observó que el vocabulario de la colección dista mucho de ser léxicamente correcto, pues, por ejemplo, *palabras como "sujerencia" y "elejir" aparecen en decenas de documentos*, sin embargo, la idea es que el corrector ortográfico sugiera palabras que conduzcan a obtener resultados

en $\{0,1\}^N$. La distancia de Levenshtein o distancia de edición considera variables antropométricas como el hecho de que escribir mal una palabra puede significar no sólo cambiar una letra por otra, sino también intercambiar dos letras u omitir una letra. Esta última distancia es la más utilizada en recuperación de la información[BYRN1999], junto con la distancia LCS (*Longest Common Subsequence*) que es una variante en que sólo se permiten eliminaciones.

Diseño e Implementación

El diseño y la implementación fueron llevadas a cabo por Gonzalo Navarro. La integración al código principal fue haciendo una llamada desde el módulo de consejos de búsqueda, tal como se mostró el diagrama de flujo 4.3. Como ahí se indica, la verificación ortográfica es invocada sobre aquellas palabras que no aparecen en ningún documento.

4.2 Extensiones del Recolector

Las extensiones del recolector tienen una importante restricción: el proceso de recolección es altamente intensivo en uso de memoria y disco, pero no particularmente de tiempo del procesador, debido a que la mayor parte del tiempo cada instancia del recolector está esperando que un sitio Web conteste y envíe la página solicitada. Todas las extensiones realizadas contemplaron este hecho.

4.2.1 Descarte de Binarios

Motivación

Se observó al generar el índice después de la primera recolección que el vocabulario resultó ser mucho mayor que el valor esperado, incluso superando la memoria disponible para la tabla de Hashing con cada palabra. Se detectó que gran parte de las nuevas palabras que iban apareciendo al recorrer la colección eran introducidas por unos pocos archivos.

Estos archivos resultaron ser binarios de 2-5Mb, codificados en 7 bits. Las codificaciones de 7 bits son usadas para transferir archivos binarios por medios que sólo aceptan texto, por ejemplo, dentro del cuerpo de un mensaje de e-mail o un *post* en un newsgroup. Los formatos más comunes son `BinHex` para Macintosh y `UUencode` en Unix; ambos formatos tienen encabezados pre-establecidos [5], sin embargo, si uno de estos archivos tiene un encabezado distinto, o el archivo está dentro de un documento de texto, el sistema no lo reconocerá.

El vocabulario debe residir en memoria para permitir búsquedas más rápidas, y por lo tanto debe mantenerse con un tamaño razonable. Esto lleva a buscar alguna heurística para reconocer porciones de un documento que estén en binario para poder descartarlas; esta heurística no debe confiar ni en la extensión del archivo, ni en su encabezado, ni en la presencia de caracteres de ASCII altos, para determinar que porciones descartar, sino sólo en características locales de cada bloque de texto.

[5] BinHex: `(This is a BinHex M.n file)`, UUEncode: `begin permisos archivo`.

Diseño

La idea, sugerida por Gonzalo Navarro, consiste en determinar un estadístico sobre las frecuencias por cáracter dentro de cada bloque estudiado. Este estadístico debe poder discriminar entre un texto en que cada carácter figure con una probabilidad uniforme y un texto en que ciertos caracteres tengan una frecuencia de aparición marcadamente mayor que la de otros.

En el idioma español, hay letras muy frecuentes (vocales y algunas consonantes como por ejemplo "c" "s" y "m") y letras muy infrecuentes (ej.: "w" "z" y "ñ"). Una codificación eficiente tendrá un máximo de información por carácter y este máximo se alcanza si cada carácter tiene la misma frecuencia, por lo que la distribución de las letras dentro del archivo codificado será similar a una distribución uniforme; los experimentos empíricos que se mencionan más abajo confirman que las codificaciones utilizadas comunmente se acercan bastante a esta descripción.

Implementación

El estadístico elegido es B tal como está definido en 4.1, sobre cada bloque del documento:

$$B = \sum_{c \in \ alfabeto} \frac{1}{f_c^2} \tag{4.1}$$

En que f_c es la frecuencia del carácter c en el bloque. Si la distribución es uniforme y considerando que el tamaño del alfabeto es 128 (2^7), se tiene $B \approx 0.006$. Para textos normales[6] se determinó $B \approx 0.065$; estos dos valores permiten discriminar fácilmente entre dos bloques.

El tamaño escogido para el bloque fue 1024 aunque cualquier otro valor que divida al tamaño de página de disco en el sistema de archivos resulta apropiado.

Se condujeron varias pruebas que llevaron a los siguientes valores de corte (empíricos): $B_{min} = 0.03$, $B_{max} = 0.11$; un bloque en que el estadístico B esté fuera de esos valores tiene una baja probabilidad de ser un documento de texto y una alta probabilidad de ser binario (la existencia de B_{max} proviene de que un texto formado sólo por unos pocos caracteres también se considera anómalo).

Prueba

En el gráfico 4.8 se observan los tamaños del vocabulario conforme se avanza en la creación del índice por la colección, comparándose un caso base con otro con el mismo caso con descarte manual de los archivos con más palabras y con descarte automático por bloques como se describió más arriba.

Tal como se observa, el descarte manual produce los mejores resultados, pero el descarte automático también es capaz de controlar aumentos excesivos en el vocabulario. Se estudió un modelo para el tamaño del vocabulario en la sección 3.3.2 que considera el vocabulario obtenido al aplicar descarte automático por bloques.

[6] Utilizando textos escogidos de 1-2Mb.

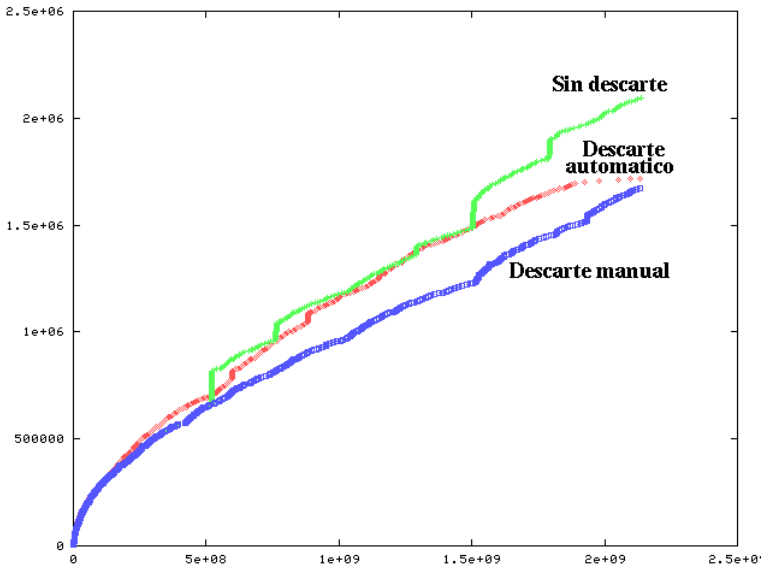

Figura 4.8: Tamaño del vocabulario, con y sin descarte de binarios.

4.2.2 Reconocimiento de Multimedios

Conceptos y Motivación

El paso de textos a hipertextos (principal causa de la virtual extinción de los sistemas gopher y auge de los sistemas Web) se manifestó no tan sólo como un cambio en la forma de establecer relaciones o enlaces entre los documentos, sino también en el contenido de los mismos. De hecho, fue esto último lo que formó gran parte del atractivo de este nuevo medio para el público general.

En este momento, uno de los aspectos más notorios y controversiales (en términos legales) de la introducción de multimedios en la Web es el uso generalizado de formatos de compresión de audio digital (principalmente MPEG capa 3) y audio streamed (principalmente RealAudio).

Muchos usuarios ya no buscan sólo documentos sino también imágenes y principalmente, música. De hecho, en los últimos 2 años la palabra "mp3" ha reemplazado a "sexo" en varias máquinas de búsqueda como la más buscada (por ejemplo, ver [METASPY]). Este fenómeno también fue observado en los registros de acceso de TodoCL.

Si bien el principal motivo de implementar un filtro que permitiera seleccionar páginas que contengan algún tipo específico de formato multimedial es atender a este requerimiento de los usuarios del sistema de búsqueda, no deja de ser importante la pregunta de qué tan masificados están cada formato multimedial en la Web Chilena, particularmente los más nuevos, como las animaciones Flash o el audio streamed de RealAudio. Estadísticas y gráficos que intentan responder a esta pregunta se presentaron en la sección 3.4.4.

Además de los formatos multimediales, se puede utilizar el mismo filtro para seleccionar páginas que contengan archivos comprimidos, archivos ejecutables, u otros tipos específicos de documento; esto permite, por ejemplo, hacer búsquedas sobre páginas que

contengan software para Windows (`zip`, `exe`), Linux (`deb`, `rpm`) y Macintosh (`hqx`, `sea`).

Diseño

Las características deseables del sistema son:

- Permitir al usuario seleccionar sólo páginas que contengan enlaces a audio, video, archivos comprimidos, animaciones o imágenes.
- Permitir al usuario seleccionar sólo páginas que contengan enlaces a un formato específico de archivo.

Se definen los siguientes grupos: `IMAGEN`, `MÚSICA/AUDIO`, `VIDEO/ANIMACIÓN`, `COMPRIMIDO`, `FUENTE`, `PROGRAMAS` y `OTRO`.

- `IMÁGEN`: Compuserve Gif, Joint Photographics Expert Group, Windows Bitmap, Portable Network Graphics, Targa Graphics
- `AUDIO`: Microsoft Waveform, RealAudio, Midi, Movie Experts Group Layer 3 (MP3), AIF, Sun Audio Format
- `VIDEO`: ShockWave Flash, QuickTime, AVI, Movie Experts Group, Macromedia Director
- `COMPRIMIDO`: ZIP, GnuZip, Redhat Package Managment, Debian Package, Macintosh BinHex, TAR, LZH
- `FUENTE`: Java Source, Basic, C
- `PROGRAMAS`: Java Class, DOS Executable
- `OTRO`: Microsoft Powerpoint, Virtual Reality Modeling Language

Implementación

Se utilizaron máscaras de bits de 14 bytes de largo, 2 bytes por grupo.

Esta extensión se desarrolló en `Perl` para el recolector y `C` para la máquina de búsqueda. Se implementaron programas que convierten las constantes en Perl en encabezados en C para mantener la consistencia.

4.2.3 Reconocimiento de Idioma

Conceptos

La mayoría de las personas se sienten más cómodos y logran niveles más altos de comprensión leyendo textos en su lengua nativa que leyendo textos en una lengua extranjera. Un grupo importante de los Chilenos no tiene dominio de otros idiomas ni es capaz de leer y entender textos, por ejemplo, en inglés.

Se infiere que dado que la mayoría de los usuarios de TodoCL son Chilenos, será más probable que quieran restringir sus consultas sólo a páginas en español respecto a que quieran restringir sus consultas sólo al inglés, por lo que sistema puede diseñarse para optimizar las consultas restringidas al español.

Si bien el usuario que no quiera o no pueda leer textos en inglés puede darse cuenta por los párrafos que son extraídos del documento por el localizador de búsqueda del idioma, es una conveniencia adicional marcar de alguna forma particular las páginas en inglés, pues permite realizar esta distinción más rápido.

Observaciones Preliminares

Se revisaron a mano 300 documentos, observándose mayoritariamente textos en español e inglés. Unos pocos documentos en portugués, alemán y francés fueron encontrados sin representar un porcentaje significativo del total.

Se observó que entre un 5-10% de los documentos estaban en inglés.

Un problema adicional se hace presente al descubrir en la colección varias páginas bilingües (en inglés y español). Las páginas bilingües son de dos tipos: aquellas en que ambos idiomas están separados (ejemplo: la primera mitad en español y la segunda mitad en inglés, o dos columnas, una para cada idioma) y aquellas en que los idiomas están mezclados párrafo a párrafo. Si bien en el papel (dípticos y manuales impresos, por ejemplo) se opta casi siempre por lo primero, en la Web no hay tanto control sobre el ancho de la página por lo que se opta por ir mezclando párrafo a párrafo.

Diseño

Un texto en inglés tendrá muchas palabras en inglés y un texto en español tendrá muchas palabras en español, ahora bien, como no es posible tener un diccionario completo en memoria de palabras en inglés y en español, se opta por tener sólo un conjunto pequeño, a saber, las 10 más usadas[7] en cada idioma:

- En inglés: “of” “the” “to”, “and”, “for”, “in”, “by”, “this”, “on”, “with”
- En español: “de”, “y”, “la”, “en”, “el”, “del”, “los”, “para”, “que”, “las”

Esta heurística aprovecha la observación de que un texto en inglés por lo general tendrá varios conectivos del inglés y un texto en español varios conectivos del español.

Implementación

Al inicializar se construyen una tabla de hashing en que las llaves son las 20 palabras del diccionario y los valores +1 para palabras en español y -1 para palabras en inglés.

Cada palabra de los primeros 400 bytes es buscada en la tabla de hashing y el valor encontrado ahí se suma a un puntaje que parte de cero. Si el puntaje final es de -2 o menos se considera que el texto está en inglés.

Pruebas

Inicialmente la condición era que el puntaje final fuera de -1 o menos, pero esto tendía a clasificar erróneamente páginas en español que tenían frases del tipo `design by NNNN` como páginas en inglés. Con puntaje de -2 o menos se es consistente con lo mencionado anteriormente respecto a marcar sólo las páginas sobre las que se tuviera mayor certeza.

[7] En la lista anterior, se descartó la palabra “a” que es usada frecuentemente en ambos idiomas y por lo tanto no es útil para discriminar.

Los textos en otros idiomas no fueron reconocidos como textos en inglés, lo cual resulta bastante positivo. Incluír más idiomas requiere un mayor análisis, pero al parecer la misma heurística podría ser usada para más de dos idiomas.

Los resultados se indican en la sección 3.4.3 de esta memoria.

Un ejemplo de cómo se marcan páginas frente a la consulta "neruda" se muestra en la figura 4.9.

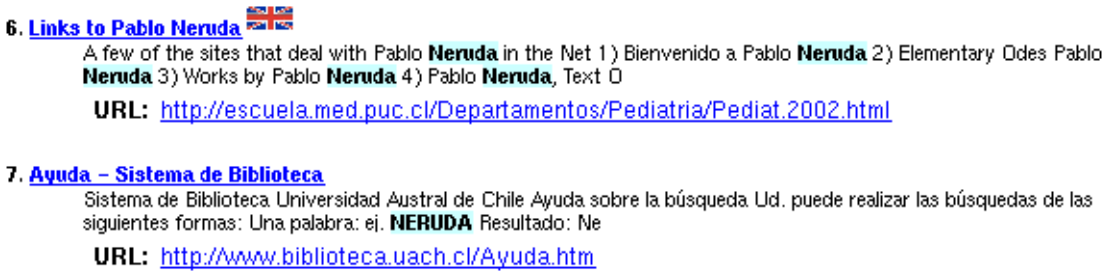

Figura 4.9: Ejemplo de cómo se marcan las páginas en inglés.

4.3 Extensiones al Buscador

4.3.1 Incorporación de un directorio de páginas

Motivación

Las estrategias de solución que un usuario puede adoptar frente a una necesidad de información son la búsqueda o la navegación, y hasta este momento no existía en TodoCL ninguna forma de que un usuario pudiera escoger páginas dentro de un cierto tópico que no pasaran por el ingreso de palabras clave, así como tampoco ninguna forma de control de calidad sobre dichas páginas que asegurara que fueran relevantes dentro del tópico al que pertenecen.

Los directorios que clasifican páginas Web usualmente son propiedad de una empresa que contrata a un equipo de editores para que recorran la Web, lo cual limita el crecimiento del directorio a los recursos económicos de la empresa patrocinante.

Para TodoCL, la alternativa de elección era utilizar la sección de páginas en Chile del Open Directory Project[ODP]. El Open Directory Project es una iniciativa cuyo objetivo es producir "el más completo directorio en la Web, confiando en un vasto ejército de voluntarios[8]" A cambio de la colaboración de voluntarios, la base de datos de páginas indexadas se licencia gratuitamente bajo un acuerdo NPL (*Netscape Public License*) que es una variante de GPL (*GNU Public License*[GNU]). Esta licencia establece permiso para reproducir e incrementar el contenido del directorio siempre que los trabajos derivados sean distribuídos bajo la misma licencia.

[8] Actualmente, se trata de más de 25000 editores en 280.000 categorías, que han clasificado cerca de 1.8 millones de sitios

Diseño

El directorio de ODP es distribuído en formato RDF (*Resource Description Framework*). Este formato está basado en SGML e incluye datos como URL, título, editor y tema de cada página clasificada; y es impulsado por [W3C].

El directorio es global, e incluye dos categorías relacionadas con Chile: *World/Regional/Chile* y *World/Español/Regional/Chile*. La primera corresponde a páginas en inglés y la segunda a páginas en español. Se utilizó la categoría con páginas en español que incluye más de 3000 direcciones, de las cuales 1000 corresponden a sitios bajo `.cl`.

La inclusión del directorio contempla además de las páginas un grado de integración muy básico con la máquina de búsqueda: frente a una consulta se desplieguen las categorías del directorio que incluyen términos de la consulta. No se provee en esta etapa de la posibilidad de priorizar las páginas en el directorio sobre las páginas en el índice al realizar una consulta, por no haberse estimado aún si correspondía hacerlo en términos de calidad de los resultados.

Implementación

El archivo de datos en RDF tiene un tamaño de aproximadamente 100Mb. Se implementó una forma de extraer la porción correspondiente a Chile y se utilizó la aplicación [RDFPARSE] con varias modificaciones para generar las páginas Web en el sitio de Todo-CL.

Un ejemplo del formato de los archivos RDF es el siguiente:

```
<Topic r:id="Top/Regional/South_America/Chile">
  <catid>26017</catid>
  <link r:resource="http://www.localaccess.com/chappell/chile/"/>
  <link r:resource="http://www.weatherhub.com/global/ci.htm"/>
  <link r:resource="http://www.visitchile.org"/>
  ...
</Topic>

<ExternalPage about="http://www.localaccess.com/chappell/chile/">
  <d:Title>Spotlight on Chile</d:Title>
  <d:Description>A comprehensive overview of the culture, society,
  and government of Chile.</d:Description>
</ExternalPage>
```

Para las consultas, se ocupó el módulo de localización de búsqueda sobre un archivo con los nombres de las categorías, ordenadas de mayor a menor número de enlaces. Frente a una consulta, se despliegan las tres categorías con más enlaces que contengan en su nombre términos de la consulta, con enlaces a dichas categorías dentro del directorio. La reutilización de código permitió que esta extensión fuera implementada en un tiempo relativamente breve, utilizando funciones del localizador de búsquedas aplicadas sobre las categorías.

Capítulo 5

Conclusiones

5.1 Sobre la Caracterización de la Web Chilena

Se verificó que varios resultados establecidos para la red global eran también válidos para la Web Chilena, así como correspondencias con el estudio realizado en la Web Brasileña. Esto indicaría que tal como otros fenómenos de Internet, la estructura de la Web es altamente autosimilar (es decir, la estructura no se modifica ante cambios de escala).

Las características locales más destacadas son:

- La tasa de utilización de los dominios inscritos es aproximadamente de un 50%
- La gran mayoría de los sitios existentes cuentan con sólo una página (estas son las llamadas "páginas de presencia en Internet" que contienen usualmente una foto, algunos párrafos de texto y la dirección de e-mail de la empresa).

A pesar de haber muchas páginas, estas en general presentan poco contenido, y están concentradas en unos pocos sitios.

Se extendieron los resultados de [B2000], estudiando características internas de las estructuras observadas en el grafo de links, en particular el hecho de que fuera de la componente fuertemente conexa principal `MAIN` la conectividad es bastante baja.

Dicho de otra forma, fuera del 25% de los sitios que forman la componente principal, la característica de *red* que se menciona profusamente tanto en la literatura especializada como en la de difusión, no existe, presentándose mayoritariamente sitios desconectados entre sí; esto está en concordancia con los modelos sobre el número de enlaces desde y hacia cada página que muestran que la mayoría de los links van hacia y desde un conjunto más bien pequeño de páginas.

Adicionalmente se mostraron características cualitativas de las componentes utilizando datos provistos por humanos, particularmente la importancia relativa de los sitios en cada componente en las preferencias de los usuarios; mostrándose discrepancias importantes entre los sitios a los que tiene mayor probabilidad de llegar un usuario que utiliza una máquina de búsqueda como TodoCL o un directorio como ODP.

5.2 Sobre las Extensiones al Buscador

Sin contar con retroalimentación de los usuarios, a través de *Focus Group* o encuestas, es difícil poder hablar de la efectividad de cada una de las extensiones por separado, pero este tipo de evaluación se encuentra fuera del alcance de esta memoria, correspondiendo a un análisis que debería considerar como mínimo los siguientes factores:

- Entrevistas en *focus-group*
- Análisis del uso de cada extensión en el log de accesos a TodoCL

Por otra parte, existen algunos aspectos que sí se pueden evaluar: el orden apropiado de implementación y el funcionamiento de algunas extensiones.

Para mejorar el tiempo de desarrollo y disminuir su complejidad, antes de cualquier otra deben implementarse las extensiones de *consejos de búsqueda* y *opciones de presentación*. Ambas proveen respectivamente de un marco a nivel lógico y de presentación para incorporar a las demás.

Además, un sistema de localización de búsqueda, o en términos más genéricos, un sistema de búsqueda multipatrón es requerido en varias ocasiones.

También fue útil desarrollar un módulo de identificación léxica básica, que permitiera buscar en un diccionario de palabras conocidas y descartar aquellas que pertenecieran a una determinada categoría. Este módulo puede ser extendido mediante el uso de algún proceso de lematización que permita reconocer variantes de las palabras en el diccionario.

En términos cualitativos, la expansión de consultas por lo general retorna palabras que tienen relación con el tema, sin embargo, resulta necesario estudiar más a fondo la forma de descartar palabras funcionales. En particular, un léxico o diccionario, por grande que sea, no es suficiente para reconocer todas las formas verbales del español.

El verificador ortográfico funciona bastante bien; una mejora interesante que puede plantearse es alguna forma de resolver el problema de que si una palabra es ingresada con faltas de ortografía y existe al menos una página en que la palabra se presenta con el mismo error, entonces no se advierte sobre la posibilidad de que se haya cometido un error ortográfico; esto es crítico puesto que resulta evidente pensar que si un error es cometido frecuentemente en las consultas, también se comete al escribir las páginas Web.

5.3 Trabajos Futuros

De entre las líneas de investigación relacionadas con el tema, destacan como fuente de posibles trabajos futuros las siguientes:

Más estadísticas que utilicen información de las componentes: en particular sería interesante conocer por ejemplo, tamaño promedio en texto y sobre todo qué tan frecuentemente son actualizadas las páginas en cada componente. Así mismo, será muy interesante conocer cómo se modifican las preferencias de los usuarios al cambiarse el algoritmo de ranking a uno que considere la estructura de links.

Del mismo modo, estudios sobre la cobertura de los buscadores internacionales respecto a páginas chilenas no se han realizado aún y podrían hacerse utilizando algunas de las técnicas disponibles (muestreo casi-uniforme de URLs o *firma léxica* de documentos).

Análisis de tópico: es necesario estudiar cómo aprovechar la información del directorio ODP y potenciarla con la aplicación de búsquedas, por cuánto lo que se hace

actualmente, buscar en los nombres del árbol de categorías, resulta demasiado simplista. Aprovechando la estructura de links y las copias locales de cada documento, podrían determinarse palabras claves de cada categoría; del mismo modo, cada categoría podría ser expandida con nuevas páginas obtenidas del buscador y que tuvieran relación de co-citación, similaridad y otra con las páginas clasificadas a mano.

En general, todas las formas de mejorar el buscador que aprovechen al máximo posible lo acotado del contexto y la gran tasa de cobertura que tiene TodoCL.

Bibliografía

[AH1999] Adamic and Huberman, *The nature of markets on the World Wide Web*, Xerox PARC Technical Report, 1999, `http://www.parc.xerox.com/istl/groups/iea/www/webmarkets.html`

[B2000] A. Broder, R. Kimar, F. Maghoul, P. Raghavan, S. Rajagopalan, R.Stata, A. Tomkins, J. Wiener, *Graph Structure on the Web*, WWW9, Mayo 2000, `http://www.almaden.ibm.com/cs/people/pragh/www9.html`.

[BYC2000] R. Baeza-Yates, C. Castillo, *Caracterizando la Web Chilena*, Reporte Técnico, Julio 2000. `http://www.todocl.cl/stats.phtml`

[BYRN1999] R.Baeza-Yates,B.Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley-Longman 1999.

[CoBWeb99] A. da Silva, E. Veloso, P. Golgher, B. Ribeiro, A. Laender, N. Ziviani, *CoBWeb: A Crawler for the Brazilian Web.* en *Proc. of the* 6^{th} *International Symposium on String Processing and Information Retrieval* (SPIRE '99) páginas 184-192. Carleton University Press, 1999

[FACEA99] Estudios Internet en Chile, *Unidad de Computación e Informática, Facultad de Ciencias Económicas y Administrativas, Universidad de Chile*, `http://www.facea.uchile.cl/uca/estudios/internet.htm`.

[G1999] G. Gulab, *Essay on Human Computer Interaction and Search Engines*, `http://www.cs.uct.ac.za/ğgulab`.

[HEAPS1978] J. Heaps, *Information Retrieval - Computational and Theoretical Aspects.* Academic Press, 1978.

[HHMN2000] M. Henzinger, A. Heydon, M. Mitzenmacher, M. Najork, *On near-uniform URL sampling*, WWW9, Mayo 2000.

[HS2000] Christoph Hölscher, Gerhard Strube, *Web seach behavior of Internet Experts and newbies*, WWW9, Mayo 2000.

[IA1998] L. Rosenfeld, P. Morville, *Information Architecture*, O'Reilly & Associates, 1998.

[JF1992] Javier Fernández, *Medios de prensa en la Web*, Memoria de Título.

[NAP99] Claudio Rutllant, *Internet, Pensando en Chile II*, Encuesta realizada por Ekhos Investigación y Consultoría e Interaccess. http://www.nap.cl/encuesta2

[PBMW1998] L. Page, S. Brin, R. Motwani and T. Winograd, *The pagerank citation ranking: Bringing order to the Web*, Technical report, Stanford University, 1998.

[PPR1999] P. Pirolli, J. Pitkow, R. Rao, *Silk from a Sow's Ear: Extracting Usable Structures from the Web*, http://acm.org/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html

[VBUSH1945] Vannevar Bush, *As we may think*, The Atlantic Monthly, Julio 1945, http://www.isg.sfu.ca/ duchier/misc/vbush/

[VMG1999] Eveline A. Veloso, E. de Moura, P. Golgher, A. da Silva, R. Almeida, A. Laender, B. Ribeiro-Neto, N. Ziviani, *Um Retrato da Web Brasileira*, Simposio Brasileiro de Computaçao, Curitiba, Brasil, Julio 2000.

[V1979] El Arte Abstracto y Figurativo, *Francesc Vicens*, Salvat Ediciones.

Software

[ALLIENDE] Análisis de Legibilidad de los Textos. *Felipe Alliende*, Departamento de Estudios Humanísticos, Universidad de Chile.

[GRAPHVIZ] GraphViz: A Graph Visualization Software, *AT&T Research Labs*, http://www.research.att.com/sw/tools/graphviz/

[OVID] Stopwords list, Ovid Technologies Inc, http://www.lib.purdue.edu/library_info/ovidweb/ref/stops.htm

[RDFPARSE] RDF Parse, *Anil Answani*, 2000. http://perlodp.cjb.net

Sitios Web

[ATW] All The Web, http://www.alltheweb.com

[AV] AltaVista, http://www.altavista.com/

[DSTATS] DomainStats.com, http://www.domainstats.com

[GNU] GNU - GNU is not Unix, http://www.gnu.org/

[GOOGLE] GOOGLE, http://www.google.com/

[METASPY] Metacrawler's MetaSpy, http://www.metaspy.com

[NETCRAFT] Netcraft server's survey, http://www.netcraft.com/survey

[NIC-CL] Network Information Center - Chile,

[NL] Northern Light Search, http://www.nlsearch.com/

[NUA] NUA internet surveys, http://www.nua.ie/surveys

[ODP] Open Directory Project, *Netscape Communications Corporation*, http://dmoz.org/about.html.

[SEWATCH] Search Engine Watch, http://www.searchenginewatch.com/reports/

[SSCHILE] Sunsite's Chile Page, http://sunsite.dcc.uchile.cl/chile

[TODOBR] TodoBR: Todo o Brasil na Internet, http://www.todobr.com.br

[TODOCL] TodoCL: Todo Chile en Internet, http://www.todocl.cl

[TODOCL2] TodoCL: Todo Chile en Internet, Laboratorio de Experimentación, http://www2.todocl.cl

[UW] Usable Web, http://www.usableweb.com/

[W3C] World Wide Web Consortium, http://www.w3.org