Babxel: Búsqueda Multilingüe

Babxel: Multilingual Search

Andrés Velasco Collado
Universidad Europea de Madrid
avc.conti@gmail.com

Resumen: La Web 2.0 ha tenido un enorme éxito gracias a la posibilidad de una interacción dinámica por parte del usuario, ya no sólo a la hora de participar en elementos colaborativos, como puedan ser los foros, sino en compartir/añadir contenido a la Web.

Dos ejemplos claros de este paradigma son YouTube y Flickr. El primero hospeda la mayor parte de los vídeos que podemos encontrar en Internet, y el segundo ha creado la mayor comunidad de fotógrafos existente en la red. Ambos servicios funcionan de una forma similar, el usuario es el que aporta contenidos junto a una información asociada al mismo. Al ser comunidades internacionales, la información añadida por el usuarios se realiza en diversos idiomas, por lo que la búsqueda de recursos multimedia en estos sitios es dependiente del idioma de la consulta.

En este artículo, presentamos Babxel, un sistema de recuperación de información multimedia y multilingüe, nacido como proyecto de fin de carrera de Ingeniería Informática, como extensión y mejora de FlickrBabel. Babxel aprovecha la capacidad de traducción multilingüe automática para generar más resultados de búsqueda relacionado con la consulta del usuario, resultados que se obtienen de las plataformas mencionadas anteriormente.
Palabras clave: Recuperación de información, multilingüe, multimedia, YouTube, Flickr, redes sociales, medios sociales, traducción…

Abstract: The Web 2.0 has been successful thank to the possibility for the user to interact dynamically, not only in collaborative environments, but also in forums and adding/sharing resources to the Web.

Two of the most important examples that apply for this paradigm are Flickr and YouTube. The first one host the majority of videos that are played on the Web, and the latter has built the biggest photographers community all over the Internet. Both services work in the same way, letting the user to upload resources with information attached to it, which help to explain or classify it. As they are international communities, the resources added by the users will be in different languages, so the query to search multimedia resources on these sites is dependable on the query language.

The purpose of this paper is to introduce Babxel, a multilingual and multimedia information retrieval system, born due to a Computer Engineering final degree project, as an upgrade and extension of FlickrBabel. Babxel takes advantages of the automatic multilingual translation to generate more results related to the users' queries, results obtained from the platforms previously mentioned.
Keywords: Information retrieval, crosslingual, multimedia, YouTube, Flickr, Social Networks, Social Media, Translation…

1 Introducción

Los Medios Sociales son herramientas tecnológicas que permiten a los usuarios compartir información y debatir sobre ella. La mayoría de los Medios Sociales son aplicaciones Web que gestionan información textual, como blog (Blogger, Wordpress), microblogging (Twitter, Pownce), wikis (Wikipedia), forums, or Social Networks (Facebook, MySpace, LinkedIn). Existen además otro tipo de aplicaciones basadas en Internet para Medios Sociales que permiten a los usuarios compartir algo más que texto, como herramientas para compartir fotografías (Flickr, Picasa), compartir videos (YouTube, Vimeo), transmisión en directo (Ustream), o compartir audio/música (last.fm, ccMixter, Freesound). Los Medios Sociales más recientes incluyen mundos virtuales (Second Life), juegos online (World of Warcraft, WarHammer Online), compartir juegos (miniclip.com) y Medios Sociales para móviles como las redes sociales nómadas donde los usuarios comparten su situación actual en el mundo real.

Los Medios Sociales han supuesto un cambio en la manera en la que la información se genera y se consume. Al principio, la información era generada por una persona y consumida por muchas otras, pero ahora es generada por muchas personas y consumida por el mismo número o mayor, cambiando las necesidades a la hora de acceder y gestionar la información. Es de destacar el gran número de usuarios y datos que gestionan los Medios Sociales: Facebook y MySpace gestionan entre 400 y 450 millones de usuarios, se estima que se generan al día 1 millón de publicaciones en blogs, servicios de microblogging como Twitter generan 3 millones de mensajes al día, Youtube gestiona más de 150.000 millones de vídeos, etc.[4]

Los recursos multimedia son parte fundamental del auge de los Medios Sociales, ya que muchos de ellos se basan en dichos recursos: YouTube y Vimeo gestionan vídeos, Flickr y Picasa gestionan fotografías, etc. Gracias a lo anteriormente comentado, el usuario pasa de poder comentar con texto un suceso, evento u opinión, a poder grabarlo o fotografiarlo y de esta manera ampliar el abanico de vías de comunicación existentes en la red. El hecho de que YouTube sea el segundo buscador más utilizado por detrás de Google no hace mas que reafirmar la idea de que el usuario prefiere que le cuenten las cosas a leerlas.

El verdadero éxito de todo esto, reside en la posibilidad de combinar diferentes aplicaciones de Medios Sociales de manera que se explote al máximo el potencial de comunicación existente en la red. Un ejemplo claro es la necesidad de buscar material multimedia en numerosas ocasiones para explicar/apoyar información escrita, ya sea en un blog, trabajo, o noticiario online, y como dato tenemos que el visualizador de YouTube se ha embebido en más de 10 millones de páginas Web[3].

De todos los servicios sociales relacionados con la gestión de información multimedia, YouTube y Flickr son los más destacados, tanto por su gran base de usuarios, cantidad de información que manejan y por la calidad de algunos de sus contenidos. A continuación se introducen algunas de sus características principales.

Flickr es una página de almacenamiento de imágenes y videos, servicios Web, y una comunidad de fotógrafos que a fecha de Octubre de 2009 posee 4 mil millones[1] de fotografías aportadas por los mismos usuarios. Aparte de ser un sitio donde compartir el contenido multimedia mencionado anteriormente, suele ser utilizado por los Bloggers para hospedar las imágenes que luego acompañaran a sus entradas escritas.

YouTube es un sitio creado parar el almacenamiento y posterior publicación de videos vía streaming, que se ha convertido en uno de las 5 primeras webs más visitadas en Internet [2]. YouTube recibe cerca de 2 mil millones de visitas al día, lo que le hace el número uno en contenido de video dedicado al público general. El último reporte estadístico de YouTube [3](con fecha primer trimestre del año 2010) muestra los siguientes datos:

- 24 horas de video subidos por minuto.
- 15 minutos es la media al día que pasa una persona viendo contenido en YouTube.

Tanto Flickr como YouTube poseen API's[5][6] (Application Programming

Interface) que ofrecen servicios Web a los desarrolladores de aplicaciones. Esta herramienta permite acceso a todo el contenido multimedia que se hospeda en ambas plataformas, el cual ha sido proporcionado y etiquetado por los propios usuarios, y que sin esta información añadida no sería posible clasificar o darle un significado.

Resulta difícil tratar con información asociada a un recurso multimedia cuando el usuario tiene libertad absoluta en proporcionar dicha información. Esto conlleva a los siguientes problemas:

- Diferentes tipos de texto asociado al contenido multimedia, por ejemplo: etiquetas, descripción título y otros campos de descripción.
- Información asociada en diferentes lenguajes.
- Etiquetado útil para un conjunto pequeño de usuarios, pero que no ayuda al tratamiento de ese contenido multimedia, ni a la búsqueda general por parte de la mayoría de los usuarios, como puedan ser las etiquetas con un propósito específico, por ejemplo: contenido etiquetado para un grupo de personas, etiquetas que hacen referencia el lugar donde fueron tomadas las fotos de las vacaciones, etc.

2 *Sistema de Babxel*

Hoy en día cuando un usuario quiere buscar un contenido lo hace en su lengua materna, ya que de esta manera estará seguro de que entenderá mejor lo que encuentre. A la hora de buscar contenido multimedia esto podría no ser cierto. Por ejemplo, si un usuario quiere buscar una foto de un perro, no le importará que una foto no esté etiquetada en su lengua materna, de manera que con que aparezca un perro y encaje en lo que buscaba, es suficiente.

Esta es una de las premisas principales de Babxel, y está construido de manera que se requiera la mínima interacción por parte del usuario.

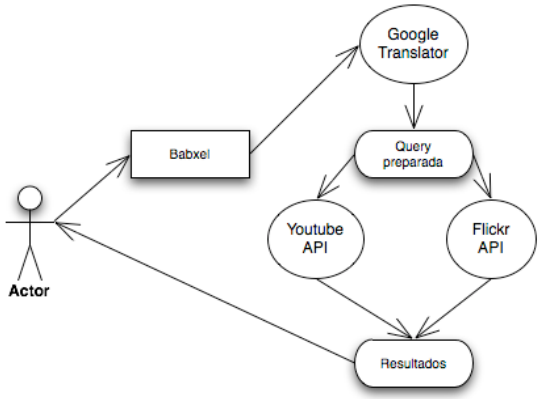

Figura 1: Funcionamiento de Babxel

La figura 1 muestra el funcionamiento básico del sistema. El usuario introduce el texto con el que quiere iniciar la búsqueda, el cuál seguramente introducirá en su lengua materna. Actualmente el usuario puede decidir el idioma de entrada, pudiendo elegir entre dos posibilidades, Inglés o Español (Español por defecto).El idioma de entrada tiene un propósito, y es que no es lo mismo traducir "Can" del Español al Inglés que a la inversa. A su vez podrá modificar los idiomas que se usarán para recuperar resultados, siendo Español, Portugués, Francés, Italiano e Inglés los idiomas configurados por defecto, y ampliables al mismo número de idiomas que proporciona la API de Google Translator, siempre y cuando el usuario lo indique expresamente.

En cualquier momento de la configuración, el usuario puede elegir que tipo de recursos se devolverán, Fotos, Videos o Ambos(por defecto).

Una vez realizado el paso previo de configuración (el cual puede saltarse y buscar con las opciones por defecto), Babxel consulta con Google Translator para traducir la entrada del usuario en los idiomas de salida seleccionados. Con esto se genera una consulta extendida/preparada para recuperar los resultados, esta consulta es preparada teniendo en cuenta el servicio Web que la va a recibir, en este caso las API's de YouTube y Flickr. Los resultados obtenidos son mostrados al usuario por orden de relevancia (por defecto).

Aprovechando el uso de las API's de Flickr y YouTube, se añaden opciones de búsqueda avanzadas, de manera que den más sentido a los resultados recolectados por el sistema para el usuario. Entre las que cabe destacar: la

Geolocalización (aportando el lugar), y la búsqueda de imágenes bajo una determinada licencia Creative Commons. Esta última sin duda es de las más útiles, ya que permite al usuario asegurarse de que las imágenes que está viendo no le van a causar ningún problema legal por el Copyright si las usa en su blog personal.

Otro pilar básico de Babxel, es el sistema de "tracking" que se ha implementado para grabar el recorrido que realiza un usuario desde que realiza la búsqueda hasta que finalmente elige una foto/video que encaja con sus pretensiones. Esto nos permitirá ver qué fotos/videos ha visitado el usuario antes de elegir el contenido final, o incluso si con la primera búsqueda ya ha conseguido el recurso que necesitaba. El objetivo es obtener patrones y tendencias para mejorar los resultados generados por Babxel . Evidentemente, para poder hacer efectiva esta funcionalidad se ha implementado un sistema de gestión de usuarios adaptada expresamente para la arquitectura de Babxel.

3 Experimentos

Se han realizado los siguientes experimentos, consistentes en elegir 10 términos y comprobar el aumento de resultados con Babxel. El objetivo es comprobar que la búsqueda con Babxel resulta ser independiente del idioma, ya que la consulta extendida da los mismos resultados tanto como si la palabra de entrada está en Español, como si está en Inglés.

	Tipo	
Término	**Normal**	**Extendida**
Perro	302,006	8,124,340
Flor	765,678	14,896,604
Coche	138,178	10,746,171
Casa	2,382,801	10,174,704
Gato	514,067	7,382,602
Luna	596,920	2,814,321
Teléfono	59,846	2,926,803
Playa	1,397,968	16,897,348
Montaña	1,039,988	9,562,534
Camiseta	106,369	579,191

Tabla 1: Experimentos en Español

	Tipo	
Término	**Normal**	**Extendida**
Dog	7,519,160	8,124,340
Flower	14,036,992	14,896,604
Car	8,924,196	10,746,171
Home	7,570,160	10,174,704
Cat	6,712,319	7,382,602
Moon	2,089,278	2,814,321
Phone	2,670,410	2,926,803
Beach	14,989,497	16,897,348
Mountain	8,315,239	9,562,534
T-shirt	481,472	579,191

Tabla 2: Experimentos en Inglés

Uno de los resultados más llamativos, es el anteriormente nombrado con la palabra "perro" (Tabla 1), si utilizamos Babxel con el Español como idioma de entrada y único de salida, obtenemos 302,006 resultados, mientras que si utilizamos la configuración por defecto(Español de entrada y Español-Inglés-Francés-Italiano-Portugués como idiomas de salida) obtenemos la cifra de 8,124,340 resultados.

Con los términos “perro”, “flor”, y “coche” se aprecia que la búsqueda extendida aporta números desorbitados con la búsqueda original. Esto se debe a que los usuarios cuando tienen que etiquetar una foto con esos términos, prefieren hacerlo con los términos en Inglés.

En términos generales se puede apreciar el gran aumento de términos en Inglés con respecto al Español. Esto se debe a que el primero es uno de los idiomas más extendidos en la red, y utilizado por la mayoría de la gente para etiquetar contenido de manera que el número de personas que puedan encontrarlo sea mayor.

A raíz de los resultados queda claro que la utilidad de la búsqueda extendida será mucho mayor para una persona de habla no inglesa que para una que si lo hable, ya que el porcentaje de mejora para esta última es inferior.

Con el término "camiseta", la suma de la búsqueda normal en Español(Tabla 1) más la búsqueda normal en Inglés(Tabla 2) supera el resultado total, esto se debe a que existe una gran cantidad de contenido etiqueta con esa palabra en Español e Inglés.

4 Conclusiones y próximos pasos

La utilización de Babxel como buscador multimedia ha resultado ser exitosa en números, ya que los resultado se ven incrementados de manera exponencial comparados con una búsqueda en la lengua materna.

Aparte de las opciones ya nombradas, Babxel contiene opciones avanzadas que permiten al usuario filtrar aún más el contenido que desea encontrar.

A pesar de los números que se consiguen, quien decide si Babxel es útil es el usuario, por ello es necesario conseguir una base importante de usuarios que utilicen la herramienta, y obtener conclusiones a posteriori mediante un análisis de los datos obtenidos por el sistema de "tracking".

En el presente y futuro cercano, las áreas de mejoras de Babxel se centrarán en:

- Mejorar la recuperación de resultados relacionados con el elegido por el usuario.
- Uso de Ontologías para mejorar la construcción de las consultas.
- Soportar más lenguajes de entrada.
- Mejorar el sistema de "tracking" en aras a construir un sistema de recomendación de contenido multimedia para el usuario.

5 Bibliografía

[1] http://en.wikipedia.org/wiki/Flickr

[2] http://www.alexa.com/siteinfo/youtube.com

[3] http://www.website-monitoring.com/blog/2010/05/17/youtube-facts-and-figures-history-statistics/

[4]http://www.socialgamingplatform.com/msm09/

[5]http://www.flickr.com/services/api/

[6]http://code.google.com/apis/youtube/overview.html

[7] J.C.Cortizo, A. Carrero, F.Carrero and B.Monsalve. FlickrBabel: Crosslingual Multimedia Retrieval. Wipley.

[8] Julio Gonzalo, Paul Clough and Jussi Karlgren. Multilingual Image Search from a user's perspective. NED.

[9] Julio Gonzalo, Paul Clough and Jussi Karlgren. Multilingual interactive experiments with Flickr. UNED.

[10] Adrián Popescu and Ioannis Kanellos. Multilingual and Content Based Access to Flickr Images.

[11] Paul Clough, Azzah Al-Maskari, and Kareem Darwish. Providing Multilingual Access to FLICKR for Arabic Users.

[12] Víctor Peinado, Javier Artiles, Julio Gonzalo, Emma Barker, and Fernando López-Ostenero. FlickLing: a Multilingual Search Interface for Flickr.

[13] Srinivasarao Vundavalli. Mining the Behavior of users in a Multilingual Information Access Task.

[14] Adrian Popescu, Gregory Grefenstette, Houda Bouamor. Mining a Multilingual Geographical Gazetteer from the Web.

[15] Paul Clough, Julio Gonzalo, Jussi Karlgren, Emma Barker and Javier Artiles, Victor Peinado. Large-Scale Interactive Evaluation of Multilingual Information Access Systems – the iCLEF Flickr Challenge