Procedimientos para la extracción de palabras clave de páginas web basados en criterios de posicionamiento en buscadores

Autores: Mari Vallez (Universitat Oberta de Catalunya, Universitat Pompeu Fabra) , Cristòfol Rovira (Universitat Pompeu Fabra) , Lluís Codina (Universitat Pompeu Fabra) y Rafael Pedraza (Universitat Pompeu Fabra)

Citación recomendada: Vallez, Mari; Rovira, Cristòfol, Codina, Lluís; Pedraza, Rafael (2010). "Procedimientos para la extracción de palabras clave de páginas web basados en criterios de posicionamiento en buscadores". *Hipertext.net*, 8, http://www.upf.edu/hipertextnet/numero-8/extraccion_keywords.html




Resumen: En este artículo se presenta un proyecto de investigación que tiene como principal objetivo el desarrollo y la exploración del potencial de una herramienta que facilite la asignación semi automática de palabras clave a documentos web. Se describen de forma sintética las principales características y prestaciones de la herramienta que se está construyendo y se analizan las bases teóricas que la justifican. La investigación que planteamos explora las posibilidades de la automatización en la asignación de palabras clave con procedimientos relativamente sencillos basados en modelos masivamente usados en las disciplinas de la recuperación de la información, el posicionamiento en buscadores y las ciencias de la documentación.

Palabras clave: Extracción de palabras clave, Metadatos, Web Semántica, Palabras clave, Metaetiquetas, Keywords, Anotación semántica, Taxonomía, Tesauros, Ontología, Posicionamiento en Buscadores

Sumario

1. Introducción

La World Wide Web representa un universo de información y de conocimiento donde a menudo resulta difícil localizar la información pertinente que necesitamos. Los algoritmos basados en el análisis de enlaces han supuesto una gran mejora en la ordenación de los resultados, sin embargo aún queda mucho camino por recorrer, en especial si se quiere automatizar una parte más amplia del proceso de recuperación de información mediante agentes de búsqueda inteligentes.

La propuesta de la Web semántica (Berners-Lee, 2001) puede representar un gran avance en este ámbito porque propone un cambio de paradigma: transformar la actual web basada casi exclusivamente en lenguaje natural a una web estructurada y organizada, donde los contenidos en lenguaje natural son etiquetados semánticamente de forma explícita para conseguir que las máquinas puedan interpretarlos. De esta forma se facilitaría el procesamiento automático de los contenidos de la web y uno de estos procesos sería precisamente la recuperación de información (Ding, 2005).

El etiquetado y la asignación de metadatos son por tanto elementos básicos del proyecto de la Web semántica, con implicaciones para cualquiera que esté relacionado con la creación y distribución de contenidos en la web. El nuevo paradigma supone una nueva forma de crear contenidos, donde sus responsables deben asumir la tarea de

su etiquetado si quieren que estos sean interpretables semánticamente por los nuevos buscadores y aplicaciones de usuario. En este contexto, surge la necesidad de herramientas que faciliten la creación automática o semi automática de esta metainformación y que asegure su calidad.

En este artículo se presenta un proyecto de investigación que tiene como principal objetivo el desarrollo y la exploración del potencial de una herramienta que facilite la asignación semi automática de palabras clave a documentos web. Esta herramienta estará basada en la extracción de palabras clave de acuerdo con las coincidencias entre el texto del documento analizado y una taxonomía predefinida (pero que siempre podrá ser editada y modificada). Los candidatos a palabras clave que se generan mediante este procedimiento se ordenarán aplicando criterios de relevancia propios de los algoritmos de posicionamiento.

A continuación se describen de forma sintética las principales características y prestaciones de la herramienta que se está construyendo y se analizan las bases teóricas que la justifican.

Esta línea de investigación viene motivada por el actual interés que suscitan las tecnologías semánticas como mecanismo para facilitar y optimizar el acceso a la información (Codina, 2009; Davies, 2009; Kiryakov, 04), contexto donde hay que situar también a el proyecto de la Web semántica del W3C.

2. La herramienta

El objetivo de la herramienta que proponemos desarrollar es facilitar la asignación semi automática de metadatos en forma de palabras clave para representar el contenido temático de documentos web.

A grandes rasgos, su funcionamiento es el siguiente:

- Se procesa el contenido textual de una página web: comparando los términos del documento con los términos de la taxonomía elaborada previamente y del mismo ámbito temático de la página analizada.
- Los términos de la página que aparecen también en la taxonomía se seleccionan: como candidatos a palabras clave para representar el contenido del documento.
- Se asigna una puntuación a cada candidato a palabras clave: en función de los criterios de relevancia habitualmente utilizados en los algoritmos de posicionamiento, como por ejemplo el número de ocurrencias del término o el hecho de estar presente en zonas preeminentes de la página web como el título (elemento <title>), los encabezados (elementos <h1>, <h2> ...), los enlaces (elemento <a ?>), la url o bien que hayan sido marcados con etiquetas de énfasis (como las negritas).
- Se ordenan los candidatos a palabras clave de acuerdo con la puntuación de relevancia obtenida por cada uno de ellos: el sistema permite asignar automáticamente las palabras clave al documento analizado a partir de una determinada puntuación, marcada como umbral, o de seleccionar manualmente las mejores palabras de la lista de candidatos.
- El conjunto de palabras clave seleccionado puede pasar a formar parte de las metadatos del documento en alguno de los formatos habituales de la web, como por ejemplo, metadatos Dublin Core como parte del código fuente del documento, en formato RDF como archivo externo, etc.

A partir de aquí, la asignación de un conjunto de palabras clave pertinentes a un documento web tiene tres consecuencias importantes:

- Facilita la representación y el acceso a la información. El conjunto de palabras clave codificado en algún formato de metadatos es una forma de representación sintética del documento de gran capacidad semántica. Ayuda a acceder a la información, ya que facilita la búsqueda por conceptos (Douglas, 2006). Las palabras clave obtenidas a partir de un lenguaje controlado, tal como una taxonomía, son una forma de conseguir que el emisor, el autor de los contenidos, y el receptor, en este caso quien busca información, estén más cerca. Es una propuesta para solucionar una parte importante del problema que implica la variedad lingüística (sinonimia y polisemia) propia de la lengua natural. En este momento no hay constancia de que los buscadores utilicen de forma generalizada los metadatos de palabras clave (keywords) presentes en los documentos web. No obstante, se considera que son un elemento que ayuda al posicionamiento si el contenido del documento está relacionado con las mismas. Además, son un instrumento de recuperación que otorgan mucha calidad en los buscadores internos, no sólo en caso de Intranets sino también en el caso de buscadores internos de sitios web abiertos con grandes volúmenes de información
- Ayuda en el posicionamiento. Cabe destacar que la herramienta que nos planteamos desarrollar resulta también interesante desde la perspectiva del posicionamiento en buscadores. Los candidatos a palabra clave con mayor puntuación serán los términos donde la página debería tener más probabilidades de quedar bien posicionada. Por lo tanto, los autores tendrán una información que les permitirá valorar si es necesario retocar los contenidos para ser procesados por los buscadores de manera más eficiente de acuerdo con sus objetivos.
- Preparación para las nuevas herramientas inteligentes. La página estará mucho mejor preparada para la Web semántica y para que en el futuro pueda ser procesada por agentes inteligentes.

Habría que destacar la importancia de nuestra investigación en relación a la comunicación social, en especial por las dos vertientes principales donde se intenta hacer una aportación: la recuperación de información y el posicionamiento. La actual sociedad de la información nos ha proporcionado nuevos canales de comunicación, un

gran volumen de fuentes de información y potentes instrumentos para el procesamiento de la información (Castells, 1997). La propuesta que se quiere llevar a cabo pretende ayudar a optimizar los procesos de comunicación en este contexto.

2.1. Limitaciones iniciales

A pesar de que la herramienta planteada tiene una vocación polivalente, en una primera fase se propone explorar su eficiencia y eficacia en un contexto limitado marcado por los siguientes elementos:

- Limitación sobre los tipos de documento. Se procesarán documentos de tipo científico o académico con gran cantidad de información textual.
- Limitación temática. Se procesarán documentos relacionados con temáticas vinculadas a las ciencias de la web.
- Limitación sobre el tipo de procesamiento. Se analizará la eficiencia en la asignación de palabras clave en función de los resultados obtenidos en procesos de recuperación de información utilizando los principales buscadores: Google y Yahoo!.
- Limitación de resultados. El planteamiento de la investigación es de tipo exploratorio con la finalidad última de valorar la eficacia y la eficiencia de la herramienta que se propone.

La limitación temática tiene una especial importancia. Una parte importante del éxito en la extracción y asignación de palabras clave vendrá motivada por la calidad del lenguaje controlado utilizado (taxonomía). Como es habitual en el campo de la representación del conocimiento, la limitación a un dominio (en este caso, Ciencias de la Web) nos permite el desarrollo de una taxonomía más completa y por tanto con mayores posibilidades de éxito.

3. Bases teóricas

En nuestra propuesta confluyen diversas disciplinas o áreas que giran en torno a las tecnologías semánticas. En los últimos tiempos este tipo de tecnologías están despertando expectativas. Sin embargo, diferentes agencias internacionales de vigilancia e investigación de las tendencias tecnológicas (Gartner: http://www.gartner.com y Forrester: http://www.forrester.com) muestran en sus informes su baja implantación en las organizaciones. Las abordan con cautela ya que después de años de desarrollo consideran que no han conseguido implantarse suficientemente. De todos modos, reconocen su gran potencial (Gartner, 07A / b) y se considera la Web semántica, tal como la define el W3C, como una tecnología emergente con una penetración de entre el 1 y el 5 por ciento en el mercado y que para su desarrollo a gran escala aún faltan más de 10 años.

En las tecnologías semánticas quedan integradas diferentes áreas temáticas, técnicas y disciplinas de orígenes muy diversos pero todas ellas vinculadas, como por ejemplo: la recuperación de información, el procesamiento del lenguaje natural, la extracción de información, los lenguajes controlados, la anotación semántica, la creación y actualización de ontologías, etc.

La herramienta que presentamos está enmarcada en el contexto de las tecnologías semánticas, y por tanto está relacionada con todos los ámbitos temáticos indicados. Sin embargo, hay tres áreas de interés que tienen una incidencia más directa en la investigación que se está realizando:

- metadatos y anotación semántica
- lenguajes controlados: taxonomías, tesauros y ontologías
- posicionamiento en buscadores

3.1. Web semántica

Uno de objetivos finales de la Web semántica es la creación de un sistema de agentes inteligentes que sean capaces de llevar a cabo inferencias de forma automatizada con la información publicada en la Web. Este objetivo es más una utopía que una realidad incluso en medio plazo (Codina 2006). Sin embargo, muchos de los desarrollos realizados gracias a impulso del nuevo paradigma han dado lugar a nuevos servicios implantados con gran éxito en la actual Web. Uno de los hitos atribuibles a la Web semántica es haber logrado instaurar el uso de diferentes estándares para la representación y el procesamiento de la información de una forma más sofisticada. Son estándares que permiten expresar los metadatos en un formato lógico al mismo tiempo que representan los lenguajes controlados (por ejemplo tesauros u ontologías) para que puedan ser procesados por programas informáticos. Estos formatos ya son utilizados de manera generalizada, como por ejemplo XML, RDF, SKOS-Core y OWL.

Después de asociar unos valores semánticos a los recursos, es necesario disponer de herramientas que faciliten la recuperación eficaz de la información. Estos instrumentos serían los llamados agentes inteligentes, que podrían interpretar y comprender la información para posteriormente facilitarla procesada a los usuarios, sin embargo estas herramientas aún están lejos de ser una realidad.

Estas tecnologías permitirían convertir la web en una infraestructura descrita de forma global donde sería posible compartir y reutilizar datos y documentos entre diferentes tipos de usuarios. Esto debería permitir que los

usuarios recuperen la información que necesiten de forma más precisa de acuerdo con las descripciones de los contenidos.

La investigación que presentamos está situada en el contexto de la migración hacia la Web semántica (Pedraza-Jimenez, 2008). Es un desarrollo para ayudar en una primera fase y tiene por objetivo facilitar la asignación de palabras clave a documentos Web. Afortunadamente no es necesario esperar a un mayor desarrollo de la Web semántica para empezar a disfrutar de las ventajas de este etiquetado. Como veremos más adelante, la asignación de palabras clave genera mejoras inmediatas tanto en la recuperación de información con los actuales buscadores, como en el posicionamiento en sus listados de resultados.

3.2. Metadatos y anotación semántica

Como ya se ha visto, uno de los elementos fundamentales de la Web semántica son los metadatos, es decir, información (datos) que describe los contenidos de los documentos a los que está asociada y representa de forma explícita el significado de estos (Aguado de Cea, 2002).

La anotación semántica realizada con metadatos es la forma de dotar de contenido semántico a los documentos y de conseguir que las máquinas puedan interpretar la información de un dominio específico.

La asignación de metadatos es un proceso complicado, lento y costoso. Una de las tareas que ayudaría en esta fase sería la construcción de herramientas para la extracción de información de forma automática y su posterior conversión en metadatos (Cunningham, 2005).

La extracción de información es el término utilizado para la actividad de extraer automáticamente información específica de textos en lenguaje natural. Existen diferentes aproximaciones para realizar este procesamiento que se pueden agrupar en dos categorías principales: los sistemas de aprendizaje automático (machine learning) y los sistemas basados en reglas y patrones (Flynn, 2007).

Las técnicas de aprendizaje automático se basan principalmente en cálculos probabilísticos a partir de colecciones de entrenamiento. Su adaptación a diferentes entornos es muy buena, aunque también hay que citar algunos de sus inconvenientes: requieren muchos ejemplos, es complicado seleccionar las fuentes adecuadas, consumen un tiempo considerable antes de obtener resultados, el rendimiento se degrada cuando crece la heterogeneidad de los documentos y la adaptación o inclusión de nuevos campos de extracción es compleja.

Los sistemas basados en reglas y patrones se sustentan en la experiencia de la persona que los desarrolla, por lo tanto son necesarios especialistas de cada dominio para definir las reglas de extracción de información. El proceso de definición conlleva mucho tiempo y la introducción de cambios en los sistemas es complicada porque en algunos casos puede suponer volver a redefinir del sistema.

Las herramientas de anotación permiten convertir en metadatos el contenido semántico extraído de las páginas web (Ureña, 2006). Estas aplicaciones se pueden clasificar en dos grandes grupos: herramientas de anotación externa y herramientas de anotación por el autor. Las aplicaciones del primer tipo permiten asociar metainformación a páginas web, pero esta no se almacena dentro de la misma página sino que se guardada de forma externa en un repositorio. Las herramientas de anotación dirigidas a los autores ayudan a incorporar los metadatos dentro o fuera de las propias páginas web siguiendo los estándares (xml, rdf...). Es en este último grupo donde se encuadra nuestra propuesta.

Existen diferentes aproximaciones para realizar la anotación semántica, que, sin embargo, se pueden agrupar en tres grandes categorías. El primer modelo se basa en la anotación lingüística, área de donde originalmente proviene el concepto de anotación ya que es propio de la lingüística de corpus. El objetivo es etiquetar los textos a partir de los diferentes niveles de la lengua. Se empieza a partir del lema para ir pasando hacia los siguientes niveles, morfosintáctico, sintáctico, semántico y discursivo (Buitelaar, 2003). Resulta de gran interés la identificación de los términos y saber cómo estos se relacionan entre sí porque esta información puede incidir en el valor de un término como palabra clave. Sin embargo, este sistema no está muy implantado en el contexto estudiado ya que es muy costoso computacionalmente.

La segunda aproximación se basa en las ontologías, las cuales son utilizadas como recurso central para extraer las conexiones entre los términos y así representar su significado (Niremburg, 01). Actualmente este sistema está teniendo una gran repercusión, sin embargo no está consolidado ya que el proceso de creación de ontologías aún no está bien solucionado (Maedche, 2001; Pedraza-Jimenez, 2007).

La tercera aproximación propone el uso de un lenguaje controlado, como puede ser un tesauro o una taxonomía, para facilitar la asignación de metadatos. Éste es el modelo que tradicionalmente ha utilizado la Biblioteconomía y Documentación para indexar de forma manual la información. Es un modelo que está directamente vinculado con la asignación de metadatos y la anotación semántica (Guyot, 2006). La propuesta que presentamos está situada en este último modelo pero aplicando una capa de procesamiento automático basada en la presencia de los términos del documento en la taxonomía y su posterior valoración en función de criterios que aplican los algoritmos de posicionamiento.

3.3. Lenguajes controlados: taxonomía, tesauros y ontología

Los lenguajes controlados son mecanismos para la representación y organización del conocimiento, con el objetivo de controlar y normalizar la asignación de palabras clave a un documento. Por lo tanto, son uno de los elementos esenciales para un uso eficaz de los metadatos, tanto si nos situamos en un contexto de trabajo

manual, como en una asignación automatizada. Las taxonomías, al igual que los tesauros y las ontologías son herramientas que permiten estructurar la información y dotarla de un mínimo de semántica (Gilchrist, 2003). El actual crecimiento de la información en la Web ha generado nuevas perspectivas para el diseño y desarrollo de los lenguajes controlados.

A continuación se describen las características principales de los lenguajes controlados que tienen una relación más directa con la propuesta presentada.

3.3.1. Taxonomía

Las taxonomías son una forma de clasificar de forma jerárquica los contenidos. El concepto tiene su origen en la biología sistemática, que estudia las relaciones entre los organismos y su historia evolutiva. Se utilizaron para establecer unos criterios para la clasificación, y así poder agrupar la diversidad de organismos en clases basándose en las propiedades que compartían.

Esta idea se extendió a otro ámbitos y una taxonomía ha pasado a ser una jerarquía semántica donde las entidades de información se relacionan a nivel de clases y subclases para organizar el conocimiento (Chris, 2007). Las taxonomías son las estructuras que vertebran las ontologías que veremos más adelante.

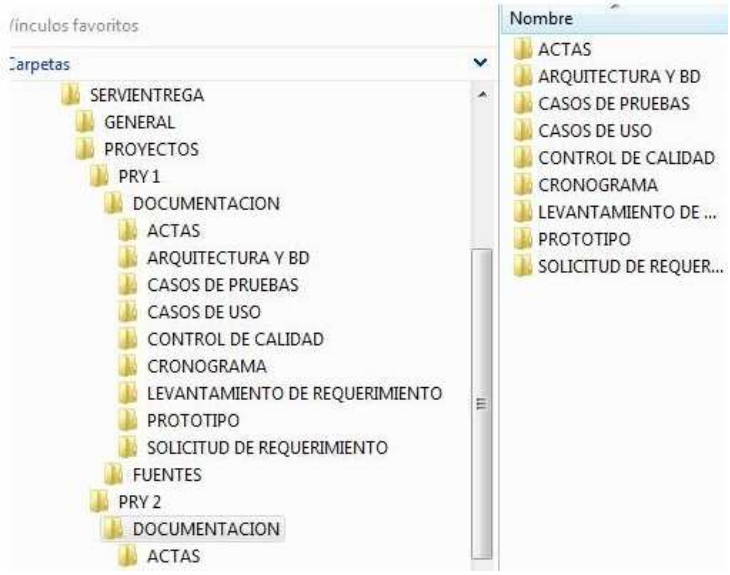

Figura 1. Ejemplo de Taxonomía

3.3.2. Tesauro

Los tesauros son unos listados de palabras o términos utilizados para organizar el conocimiento de un dominio con el objetivo de controlar la descripción temática de un documento. Se trata de un tipo de lenguaje documental formado por términos normalizados, los descriptores, y por las relaciones semánticas y funcionales que se establecen entre estos términos. Las relaciones semánticas utilizadas son: equivalencia, asociación y jerarquía (López-Huertas, 99).

Los tesauros tienen un elevado control terminológico y una gran capacidad de especialización. Son muy útiles para optimizar la recuperación de información en entornos cerrados, ya que ayudan a la desambiguación y la normalización semántica en la expresión del contenido de los documentos. Su utilización es habitual en los entornos de las bibliotecas, centros de documentación, bancos de imágenes y bases de datos científicas, sin embargo en otros ámbitos relacionados con la recuperación de la información no están tan extendidos.

Figura 2. Ejemplo de Tesauro

3.3.3. Ontologías

Las ontologías tienen sus antecedentes en la metafísica, rama de la filosofía que se centra en la naturaleza de la realidad. La Ontología fue una iniciativa para describir las relaciones básicas del ser, de su existencia y para definir los entes y sus tipos (Echeverría, 1998).

A partir de los 80, las ontologías son utilizadas por la inteligencia artificial como un instrumento para representar el conocimiento en un área determinada. Las ontologías son especificaciones formales y explícitas que representan los conceptos y las relaciones en un determinado dominio (Gruber, 1993) .

La espectacular evolución de la Web y el gran interés existente por el desarrollo e implantación de la Web semántica han llevado a las ontologías a desempeñar un papel muy destacado, a pesar de ser más simbólico que real. En teoría, son una de las piezas clave para la comunicación entre organizaciones, personas y aplicaciones y así facilitar la interoperabilidad entre sistemas. Gracias al conocimiento almacenado en las ontologías, los agentes inteligentes podrían extraer directamente datos de las páginas web, procesarlos y hacer inferencias. Sin embargo, esta funcionalidad actualmente aún no está disponible fuera de dominios reducidos.

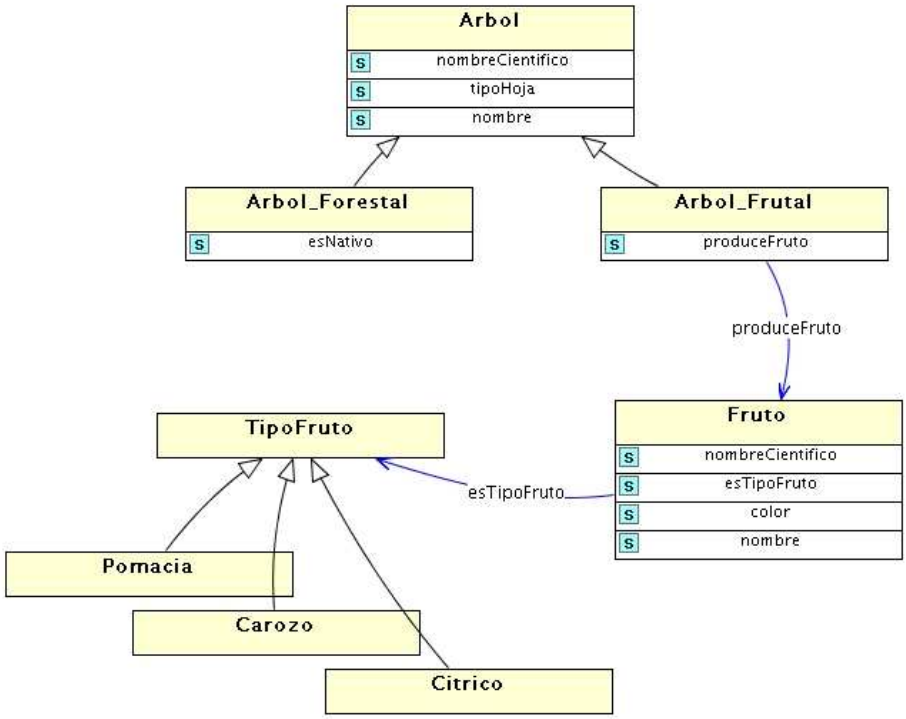

Figura 3. Ejemplo de ontología

Ejemplo de Ontología La formalización del conocimiento en las ontologías es una de las barreras iniciales para la implantación de la Web semántica, ya que la construcción de ontologías es un proceso extremadamente lento, costoso y propenso a errores. Requiere un gran esfuerzo y especialización que muchas organizaciones no tienen a su alcance.

Existen diferentes métodos y herramientas que ayudan a la creación y desarrollo de las ontologías de una forma (semi) automática. La ingeniería de ontologías es el nombre de la disciplina encargada del estudio y construcción de las diferentes herramientas que tienen por objetivo diseñar mecanismos que agilicen el proceso de construcción de ontologías de un determinado dominio. Sin embargo, no existe un consenso en la comunidad científica para concretar las diferentes fases implicadas en su desarrollo.

Como ya se ha dicho, la herramienta que planteamos desarrollar estaría basada en una taxonomía de construcción manual, de temática muy específica y con la mayoría de las relaciones propias de un tesauro como: equivalencia, asociación y jerarquía.

3.4. Posicionamiento en buscadores

El posicionamiento en buscadores es el conjunto de técnicas para conseguir que una página web aparezca en las primeras posiciones en los listados de resultados de los buscadores cuando los usuarios ejecutan unas determinadas ecuaciones de búsqueda.

Para los autores de los contenidos es primordial el buen posicionamiento de sus páginas web, ya que cada vez es mayor la proporción de tráfico proveniente de los motores de búsqueda. Estudios recientes aseguran que entre el 50 y el 70 por ciento del total del tráfico de un sitio web puede proceder de motores de búsqueda (Valentine, 2007) y no son extraños los casos donde el porcentaje llega al 90 por ciento.

Una de las fases principales a la hora de mejorar el posicionamiento de una determinada página es determinar las palabras clave para las que se desea estar bien posicionado. Las palabras clave deben ser seleccionadas en función de los contenidos, los objetivos y el público de la página web. En este contexto resulta útil identificar tres o cuatro palabras clave principales teniendo en cuenta los siguientes aspectos (Gonzalo, 04):

- Relación con el contenido. Las palabras clave seleccionadas deben reflejar los contenidos de la web y deben coincidir con las que utilizarían los usuarios para localizar la página web a posicionar.
- Popularidad y competencia. Los términos individuales más utilizados, suelen tener mucha competencia y por tanto resulta difícil posicionarse entre los primeros resultados para ellos. La solución suele estar en seleccionar ?frases clave?, formadas por dos o tres palabras que no sean muy populares y optimizar las páginas web por las mismas.
- Para valorar la efectividad de las palabras clave seleccionadas existe el cálculo del Índice de Efectividad de una Palabra Clave (Keyword Effectiveness Index). Es un indicador que muestra la oportunidad de una determinada palabra, en base a su popularidad, número de búsquedas mensuales realizadas con el término, y a su competitividad, número de resultados obtenidos cuando se realiza una búsqueda por esa palabra.

Por otra parte, para mejorar el posicionamiento hay que tener en cuenta como actúan los algoritmos de posicionamiento de los buscadores en relación al texto de las páginas Web. Es sabido que los principales buscadores colocan antes una página en la que las palabras usadas en las búsquedas están en zonas de especial relevancia, como por ejemplo en el título (elemento <title>), los encabezados (elementos <h1>, <h2> ..), los anclajes (atributo href), las negritas, los títulos (atributo title) de los gráficos... o incluso en el texto del principio del documento o en los anclajes de los enlaces de otras páginas que apuntan hacia la página que queremos posicionar.

El ámbito del posicionamiento tiene dos implicaciones importantes en la investigación que planteamos.

- Creación de la taxonomía. En el proceso de selección de los términos de la taxonomía deberá considerarse de forma prioritaria el Índice de efectividad. Para priorizar la terminología más utilizada por los usuarios de la red y facilitar su recuperación.
- Ordenación de los candidatos a palabras clave. Como ya se ha indicado, nuestra herramienta proporcionará a los usuarios un listado de candidatos a palabras clave en función de la taxonomía y del contenido de la página analizada. Este listado estará ordenado en función de mayor o menor presencia de los términos en las zonas relevantes que aplican los algoritmos de ordenación de resultados. La consecuencia de esta ordenación es que el usuario sabrá cuáles son los términos más importantes para la descripción de su contenido de acuerdo con un conjunto de criterios ampliamente utilizados por los buscadores. Además habrá que explorar si la asignación automática a partir de una determinada puntuación resulta efectiva.

4. Conclusiones

A pesar de que la Web semántica es todavía una utopía, la migración hacia la Web semántica es ya una realidad. El etiquetado de los contenidos de la web por medio de metadatos expresados con formatos estándares se está generalizando. Poco a poco los contenidos disponibles en Internet se están preparando para un futuro que no sabemos si finalmente acabará siendo exactamente como se ha estado prometiendo por el W3C. No importa, el nuevo paradigma ha impulsado interesantes pasos hacia una Web más procesable.

El proyecto presentado se sitúa en esta fase de migración hacia la Web semántica con un cierto escepticismo con el resultado final y un gran entusiasmo por los resultados parciales. La extracción semiautomática de palabras clave de las páginas web podría ser un nuevo elemento parcial que tendría inmediatas e interesantes repercusiones y que al mismo tiempo representaría un paso más hacia los objetivos finales. La investigación que planteamos explora las posibilidades de la automatización en la asignación de palabras clave con procedimientos relativamente sencillos basados en modelos masivamente usados en las disciplinas de la recuperación de la información, el posicionamiento en buscadores y las ciencias de la documentación.

5. Bibliografía

AGUADO DE CEA, Guadalupe; ÁLVAREZ DE MON, Inmaculada; PAREJA, Antonio (2002): "Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de web semántica: OntoTag", Revista Iberoamericana de Inteligencia Artificial, num. 17, p. 37-49.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora (2001): "The semantic Web", Scientific American, vol. 284, num. 5 (May), p. 34-43.

BUITELAAR, Paul; DECLERCK, Thierry (2003): "Linguistic Annotation for the Semantic Web", en Siegfried Handschuh and Steffen Staab: Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications Series, vol. 96, IOS Press, 2003.

CASTELLS, Manuel (1997): La era de la información. Economía, sociedad y cultura. (3vols.). Madrid: Alianza

CODINA, Lluís; ROVIRA, Cristòfol (2006). "Web Semántica: visión global y análisis comparativo", en Tendencias en documentación digital. Gijón: Trea. http://eprints.rclis.org/archive/00008637/

CODINA, Lluís; MARCOS, Mari-Carmen; PEDRAZA, Rafael (COORDS.). Web semàntica y sistemas de información documental. Gijón: Trea, 2009

CUNNINGHAM, Hamish; BONTCHEVA, Kalina; LI, Yaoyong (2005): "Knowledge management and human language: crossing the chams", Journal of Knowledge Management, vol. 9, num. 5, p. 108-131.

GILCHRIST, Alan (2003): "Thesauri, taxonomies and ontologies : an etymological note", Journal of documentation, Volum 59, Num. 1, p. 7-18.

GUYOT, Jacques; RADHOUANI, Saïd; FALQUET, Gilles (2006): "Conceptual Indexing for Multilingual Information Retrieval", Accessing Multilingual Information Repositories, Lecture Notes in Computer Science, vol. 4022, p. 102-112.

FLYNN, P.; ZHOU, L.; MALY, K.; ZEIL, S. ; ZUBAIR, M. (2007): "Automated template-based metadata extraction architecture", Lecture notes in computer science, 4822, p. 327-336.

DAVIES, John; GROBELNIK, Marko; MLADENI?, Dunja (2009): "Challenges of Semantic Knowledge Management", Semantic Knowledge Management, p.245-247.

DING, Li, et al. (2005): "Search on the Semantic Web", IEEE Computer, vol. 38, num. 10 (October), p. 62-69.

DOUGLAS, T.; CERI, B.; DOROTHEE, B. ;DANIEL, C. (2006): "Query expansion via conceptual distance in thesaurus indexed collections", Journal of Documentation, 62, p. 509-533.

ECHEVERRÍA, Rafael (1998): La ontología del lenguaje, Editorial Dolmen, 5ª Ed, Palma de Mallorca.

GARTNER RESEARCH GROUP (2007a): "Taking stands on the Semantic Web", Id Number: G00148696.

GARTNER RESEARCH GROUP (2007b): "Finding and exploting value in Semantic Technologies on the web", Id Number: G00148725.

GONZALO PENELA, Carlos (2004): "La selección de palabras clave para el posicionamiento en buscadores", [on line]. Hipertext.net, núm. 2, 2004. http://www.hipertext.net

GRUBER, Thomas (1993): "A translation approach to portable ontologies", Knowledge Acquisition, vol. 5, num. 2, p. 199-220.

KIRYAKOV, Atanas; POPOV, Borislav; TERZIEV, Ivan; Manov, Dimitar Ognyanoff, Damyan (2004): "Semantic annotation, indexing, and retrieval", J. Web Sem. 2 (1), p. 49-79.

LÓPEZ-HUERTAS, M. J. (1999): "Potencialidad evolutiva del tesauro: Hacia una base de conocimiento experto", en La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de la información, actas del IV Congreso ISKO, p. 133-140.

MAEDCHE, A.; STAAB, S. (2001): "Ontology Learning for the Semantic Web", IEEE Intelligent Systems, Special Issue on the Semantic Web, vol. 16, num. 2, p. 72-79.

CHRIS, M. (2007): "Taxonomy development: assessing the merits of contextual classification",Records Management Journal, 17, p. 7-16.

MILLER, George A. (1995): WordNet: a lexical database for English en Communications of the ACM, Vol. 38, Num. 11, p. 39-41.

NIRENBURG, Sergei; RASKIN, Victor (2001): "Ontological semantics, formal ontology, and ambiguity", en Proceedings of the international Conference on Formal ontology in information Systems, ACM, vol. 2001, p. 151-161.

PEDRAZA-JIMENEZ, Rafael; CODINA, Lluís; ROVIRA, Cristòfol (2007): "Web semántica y ontologías en el procesamiento de la información documental", en El profesional de la información, 2007, noviembre-diciembre, v. 16, n. 6, p. 569-578.

PEDRAZA-JIMENEZ, Rafael; CODINA, Lluís; ROVIRA, Cristòfol (2008): Semantic Web adoption: online tools for web evaluation and metadata extraction. En Da Ruan et al. (ed) Computational Intelligence In Decision And Control. Proceedings Of The 8Th International Flins Conferenc New Jersey: World Scientific Publishing Co Pte Ltd, 2008. ISI Document Delivery No.: BIF16.

ROVIRA, Cristòfol; MARCOS Mari-Carmen (2006): "Metadatos en revistas-e de Documentación de libre acceso", El profesional de la información; 15(2): p.136-143.

ROVIRA, Cristòfol; MARCOS Mari-Carmen (2007): "Repositorios de publicaciones digitales de libre acceso en Europa: análisis y valoración de la accesibilidad, posicionamiento web y calidad del código digital", El profesional de la información; 16(1): p. 24-38

UREN, Victoria, et al. (2006): "Semantic annotation for knowledge management: Requirements and a survey of the state of the art", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 4, num. 1 (January 2006), p. 14-28.

VALENTINE, Mike (2007): Search Engine Optimism. En http://searchengineoptimism.com/Google_refers_70_percent.html

VOSSEN, Piek (2004): EuroWordNet: "A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index", en International Journa l of Lexicography, Vol. 17, Num. 2, p. 161-173.

Última actualització 08-11-2010