LAS TECNOLOGÍAS DE LOS MOTORES DE BÚSQUEDA DEL FUTURO

TECHNOLOGIES FOR THE NEXT GENERATION SEARCH ENGINES

Recibido: 19/07/07
Aceptado: 24/09/07

José María Gómez Hidalgo
Dr. en Ciencias Matemáticas

José Carlos Cortizo Pérez
Ingeniero Informático

Francisco Carrero García
Ingeniero Informático

Borja Monsalve Piqueras
Ingeniero Informático
Departamento de Sistemas Informáticos

Universidad Europea de Madrid

RESUMEN

Es indudable que Internet en general, y la Web en particular, tienen una influencia creciente en nuestras vidas y se han convertido en un medio de comunicación y un recurso informativo de primer orden. La gran cantidad de información disponible en la Web se hace accesible primordialmente a través de los motores de búsqueda como *Google, Yahoo!* o *Altavista*. Las empresas que operan estos motores son ahora multinacionales con enormes ingresos financieros obtenidos a través de la publicidad que logran por el tráfico de usuarios que acumulan. Su supervivencia depende de seguir siendo útiles, y cada vez más, para los usuarios, algo que sólo pueden lograr a través de la innovación e implantación de tecnologías y funcionalidades cada vez más avanzadas. En este artículo presentamos una revisión de algunas de las tecnologías que creemos clave para los motores de búsqueda del presente y del futuro, centrándonos en la personalización y la localización, la búsqueda social, la búsqueda en la Web semántica, la búsqueda translingüe, y el control de fraude en buscadores.

Palabras clave: Motores de búsqueda, *Google*, personalización, localización, redes sociales, Web semántica, búsqueda translingüe, *spam*.

ABSTRACT

With no doubt, Internet in general ant the Web in particular have an increasing influence in our lives, and they have become a mass media and a first class information resource. The huge amount of information available in the Web is primary accessible to its users via search engines like Google, Yahoo! or Altavista. The companies operating these search engines are now multinational corporations with enormous financial incomes generated by their user traffic. But they will only survive if they keep being more and more useful for their users, something they can only achieve by innovating and implementing increasingly advanced technologies and functionalities. In this article, we review some of the key technologies we believe that make the search engines of the present and the future, focusing on personalization and ***localization, social search, search in the Semantic Web, cross lingual search, and fraud detection and control in search engines.***

Key words*: Search engines, Google, personalization, localization, social networks, Semantic Web, cross lingual search, search engine spam.*

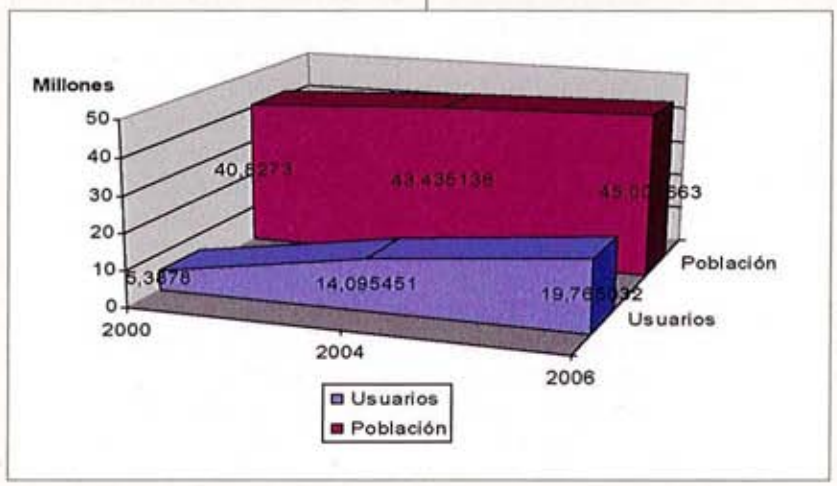


Figura 1. Crecimiento de la población y de los usuarios de Internet en España durante los últimos años.

1. INTRODUCCIÓN

En los últimos años, Internet ha dejado de ser una red de comunicación empresarial y universitaria para convertirse en un medio de comunicación de masas. Más de 1,000 millones de usuarios se conectan a Internet para enviar correos a sus amigos, compañeros y clientes, para adquirir productos como los DVD y libros, para buscar información sobre

su compañía o la competencia, o sobre salud, para reservar billetes de avión y para innumerables cosas más. Internet y la información almacenada en ella se ha convertido en un instrumento de uso común, casi imprescindible en el trabajo y cada vez más usado para el ocio. Muestra de ello es el importante crecimiento del número de usuarios de Internet en el ámbito español, que se puede observar en la figura 1, comparativamente con el crecimiento de la población. Según estadísticas de distintas consultoras[1], en el año 2000 contaba España con una población de aproximadamente 41 millones de habitantes, de los cuales cinco millones y medio eran usuarios de Internet (13,2%). En 2006, la población alcanzó los 45 millones, de los cuales casi 18 son usuarios (43,9%).

Si el correo electrónico fue la primera aplicación que se extendió en Internet, sin duda su llegada al gran público se debe a la invención de la World Wide Web por parte de **Tim Berners-Lee**, en 1991. La simplicidad de la consulta y de la edición de contenidos, junto con su creciente interactividad, han convertido a la Web en la aplicación por excelencia en Internet. La cantidad de información y recursos disponibles en es tan asombrosa, que los motores de búsqueda o buscadores como *Google!*[2] se han erigido en el punto de entrada en la misma por excelencia y, lógicamente, en las páginas más visitadas. Hoy por hoy, los buscadores son grandes corporaciones con ingresos millonarios en publicidad, que incluso aspiran a "*organizar y dar acceso a toda la información disponible*", como es el caso de *Google!* [Google07].

Los buscadores tienen la difícil misión de ayudar a localizar la información disponible en la Web (e incluso en Internet) a sus usuarios. Tan compleja misión presenta la dificultad adicional de que el modo en que los usuarios utilizan la Internet y la Web no ha dejado de evolucionar desde su inicio. Los nuevos usos representan nuevos retos para los buscadores, que deben incorporar tecnologías cada vez más avanzadas para satisfacer las demandas de sus usuarios.

En este trabajo pretendemos revisar algunas de las tecnologías emergentes en los buscadores y en la Web, aunque la extensión del mismo impide que seamos todo lo exhaustivos que desearíamos. Hemos optado por tratar con cierto detalle algunos de los temas que consideramos más interesantes, empezando por la personalización y la localización de las búsquedas, continuando con las redes sociales y la Web semántica y su relación con la búsqueda de información, y finalizando nuestra revisión con la búsqueda translingüe y con la detección del fraude en buscadores.

Obviamente, nos dejamos en el tintero temas indudablemente importantes, como los aspectos relativos a la eficiencia en un entorno de incesante crecimiento en los datos disponibles, las técnicas de soporte a las nuevas formas y objetivos de búsqueda (como buscar para aprender), el tratamiento de la información multimedia crecientemente presente en la Web, la gestión de la privacidad de los usuarios y de los derechos de autor de los suministradores de contenidos, etc. Algunos de estos temas se revisan en la reciente monografía centrada en la búsqueda en la Web del futuro de la revista *Novática* [Pages07].

2. PERSONALIZACIÓN Y LOCALIZACIÓN

La cantidad de información disponible en la Red aumenta de forma exponencial, creando nuevos retos para una búsqueda más efectiva [**Liu**04]. Cuándo se envía la misma consulta a un motor de búsqueda que no admita personalización, devolverá los mismos resultados sea cuál sea el usuario que ha realizado la consulta. Sin embargo, las necesidades de información de los distintos usuarios no son las mismas. Por ejemplo, para la consulta "*apple*", algunos usuarios estarán buscando información sobre manzanas, mientras que otros estarán buscando información sobre la Compañía **Apple** Inc. Una opción para concretar las consultas es restringirlas al ámbito de una categoría temática jerárquica (como las existentes en *DMOZ*[3] o *Yahoo!*[4]), pero existe una gran oportunidad de aclaración utilizando el conocimiento que los buscadores tienen sobre el usuario.

Figura 2.1. Página principal de Google para dispositivos móviles personalizada para un usuario determinado.

Los buscadores como *Yahoo!* y *Google* almacenan todos los días Terabytes de información acerca de los usuarios de sus sistemas [**BaezaYates**06]. Además, gracias a la gran integración de nuevos servicios, los datos disponibles son de muy diversa índole (correos electrónicos, búsquedas realizadas, páginas visitadas, fotografías, noticias, etc.) lo cuál ofrece gran cantidad de información de gran

1 Fuente: Internet World Stats, http://www.internetworldstats.com/.
2 http://www.google.es.
3 http://www.dmoz.org.
4 http://www.yahoo.com.

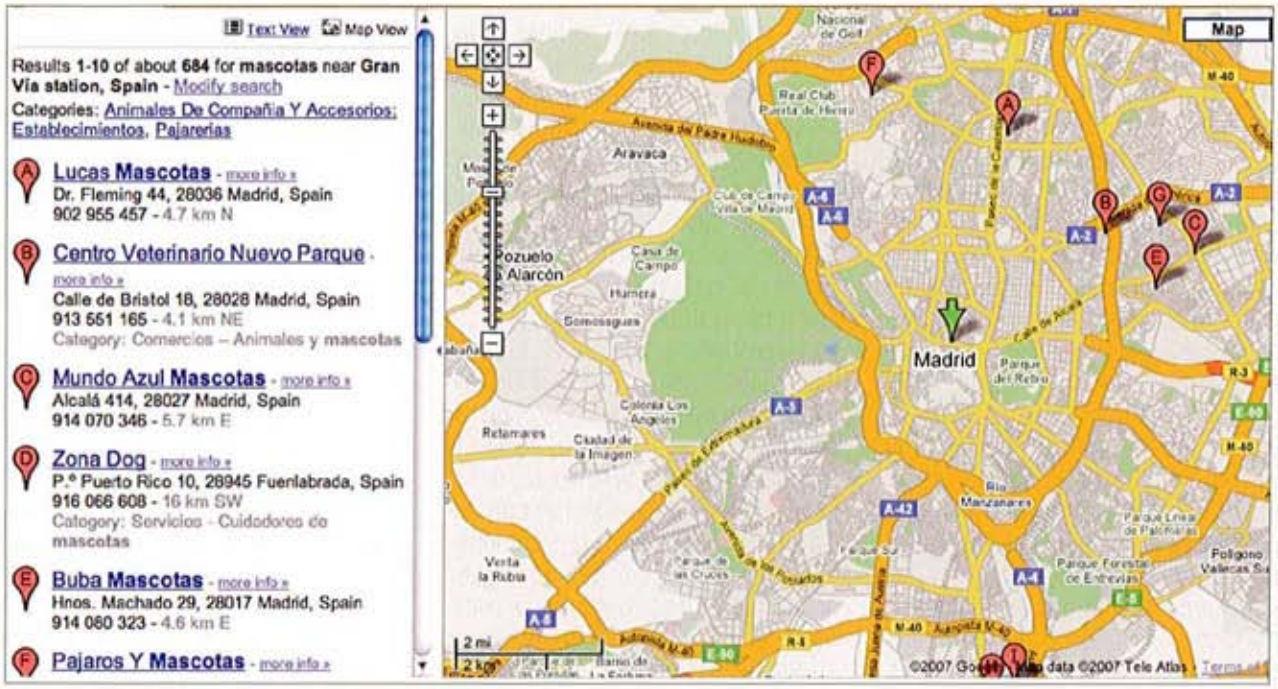


Figura 2.2. Búsqueda por tiendas de mascotas localizada a la ubicación del usuario (en la Gran Vía madrileña).

calidad, que permite obtener una visión general del usuario bastante acorde con sus gustos e intereses (perfil). De esta manera, a un usuario que recientemente ha buscado términos como *iPhone*, *MacBook* y ha estado viendo fotografías de un monitor de **Apple**, y habitualmente consulta noticias tecnológicas relacionadas con la empresa Apple Inc, cuando consulte *apple* en el buscador, se le devolverán los documentos que tengan que ver con la empresa **Apple** Inc. Sin embargo, a una persona que habitualmente busca información acerca de frutas, se le devolverá información sobre manzanas y no sobre la compañía de ordenadores.

La personalización no sólo se basa en el historial de navegación del usuario (sus gustos) sino que también puede tener en cuenta otros aspectos importantes como los dispositivos utilizados para navegar. En la Figura 2.1 se muestra la página principal de *Google* para dispositivos móviles personalizada para un usuario. Si el usuario realiza una consulta a un buscador desde su teléfono móvil, es preferible devolverle las webs que están especialmente diseñadas para visionarse con dispositivos móviles, de forma que se mejore su experiencia en la Red.

La localización también es relativa a la personalización pero se queda a un nivel más general que el usuario pues no tiene en cuenta los gustos concretos de éste sino que trata de adaptar los contenidos a la ubicación física de los usuarios. Para el caso de un usuario que busque "tienda de mascotas" y realice una consulta desde Madrid, la localización premia aquellos resultados que tengan una vinculación con la localización física del usuario, Madrid. Así pues, las páginas web de las tiendas de animales que estén en Madrid deben *salir* antes que otras pues son de mayor interés para el usuario que las que estén en otras ubicaciones. Este aspecto es todavía más importante cuando la búsqueda se realiza desde un dispositivo móvil ya que si un usuario busca un restaurante desde su móvil, agradecerá que la primera información que se le ofrezca sea la de aquellos restaurantes que estén más cerca de su ubicación actual. En la Figura 2.2 se muestran los resultados de una búsqueda por "tiendas de mascotas" estando el usuario en la Gran Vía de Madrid. En los resultados se puede apreciar cómo las primeras entradas son aquéllas que más cerca se encuentran de la ubicación del usuario.

Tanto la personalización como la localización son tecnologías bastante cercanas y en creciente uso, sobre todo por la gran variedad de dispositivos con los que podemos conectarnos, la siempre presente ubicuidad y la gran cantidad de información existente que hace que cualquier manera de restringirla, sin perder información de interés, sea de gran beneficio para el usuario.

A su vez, además de beneficiar al usuario, la personalización y localización permiten mejorar el modelo de negocio de los buscadores, ya que permite que estos muestren publicidad más efectiva al usuario, bien sea por proximidad geográfica o bien por afinidad en cuánto a gustos y costumbres.

Las técnicas utilizadas para la localización y personalización de las búsquedas van desde el aprendizaje por refuerzo [**Sutton**98], agrupamiento [**Ferragina**07], categorización automática [**Gómez**05] y generación au-

tomática de perfiles [**Sugiyama**04]. Además de estas técnicas de aprendizaje, es muy importante contar con una buena organización de los datos y una gran capacidad de almacenamiento y computo, ya que debido al gran número de usuarios de Internet, se manejan Terabytes de información a diario, por lo que si no se hace de una forma óptima, se corre el riesgo de no ser capaces de explotar los datos al ritmo al que se producen.

Ahora bien, tanto la personalización como la localización tienen el riesgo añadido de poder invadir la privacidad del usuario [**Sullivan**03]. ¿Dónde se puede poner el listón de lo que es tolerable o no por parte del usuario en cuánto a la utilización de sus datos? Los buscadores corren el riesgo de convertirse en el *Gran Hermano* de **Orwell** si no son capaces de imponerse límites a la hora de aprovechar la información personal de los usuarios.

3. LA BÚSQUEDA SOCIAL

A medida que el número de usuarios de la Web ha ido aumentando, lo ha hecho también la cantidad de información disponible. Al principio, poseer una página Web era algo reservado a unos pocos: prácticamente sólo grandes empresas, Organismos oficiales o usuarios avezados podían presumir de ello. Inicialmente esta presencia era meramente informativa. Básicamente se reducía a ofrecer información estática al internauta; información que terceras partes consideraban que podía ser interesante para el navegante, pero no demandada directamente por este. En resumidas cuentas: no existía interacción con el usuario.

Poco a poco esta tendencia ha ido cambiando, y la propia Red se ha ido volviendo más dinámica e interactiva. Se ha pasado de tener páginas meramente informativas a sitios creados por y para los usuarios. Ya no es necesario poseer conocimientos técnicos para convertirse en creador de un vistoso y exitoso sitio Web; basta con tener algo que contar y ser capaz de rellenar un formulario. Han llegado los *blogs*, las *wikis* o las comunidades *online* (de fotos, como *Flickr*[5], o de videos, como *YouTube*[6]). En definitiva, lugares en la Red que facilitan y fomentan que los usuarios pudieran colaborar y compartir. Han nacido la llamada Web 2.0 o Web social, denominada así no porque se haya producido una evolución en la tecnología, sino por el cambio que se estaba dando entre desarrolladores y usuarios a la hora de utilizar Internet.

Esta variación ha originado, a su vez, nuevas necesidades y enfoques a la hora de localizar información interesante para el usuario. Han comenzado a quedarse pequeños los buscadores "tradicionales", basados en técnicas de análisis del contenido de las páginas o de los enlaces de las mismas, como los primeros *Altavista*, *Yahoo!* o *Google*. Se hace necesario tener en cuenta nuevos factores, derivados precisamente de ese nuevo uso que se estaba dando a la Web. Comienzan a surgir buscadores que tienen también en cuenta aspectos tales como la colaboración entre usuarios, la inteligencia colectiva y otros factores que permiten obtener mayor rendimiento de la información. Algunos ejemplos de estos nuevos buscadores, llamados buscadores sociales o buscadores 2.0, son *Swicki*[7] (un buscador regido por una comunidad, y que aprende de las búsquedas que hacen sus miembros para ofrecer resultados más concretos), *Rollyo*[8] (que también realiza búsquedas dentro de una comunidad), *Clusty*[9] (un motor *cluster*[10], que agrupa los elementos similares, organizando así las búsquedas por categorías) o *Lexxe*[11] (diseñado para responder brevemente a las consultas, en lugar de localizar la página donde podría encontrarse la respuesta). Un ejemplo paradigmático es *ReferralWeb* [**Kautz**97], un sistema que saca partido de la red social para realizar filtrado colaborativo y recomendación de recursos usando como evidencia las opiniones de los vecinos en la comunidad social.

Los buscadores tradicionales son muy buenos encontrando información, y lo hacen cada vez con más precisión y analizando más información. Sin embargo, no tienen en cuenta las ideas asociadas a las palabras que están buscando. Los buscadores sociales tratan de avanzar en esa línea, haciendo búsquedas más significativas y hasta basadas en preguntas (como *Lexxe*). Exploran una vía alternativa, ya que profundizan en los intereses de los usuarios, y han dado lugar a nuevas maneras de representar la información, como por ejemplo las *tag clouds* (nube de etiquetas), que se utilizan como representaciones visuales de las palabras clave (o etiquetas) utilizadas en un sitio Web, y donde los usuarios pueden ver fácilmente qué conceptos son los más relevantes, ya que aparecen destacados con respecto los demás por su mayor tamaño. En la figura 3.1 se muestra una nube de etiquetas característica de un sitio Web social.

En la actualidad algunas las empresas más representativas del mundo de las búsquedas en Internet ya comienzan a mover ficha hacia la búsqueda social. *Yahoo!*, por ejemplo, posee un buscador (*MyWeb* 2.0[12]) que cada vez integra más aspectos sociales en sus búsquedas.

5 http://www.flickr.com/
6 http://www.youtube.com/
7 http://swicki.eurekster.com/
8 http://www.rollyo.com/
9 http://clusty.com/
10 Un cluster es, básicamente, un grupo de ordenadores que se comportan como si fueran uno solo.
11 http://www.lexxe.com/
12 http://myweb2.search.yahoo.com/

Google por su parte también está tomando cartas en el asunto, añadiendo una sección de favoritos[13] al historial de búsqueda, que permite crear etiquetas y comentarios para asociarlos a las páginas que ya se han visitado. Esto podría permitir a los usuarios, en un futuro, ver las etiquetas y comentarios de otras personas e incluir esta información en las búsquedas. Mientras, **Microsoft**, que estima que el 50% de las consultas realizadas no son del todo resueltas por los buscadores tradicionales, comienza a trabajar en un proyecto de buscador social donde, entre otras características, pretenden incorporar en los procesos de búsqueda consultas en círculos de amigos.

4. LA BÚSQUEDA SEMÁNTICA

El rápido crecimiento del número de páginas Web ha ido forzando las necesidades en cuánto a herramientas y tecnologías que permitan un acceso más cómodo y eficiente a las mismas. En un principio, la única forma de navegar por la Web era conocer de antemano las direcciones (URLs) de las páginas a consultar, lo cual limitaba en gran medida la capacidad de recuperar información relevante, así como condicionaba el acceso a nuevos contenidos. Posteriormente surgieron los directorios donde se *categorizaban* las páginas de forma manual en una taxonomía de temas de interés y, finalmente, surgieron los buscadores como *Lycos* y *AltaVista* que descubrían automáticamente las nuevas páginas Web y las añadían a su base de datos indexándolas para permitir recuperarlas al buscar ciertos términos.

Figura 3.1. Nube de etiquetas característica de un sitio Web social. Las etiquetas de mayor tamaño son las más frecuentemente usadas por los usuarios en sus contenidos.

Como nota final, conviene resaltar el importante impacto que las redes sociales tienen en el modelo económico basado en publicidad de los buscadores. Si éstos han conseguido adaptar la publicidad al contenido de las páginas con el fin de aumentar el éxito de la misma, este éxito se potencia enormemente con la adaptación adicional que supone el que los lectores de la página no son casuales, sino miembros de una comunidad con intereses comunes que ha emergido en el seno de una red social, y que dan un valor especial a cuanto encuentran en la página. Ello implica que las tasas de éxito se pueden incrementar espectacularmente, haciendo aún más valioso el negocio publicitario.

La visión de la Web semántica de **Tim Berners-Lee** [BernersLee01] permite un mayor grado de expresividad en las páginas, ya que en las páginas Web no solo se codifican datos (palabras), si no que también se introduce conocimiento (conceptos y reglas de inferencia). Este conocimiento adicional proporciona información extra que seguramente no sea útil al que navega por la página, pero sí que resulta muy útil para que las máquinas sean capaces de extraer conocimiento de una forma más simple y estandarizada. A este respecto, sirva el ejemplo de una página Web de un trabajador de una empresa, donde se incluye su información de contacto. Para un humano resulta muy simple extraer la información de contacto ya que asocia patrones de texto (una dirección junto a un teléfono y un *mail*, por ejemplo) y es capaz de extrapolar del mundo de los datos al mundo de los conceptos. Sin embargo para automatizar el proceso de forma que un programa pueda extraer esta información, se necesita un *software* muy elaborado, ya que hay que saber identificar patrones de direcciones, teléfonos, correos, así como reconocer de todas las posibles direcciones que aparezcan en una web, cuál es la referente al contacto. Ahora bien, para facilitar la comprensión de estos datos, se puede etiquetar la web con conocimiento acerca de la información que existe en la misma, por ejemplo que estamos hablando de una persona, cuyo nombre es "Fulanito", su dirección de correo electrónico es "fulatino@miempresa.com" y su dirección postal es "C/Perdida, 32 (Madrid)".

La Web Semántica se puede definir como un marco que permite publicar, compartir y reutilizar datos y conocimiento tanto en la red como a

13 http://google.blognewschannel.com/index.php/archives/2005/10/10/google-adds-tagging/

través de aplicaciones [Ding05]. Dentro de este marco, se encuentran

1- XML, que provee una sintaxis elemental para estructura el contenido dentro de los documentos, pero sin asociar ningún tipo de semántica al mismo.

2- RDF, un lenguaje que permite expresar modelos de datos, tanto las descripciones de los objetos (recursos) como las relaciones entre los mismos. Los modelos basados en RDF se pueden representar en sintaxis XML.

3- OWL, que añade más vocabulario para describir propiedades y clases como las relaciones entre clases (por ejemplo que sean disjuntas), cardinalidad (por ejemplo, que sean exactamente 3), igualdad, características de las propiedades (por ejemplo, simetría) y clases enumeradas.

4- SPARQL, un protocolo y lenguaje de consultas para recursos de la Web semántica.

5- Ontologías, que definen conceptos y relaciones entre los mismos, como FoaF (*Friend of a Friend*) [Foaf07], que es una ontología basada en RDF que permite modelar la información de personas y las relaciones entre las mismas.

Conociendo estas herramientas, ya podemos modelar la información de contacto del ejemplo anterior. La Figura 4.1 muestra un diagrama que muestra de una forma visual las relaciones entre los distintos conceptos que forman parte de esta información.

Con todo esto, ya se pueden generar documentos anotados semánticamente pero, ¿cómo afecta esto a la búsqueda de documentos? El funcionamiento en los buscadores convencionales comienza con la introducción de una serie de palabras clave a buscar, sin embargo, un buscador para la Web semántica debe aprovechar la información conceptual para sacar mayor partido de la consulta, lo cuál implica una mayor precisión a la hora de realizar las consultas. Siguiendo el ejemplo anterior, si se quiere conocer la dirección de contacto de una persona que se llama Fulanito, se tendrá que indicar que vamos a buscar un campo conceptualmente etiquetado como dirección de contacto asociado a un individuo cuyo nombre es "Fulanito".

bemos olvidar que tan importante como estos dos aspectos es la identi-

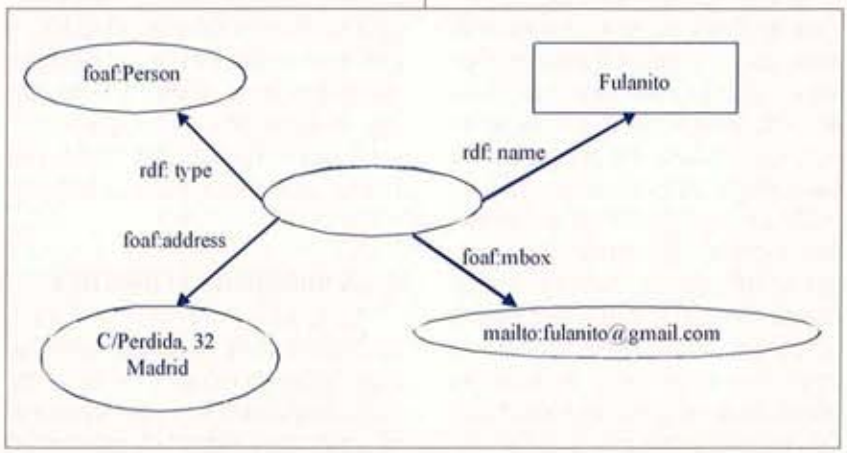


Figura 4.1. Representación conceptual de los datos de contacto de una persona (Fulanito) haciendo uso de las relaciones entre elementos que nos ofrecen rdf y foaf.

Puede parecer intrincado el tener que desarrollar consultas tan elaboradas, pero conviene resaltar que la Web Semántica está orientada a la comunicación entre máquinas [**BaezaYates**07a]. La Web Semántica promueve la creación de Agentes, software autónomo y seudo-inteligente capaz de procesar la información proveniente de diversas fuentes, mezclarla e intercambiar los resultados con otros programas. Imagínese que tiene una cita con Fulanito dentro de cuatro horas, que está correctamente anotada en su agenda electrónica, pero no tiene la dirección de contacto. Cuándo le pida a su agenda electrónica ver a donde se tiene que dirigir, será el agente inteligente programado en la misma la que se encargará de realizar esta intrincada consulta mostrando, posteriormente, los resultados (por ejemplo, un callejero donde se señala la dirección a la que tiene que acudir), de forma totalmente transparente al usuario, y sin que este tenga que aprender a utilizar ninguno de los lenguajes que están dentro del marco de la Web Semántica.

5. LA BÚSQUEDA *TRANSLINGÜE*

Es un hecho ampliamente aceptado que la globalización ha provocado cambios a niveles político y económico. No obstante, en ningún caso dedad cultural, dentro de la cual el lenguaje juega un papel esencial. En este sentido, la mayor penetración inicial de Internet en la cultura anglosajona llevó consigo una inicial predominancia de la lengua inglesa, tanto a nivel de usuarios como de contenidos. Sin embargo, estudios recientes demuestran que la preponderancia del inglés se ha visto reducida, en detrimento de una mayor diversidad lingüística. El estudio "*Lenguas y culturas en la red 2005*", realizado por **Funredes** y **Unión Latina** [UnionLatina05], concluye que el idioma predominante en Internet sigue siendo el inglés, aunque las estadísticas muestran un constante aumento del resto de las lenguas en la red. Entre 1998 y 2005, el porcentaje de internautas de habla inglesa bajó de un 60,5% a un 28,6%, mientras la presencia las páginas web en inglés bajó del 75% al 45%. Si bien es cierto que cada vez más personas en el mundo hablan otros lenguajes además de su lengua materna, y que el inglés es el idioma más extendido como segunda lengua, un usuario que, como ejemplo, hable sólo español, en 2005 tenía acceso únicamente al 4,6% del total de las páginas Web disponibles en Internet.

En la actualidad, la mayoría de motores de búsqueda están limitados a devolver documentos en el mismo idioma de la consulta. Algunos, como

Google, utilizan sistemas de traducción automática para traducir los documentos encontrados, pero, como ellos mismos reconocen en su web, "incluso el *software* sofisticado de hoy en día no se aproxima a la fluidez de un hablante nativo". En este punto, se hace necesario integrar las capacidades de búsqueda con esta creciente diversidad lingüística, algo que no ha pasado desapercibido para investigadores de todo el mundo. La recuperación translingüe de información tiene como objetivo proporcionar a un usuario información en un lenguaje diferente al lenguaje utilizado en la consulta (generalmente, su lengua materna).

Desde que se creara un taller dedicado específicamente a la recuperación translingüe de información, en la conferencia ACM SIGIR de 1996 [**Grefenstette**96], han aparecido programas internacionales de investigación, talleres, conferencias y campañas centradas en el tema, como el *Foro de Evaluación Translingüe* en 2000 [**Peters**01]. Estos programas han impulsado la investigación en la recuperación de información translingüe, obteniéndose unos resultados esperanzadores, pero de momento poco eficientes para ser implementados en motores de búsqueda reales.

En general, se aplican tres estrategias en la mayoría de los sistemas desarrollados: traducción de consultas, traducción de documentos *interactiva* y traducción *en segundo plano* con indexación de documentos. La primera convierte el texto de la consulta en el lenguaje en que se desea recuperar la información, lo que plantea tres retos:

1- Encontrar una traducción para cada uno de los términos escritos en el idioma original. Algunos términos no tienen una traducción directa, y otros son extranjerismos (expresiones lingüísticas tomadas de un idioma extranjero y usada en la lengua propia).

2- Seleccionar las traducciones adecuadas para cada término de entre las posibles, en función del contexto.

3- Un sistema de recuperación translingüe debe ser capaz de asignar diferentes pesos a las diversas traducciones posibles.

La segunda estrategia consiste en realizar una traducción *on line* de los documentos recuperados. En la práctica, la eficiencia de este enfoque es muy baja, debido a la complejidad computacional que requiere la traducción automática y al gran tamaño que presenta generalmente la colección de documentos. Como alternativa, se puede realizar una traducción menos costosa y menos precisa que permita aplicar técnicas de recuperación de información.

Figura 5.1. Resultado de una búsqueda experimental translingüe en Google. A la izquierda, los resultados en español, y a la derecha en inglés.

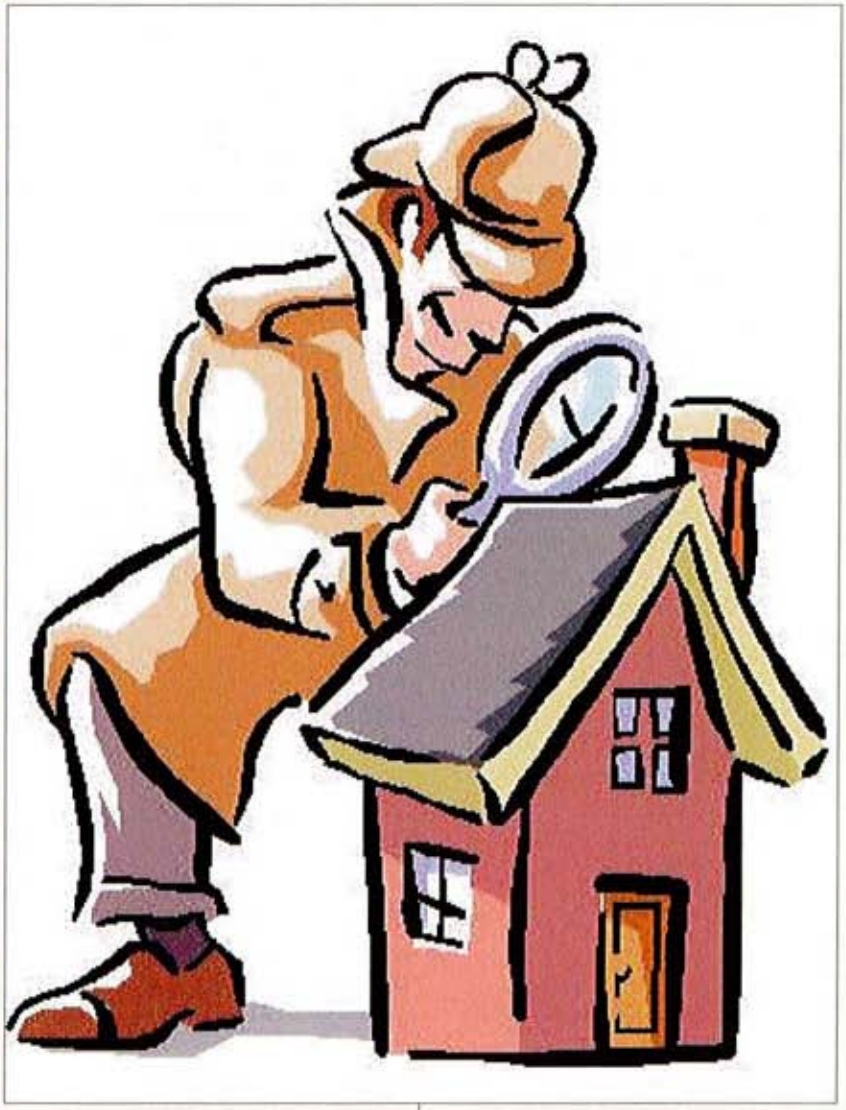
Por último, una tercera estrategia consiste en traducir toda la colección de documentos al lenguaje del usuario, reduciendo la búsqueda translingüe a una búsqueda monolingüe en los documentos traducidos. Este enfoque puede ser demasiado costoso en espacio de almacenamiento si todos los documentos deben ser previamente traducidos a todos los lenguajes disponibles.

Hasta el momento, *Google* es el motor de búsqueda que está obteniendo mejores resultados en la implantación de sistemas de recuperación translingüe. Recientemente, ha puesto a disposición de los usuarios de Internet un prototipo[14] que no se limita únicamente a traducir las páginas devueltas en sus búsquedas, permitiendo realizar consultas en varios idiomas. Como ejemplo, la consulta para la frase "historia de alemania", expresada en español, puede devolver resultados en español y en inglés, según se puede observar en la figura 5.1.

6. DETECCIÓN DE FRAUDE EN BUSCADORES

Los buscadores tienen como principal modelo de negocio la publicidad. La presentación de enlaces patrocinados relevantes a las búsquedas, y en creciente medida, los programas de afiliación de los creadores de conteni-

14 http://translate.google.com/translate_s?hl=es.



dos, les permiten servir publicidad personalizada, adaptada a los gustos e intereses de los usuarios, con tasas de retorno razonablemente altas.

Este modelo de negocio, que genera ingresos millonarios, depende de manera crítica en la confianza que los usuarios tengan en la "pureza" del *ranking* u ordenación presentada ante una consulta. La pureza o calidad del ranking consiste en que el mismo dependa exclusivamente de criterios basados en la calidad y relevancia de los recursos presentados.

Por otra parte, los primeros puestos de un ranking ante una búsqueda popular o relacionada con un negocio son muy codiciados.

Los estudios demuestran que los usuarios no suelen examinar recursos más allá de las dos primeras páginas de resultados[15], lo que hace que un buen posicionamiento atraiga tráfico e ingresos y un mal posicionamiento los haga perder. Y, cuando hay ingresos de por medio, siempre aparecen oportunistas poco éticos.

El fraude más común en los buscadores es el *spam* de buscadores, denominación que toma del problema del correo electrónico basura [**Gómez**02]. El *spam* de buscadores consiste en la obtención de una posición inmerecidamente alta en el *ranking* asociado a una consulta [**Baeza-Yates**07b].

La expresión "inmerecidamente" es suficientemente flexible para acomodar cierto grado de ambigüedad, pero es claro que hace referencia a la temática del recurso o página y su relevancia a la consulta, y al mismo tiempo, al método obtenido para alcanzar la posición.

El ejemplo más obvio es cuando se recupera un recurso pornográfico ante una consulta no relacionada con el sexo. Ello se debe al uso sistemático de palabras clave populares en las páginas, con la esperanza de que el buscador las recupere ante consultas que utilizan dichas palabras clave. Las palabras clave populares están a veces disponibles, y otras son perfectamente predecibles por simple "ingeniería social". Por ejemplo, parece una buena idea utilizar "juegos olímpicos" como palabra escondida en una página desde unos meses antes de la celebración de los mismos, aunque dicha página no trate de los Juegos Olímpicos. De manera similar se pueden usar nombres de celebridades, marcas famosas, etc. Estas palabras son fáciles de ocultar dentro de las etiquetas META del código HTML, o bien en el cuerpo de la página usando colores de bajo contraste con el fondo.

Este tipo de *spam* se denomina spam de palabras clave (*keyword spam*). **Ntoulas** *et al.* [**Ntoulas**06] han investigado las propiedades de este tipo de *spam*, propiedades que pueden ser usadas para detectarlo. Por ejemplo, estos investigadores han detectado una correlación clara entre el hecho de que la página sea *spam* y el número de palabras clave de la página (cuantas más palabras clave, es más probable que sea *spam*). Existen otras propiedades útiles de este tipo de páginas, como el número de palabras en el título de la página, la longitud media de las palabras, la cantidad de texto dentro de las etiquetas "<A>" que indican hiperenlaces, o la fracción de contenido visible.

15 Un estudio realizado por IProspect (empresa consultora de servicios de publicidad interactiva), concluye que el 81% de los usuarios de motores de búsqueda examinan solamente las dos primeras páginas de resultados antes de pulsar alguno de los enlaces a recursos presentados en dichas página

El *spam* en buscadores se basa en la utilización de ingeniería inversa del criterio de ranking de los mismos. Desde la aparición de *Google* y su *PageRank* en 1998 [**Brin**98], la mayoría de buscadores ha ido incorporando en su *ranking* algún criterio de calidad basado en enlaces. El principio genérico que subyace a esta idea es que una página que es enlazada desde muchas otras, es considerada de calidad por sus autores. De esta manera, estos autores transfieren su prestigio a la misma, que lo acumula y transmite a aquellos que en ella se citan, en un proceso iterativo. Este concepto proviene del análisis de citas bibliográficas y el factor de impacto, y es la aplicación primera y más inmediata de las redes sociales (de creadores de contenido) en la búsqueda.

El *spam* de enlaces o *link spam* consiste en la obtención fraudulenta de enlaces entrantes que transfieran prestigio a la página objetivo, lo que la posiciona de manera inmerecidamente alta en el *ranking* de un buscador. La manera más habitual de hacer esto es crear una serie de páginas enlazadas entre sí usando topologías específicas, adaptadas a transferir prestigio a una dada. Estos grupos de páginas se denominan "granjas de enlaces" (*link farms*) [**BaezaYates**05], y las topologías más efectivas (con más capacidad de canalizar prestigio) han sido estudiada con detalle en [**Gyöngyi**05]. Las páginas se pueden crear en dominios donde los abusadores tienen pleno control (en terceros países o proveedores poco éticos, o con poco o ningún control sobre sus recursos de Internet), o bien en entornos interactivos como las bitácoras o *blogs*, sus comentarios, etc. Es crecientemente habitual ver *blogs* plagados de comentarios como "Un *blog* muy interesante. Visita mi página http://...". Es más, estos comentarios no se introducen de manera manual, sino usando programas automáticos que permiten extenderlos de manera masiva.

La detección del *spam* de enlaces no es nada sencilla, porque las páginas creadas para canalizar prestigio hacia una tercera pueden tener un aspecto totalmente inocente y, de hecho, incluir texto legítimo extraído de páginas de terceros. Por ejemplo, en [*BaezaYates*05] se presenta un método basado en técnicas de inteligencia artificial (concretamente de aprendizaje automático) que logra detectar cerca de un 80% de las páginas spam usando una serie de atributos de las mismas que pueden ayudar a distinguirlas de las legítimas, incluyendo el número de enlaces entrantes y salientes, la fracción de páginas que citan una página y son, a la vez, citadas por ella, el valor máximo de *PageRank* de la página, la desviación típica del *PageRank* de las páginas vecinas, etc. Una manera alternativa de evitar el *spam* de enlaces es penalizar a las páginas que parecen ser de *spam*, por lo que se evita que propaguen el

prestigio. Por ejemplo, el sistema *SpamRank* [**Benczúr**05] consiste en detectar para cada página aquéllas que canalizan más prestigio hacia ella, y luego estudiar su regularidad en relación con una distribución estadística estándar. Si son irregulares, su prestigio se penaliza en proporción a su irregularidad.

Una última forma de fraude muy importante en los buscadores es el fraude de *clicks* [*Jansen*06]. Los buscadores estructuran su negocio publicitario en forma de programas de afiliación como los conocidos *AdWorks* y *AdSense* de *Google*. Por ejemplo, una página inserta publicidad gestionada por *Google*, que determina que anuncios son los más relevantes según el contenido de la páginas y por tanto el perfil de sus visitantes. Cuando los visitantes pulsan sobre los anuncios, el dueño de la página recibe dinero de *Google* (que, a su vez, lo recibe del anunciante). Algunos dueños de páginas poco éticos esparcen virus de tipo troyano que hacen que los ordenadores de miles de usuarios de Internet pulsen periódicamente, sin que los usuarios lo sepan, sobre los anuncios, creando *clicks* ficticios para ingresar dinero fraudulentamente. Este problema está muy abierto y es posiblemente el más grave al que se enfrenten los buscadores modernos.

7. CONCLUSIONES

A estas alturas, nadie puede negar el creciente papel que Internet en general, y la Web en particular, tiene en nuestras vidas, tanto a nivel laboral como particular y social. Dada la gran cantidad de información existente en una Web crecientemente interactiva, los motores de búsqueda y los directorios siguen y seguirán durante bastante tiempo siendo la puerta de entrada a la Web para muchos usuarios, y la principal vía para encontrar información para la mayoría.

En este contexto, es muy sensato reconocer que las tecnologías y funcionalidades implementadas en los buscadores van a tener un impacto crítico en nuestras oportunidades de beneficiarnos realmente de la Sociedad de la Información y del Conocimiento. Los autores de este artículo pretendemos repasar algunas de las tecnologías que nuestra experiencia nos permite reconocer como claves en los motores de búsqueda de hoy y mañana, inevitablemente obviando algunas por cuestiones de espacio. Creemos que hemos conseguido ofrecer una puerta de entrada a estas tecnologías para los lectores de este artículo, puerta de entrada que debe permitir ahondar en cada tema a los más interesados. Para aquellos temas que han quedado en el tintero, recomendamos de nuevo la lectura de la reciente monografía de la revista *Novática* "*Buscando en la Web del futuro*" [Pages07].

8.-BIBLIOGRAFÍA

- [BaezaYates05] BAEZA-YATES, R., CASTILLO, C., LOPEZ, V. *Pagerank increase under different collusion topologies. Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

- [BaezaYates06] BAEZA-YATES, R., *Queries and Clicks as a Source of Knowledge.*I Jornadas MAVIR; Tecnologías de la Lengua en la WWW: retos y mercados potenciales.

- [BaezaYates07a] BAEZA-YATES, R., BOLDI, P., GOMEZ HIDALGO, J. M. *Presentación: buscando en la Web del futuro. Novática* (Revista de la Asociación de Técnicos de Informática), nº 185, enero-febrero 2007, año XXXIII, pág. 3-4.

- [BaezaYates07b] BAEZA-YATES, R., BOLDI, P., *GOMEZ HIDALGO, J.M. Recuperación de información con adversarios en la Web. Novática*, nº 185, enero-febrero 2007, año XXXIII, pág. 29-35.

-[Benczúr05]. BENCZUR, A. CSALOGANY, T. SARLOS, UHER, M. *Spamrank—fully automatic link spam detection work in progress.* Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, 2005.

- [BernersLee01] BERNERS-LEE, T., HENDLER, J., LASSILA, O. *The Semantic Web.* Scientific American 284 (5), 35-43, 2001.

- [Brin98] BRIN, *Sergey. Lawrence Page. The Anatomy of a Large-Scale Web Search Engine.* Proceedings of the Seventh World Wide Web Conference, April 14-18, Brisbane, Australia, 1998.

- [Ding05] DING, L., FININ, T., JOSHI, A., PENG, Y., PAN, R., REDDIVARI, P., 2005. *Search on the Semantic Web*, Technical Report TR CS-05-09, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore.

- [Ferragina07] FERRAGINA, P., GULLI, A., *Snaket: A Personalizad Search-result Clustering Engine*, Upgrade: European Journal for the Informatics Professional, Vol. VIII, nº. 1, February 2007.

- [Foaf07] FOAF, 2007. *The Friend of a Friend (FOAF) Project.* Sitio Web accesible en: http://www.foaf-project.org/. [Último acceso: 04/07/2007].

- [Gómez02] GOMEZ HIDALGO, J.M. *Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization.* ACM Symposium on Applied Computing, Universidad Carlos III de Madrid, Spain, March 11 - 14, 2002.

- [Gomez05] GÓMEZ-HIDALGO, J.M., BUENAGA, M., CORTIZO-PÉREZ, J.C., "*The Role of Word Sense Disambiguation in Automated Text Categorization*". Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005. Pp. 298-309, 2005.

- [Google07] Google Inc., 2007. Google Corporate Information: Company Overview. Sitio Web accesible en: http://www.google.com/intl/en/corporate/index.html. [Último acceso: 04/07/2007].

- [Grefenstette96] GREFENSTETTE, G. 1996. *Cross-linguistic information retrieval workshop.* In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in information Retrieval (Zurich, August 18 - 22, 1996). SIGIR '96. ACM Press, New York, NY, 344.

- [Gyöngyi05] GYÖNGYI, Z., GARCIA-MOLINA, H.. *Web spam taxonomy.* Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, 2005.

- [IProspect06] IProspect, 2006. *IProspect Search Engine User Behavior Study*

(April 2006). IProspect White paper, accesible en: http://www.iprospect.com/about/whitepaper_seuserbehavior_apr06.htm. [Último acceso: 04/07/2007].

- [Jansen06] JANSEN, B. J. 2006. *Adversarial Information Retrieval Aspects of Sponsored Search.* Second International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2006). The 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR2006). 6-11 August. Seattle, Washington.

- [Kautz97] KAUTZ, H., SELMAN, B., and SHAH, M. 1997. *Referral Web: combining social networks and collaborative filtering. Common.* ACM 40, 3 (Mar. 1997), 63-65.

- [Liu04] LIU, F., YU, C., MENG, W., "*Personalized Web Search for Improving Retrieval Effectiveness*", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No 1, 28-40, January 2004.

- [Ntoulas06] NTOULAS, A., NAJORK, M., MANASSE, M., FETTERLY, D. 2006. *Detecting spam web pages through content analysis.* In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, NY, 83-92.

- [Pages07] Pages, Llorenç (editor), 2007. *Búsqueda en la Web del futuro. Novática* (Revista de la Asociación de Técnicos de Informática), nº 185, enero-febrero 2007, año XXXIII.

- [Peters01] PETERS, C. (Ed.). *Cross-Language Information Retrieval and Evaluation. Workshop of Cross-Language Evaluation Forum*, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers. Lecture Notes in Computer Science 2069, Springer 2001.

- [Sugiyama04] SUGIYAMA, K., HATANO, K., YOSHIKAWA, M., "*Adaptive Web Search Based on User Profile Constructed without Any Effort from Users*". Proceedings of the 13th International Conference on World Wide Web. Pp. 675-684.

- [Sullivan03] SULLIVAN, D., "*Search Privacy at Google & Other Search Engines*" de Danny Sullivan. Search Engine Watch, April 2003.

- [Sutton98] SUTTON, R., BARTO, A., "*Reinforcement Learning: An Introduction*", MIT Press, Cambridge, 1998.

- [UnionLatina05] Unión Latina, 2005. Lenguas y culturas en la red 2005. Documento Web accesible en: http://dtil.unilat.org/LI/2005/index_es.htm. [Último acceso: 04/07/2007]. ■