Resumen Tema 3: Búsqueda

José Alberto Benítez Andrades

Enero 2011

En este trabajo se resumen las conclusiones obtenidas después de haber realizado la lectura de los artículos propuestos de Steve Lawrence y C. Lee Giles, *Searching the World Wide Web*, Nick Craswell, David Hawking y Stephen Robertson, *Effective Site Finding using Link Anchor Inofrmation* y Dunja Mladenic, *Text-Learning and Related Intelligent Agents: A survey.*

1. Características propias de la web que afectan a la búsqueda.

En primer lugar, debemos destacar que Internet, y concretamente la web, es una fuente de información joven, distribuida, de carácter dinámica, y que crece de una manera muy rápida. Para poder obtener y recuperar la información, no podemos mirar atrás y utilizar las tecnologías antiguas, debido a que, fueron creadas en su momento, para poder indexar colecciones de documentos estáticos, no dinámicos, directamente accesibles.

La web posee una naturaleza que hace cuestionarse si la arquitectura que poseen los buscadores, centralizada, puede mantener la cantidad tan grande de documentos que hay actualmente en la red, y sobre todo, si son capaces de actualizar sus bases de datos, de forma que, detecten la información que es modificada, borrada o insertada. Las respuestas a estas preguntas impactan en la mejor metodología de búsqueda a seguir y en el futuro de la tecnología de búsqueda en Internet. Los buscadores poseen una cobertura bastante limitada, ya que, ningún buscador llega a indexar una tercera parte del total de las páginas que existen realmente en Internet.

Existen muchos problemas que afectan a los buscadores a la hora de poder convertirse en motores de búsqueda potentes, caben destacar los siguientes puntos principalmente:

- Gran número de webs que existen actualmente en Internet: Hace 10 años, había un número bastante elevado de webs en todo Internet, pero en la época actual, ese número se ha elevado de manera exponencial. Hay demasiada información, de la cual, una parte de ella, es además SPAM, y no interesa en los buscadores porque no proporciona la información que necesita el usuario.

- Crecimiento continuo de las webs: A diario, crece el número de páginas web que crean distintos tipos de personas. Ya sean, webs particulares, de negocios, blogs de opinión, etcétera. Siempre se recomienda que mediante distintas herramientas, se introduzcan las webs en distintos motores de búsqueda de forma manual, para facilitar el trabajo a estos. No obstante, muchas personas desconocen este tipo de herramientas.

- Velocidades de los rastreadores que se utilizan, junto con el hardware necesario para poder ejecutar las diferentes herramientas necesarias para indexar las webs.

- Duplicidad de contenido web: Cuesta bastante tiempo, descargar las páginas, comprobar si es contenido duplicado, y posteriormente borrarlo, pensando en que además de esa acción, se deben insertar URLs nuevas, y cambiar algunas que ya estaban, porque se han modificado.

2. Tipos de información a considerar en la búsqueda en web:

2.1. Contenido Textual.

La manera más común de expresar y comunicar la información o el conocimiento sobre alguna materia en la red, es mediante el texto. El texto, puede codificarse en bits de dos formas principalmente:

- EBCDIC y ASCII, al principio con 7 bits y posteriormente 8
- Unicode: Posee 16 bits y se utiliza para acomodar los lenguajes orientales.

Los sistemas de recuperación de información, deben recuperar la información en varios formatos debido a que, no existe un formato único, sino que existen varios formatos de texto. Hace años, se convertían los documentos necesarios, pero en la actualidad se utilizan una serie de filtros para evitar estas conversiones. Algunos de los formatos que se utilizan en los documentos son: formato para intercambio de documento (RTF), formato para mostrar (PDF, PostScript), formato para codificación de correo (MIME), ficheros comprimidos, uuencode/uudecode, binhex. La cantidad de información tiene una relación con la distribución de símbolos en el documento.

La entropía, en la teoría de la información, es una magnitud que mide la información provista por una fuente de datos, es decir, lo que nos aporta sobre un dato o hecho concreto.

Por ejemplo, si nos dicen que todos los comercios están cerrados, en un domingo, no nos aporta nada nuevo porque ya sabemos que los domingos se descansa, sin embargo, si nos dicen un día entre semana, que los comercios están cerrados, significa que nos encontramos en un día festivo.

La medida de la entropía puede aplicarse a fuentes de información de cualquier naturaleza, y nos permite codificarla adecuadamente, indicándonos los elementos de código necesarios para transmitirla, eliminando toda redundancia. (Para indicar el resultado de una carrera de caballos basta con transmitir el código asociado al caballo ganador, no hace falta contar que es una carrera de caballos ni su desarrollo).

La entropía nos indica el límite teórico para la compresión de datos.

Su cálculo se realiza mediante la siguiente fórmula:

$$H = \sum_{i=1}^{m} p_i log_2 p_i$$

donde H es la entropía, las p son las probabilidades de que aparezcan los diferentes códigos y m el número total de códigos. Si nos referimos a un sistema, las p se refieren a las probabilidades de que se encuentre en un determinado estado y m el número total de posibles estados.

Se utiliza habitualmente el logaritmo en base 2, y entonces la entropía se mide en bits.

Por ejemplo: El lanzamiento de una moneda al aire para ver si sale cara o cruz (dos estados con probabilidad 0,5) tiene una entropía:

$$H = \frac{1}{2} log_2 \frac{1}{0,5} + \frac{1}{2} log_2 \frac{1}{0,5} = (0,5 + 0,5)\, log_2 2 = 1bit$$

A partir de esta definición básica se pueden definir otras entropías.

Modelado del lenguaje natural: Los símbolos son los encargados de separar a las palabras o son parte de ellas, no poseen una distribución uniforme, se trata de un modelo binomial y dependen de los símbolos previos, el modelo Markov de orden K. Las palabras pueden ser tomadas por símbolos.

La llamada Ley de Zipf, formulada en la década de los cuarenta por el lingüista de Harvard George Kingsley Zipf, es una ley empírica según la cual, por ejemplo, en una lengua la frecuencia de aparición de las distintas palabras sigue una distribución que puede aproximarse por

$$P_n \sim 1/n^a$$

donde Pn representa la frecuencia de una palabra ordenada n-ésima y el exponente es próximo a 1. Esto significa que el segundo elemento se repetirá aproximadamente con una frecuencia de $1/2$ de la del primero, y el tercer elemento con una frecuencia de $1/3$ y así sucesivamente. Una ley no empírica, pero más precisa, derivada de los trabajos de Claude Shannon fue descubierta por Benoît Mandelbrot.

De manera similar a la ley de Zipf, existe otra ley empírica que describe el comportamiento de los términos dentro de un texto escrito denominada ley de Heaps. En esta ley, se plantea una relación entre el tamaño del texto (cantidad de palabras) y el crecimiento del vocabulario (cantidad de palabra únicas). En particular, postula que el tamaño del vocabulario (y su crecimiento) es una función del tamaño del texto.

$$V = Kn^\beta$$

donde:
N: Es el tamaño del documento (cantidad de palabras)
K: Constante que depende del texto, típicamente entre 10 y 100.
β: También es una constante que depende del texto, donde 0 <>

- 10 <= K <= 20
- 0.5 <= β <= 0.6

Por lo tanto, si K = 20 y beta = 0.5, resulta:

N	V
10000	6325
25000	10000
40000	12649
80000	17889
100000	20000

Nótese que el tamaño del corpus creció 10 veces, mientras que el vocabulario apenas superó las 3 veces su tamaño inicial.

Los resultados de la ley de Heaps plantean que a medida que se incorporan documentos a una colección, cada vez se descubrirán nuevos términos para el vocabulario.

Su aplicación es directa ya que permite estimar el tamaño del vocabulario con lo cual se puede determinar – por ejemplo – la escalabilidad de las estructuras de datos necesarias para almacenar los índices que soportan el SRI. Esto es altamente útil si se utilizará una tabla de hash en memoria para el índice.

2.2. Información en los enlaces (link anchor information).

Un hipervínculo, según el documento *Effective Site Finding using Link Anchor Information*, se trata de una relación entre dos documentos o dos partes del mismo documento. El documento origen es el que contiene el enlace. En una web, el documento fuente contendría un texto como: < a href="http://www.indipro.es"> INDIPRO Site

Llamaremos documento objetivo a aquel al que se refiere el enlace, es decir, a http://www.indipro.es. Los métodos que utilizan el ranking basado en hipervínculos se dividen principalmente en tres tipos:

1. La suposición por **recomendación** : Al vincular un objetivo, un autor de la página es recomendado. Acorde con esto, una página es más recomendable, cuanto más alto es el grado que tiene y debe estar mejor clasificada. Esto se puede basar en un conteo de enlaces simples o en el cálculo de un peso por la propagación de la página iterativa. Un estudio reciente encontró que los métodos de recomendación de enlaces hacen un buen trabajo recolectando objetos muy interesantes, dicho por expertos. Los jueces de calidad son muy necesarios para establecer las recomendaciones. Sin embargo, hay muchos criterios de evaluación de enlaces, para unos usuarios pueden ser unos enlaces mejores que otros.

2. La suposición de "**localización de tema**" : las páginas conectadas por enlaces tienen más probabilidad de ser del mismo que tema que las que no lo están. Distintos estudios del año 2000 encontraron que esto era cierto. Usando estos métodos, una página que está enlazada a páginas relevantes, puede ser rankeada de mejor manera.

3. La suposición de **descripción de anclaje** : El texto de anclaje de un enlace describe su objetivo. Usando el enlace mencionado anteriormente, el texto de anclaje "INDIPRO Site" está describiendo http://www.indipro.es/. Este método produce que se indexe por la descripción de anclaje, el enlace de destino, de forma razonable.

Los métodos de rankeado que utilizan muchos motores de búsqueda comerciales, incluido Google, incorporan la recomendación de enlaces. Google también asocia textos de anclaje con sus documentos de destino.

2.3. Estructura de enlace entre páginas.

Los hipervínculos que existen en cada página web, son muy importantes. Porque, el significado que interpretan los motores de búsqueda de los hipervínculos, es que, el autor de la página web, está recomendando las páginas a las que él enlaza desde su web. Es por eso que, una página que tiene un mayor número de enlaces apuntando hacia ella, tiene un ranking mayor, que una web que posee pocos enlaces. Este tipo de ranking se basa en el cálculo del número de enlaces a cada página junto con el cálculo de una serie de valores que llamamos pesos de cada página por propagación iterativa. También en muchas ocasiones son necesarios los métodos de recomendación de enlaces para evaluarlo de manera correcta.

Uno de los experimentos que realizan Nick Craswell y David Hawking en su estudio, consiste en comparar los métodos de rankeo por texto de anclaje y por contenido. La manera de determinar en qué posición se encuentra una página web, por un término concreto, en el método de texto de anclaje, es bastante clara. Realizan la prueba teniendo una colección de enlaces vasados en un conjunto VLC2. Para determinar qué opción es la más correcta buscan el texto "excite" y encuentran que hay 11.000 enlaces a la página http://www.excite.com, de los cuales, 7332 tienen como texto de anclaje "excite", 910 "excite netsearch", 294 enlaces "http://www.excite.com".

La página www.excite.com ha ganado su posición por el término "excite" porque tiene muchos enlaces a ella con ese término como palabra de anclaje descriptiva.

También es cierto, que a lo largo de los años se ha jugado con este tipo de rankeo y se han conseguido hacer, con la ayuda de comunidades de webmasters, cosas verdaderamente curiosas. Existe el caso de que hace unos años, la Sociedad General de Autores Españoles (SGAE) tuvo enlaces a su web teniendo como término de anclaje "ladrones", de tal forma que, al haber un número tan elevado de enlaces a su site con ese término, si buscábamos en Google "ladrones" aparecía su página posicionada en la primera posición.

Es por ello, que los buscadores no tienen en cuenta únicamente este método a la hora de realizar el ranking de páginas, debido a los webmasters que utilizan trucos, de hecho, en la actualidad se penalizan este tipo de webs en cuanto son detectadas.

2.4. Otros.

La suposición de la localización de tema - topic locality - surge del intercambio de enlaces entre páginas de la misma temática, de forma que, los enlaces que haya en una web que trate una temática, supuestamente, tratarán sobre la misma temática. Utilizando estos métodos, una página que es adyacente por medio de un enlace a páginas probablemente relevantes pude tener una posición más alta en el ranking.

3. Proceso de indexación de la información en web.

Es el encargado de la caracterización de documentos para la tarea de recuperación de información. De forma ideal, un documento indexado debería funcionar como una representación de los contenidos semánticos de un documento original (o consulta).

Por lo general, la indexación tiene como objetivo conseguir una lista de términos con significado (conceptos) con información asociada (frecuencia en el documento, frecuencia en la bbdd, concurrencia). Un término es una palabra, en ocasiones reducida a su forma raiz por algún algoritmo de lematización, pero puede ser una frase, un nombre propio o incluso expresiones especiales tales como fecha, lugares, etc.

Los términos se reconocen con técnicas relacionadas con el lenguaje. En la recuperación de información, un documento es un conjunto de cadenas concatenadas sin tener en cuenta las propiedades del lenguaje natural de los documentos. Pero este enfoque tiene una serie de inconvenientes.

- Los algoritmos de lematización no extraen la forma base de las palabras vía análisis morfológico, así que pueden fallar en identificar variaciones de los términos en lenguajes con una morfología solo ligeramente más compleja que el Inglés.
- La ambigüedad léxica es ignorada en general, lo que supone un problema al no poder distinguir diferentes significados de una misma palabra.
- Tiene errores a la hora de relacionar palabras sinónimas.

La indexación es un proceso lento y costoso, pero sólo debe ejecutarse en el momento de crear el índice invertido, aunque es muy importante actualizarlo cuando haya modificaciones. La indexación consta de los siguientes pasos:

- Recorrer todos los documentos sobre los que queremos buscar. Este puede ser un conjunto finito y conocido, por ejemplo: las páginas HTML de una carpeta, o puede ser desconocido: todas las páginas de Internet. En este último caso, los buscadores utilizan los que se denominan robots, crawlers o también spiders: Son pequeños programas que van rastreando la estructura de la web en busca de nuevas páginas. Este tipo de programas van recolectando páginas a través de los enlaces a otras páginas, así en un ciclo interminable. Una vez que el robot localiza una página, puede procesarla a través del siguiente paso.

- Procesar el documento a indexar: Se descompone hasta obtener la lista de palabras que lo forman. Este proceso puede ser muy sencillo, como en los archivos de texto plano, que tiene todas las palabras separadas por espacios u otros caracteres, o muy complejo, como un documento PDF que debe ser decodificado, separando el formato y las imágenes, extrayendo solamente el texto plano.

- Con la lista de palabras de un documento, creamos el índice invertido: Guardamos la lista apuntando en qué documento hemos encontrado cada palabra. Habrá algunas palabras que solo aparezcan en un documento (la búsqueda de esa palabra nos dará un solo resultado), mientras que otras palabras más comunes aparecerán en muchos documentos.

- Opcionalmente podemos almacenar el documento, en su estado actual, en la propia base de datos. De esta forma, podemos consultar cualquier documento aunque el original deje de estar disponible (como hace Google con su caché).

Lematización

El **stemming o la lematización** es un método para reducir una palabra a su raíz o (en inglés) a un stem o tema. Hay algunos algoritmos de stemming que ayudan en sistemas de recuperación de información. Stemming aumenta el recall que es una medida sobre el número de documentos que se pueden encontrar con una consulta. Por ejemplo una consulta sobre "bibliotecas" también encuentra documentos en los que solo aparezca "bibliotecario" porque el stem de las dos palabras es el mismo ("bibliotec").

El algoritmo más común para stemming es el algoritmo de Porter. Existen además métodos basados en análisis lexicográfico y otros algoritmos similares (KSTEM, stemming con cuerpo, métodos lingüísticos...).

Snowball es un pequeño lenguaje de programación para el manejo de strings que permite implementar fácilmente algoritmos de stemming. Se puede generar código en ANSI C y Java. Las páginas de Snowball contienen stemmers para 12 idiomas (incluido el castellano , [[|idioma catalán|catalán]] y euskera). Todas las explicaciones, sin embargo, son dadas en inglés.

Desde hace poco tiempo Google utiliza stemming al igual que MSN search (donde tiene que activarse explícitamente). En general, los buscadores comerciales no dan muchas explicaciones sobre los algoritmos utilizados.

Extracción de palabras clave

Los términos que nos proporciona el lematizador son utilizados como términos para la indexación de los documentos. Se asigna un peso acorde con la frecuencia a cada término en cada documento y en la colección.

Existen dos tipos de frecuencia:

- Frecuencia del término tf(t): Número de ocurrencias de la palabra (t) en el documento. Cuanta más veces se repita un término, más relación va a tener el documento con ese término concreto.

- Frecuencia en el documento f(t): Número de documentos que se indexan por t, es decir, el total de documentos que contienen el término. Cuanto menos aparezca un término en los documentos, más discriminatorio es.

Para valorar el peso que tiene cada término se usa la siguiente fórmula:

$$tf(t)*log(N/f(t))$$

N: Representa el número total de documentos que hay en la colección.

4. Interfaces, browsing y visualización de la búsqueda.

La búsqueda de información es un proceso poco preciso debido a que los usuarios no saben exactamente cómo realizar la búsqueda necesaria para encontrar lo que desean con los términos que quieren. Es por ello que se necesitan interfaces de usuario que ayuden al usuario a solventar esta falta de conocimiento. Una de las infercaes más importantes para la recuperación de información es la interfaz hombre-máquina.

Ben Shneiderman dijo unas sabias palabras: "*Cuando un sistema interactivo esta bien diseñado, el interfaz casi desaparece permitiendo a los usuarios que se concentren en su trabajo, exploración o placer*".

Existen unos principios de diseño, comenzando por el **feedback o la retroalimentación**.

La retroalimentación es importante para los interfaces de acceso a la información. Gracias a ella, se realizan elecciones importantes de diseño como por ejemplo, qué operaciones van a realizarse de forma automática por el sistema y cuáles deben iniciarlas y controlarlas los usuarios.

También se consigue reducir la carga de la memoria. Pero, ¿ cómo podemos reducir el acceso a la información ? Hay dos métodos importantes:

- Proporcionando mecanismos para guardar constancia de las decisiones que se realizan durante el proceso de búsqueda.

- Proporcionando información navegable que sea relevante al estado actual del proceso de acceso a la información.

Por otra parte, hay una serie de interfaces alternativos para usuarios expertos y noveles, que enfrentan la simplicidad contra la potencia y que ofrecen puentes que son más o menos avanzados dependiendo del usuario al que vayan dirigidos.

Otra decisión importante que debe tomarse a la hora de realizar un diseño, es la cantidad de información que queremos mostrar al usuario mediante el sistema de acceso a la información.

En cuanto al papel de la **visualización** tiene las siguientes características:

- Contiene la interfaz de objetos : las ventanas, los menús, iconos y cajas de diálogo que utiliza estos sistemas.

- La visualización de información proporciona descripciones visuales de espacios de información que son bastante grandes.

- Las personas, por lo general, tenemos una gran costumbre a los objetos visuales y a las imágenes en general.
- En lo que se refiere a la visualización de información científica, es posible.
- Sin embargo, en relación a la visualización abstracta de información, es imposible.
- Para representar el proceso de acceso a la información, se utilizan técnicas de visualización de información como por ejemplo el buscador http://www.kartoo.com/, que devuelve resultados como una serie de mapas en flash, en los cuales, el documento más relevante, es el más grande.

Algunas de las técnicas de visualización existentes en la red y en la web 2.0 son:

- La muestra de iconos y cambios en el colorido.
- Mostrar enlaces entre distintos documentos.
- Capacidad de poder realizar zooms.
- Lentes mágicas.

En el artículo de Alan J.Dix, Janet E. Finlay, Gregory D. Abowd, Russell Beale y Prentice Hall, *Human Computer Interaction*, se ofrece una métrica de evaluación, no para la web sino para todo el sistema de interacción, formado por 10 elementos (p. 414):

1. Visibilidad del estado del sistema. El sistema debe mantener al usuario informado de lo que está pasando.
2. Correspondencia entre el sistema y el mundo real.
3. Control del usuario y libertad. Soportar deshacer y rehacer.
4. Consistencia y estándares.
5. Prevención de errores.
6. Reconocimiento mejor que tener que confiar en la memoria.
7. Flexibilidad y eficiencia de uso.
8. Diseño estético y minimalista.
9. Ayudar a los usuarios a reconocer, diagnosticar y recuperarse de sus errores. Los errores deberían ser expresados en lenguaje sencillo - Sin códigos, indicar el problema con precisión y sugerir constructivamente una solución.
10. Ayuda y Documentación. Incluso si es mejor que el sistema pueda ser usado sin documentación, puede ser necesario proporcionar ayuda y documentación. Esta información debería ser fácil de buscar y estar enfocada a la tarea del usuario, listar pasos concretos y no ser muy grande.

5. Metabúsqueda.

Un metabuscador es un motor de búsqueda que envía una solicitud de búsqueda a otros múltiples buscadores o bases de datos, retornando un listado con los resultados de búsqueda o un listado de enlaces para acceder a los resultados individuales de cada buscador de forma fácil.

Los metabuscadores permiten a sus usuarios ingresar criterios de búsqueda una sola vez, y acceder a múltiples buscadores de forma simultánea.

Los metabuscadores no suelen tener una base de datos propia, sino que simplemente emplean los resultados de otros buscadores, generalmente unificándolos empleando algoritmos propios para ordenarlos en relevancia (por lo general, eliminando aquellos resultados idénticos).

Los metabuscadores suelen entregar resultados de páginas web de la WWW, pero también existen algunos específicos que buscan en foros de discusión, grupos de noticias, weblogs, imágenes en la web, documentos gratuitos o libres en la web, etc.

Los pasos del funcionamiento de un metabuscador:

- El usuario realiza su petición al metabuscador.
- El metabuscador formatea dicha petición de acuerdo a la interfaz de cada uno de los buscadores y les pasa la petición.
- Los buscadores realizan la búsqueda utilizando sus medios habituales a partir de los sitios web en internet.
- Éstos devuelven la información obtenida al metabuscador, el cual analiza los datos.
- El metabuscador organiza la información de acuerdo a los criterios del mismo y se la muestra al usuario.

Algunos metabuscadores que hay en la actualidad son los siguientes:

Vivisimo (http://www.vivisimo.com)

Sus fuentes son los principales buscadores internacionales, Alltheweb, Yahoo y MSN entre otros y presenta los resultados agrupados automáticamente por categorías. A pesar de estar en inglés es muy fácil de utilizar.

IxQuick (http://www.ixquick.com)

Combina los resultados basándose en los 10 primeros sitios web recibidos de los diferentes buscadores. Sus principales fuentes son Alltheweb, ODP (Open Directory Project) y MSN, entre otros. Este buscador se encuentra en Español.

Lomejor (http://www.lomejor.com.ar)

Este metabuscador orientado para la búsqueda de contenidos en español dispone entre sus fuentes los mejores buscadores internacionales, españoles y argentinos. Como son Google, Terra Argentina, Yahoo, Alltheweb y Altavista, entre otros.

1Blink.com (http://www.1blink.com/)

Multibuscador de páginas rápido y eficaz. Las búsquedas proceden de AltaVista, Thunderstone, Wisenut, Lycos, Yahoo!, Fast, Looksmart.

Botfeeder (http://www.botfeeder.com/)

Multibuscador que proporciona la opción de búsquedas temáticas (imágenes, noticias, multimedia, referencia, bitácoras o compras).

Clusty (http://clusty.com/)

Potente metabuscador, agrupa los resultados por categorías.

Creative Search (http://creativesear.ch/)

Servicio que combina las búsquedas de Google, Wikipedia, Noticias, Blogs, Flicker, la librería Amazon e iTunes.

ez2Find (http://ez2www.com/)

Metabuscador potente con eficacia y depuración en la presentación de resultados, incorpora también servicio de noticias.

GoingInto.Com (http://www.goinginto.com)

Servidor que facilita consulta de metabúsqueda.

Ilectric (http://ilectric.com/)

Servidor que concentra en unos pocos recursos amplias capacidades de búsqueda. Incorpora un metabuscador que combina la búsqueda de texto e imágenes entre los buscadores genéricos más importantes.

Metacrawler (http://www.metacrawler.es)

Buscador simultáneo en Google, Altavista, Yahoo, Teoma, Lycos, etc.

Infonetware (http://www.infonetware.com/)

Metabuscador que interroga a los principales buscadores y bases de datos de noticias que muestran los resultados estructurados y sin duplicaciones. Ofrece la opción de acotar los resultados propuestos agrupados por términos o palabras clave.

Otros metabuscadores

- Biwe (http://multibuscador.biwe.com)
- Buscamultiple (http://www.buscamultiple.com)
- Dogpile (http://www.dogpile.com)
- I-Une (Fuente: Consoft http://www.consoft.es)

6. Agentes web.

Es una realidad, que cada día, a un gran número de personas, nos falta tiempo para hacer todas las cosas que queremos hacer, proyectos, aprendizaje de materias nuevas, lectura de novedades en el mundo, etc. Internet se ha convertido en una gran herramienta, que facilita enormemente estas tareas y consigue que tengamos toda la información que necesitamos, a golpe de ratón. La capacidad de poder leer periódicos online, comprar billetes de avión, reservar hoteles, restaurantes, etc.

También es cierto, que cada usuario tiene una serie de intereses particulares y organizan la información que necesitan en la red, a su manera. Es por ello, que muchas veces se hace necesaria la existencia de una herramienta que nos facilite obtener la información que realmente necesitamos, sin perder el mayor tiempo posible realizando búsquedas.

Para poder conseguir esto, existen las *técnicas de aprendizaje automático* y los *agentes inteligentes,* o *agentes web* que utilizan estas técnicas para conseguir su objetivo. En el texto de *Dunja Mladenic* y *J. Stefan Institute*, se estudian estas herramientas y aplicaciones.

Principalmente, existen dos métodos para poder conseguir ayudar al usuario a encontrar la información deseada, con los asistentes de usuario o sistemas de recomendación: **basados en contenido** y **colaborativos**.

El enfoque de los agentes que utilizan el método **basado en contenido**, es el siguiente:

- Para clasificar el texto, el sistema busca objetos similares a los que tiene el usuario comparándolo completamente por el contenido de cada uno. El objetivo tiene sus raíces en la recuperación de información. Un problema principal de los agentes que usan este método, es encontrar contenido que ya ha sido visto anteriormente y descartarlo.
- Este tipo de agentes, utilizan diferentes métodos de **text-learning**.
- El método basado en contenido es muy popular en los sistemas que trabajan con datos que son texto, por ejemplo, en documentos web o noticias.
- Existen distintos tipos de agentes que utilizan este método, como por ejemplo:
    - WebWatcher: ayuda al usuario a encontrar documentos similares en la web, en función de una serie de palabras clave que escribe el usuario en sus búsquedas comunes. Las rastrea, salva el contenido que ve el usuario, y busca información relacionada.
    - Lira: el sistema Lira, aprende mientras navega el usuario por internet. Sus búsquedas son almacenadas y el agente selecciona las mejores páginas y recibe información por parte del usuario, que evalúa los enlaces que ve. Utiliza métodos heurísticos.
    - Musag: este sistema, funciona creando un diccionario con las palabras más relevantes en las búsquedas que realiza el usuario, y busca las páginas relacionadas en función de la repetición de estas palabras en el contenido de los artículos que ha leído el usuario.
    - Letizia: el sistema ofrece nuevos enlaces al usuario, habiéndose fijado en la búsqueda que se ha realizado, mostrando la información de los enlaces sugeridos, en una nueva pestaña.
    - Personal WebWarcher: el usuario realiza una búsqueda, y cuando accede a un artículo, el agente examina la web y los enlaces que posee, marcando los que son más relevantes, incitando al usuario a que los visite.
    - Algunos otros son CiteSeer, NewsWeeder, ContactFinder, FullFinder e Internet Fish.

Además de estos agentes anteriores que se basan en el contenido, existen otros que utilizan el método **colaborativo**:

- En el enfoque colaborativo, hay un conjunto de usuarios que usan el sistema.
- Está basado en el *Social learning.*
- No analiza el contenido, analiza la evaluación que realizan los usuarios con cada enlace, en función de los intereses de cada uno.
- Este tipo de enfoque, se utiliza sobre todo para documentos no textuales, como por ejemplo, vídeos, imágenes, películas, etc.
- Algunos ejemplos de agentes que utilizan este método son los siguientes:
    - Firefly y Ringo: se utiliza para la recopilación de música, teniendo en cuenta el tipo de música que le gusta a cada usuario y las votaciones.
    - Siteseers: Sistema de recomendación de páginas web, que posee marcadores individuales.
    - Phoaks: Sistema que reconoce automáticamente y distribuye recomendaciones de recursos Web minados desde nuevos mensajes desde Usenet.

- GroupLense: Tiene una segunda base de datos para almacenar puntuaciones que los usuarios han dado a los mensajes y correlaciones entre parse de usuarios basados en sus puntuaciones.
- Referral Web: Sistema interactivo para reconstrucción, visualización y búsqueda de redes sociales en Internet. Es parecido a ContactFinder.
- Fab sytem: Combina los dos métodos realmente, basado en contenido y colaborativo. Usa el método basado en contenido para crear un perfil de usuario y además cada usuario puntúa los diferentes enlaces.
- WebCobra: también combina los dos métodos como el agente anterior.
- Lifestyle Finder: funciona generando un perfil de usuario, y en función de los intereses de cada uno, se crean grupos de usuarios con los mismos intereses.

En el documento de *Text-Learning and Related Intelligent Agents: A survey* sus autores realizan un análisis más exhaustivo del Personal WebWatcher:

- Conecta al navegador como un proxy.
- Marca mediante variables booleanas los enlaces que tienen más relevancia de la página que va a ver.
- Guarda el contenido destacado y lo usa para sugerir webs.
- Su estructura es la siguiente:
  - Tiene un proxy que recuerda las búsquedas.
  - Mediante el learner, aprende lo que quiere el usuario.
  - Muestra la web que quería visualizar el usuario, marcando los enlaces más interesantes que posee la web.

7. Áreas de investigación relacionadas con la búsqueda en web.

La búsqueda en la web, esta relacionada con otra serie de áreas de investigación, cuya temática principal está unida al funcionamiento de los motores de búsqueda principalmente:

- Recuperación de la información
- Procesamiento del lenguaje natural.

Además de las dos anteriores, basándonos en el estudio realizado en este trabajo, y teniendo en cuenta las distintas fases por las que se debe pasar para tener una buena búsqueda en la web, está también relacionado con las siguientes áreas de investigación:

- Contenidos textuales: leyes de Zipf y Heaps entre otras, que plantean que a medida que se incorporan documentos a una colección, cada vez se descubrirán nuevos términos para el vocabulario.
- Información en los enlaces y sus métodos: Métodos de recomendación, localización de tema y descripción de anclaje. Técnicas que hay para "rankear" webs.
- Indexación web: y lo que supone con ella, lematización y extracción de palabras clave.

- Navegabilidad por la red.
- Metabuscadores: buscadores que utilizan otros buscadores para poder encontrar contenidos específicos.
- Agentes web, o asistentes del usuario: aplicaciones que ayudan a los usuarios a encontrar en la red todo lo que es de su interés, de una manera más rápida y totalmente automatizada.

8. Conferencias internacionales donde se aborda la búsqueda en web.

Algunas de las conferencias internacionales que abordan el tema de la búsqueda web, indexación y métodos de indexación son las siguientes:

- International World Wide Web Conference(IW3C2).
- International journal of Computer Networks & Communications (IJCNC)
- International Conference on Internet and Web Engineering
- Interlink Web Design Conference
- International Conference on Web Intelligence, Mining and Semantics
- International Conference on Web-based Learning (ICWL 2010)
- International Conference on Machine Learning (ICML97)
- International Conference on Autonomous Agents (Agents '98)
- International Conference on Web Information Systems and Technologies

Referencias

[1] Steve Lawrence and C. Lee Giles. Searching the World Wide Web. Science vol. 280, 1998.

[2] Nick Craswell, David Hawking and Stephen Robertson. Effective Site Finding using Link Anchor Information. Research and Development in Information Retrieval, SIGIR 2001.

[3] Dunja Mladenic. Text-Learning and Related Intelligent Agents: A survey. IEEE Intelligent Systems, 1999

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information retrieval. ACM Press. Addison Wesley, 1999

[5] Lovins J,B. "Development of a stemming algorithm, M e c h a n i c a l T r a n s l a t i o n a n d C o m p u t a t i o n a l Linguistics".1968; 11(1-2): 22-31.

[6] Porter M. "Analgorithm for suffixstripping , 1980;Program 14(3): 130-137.

[7] Maristella Agosti, Alan Smeaton: Information Retrieval and Hypertext Kluwer Academic Publisher, 1996.

[8] Pertti Vakkari: Relevance and Contributing Information types of Searched Documents in task performance Proc. of SIGIR 2000.

[9] Cristoph Hölscher, Gerhahrd Strube: Web search behavior of Internet experts and Newbies Proc. of WWW9, 2000.

[10] Jon Kleinberg: Authoritative sources in a Hyperlinked environment Proc. of ACM-SIAM Symposium on Discrete Algorithms, 1998.

[11] Codina, Lluis. Marcos, Mari Carmen. "Posicionamiento web: conceptos y herramientas". El profesional de la información, v. 14, n. 2, marzo-abril, 2005. http://www.mcmarcos.com/pdf/2005_posicionamiento-epi-maq.pdf

[12] Lopez Ureña, L. Alfonso Resolución de la ambigüedad léxica en Tareas de Clasificación Automática de Documentos. http://www.sepln.org/monografiasSEPLN/monografiaUrena.pdf

[13] Peñas Padilla, Anselmo. Técnicas lingüísticas aplicadas a la búsqueda textual bilingüe: Ambigüedad, variación terminológica y multingüismo. http://www.sepln.org/monografiasSEPLN/monografiaAnselmo.pdf

[14] Luhn, H.P., "The automatic creation of literature abstracts", IBM Journal of Research and Development, 2, 1pags. 59-165. 1958.

[15] Peña, R., Baeza-Yates, R., Rodriguez, J.V. "Gestión Digital de la Información". Alfaomega Grupo Editor. 2003.

[16] Zipf, G. K. "Human Behaviour and the Principle of Least Effort" Reading, MA: Addison- Wesley Publishing Co. 1949.

[17] Blog de Javier Espadas: http://www.jesba.com/javier/espadas/topologia-de-internet-y-ley-de-zipf/

[18] Blog de Ferbor: http://ferbor.blogspot.com/2007/05/recuperacin-de-informacin.html

[19] SIGNUM. Motores para el procesamiento del español. http://www.lenguaje.com/desarrollo/desarrollo.php

[20] Sociedad Española para el Procesamiento del Lenguaje natural (SEPLN). http://www.sepln.org

[21] TACTWeb 1.0 Home Page. http://tactweb.humanities.mcmaster.ca/tactweb/doc/tact.htm

[22] Alvis. Home Page. http://www.alvis.info/alvis/