Trabajo final: Búsqueda en portales web

José Alberto Benítez Andrades

71454586A

Motores de Búsqueda Web
Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web
UNED
01/09/2011

ÍNDICE

0. Resumen

A continuación, en este pequeño estudio, se intentarán explicar las semejanzas y diferencias entre la búsqueda web y la búsqueda en un portal, destacando los distintos retos diferenciales que supone un buscador de portal. Además de esto se analizarán cuáles son las aportaciones más importantes en lo que se refiere a la búsqueda en portales web existentes en la literatura científica. También mostraré una serie de mejoras propuestas, con componentes actuales, sobre los buscadores que existen en la actualidad. Se diferenciarán las fuentes de información disponible para hacer un raking. principales aportaciones al tema de la búsqueda en portales web existentes en la literatura científica. También se propondrán mejoras -con algún componente novedoso- sobre los buscadores actuales, diferenciando cuáles son las fuentes de información disponibles (para hacer un ranking) y cómo se propone utilizarlas.

1. Introducción.

En este apartado se hace una breve descripción de qué es un portal web y qué funcionalidades ofrece a los usuarios a diferencia de una web normal para entender mejor los aspectos que se considerarán en los siguientes apartados. Además se hace una clasificación de los tipos de portales web más utilizados actualmente.

Qué es un portal web

Los portales de Internet son puntos de entrada para la presentación y el intercambio de información a través de la Web dirigidos principalmente a una comunidad de usuarios con un interés común. Los portales incluyen enlaces, buscadores, foros, documentos, aplicaciones, tiendas on-line, correo electrónico, alojamiento web, motor de búsqueda etc. que sirven como apoyo para realizar el intercambio de información de forma eficaz.

Un portal debe tener necesariamente una temática que sea de interés para un grupo amplio pero a la vez especializado de usuarios, por tanto debe ofrecer servicios que capten el mayor número de internautas posibles y para que los usuarios cumplan sus expectativas de información el portal debe seleccionar y presentar la información requerida de forma rápida y eficaz.

Clasificación de los portales web

Una clasificación de portales en función de sus usuarios sería la siguiente:

- Portales horizontales: también llamados portales masivos o de propósito general, se dirigen a todos los usuarios en general e intentan captarlos a través de los diversos servicios que ofrecen. Su principal objetivo es que los usuarios que accedan a Internet lo hagan siempre a través del portal, por tanto los servicios que ofrece suelen ser gratuitos y su principal ingreso se basa en la publicidad además del número de visitas, lo cual puede resultar a ser molesto para el usuario ya que si abusan de la publicidad los servicios que ofrecen quedan relegados a un segundo plano. Portales muy populares fueron Terra, Yahoo o MSN, pero actualmente ese tipo de portales han caído en desuso dando paso a las páginas de inicio personalizadas, como myYahoo, iGoogle, netVibes, en las que el usuario establece una de ellas como página de inicio y elige los contenidos que quiere que se le muestren al abrir la página.

- Portales verticales: Son portales especializados en determinados temas y su público son usuarios interesados por dichos temas. La temática de estos portales puede variar desde la música, el empleo, la inmobiliaria, las finanzas, el arte, la educación o los deportes. Ejemplos de portales de este tipo son: Idealista, Fotocasa (portales inmobiliario), Infojobs o Tecnoempleo (portales de empleo), Invertia o Finanzas. com (portales de _nanzas), coches.net y motos.net (venta de coches y motos), segundamano.es (portal de anuncios de productos de segunda mano), Booking.com (portal para buscar y reservar habitaciones de hotel),etc.

A día de hoy quizá carece de sentido hablar de portales de la manera que se concebían años atrás. Actualmente existen portales verticales, es decir, de un tema determinado, pero no predominan los portales horizontales o de propósito general ya que es una idea desfasada. Hoy en día existen sitios web de gran tamaño, con gran cantidad de información y contenidos, pero no dejan de ser sitios web. Por tanto el presente trabajo se centrará en la búsqueda interna en sitios web en general y en portales verticales.

2.Búsqueda en portales y sitios web

Introducción

Una investigación de IDC () patrocinada por EMC, midió y predijo la cantidad y el tipo de información digital disponible en 2007.

El estudio, denominado The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010 (El universo digital en expansión: un pronóstico del crecimiento de la información mundial hasta 2010), revelaba la cantidad de información que el mundo crea y copia en un determinado año. Además, realizaba predicciones sobre todo este "universo digital" hasta el año 2010. Los resultados del informe arrojaron las siguientes conclusiones:

- En 2006, el universo digital tenía un tamaño de 161.000 millones de gigabytes (161 exabytes)
- IDC proyectaba una sextuplicación anual de la información de 2006 a 2010
- Mientras que casi el 70% del universo digital sería generado por particulares para 2010, las organizaciones serían responsables de la seguridad, la privacidad, la con_abilidad y el cumplimiento con las normas de al menos el 85% de la información
- Las imágenes, captadas por más de mil millones de dispositivos en el mundo, desde cámaras digitales y teléfonos con cámara hasta escáneres médicos y cámaras de seguridad, representan el mayor componente del universo digital.
- Información no estructurada - Más de 95% del universo digital se compone de información no estructurada. En las organizaciones, la información no estructurada representa más de 80% de toda la información.

A la vista de estos resultados estamos ante un crecimiento exponencial de la cantidad de información disponible en la web, por tanto es necesario el desarrollo adecuado de las infraestructuras de la información para gestionar este crecimiento. La recuperación de información se hace realmente difícil, ya que la capacidad humana para la lectura se mantiene constante y los usuarios tienen la limitación de leer sólo aquello que se ajusta a sus necesidades ante el volumen de información disponible que supera ampliamente su capacidad de captación y entendimiento.

Además la mayoría de la información disponible en la web no está estructurada y no existe ninguna norma general para su creación. Los sitios web crecen y se desarrollan sin un mecanismo de control. También la facilidad de edición por parte de los usuarios hace que la

calidad del contenido sea en ocasiones dudosa y no hay forma de distinguir información de calidad de aquella que no lo es.

A pesar de las dificultades se han realizado esfuerzos por parte de los creadores de sitios web, creadores de herramientas, profesionales de la información y usuarios para facilitar la recuperación de contenidos en la web:

- herramientas de búsqueda y algoritmos en constante perfeccionamiento
- estrategias para optimizar el ranking de resultados de esas herramientas
- directorios generales y especializados para reunir y seleccionar la información de la web
- metadatos para añadir información al sitio web y no hacerlo únicamente dependiente de su texto
- pautas para la normalización de los sitios web
- creciente interés del ámbito institucional, universitario e investigador por una mayor optimización

Además de realizar esfuerzos para recuperar la información que el usuario busca es necesario trabajar también dentro de los sitios web y los portales para que una vez que el usuario llega a él, bien porque conoce su dirección URL o bien porque un buscador le lleva a él, pueda encontrar la información que busca. Por tanto es interesante conocer el comportamiento de los usuarios dentro del sitio o el portal para poder desarrollar buscadores internos que faciliten la tarea de encontrar la información.

Uso de los buscadores internos

Muchos usuarios que acceden a un sitio web utilizan un buscador interno para encontrar información, otros recurren a la navegación jerárquica, es decir, navegan por la estructura de la web siguiendo enlaces, menús, barras de navegación, etc. y otros utilizan ambas cosas.

Algunos usuarios recurren a los buscadores internos cuando la página no está estructurada correctamente, por tanto los buscadores internos son una herramienta que ayuda a encontrar la información aunque no son herramientas 100% eficaces puesto que también hay que tener en cuenta que los usuarios pueden utilizar términos inadecuados para realizar las búsquedas.

Mientras que los buscadores web son una herramienta de uso generalizado por los usuarios cada vez que acceden a la web, los buscadores internos son más bien un complemento a la navegación jerárquica y su uso depende del usuario, su experiencia, la situación de búsqueda y el tipo de contenido a buscar.

Los buscadores internos permiten el acceso de forma fácil y rápida a un contenido puntual y son esenciales, como se ha mencionado anteriormente, cuando un sistema de navegación está mal estructurado, ya que en este caso la informaci ón no se encontrará a través de una búsqueda intuitiva por la estructura del sitio, ni a través del mapa del mismo, sino a través del buscador interno.

Buscadores internos

Un buscador es una herramienta que examinando la información existente de dominio público en la red, la indexa y la almacena, de forma que posteriormente dicha información se puede recuperar a través de consultas por palabras o términos clave. Los buscadores internos o locales examinan únicamente la información contenida en un sitio web, a diferencia de los buscadores web que examinan e indexan toda la información contenida en la Web.

Algunos autores como Yusef Hassan aconsejan utilizar un buscador interno cuando el sitio supere las 150 páginas, ya que de lo contrario es bastante probable que no haya resultados para la mayoría de las consultas que se realicen. En cambio otros establecen esa pauta por encima de las 200 páginas. Toni Vicens apunta que los buscadores internos influyen de forma decisiva en la primera impresión de los usuarios y son cruciales para que éstos encuentren los contenidos que buscan, por lo tanto hay que prestar especial atención a su usabilidad.

Los usuarios se dirigen de manera casi exclusivamente centrada a encontrar lo que buscan en Internet. No prestan mucha atención a otros temas diferentes del buscado y si un sitio web no parece relevante para sus objetivos, el usuario vuelve al anterior en dos o tres segundos.

Los usuarios confían casi ciegamente en los buscadores como herramienta principal de sus búsquedas dentro de un sitio, es decir, si un buscador interno no encuentra una determinada información el usuario considerará que la información no está disponible en este sitio. Ello supone que cualquier error en el funcionamiento de un buscador puede tener efectos fatales.

Algunos autores niegan este comportamiento y afirman la preponderancia de la navegación jerárquica sobre el uso del buscador. Este comportamiento también podría

explicarse por la pésima calidad de los resultados de los buscadores internos existentes en la mayoría de sitios web. Después de varias interacciones los usuarios descubren que solo los buscadores internos de ciertos sitios son fiables y en el resto no se molestan en realizar búsquedas, pero eso no significa que no los prefieran a la navegación por categorías, un ejemplo de ello es el imprescindible buscador de Amazon.com.

Cuando los usuarios tropiezan con alguna dificultad en el manejo o navegación de algún sitio web, no tratan de aprender su funcionamiento, continúan buscando en otros sitios. Los usuarios se muestran muy poco tolerantes a la dificultad porque saben que siempre existen muchos otros sitios web donde obtener la misma información y están a un solo click de distancia.

Para proporcionar un buscador interno en un sitio web existen varias posibilidades:

- crear un motor de búsqueda propio. Esta opción es compleja y requiere de conocimientos específicos de programación para llevarlo a cabo pero permite desarrollar soluciones a medida. En este caso es el administrador del sitio quien encargará de su mantenimiento.

- utilizar las herramientas gratuitas o de pago que facilitan otros proveedores, como Google Site Search1 o Atomz2, de esta manera tanto el índice como el mantenimiento pasan a ser responsabilidad de esa tercera parte implicada y no del administrador del sitio. La consulta del usuario se realiza en el propio sitio y la búsqueda se realiza en índice que reside externamente, el usuario recibe la respuesta e interactúa con el buscador de forma transparente, sin notar que la búsqueda se realiza en otro servidor.

La segunda opción tiene grandes ventajas ya que suelen tratarse de herramientas de calidad, no hay que ocuparse de la configuración, ni de la instalación, y ahorran tiempo, pero también conlleva desventajas ya que tanto la herramienta como la interfaz es estándar y no personalizada y al depender de un servidor externo todo el mantenimiento y las actualizaciones quedan a expensas del mismo no pudiendo controlarlo.

Factores importantes en un buscador interno

En el primer factor que hay que pensar a la hora de construir un buscador interno para un sitio web es el usuario. Hay que pensar en sus necesidades, en cómo harán las búsquedas y el tipo de información que esperan encontrar. Dependiendo del sitio web los usuarios pueden ser muy diferentes y tener distintos intereses, por tanto hay que considerar de forma prioritaria este aspecto.

La calidad de un buscador interno no depende exclusivamente del programa o motor de búsqueda seleccionado sino que también hay que prestar atención a las decisiones tomadas en su implementación.

La localización del buscador dentro del portal es muy importante, debe estar a la vista preferiblemente en la parte superior de la página. El usuario debe localizarlo fácilmente, sin necesidad de buscarlo, y elegir un lugar estándar para su colocación facilitará que el usuario asocie la herramienta con un lugar concreto dentro de los sitios web. Es recomendable que aparezca no sólo en la página principal sino que al navegar por el sitio y cambiar de página se mantenga a la vista, para poder utilizarlo en cualquier momento. En algunos sitios aparece la caja de búsqueda directamente y en otros aparece un enlace que lleva a otra página que contiene el buscador.

La interfaz también es un aspecto clave para que el uso del buscador resulte amigable para el usuario. No sólo importante es que exista un buscador interno y que se pueda acceder a él localizándolo fácilmente, es esencial que pueda ser identificarlo como tal. Una norma básica de usabilidad es que la herramienta tenga una interfaz simple y clara que los usuarios identifiquen fácilmente. Lo más estándar en materia de buscadores internos es un campo de texto con un botón al lado que diga "Buscar" u otra forma similar. Es deseable que el tamaño de la caja donde se incorporará el texto de la consulta sea el suficiente como para ver la expresión de búsqueda que se formula o al menos la mayor parte de ella. Esto facilitará al usuario ver y corregir los posibles errores ortográficos o de tecleo, lo cual influirá en la calidad de las respuestas.

La presentación de los resultados debe ser clara y concisa, ya que una mala presentación arruinaría el trabajo del motor de búsqueda. Se debe mostrar el número total de resultados encontrados y la consulta que los ha generado. Deben ser resultados con la suficiente información para que el usuario sepa si le interesan o no, ni tan extensos que requieran mucho tiempo para decidir ni tan breves que no lo permitan. Si no hubiera

resultados es recomendable mostrar un mensaje como “No se encontraron resultados” y si es posible sugerir una consulta alternativa.

A diferencia de un buscador que rastrea toda la Web, un buscador interno puede ofrecer mayor claridad en la presentación de los resultados y en su forma de ordenarlos ya que las páginas del sitio se pueden diseñar pensando en su posterior recuperación y de la misma manera el buscador se puede implementar pensando en las peculiaridades de sitio web y en su estructura en general. No obstante en caso de elegir como buscador interno una herramienta ofrecida por otro proveedor no se dispondrá de esta ventaja.

Un aspecto de importancia común para buscadores internos y generales es la forma de ordenar los resultados. En ambos casos deben ordenarse en función de la relevancia de los mismos ya que si los primeros resultados no son acertados el usuario tendrá una mala impresión de la herramienta y probablemente dejará de utilizarla sin realizar más consultas. Algunos buscadores ofrecen al usuario la posibilidad de elegir el criterio para ordenar: por término, por fecha, por ubicación, etc.

3.Buscadores web vs. Buscadores de Portales

El funcionamiento general de ambos buscadores es el mismo, ambos utilizan robots o crawlers para rastrear la web y un índice en el que almacenan los datos. La principal diferencia entre un buscador web genérico y un buscador de portal o vertical es que el primero trabaja sobre toda la web y el segundo sobre un sector específico de páginas. Al trabajar sobre un conjunto de páginas concreto y más reducido, los buscadores de portales contienen información más especializada sobre el tema que traten y además pueden actualizar esa información con más facilidad y mayor frecuencia. Como consecuencia, son más valiosos que los genéricos para los usuarios con interés en una determinada materia y, por tanto, tienen una audiencia muy específica. Así, existen buscadores verticales para medicina, ciencia, educación, viajes, compañías comerciales, buscadores de trabajo, etc. Estos buscadores devuelven resultados más precisos, más rápidamente y mediante la realización de consultas más simples que las que se necesitarían en los genéricos.

Adicionalmente, los buscadores verticales pueden ofrecer herramientas de búsqueda avanzada al usuario diseñadas especí_camente para el sector que se esté tratando. Si hablamos de un buscador de alojamientos éste permite elegir el periodo de fechas de la

estancia o el régimen de alojamiento y devuelve información sobre la disponibilidad de cada hotel, mientras que un buscador general permitiría buscar los hoteles pero no daría información añadida.

Un ejemplo de buscador especializado es Scirus3, dedicado a la búsqueda en millones páginas web dedicadas a Ciencia.

Con el creciente uso de los buscadores genéricos, el volumen de información que ofrecen ha aumentado sustancialmente. Esto ha resultado cada vez más frustrante para los usuarios que han intentado buscar información en un tópico especializado, como información local, destinos de viaje o canales específicos para empresas.

En un buscador como Google la cantidad de resultados devueltos para un término concreto supera con facilidad el millón de resultados. Tómese como ejemplo un taller mecánico cuyo propietario quiere encontrar información sobre piezas de coches, ya sean proveedores, recambios, desguaces, etc. Si introdujera el término de búsqueda _radiador_ en un buscador genérico gran parte de los resultados que obtendría no estarían relacionados con la mecánica del automóvil sino con radiadores de calefacción. En cambio esta misma búsqueda realizada en un buscador especializado ofrecería resultados mucho más concretos y útiles para el propósito del usuario.

También hay que señalar que en ocasiones los buscadores generales no pueden acceder a la información de los portales, por tanto serán incapaces de devolver como resultado la información contenida en los mismos. Esto ocurre cuando la información de dichos portales se almacena en bases de datos a las cuales los crawlers de los buscadores generales no pueden acceder y por tanto no podrán indexar esos contenidos.

Por otro lado a la hora de mostrar el ranking de resultados también existen diferencias en los factores que se evalúan para componer dicho ranking. En un buscador general se tiene en cuenta la estructura de la web, es decir, las interconexiones del grafo que componen los enlaces entre páginas, pero no se tiene en cuenta el contenido de las páginas. En cambio, en un buscador vertical se debe tener en cuenta el contenido de la página, ya que los mejores resultados serán aquellos que contengan información más específica para el tema que se trate.

Resumiendo, las ventajas que proporcionan los buscadores específicos o de portal al trabajar sobre contenido concreto son las siguientes:

- Mayor precisión debido a un alcance limitado
- Aprovecha el conocimiento del dominio incluyendo taxonomías y ontologías.
- Apoyo a las tareas específicas de usuario único.
- Economiza el tiempo del usuario, reduciendo la navegación innecesaria.

"Las soluciones de búsqueda de dominio-específico se centran en un área del conocimiento, creando experiencias de búsqueda personalizadas, que a causa del limitado cuerpo del dominio y las claras relaciones entre los conceptos, proporcionan resultados muy relevantes para los buscadores." John Battelle.

4. Aportaciones bibliográficas

Philip O'Brien y Tony Abou-Assaleh proponen en una herramienta de crawling enfocada u orientada a determinar aquellas páginas que pertenecerían a una categoría concreta dentro de una clasificación para así conseguir un conjunto de páginas más específico y relevante para un tema particular y a la vez mejorar el ancho de banda y la cantidad de datos almacenados. Además proponen un algoritmo de clasificación enfocado u orientado a distribuir los documentos de distintas categorías en subtemas, el cual les permite generar un índice que comparado con el algoritmo PageRank y el algoritmo TSPR (Topic Sensitive PageRank) ofrece mejores datos de precisión que el primero de ellos y equivalentes para el segundo, pero permite reducir el coste computacional y de almacenamiento.

Ismail Sengor Altingovde et al. describen en la arquitectura de un sistema automático de construcción de portales web especializados. Consta de tres componentes principales: un crawler enfocado que recoge las páginas específicas del dominio, un motor de extracción de información que recoge los campos útiles de esas páginas y un motor de búsqueda que permite las típicas consultas basadas en palabras clave y búsquedas avanzadas en los campos extraídos. El sistema aumenta y actualiza con la cantidad de páginas recolectadas haciendo rastreos periódicos. Además presentan como prototipo un buscador de cursos y realizan un estudio en el que muestran como su sistema produce resultados de mejor calidad y logra una precisión mayor que los sistemas de búsquedas basados en palabras clave.

Diligenti et al. proponen un marco general para la definición de sistemas de puntuación de páginas web tanto para buscadores horizontales o generales como buscadores verticales o específicos que incorpora y extiende muchos de los modelos establecidos en la literatura pero además contiene importantes características orientadas sobre todo a los buscadores específicos o verticales. En particular, la estructura topológica de la web, así como el contenido de las páginas web de juegan de forma conjunta un papel crucial para el cálculo de la puntuación. Los resultados experimentales apoyan la eficacia de la propuesta que emerge claramente en especial para la búsqueda vertical. Por último, vale la pena mencionar que el modelo descrito en su trabajo es muy adecuado para la construcción de sistemas de puntuación de páginas web basados en el aprendizaje, lo que puede, en principio, incorporar la información del usuario mientras navega por la Web.

5.Propuesta de mejoras

Dentro del crecimiento de la Web y de la cantidad de información contenida en ella, los recursos de audio y vídeo cada vez son más abundantes. La búsqueda en portales compuestos por estos recursos se podría mejorar notablemente si se incluyera el reconocimiento de voz para indexar el contenido de audio de los recursos. Normalmente en estos portales se utilizan palabras clave o tags que describen el contenido de los recursos y la búsqueda se realiza sobre dichas palabras clave, por ejemplo en Youtube4. En cambio si se indexara el contenido del audio de los recursos se podría devolver como resultados de la búsqueda aquellos recursos en cuyo audio apareciera exactamente lo que el usuario está buscando, ya que el éxito de la búsqueda sobre palabras clave depende de lo apropiados que sean los tags que describen al vídeo, los cuales suelen ser establecidos por el usuario que carga el vídeo, en el caso de Youtube, o por el administrador del portal.

La web semántica es otra de las mejoras que pueden hacer que el usuario esté más satisfecho con los resultados que le presenta un motor de búsqueda. La web semántica trata de entender al usuario y ofrecerle los resultados más apropiados para sus necesidades. Aunque la web semántica es transparente para el usuario algunos sitios han empezado a implementarlo. Amazon5 fue uno de los primeros portales que lo pusieron en práctica sugiriendo productos relacionados con los intereses del usuario basándose en sus preferencias. Cuando un usuario compra un libro Amazon sugiere otros libros que podrían ser de su interés en función de las compras realizadas por otros usuarios que también compraron ese mismo libro. El resultado para Amazon fue un aumento masivo de las ventas así como de las ganancias. Google también está mejorando sus resultados de búsqueda haciendo reordenamientos basándose en el historial de navegación del usuario.

Bibliografía

[1] Battelle, J., The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture. New York: Portfolio. 2005

[2] Nielsen, J., Search and You May Find,1997 Disponible en: http://www.useit.com/alertbox/9707b.html

[3] Atlingovde, I., Ozcan, R., Cetintas, S., Yilmaz, H, Ulusoy, O., An Automatic Approach to Construct Domain-Speci_c Web Portals, CIKM'07, pp. 849-852, 2007

[4] Hassan, Y., Buscador Interno. En: No Solo Usabilidad, nº 1, 2002. <nosolousabilidad.com>. ISSN 1886-8592.

[5] Vicens, T., Normas básicas de usabilidad para buscadores internos. 2002 Disponible en:

http://www.evolucy.com/esp/columns/20021024_usabilidad_buscadores.html

[6] O'Brien, P., Abou-Assaleh, T.. Focused ranking in a vertical search engine.

In SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, pages 912_912, New York, NY, USA, 2007

[7] Gantz, J. F., Mcarthur, J., Minton, S. The Expanding Digital Universe. Director, 285(6). IDC. 2007. Disponible en: http://www.emc.com/collateral/analyst-reports/expanding-digital-idcwhite-paper.pdf.

[8] Diligenti, M., Gori, M., Maggini, M., Web Page Scoring Systems for Horizontal and Vertical Search, In 11th International World Wide Web Conference, WWW 2002