Indización Web con identificadores geográficos para zonas específicas

Jordi Corvillo Martínez[1]
jordicm@mx3.redestb.es

RESUM

Es proposa un model d'indexació Web basat en els codis postals, com a mètode per a una millor recuperació per matèries i que resoldria el problema de la recerca geogràfica a Internet.

RESUMEN

Se propone un modelo de indización Web basado en los códigos postales, como método para una mejor recuperación por materias y que resolvería el problema de la búsqueda geográfica en Internet.

Introducción

Es un hecho que las soluciones existentes al problema del elevado nivel de congestión de la red Internet desde la perspectiva de la recuperación de información y más concretamente desde el campo de los motores de búsqueda, no cumplen las expectativas que el usuario espera de ellas.

De hecho, el usuario suele buscar, inconscientemente o no, una fuente local de información, esto es, información de la zona en que reside o desarrolla su profesión, pero como que no suele disponer de herramientas fiables de búsqueda en su zona, se ve obligado a buscar más allá de su estricto interés. Y aún más: los motores de búsqueda convencionales no permiten por ahora acotar una consulta mediante limitadores geográficos, más allá de las consabidas palabras clave sobre el texto completo.

Por ejemplo, cuando efectuamos una búsqueda simple en un motor de búsqueda de texto completo, obtenemos un número de resultados demasiado grande y muy frecuentemente poco o nada adecuado. Esto se produce como resultado de la siguiente situación: una palabra concreta aparece cierto número de veces en un documento, pero esta palabra no tiene por qué ser necesariamente significativa en el contexto del tema central del documento.

Podemos introducir la siguiente búsqueda simple en Altavista:

Potatoes+Mataró o bien Potatoes and Mataró

y esperar que aparezca un documento que hable de las reconocidas patatas de Mataró. En definitiva: realizar una búsqueda geográfica más profunda puede llegar a ser una tarea realmente difícil.

Interrogando motores de búsqueda como Infoseek, más basados en la recuperación de las etiquetas META de las páginas Web que Altavista, más centrado en la indización Web a texto completo, probablemente no obtendremos ni tan siquiera los resultados que el segundo nos ofrece debido a que la indización Web basada en las etiquetas META no garantiza la calidad de la recuperación porque no existen todavía modelos normalizados de descripción en uso. Se da el problema de que, en muchas ocasiones, dentro de las etiquetas META aparecen palabras clave que no tienen ninguna clase de relación con la temática del documento.

Actualmente, un motor de búsqueda no «sabe» distinguir si un documento que contiene las palabras clave «sex» o «xxx» en la etiqueta META correspondiente, en realidad es un documento sobre excursionismo en Transilvania. De cualquier modo, los autores que se sirven de esta picaresca quedan así desacreditados por completo delante del resto de la comunidad.

1. Quiero agradecer de todo corazón su ayuda a las personas que han facilitado la elaboración y publicación de esta comunicación: Jaume Baró, Cristina Barragán, Sergi Brualla, Lluís Codina, Francisco Núñez y Ernesto Spinak. Estoy realmente orgulloso de contar con amigos como ellos. También quiero dedicar especialmente este trabajo a mi familia.

1. Indización distribuida versus indización centralizada

El modelo centralizado de búsqueda e indización utilizado por servicios como Altavista, Infoseek, Excite y otros, crea un enorme índice de la gran mayoría de los recursos disponibles en Internet y lo ofrece a través de un único servidor. Puede haber copias de este índice en otras partes de la Red, los llamados servidores espejo o *mirrors*, pero cada uno de ellos también es accesible a través de una sola máquina. El modelo distribuido, usado por servicios como Harvest, ALIWEB o Glimpse, crea pequeños índices junto a los documentos que indiza.

Ambos modelos tienen ventajas y desventajas pero hay algunos puntos a favor del modelo distribuido: el primero es que, muy probablemente, la centralización no va a poder absorber como hasta ahora el crecimiento exponencial de la información en Internet y, a la vez, el incremento de su complejidad.

El segundo es que un sistema centralizado será cada vez más difícil de mantener actualizado, debido a que estos sistemas pretenden indizar la Red entera. Un modelo de trabajo como éste no dispondrá ni del suficiente ancho de banda ni de la capacidad de cálculo de unos procesadores cada vez más potentes pero a los que progresivamente se les añadirán más y más cargas.

El tercero es que esta centralización fuerza al usuario a realizar la búsqueda sobre toda la base de datos del sistema, con lo que es bastante difícil limitar la búsqueda a los documentos sobre o de Túnez, por ejemplo. Los sistemas de indización distribuida, en cambio, permiten la creación de índices geográficos fácilmente. Esta característica podría ser un instrumento muy útil para refinar las búsquedas en la Red.

2. Sistemas de indización distribuida

2.1. Harvest

Harvest es un sistema integrado de herramientas para localizar, almacenar y recuperar información a través de Internet. Harvest presenta un rendimiento muy elevado en los aspectos de tráfico de red, gestión de servidores remotos y espacio en disco: puede reducir la carga de los servidores HTTP y FTP por un factor de 4 extrayendo datos de los índices y del orden de un factor de 6.600 en la conexión con los servidores remotos; tráfico de red por un factor de 59 y requerimientos de espacio para los índices por un factor de 43.

Este sistema posee un formato de indización estructurado denominado Summary Object Interchange Format (SOIF) que permite consultas estructuradas (por ejemplo, consultas por campos de autor o título). Los responsables de Harvest están desarrollando un sistema de cachés regionales, para llegar a elaborar conjuntos de jerarquías nacionales, más o menos en la línea que estamos presentando aquí.

2.2. ALIWEB

ALIWEB propone que sea el proveedor de contenidos el responsable inicial de la descripción de los recursos, de tal manera que los programas recojan de forma automática estas descripciones y las introduzcan directamente en una base de datos. Debido a que los ordenadores y sus programas todavía no pueden interpretar el lenguaje natural, estas descripciones deben ser redactadas en un formato conciso y normalizado.

El sistema de indización ALIWEB trabaja de la siguiente manera:

a. El usuario redacta la descripción de sus recursos en un formato normalizado en un fichero en el Web, bien a mano o bien usando programas específicos.

b. El mismo usuario avisa a ALIWEB de ese nuevo fichero.

c. ALIWEB recupera de forma regular todos los nuevos ficheros y los introduce en una base de datos.

d. Cualquier navegante de Internet puede visitar ALIWEB y realizar una búsqueda en su base de datos.

La base de datos se actualiza frecuentemente (casi cada día) y esto hace que la información sea muy reciente. El usuario sólo tiene que preocuparse de realizar las descripciones de sus recursos, dado que ALIWEB realiza todo el trabajo de recuperar e introducir los ficheros en la base de datos. De este modo, se obtienen dos ventajas muy importantes: por un lado, con este sistema la información sobre los recursos debería ser correcta, adecuada y suficiente y, por otro lado, al manejar ficheros de descripciones pequeños se reduce mucho la sobrecarga en los servidores.

Por lo que se acaba de comentar, puede leerse que ALIWEB también ofrece posibilidades al usuario desde el punto de vista de las fuentes locales de información.

3. Herramientas para la indización

3.1. Robots, agentes o arañas

Un robot es un programa que analiza el World Wide Web, recuperando información de cada servidor que visita. El recorrido se inicia visitando una página Web determinada, siguiendo sus enlaces y así hasta indizar el Web entero.

El problema de las diferentes denominaciones de los robots se basa en que hay ciertas diferencias en el comportamiento de las arañas o *spiders*, también conocidas con el nombre de vagabundos o *wanderers*, y los robots. Mientras que las primeras actúan más bien como sofisticados navegadores transparentes al usuario y no «viajan» por la Red, los robots abandonan el servidor del que parten y se mueven por la Red recuperando documentos, es decir, «viajan» por la Red.

Para colaborar con el razonable propósito de evitar la sobrecarga de los servidores, los robots suelen seguir actualmente el protocolo *Proposed Standard for Robot Exclusion*, también conocido con el nombre de protocolo de exclusión de robots, ideado por Martijn Koster en 1994. La razón por la que pensó en tal proyecto, según sus propias palabras, fue la siguiente: «En 1993 y 1994 había ocasiones en que los robots no eran demasiado bien recibidos en los servidores que visitaban por varias razones. A veces era porque solicitaban demasiado rápido los documentos o porque recuperaban el mismo documento reiteradamente. En otras ocasiones los robots accedían a partes de los servidores nada susceptibles de ser rastreadas por un robot, como información duplicada, información temporal o scripts CGI, con todos los problemas de seguridad y desconfianza hacia los robots que esto entraña.» Por lo tanto, Koster decidió redactar una lista de directrices sobre cómo los robots deberían interactuar con los servidores Web.

Con el sistema de indización propuesto por el autor, los robots no saturarán el ancho de banda en exceso ya que no es necesario realizar un análisis de texto completo para extraer la información que se necesita de la etiqueta META específica.

3.2. HTML: la etiqueta META

De acuerdo con la especificación HTML 4.0, ya aprobada por el W3C (World Wide Web Consortium), la etiqueta META puede ser usada para describir propiedades de un documento y para asignar valores a esas propiedades. El objetivo básico de esta etiqueta es albergar metainformación o metadatos dentro de un documento HTML. Un uso hasta ahora bastante extendido de la etiqueta es el de contener palabras clave que describan con la mayor exactitud posible el contenido del documento para que un motor de búsqueda (su robot) pueda elaborar su índice. La propiedad que pretendemos describir es la localización geográfica mediante un código postal. La sintaxis de esta etiqueta META podría ser como la que sigue:

<META NAME=«geogloc» COUNTRY=«es» POSTALCODE=«17200»>

La información que el robot podría extraer de esta etiqueta es que ese documento es de Palafrugell, comarca del Baix Empordà, provincia de Girona, Cataluña y, como dominio de país o estado (lo que ustedes prefieran), España.

Así, con sólo dos atributos dentro de la etiqueta (dominio de país o estado y código postal), podemos extraer la localización geográfica exacta del documento. Los nombres de país o estado podrían ser los mismos que los usados como nombres de dominio nacionales en Internet.

4. Indización Web con códigos postales

Probablemente, un sistema más racional para la indización Web geográfica podría estar basado en clasificaciones ya establecidas y normalizadas como, por ejemplo, los códigos postales.

Hay suficientes códigos postales establecidos para cada zona de cada país, de forma que los servicios postales han organizado su territorio perfectamente para conseguir la mayor eficiencia posible en la distribución del correo. Todos conocemos nuestro código postal: se trata de un sistema fácil, útil y comprensible.

El modelo de indización propuesto por el autor debería tener las siguientes características.

El proceso de indización, basado en el robot o agente, realiza una indización geográfica de una manera transparente para el usuario. Tres opciones serían posibles:

1) La primera consistiría en crear un índice por zona: éste sería un modelo distribuido, en el que la cobertura de la zona geográfica podría variar en función de las necesidades del servicio.

La creación de índices de zona permitiría establecer motores de búsqueda locales muy fiables: en este caso, el robot buscaría etiquetas META que contuviesen códigos postales predeterminados por país. Una vez que los índices fuesen creados se enviarían al servidor de búsqueda local. Esto sería una especie de indización centralizada en tanto que el robot realiza una búsqueda masiva, sin restricciones muy concretas.

Pero el resultado final es más parecido al de un sistema distribuido porque cada zona dispondría de sus páginas indizadas, que serían así más fácilmente accesibles, tanto para los navegantes de esa zona como para los del resto de Internet.

En esta opción el mantenimiento y la actualización del sistema serían complejos.

2) La segunda opción consistiría en que cada servidor de búsqueda local tendría un robot que buscaría exclusivamente documentos con los códigos postales de esa zona. Éste es un sistema más preciso, rápido y fácil de actualizar y mantener. Será un sistema aplicable cuando se solucionen los problemas de ancho de banda de la Red.

3) La tercera opción evita el uso de robots: de forma similar a ALIWEB, el propio usuario registraría sus páginas en su servidor de búsqueda local, clasificado al estilo de Local Yahoo <http://local.yahoo.com>. Este sistema evitaría por completo los problemas de saturación del ancho de banda de los robots. Sin embargo, el mantenimiento, la actualización y la precisión de los datos también serían problemáticos en este caso.

Por lo que respecta a la interfaz de usuario, el modelo propuesto aquí podría estar basado en las siguientes opciones:

1) El usuario navegaría por un índice general al estilo de Local Yahoo.

2) Uso de mapas sensibles, que permitirían buscar rápidamente en una zona determinada.

Desde el punto de vista del que suscribe, esta solución basada en códigos postales facilitaría enormemente la recuperación de información, sobre todo porque resuelve en cierta forma los problemas de masificación de la búsqueda en Internet. Y lo mejor es que lo hace con herramientas existentes y simples como la etiqueta META del HTML o los códigos postales. Quizá otros sistemas distribuidos como Harvest o ALIWEB no han alcanzado suficiente éxito porque son difíciles de asimilar para el usuario común.

Este sistema podría ser implantado dentro del Dublin Core Element Set o alguna iniciativa similar. Dublin Core ya dispone de un elemento de cobertura geográfica, pero no es tan sencillo ni descriptivo como el que se ha explicado aquí.

Para finalizar, desearía hacer énfasis en la idea de que la indización centralizada contradice la filosofía descentralizada de Internet. Y que en la medida que la propia estructura hipertextual del Web dispersa y disgrega la información, es necesario desarrollar modelos de indización efectivos, sin renunciar a la distribución de los recursos porque esto hace que sean accesibles de forma rápida, fácil y democrática.

El sistema aquí propuesto es únicamente un paso más hacia una mejor recuperación por materias, pero creo que resuelve de forma bastante satisfactoria el problema de la búsqueda geográfica en Internet.

Bibliografía

Ad Hoc Working Group: coverage element. *Dublin Core element: coverage*. Accesible desde: http://www.sdc.ucsb.edu/~mary/coverage.htm [Consultado el 28-11-1997].

CHEONG, Fah-Chun. (1996). *Internet agents: spiders, wanderers, brokers and bots*. Indianapolis: New Riders.

CODINA, Lluís. (1997). «Sortear el laberinto». *Byte España*. Nº 5, (diciembre), p. 70-78.

EICHMANN, David. *Ethical Web agents.* Accesible desde: http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Agents/eichmann.ethical/eichmann.html [Consultado el 20-12-1997].

FISCHER, Keith D. (1997). *The WWW robot and search engine FAQ.* Accesible desde: http://sunsite.unc.edu/boutell/faq/www_faq.HTML [Consultado el 6-1-1998].

Guía código postal 96-97. (1996). Madrid: Correos y Telégrafos.

Hypertext Markup Language Specification 4.0. Accesible desde: http://www.w3.org/TR/REC-html40/ [Consultado el 20-12-1997].

KOSTER, Martijn. (1993). *Guidelines for robot writers.* Accesible desde: http://webnexor.co.uk/mak/doc/robots/guidelines.html [Consultado el 8-12-1997].

KOSTER, Martijn. (1993). *Proposed standard for robot exclusion.* Accesible desde:http://info.webcrawler.com/mak/projects/robots/norobots.html [Consultado el 8-12-1997].

MANIEZ, Jacques. (1993). *Los lenguajes documentales: concepción, construcción y utilización en los sistemas documentales.* Madrid: Pirámide; Salamanca: Fundación Germán Sánchez Ruipérez.

MUSELLA, Davide; PADULA, Marco. *The authors catalogue their documents for a light Web indexing.* [Milan]: Istituto per le Tecnologie Informatiche Multimediali. Accesible desde: http://gea01.pangea.org/inet96/a2/a2_4.htm [Consultado el 19-12-1997].

TITEL, Ed. (1996). *Fundamentos de programación con HTML y CGI.* Madrid: Anaya Multimedia.

WEIDER, Chris. *The future of search on the Internet.* Bunyip Information Systems. Accesible desde: http://gea01.pangea.org/inet96/a2/a2_1.htm [Consultado el 18-12-1997].