Reseñas de las conferencias organizadas por el Máster en Información y Comunicación Científica y el grupo de investigación HUM-466 de la Universidad de Granada durante el mes de mayo de 2011

Roberto Fernández Pérez
robertomafp@gmail.com
5º Licenciatura de Documentación
Universidad de Granada

Primera sesión: 19 de mayo.

10:00 h. Louise Edwards

Europeana, The European Library y las licencias Creative Commons

Lousie Edwards, responsable principal de la European Library y coordinadora de Europeana, estuvo el pasado 19 de mayo en el salón de grados de la facultad de Odontología de la UGR en la primera sesión de las conferencias del Máster en Información y Comunicación Científica de esta misma universidad.

La charla supuso una introducción al paisaje de la información en Europa, panorama que está fuertemente influido por los dos proyectos mencionados en colaboración (sobre todo la European Library) con las principales bibliotecas nacionales europeas, unificadas en el CENL (Conference of European National Libraries).

The European Library y Europeana son dos proyectos ambiciosos que pretenden equilibrar el acceso a la información con respecto a la hegemonía de Google al otro lado del charco. Los factores de éxito de estas plataformas son la orientación al usuario, la robustez del servicio, el acceso al contenido completo de los documentos y la expansión continua mediante nuevas relaciones estratégicas. Entre las prioridades de la European Library para el plan 2010-2012 está dar cabida a las comunidades de investigación y aprendizaje y rediseñar el sistema de búsqueda mejorando herramientas técnicas como el *central index* que complemente poco a poco con la actual búsqueda federada.

Gran parte de la charla derivó hacia la polémica al solicitarse desde el público una explicación del tratamiento de las licencias Creative Commons en Europeana. Al final no quedó muy clara la posición por parte de Edwards quien aseguró que el portal no vende datos ni metadatos de los recolectores pero tampoco garantiza que la información que estos depositan en la plataforma no sea usada con fines comerciales. Desde el Repositorio Institucional de la UGR, DIGIBUG, se reprendió a Europeana por modificar las licencias base "*Reconocimiento-No comercial-SinObraDerivada*" que se aplican en este repositorio de nuestra universidad.

12:00 Luis Ureña López

Procesamiento del Lenguaje Natural para Sistemas de Información Geográfica

En el marco de conferencias ofrecidas por el Máster en Información y Comunicación Científica de la UGR hemos podido disfrutar de una interesante charla a cargo de Luis Alfonso Ureña López, profesor del Departamento de Informática de la Escuela Politécnica Superior de la Universidad de Jaén y Presidente de SEPLN (Sociedad Española de Procesamiento del Lenguaje Natural).

Ureña comenzó su intervención demostrando que los buscadores convencionales, cuyos antecedentes encontramos en Salton y más tarde serían emulados por Google y otros motores similares, no identifican correctamente los términos geográficos de una consulta. Partiendo de este inconveniente los Sistemas de Información Geográfica (GIR) deben solventar algunos problemas inherentes a la ambigüedad e imprecisión de las palabras. De entrada, términos ambiguos como *Reading*, ciudad del condado de Berkshire en Inglaterra y *reading*, gerundio del verbo leer en inglés, están sujetos a la "interpretación" del buscador. Otros términos o conjuntos de términos, bien sea por ambigüedad (*Córdoba*) o por su imprecisión (*north of Spain*) también son elementos a tener en cuenta a la hora de recuperar información.

Estimulada por esas carencias la investigación debe apoyarse en técnicas de Procesamiento del Lenguaje Natural con el fin de reconocer y tratar las referencias geográficas en cualquiera de sus variantes. El trabajo de Ureña y su equipo pretende determinar la idoneidad de las técnicas de PLN a través de la evaluación de las mismas. Para ello han desarrollado un sistema de GIR que mediante la indización geográfica y textual de una base de conocimiento recolectada a través de Wikipedia y fuentes similares permite la detección y reconocimiento de entidades geográficas en las consultas. Implementa además una herramienta de traducción de modo que los resultados se amplíen al inglés.

Aspectos técnicos

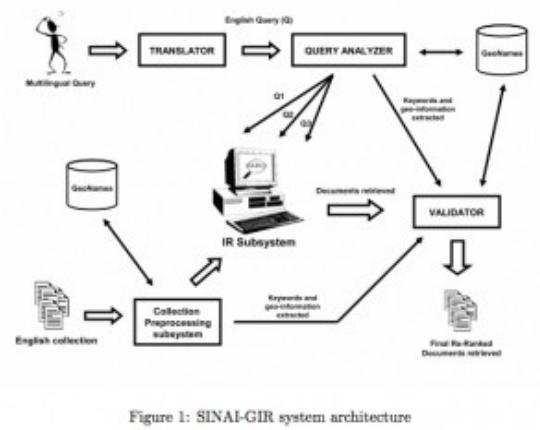

Figure 1: SINAI-GIR system architecture

Arquitectura del sistema (Fuente)

El GIR propuesto por el profesor Ureña consta de dos etapas en su construcción: una etapa de indexado y una etapa de consulta. En la etapa de indexado se confeccionan el índice textual y el índice geográfico. El primero sigue los pasos habituales de preparación de los textos (parseado, eliminación de stop words, steaming, etc.), ponderación de los términos, fichero inverso, etc.
El segundo índice recoge dedicadamente los términos geográficos para prepararlos de cara a

la etapa de consulta. Será en esta fase donde se lleve a cabo un análisis sintáctico de la consulta geográfico-textual descomponiendo la misma en una sintaxis adecuada (del tipo *qué-relación espacial-dónde*. A continuación la pregunta es traducida para mejorar la precisión en el matching y realizar un reconocimiento de entidades y georrelaciones. Para lograr una expansión geográfica de las consultas se lleva a cabo una reformulación de las mismas normalmente mediante la adición de partes al conjunto. Para finalizar se realiza un filtrado de los documentos en base a esta última operación y se reordenan de nuevo los documentos

Al sistema resultante de estos procedimientos se le ha llamado SINAI-GIR. La herramienta alberga un módulo especialmente efectivo (GeoNer) en la detección y reconocimiento de entidades geográficas a través de recursos externos como Wikipedia o el diccionario geográfico Geonames

Más información:

SINAI-GIR System. University of Jaén at GeoCLEF 2008. Participación del grupo de investigación SINAI en el foro de Cross-Language Evaluation Forum GeoCLEF y descripción de la herramienta SINAI-GIR
http://www.clef-campaign.org/2008/working_notes/Perea-Ortega-paperGeoCLEF2008.pdf

GeoNer: un reconocedor de entidades geográficas para inglés basado en GeoNames y Wikipedia
http://www.sepln.org/revistaSEPLN/revista/43/articulos/art4.pdf

Segunda sesión: 20 de mayo

12:00 Luis Collado

Charla de Luis Collado, director de Google Books y Google News en España y Portugal

Siguiendo con las charlas/conferencias organizadas por el Master de Información y Comunicación Científica hoy ha sido el turno de Luis Collado, director de Google Books y Google News en España y Portugal.

El representante en España de estos servicios de la empresa estadounidense ha venido a la UGR a despejar algunas dudas sobre la política de Google respecto a la digitalización de libros y, en menor medida, sobre el servicio de recolección de noticias que también ofrece el buscador.

Prácticamente hasta la mitad de la intervención Collado expuso una síntesis de los principales cambios que está experimentando la sociedad bajo la influencia de internet: de la plaza del pueblo hemos pasado a encontrarnos en la nube donde 2000 millones de usuarios desarrollan un papel activo de consumo, selección, edición y producción de contenidos frente al rol pasivo que se ha venido desarrollando hasta hace algunos años. Precisamente ahora es la gran internet o *la nube* el elemento observador -aunque dinámico- de todos estos acontecimientos. Frente a los que la conciben como un enorme depósito de basura de contenidos Collado compara internet con una imagen absolutamente llena de detalles donde cada uno de los usuarios puede focalizar su atención y consumir los contenidos más satisfactorios a sus intereses o necesidades. Debajo de la foto que supone internet hay, sin embargo, mucha más información: la llamada internet invisible que alberga contenidos de todo tipo. Google, según su representante, se erige como un herramienta para filtrar todo ese contenido, extraerlo del fondo de la charca oscura cuyo fondo no vemos y convertirlo en accesible para todo el mundo (¿nos suena a algo a los bibliotecarios?) Esta misión de filtrado adquiere su máxima expresión en muchas de las herramientas más o menos implementadas que se han diseñado en el seno de Google, como es el caso de Google Fast Flip, un buscador de contenido de prensa. Pero la política de altruismo cultural se percibe mucho más significativamente en el depósito de libros que supone Google Books.

Este querer llegar a todo y a todos que persigue Google implica que el todo sea *todo*, es decir, lo que está y lo que no está en internet. Lo que no está en internet es, sin ir más lejos, toda la producción escrita que la humanidad lleva almacenando durante siglos, el contenido no nativo: libros, revistas, documentos antiguos, etc., que supone una fuente de conocimiento importantísimo parte del cual el buscador no ha querido dejar de ofrecer en parte alentado por su ingente aparato tecnológico. Mediante Google Books el usuario puede buscar, descubrir, hojear, comprar y leer: una emulación de las actitudes que adopta el cliente que entra en una librería. Llegado a este punto Collado echa por tierra la polémica nº 1 con respecto a este servicio: Google Books no permite el acceso gratuito y completo a los libros que digitaliza sino que tan sólo ciertas partes del libro están disponibles para su consulta. Se está por tanto

persiguiendo un equilibrio entre el cumplimiento de los derechos de autor (Google da opción de comprar el libro a la editorial correspondiente) y el acceso libre a los contenidos. Collado insiste para despejar sospechas más oscurantistas: los ingresos que genera Google llegan exclusivamente vía publicidad y nunca por la venta directa de los libros digitalizados.

Otros *lab* sobre los que trabaja la empresa en este campo revelan las inquietudes del buscador en la investigación sobre libros. Ejemplos: Books Ngram Viewer, que realiza un análisis de cualquier palabra a través de su frecuencia de aparición cronológica en los libros; el citado Google Fast Flip, etc.

A la hora de hablar de digitalización Luis Collado hace referencia a la potencialidad de su OCR para poder reconocer tipografías antiguas, a la indexación y a otras operaciones que hacen más complejo el proceso. También se alude a la conservación digital, un problema que preocupa enormemente a la empresa. Siguiendo con esta temática Luis alude a la profecía sobre el final del papel, la cual niega rotundamente en favor de una complementación del libro digital, libro en papel, dispositivos, etc. Lo que sí cambia, sin embargo, es el proceso de lectura, afectada por la maldad del hipervínculo y caracterizada por ser menos secuencial (más picoteo, más microcontenidos) en relación a su análoga tradicional. La lectura de hoy permite discriminar contenidos y diferenciarlos escogiendo lo que se quiere y lo que no se quiere leer.

Casi en la finalización de la charla se habló del entorno actual de la web: el exceso de información al que nos sometemos actualmente contrasta con la brecha digital que aumenta entre quienes tienen y no tienen acceso a la tecnología. En ese sentido la posición de Collado es antideterminista: la tecnología es un medio, nunca un fin que no deje ver el contenido, la verdadera riqueza de la web.

Tercera sesión: 31 de mayo

11:00 Ricardo Baeza-Yates

Charla de Ricardo Baeza-Yates: exprimiendo las folksonomías

El pasado martes 31, en la tercera y última sesión del ciclo de conferencias del Master de Información y Comunicación Científica hemos tenido el privilegio de contar con la presencia del chileno Ricardo Baeza-Yates, director de los laboratorios de Yahoo Search y, hoy por hoy, una de las máximas autoridades en materia de recuperación de información.

La presentación ha girado en torno al fenómeno del *crowsourcing* centrado sobre todo en el uso de folksonomías como herramienta de trabajo y también como materia prima para llevar a cabo experimentos de RI. La minería de etiquetas, de texto y de consultas supone un campo de trabajo enorme y una oportunidad para aprovechar y reutilizar el contenido generado por la web 2.0.

Ricardo defiende la inteligencia colectiva o *crowdsourcing* apoyándose en la frase del periodista americano James Surowiecki: "bajo las circunstancias correctas los grupos son muy inteligentes". Prueba de ello es la cola larga o *long tail* que representa el comportamiento y los intereses de las personas tanto masiva como minoritariamente (contenidos populares y no populares, respectivamente) Este patrón, también reconocible en otras leyes (Zipf, Lokta, etc.), demuestra que el contenido no popular -los intereses o gustos extraños, los lugares no comunes- en la web es tan grande como el contenido popular (p.e., Shakira) y que hay un cierto equilibrio entre las pocas cosas que hacen muchas personas y las muchas cosas que hacemos todos. Un sistema de información que no se ocupe de la cola larga está olvidando los intereses de muchos usuarios. Por ello la cobertura inicial debe partir de la diversidad. Al respecto de la calidad no hay motivo de alarma: mucha cantidad de información buena o mala hace que el nivel de información buena sea alto.

Partiendo de estas premisas Baeza-Yates nos ha enseñado ejemplos, tanto en fase de experimentación como consolidados, de proyectos que se basan en la minería de etiquetas (Flickr, TagExplorer), de texto (Time Explorer) o de consultas para la recuperación de información enriquecida y/o para la toma de decisiones.

La minería de etiquetas utiliza herramientas semánticas como Wordnet o Wikipedia para extraer conocimiento de los metadatos de las fotografías (entre otros documentos). En estos trabajos se utiliza el etiquetado visual por parte del usuario (folksonomía pura) que incrementa la precisión de los resultados o la búsqueda facetada mediante datos de geolocalización de las fotos o rutas inferidas de conjuntos de fotografías.
En todo este proceso hay una contraposición/diálogo constante entre la inteligencia de la folksonomía y el algoritmo subyacente del sistema.

La minería de texto pretende llegar a niveles semánticos partiendo de niveles textuales
(Texto ——> IR ——> NLP ——> Semántica) Esto se logra cruzando los metadatos de

fuentes como Wikipedia con los datos textuales que componen los documentos y con apoyo de herramientas de procesamiento de lenguaje natural. Es decir, se identifican entidades a través de estos cruces de datos así como las relaciones que hay entre ellas (personas, lugares, fechas, etc.) Ejemplos de aplicaciones basadas en minería de texto son Correlator y Time Explorer (un sorprendente sistema que busca entidades en base a predicciones realizadas por las personas y extrae información enriquecida para la toma de decisiones)

Por último, la minería de consultas usa las transacciones y las relaciones entre consultas de usuarios para modificar, reorientar o corregir las preguntas y también para inferir especializaciones, generalizaciones, paralelismos (tesauros), distintas rutas de respuesta, desambiguación, definición de término (diccionarios) y otras operaciones. Estas relaciones se visualizan mediante grafos.

Para Ricardo el tratamiento de los datos populares (en el sentido de *folk*), aunque supone un gran potencial de uso, no está exento de problemas:

-el rendimiento de estos sistemas se puede mejorar en el sentido de llegar a más información aunque con ello se pierda calidad

-la *long tail* implica un problema de evaluación que es generado por el propio carácter *raro* de los contenidos y su situación en la web

-agregación y personalización: se debe equilibrar la investigación de los comportamientos de las personas desde estas dos perspectivas.

-el futuro pasa por equilibrar los problemas de privacidad, la contextualización y el uso de datos a gran escala

En el turno de preguntas han surgido cuestiones muy interesantes con respecto al papel de los motores de búsqueda: éstos, en ocasiones, sesgan los resultados, por ejemplo por las limitaciones visuales de la interfaz (diez resultados por pantalla de alguna forma ocultan el resto de documentos recuperados) El sistema alimenta ciertos comportamientos en las consultas que en general siguen aislando a esos usuarios de la *long tail* y favoreciendo a los consumidores de contenidos mayoritarios. Se ha de perseguir el equilibrio entre ambos sectores.

Otro factor de sesgo de los buscadores es el hecho de replicación de resultados que provoca una retroalimentación de los contenidos de forma que, dada una serie de consultas distintas aunque similares, los resultados son prácticamente siempre los mismos. En este aspecto también se intenta equilibrar la balanza

La reflexión final ha pasado a centrarse en el otro lado, en el del usuario: las consultas mejorarían si éste mejorara los planteamientos, porque el funcionamiento del sistema se regula tanto desde el sistema como desde la actuación de los usuarios como un conjunto.