UABDIVULGA
BARCELONA RECERCA I INNOVACIÓ

09/2009

Aly Conteh, director del Proyecto de Digitalización Masiva de la "British Library"

"Con las técnicas de reconocimiento y análisis de documentos se podrá hacer investigación con una cantidad inmensa de datos"

Aly Conteh es director del Programa de Digitalización Masiva de la "British Library" (Biblioteca Británica), una de las bibliotecas más grandes del mundo, con unos 150 millones de documentos en todas las lenguas y formatos. Actualmente, coordina un proyecto para digitalizar 23 millones de páginas de libros del siglo XIX, 4 millones de páginas de diarios de antes del 1900 y centenares de manuscritos que se pondrán a disposición de investigadores, estudiantes y público general a través del web. El mes de julio pasado fue uno de los invitados del X Congreso Internacional sobre Reconocimiento y Análisis de Documentos (ICDAR), organizado por el Centro de Visión por Computador de la UAB. Dentro de este campo, el análisis y reconocimiento de documentos combina técnicas de procesamiento de imágenes, reconocimiento de formas y

visión por computador para la extracción automática de contenidos textuales o gráficos de documentos digitalizados.

Aly Conteh dirige, desde el año 2003, el Programa de Digitalización Masiva de la "British Libray". Forma parte de la Dirección Ejecutiva del Proyecto Impact, un proyecto de digitalización masiva creado por la Comisión Europea dentro del 7º Programa Marco. Es miembro del grupo de expertos en Digitalización y Preservación Digital de los Estados Miembros de la Comisión Europea y asesora al gobierno británico en materia de digitalización.

- Qué es la tecnología para el análisis y reconocimiento automático de documentos?

- Cuando hablamos sobre análisis y reconocimiento de documentos desde el punto de vista de una biblioteca nacional, como es la “British Library”, estamos considerando una actividad clave que es la digitalización. Esta actividad contempla la manera como manipulamos material histórico, como por ejemplo diarios, libros, manuscritos... y los hacemos accesibles para el web. Lo que nos permiten estas tecnologías es añadir un valor a esta documentación. Por ejemplo, la investigación tradicional con diarios implica tener el diario físicamente o en microfilms y rastrear cada página para encontrar la información que necesitamos. Esto está bien cuando buscas una fecha o un tema concretos, pero si buscas información más general, es más útil y rápido utilizar tecnologías como el reconocimiento óptico de caracteres (OCR-Optical Character Recognition), que nos permite escanear las páginas y detectar los caracteres de cada palabra de manera individualizada. Estos caracteres se incluyen en una base de datos del software, para que las pueda reconocer y detectar. Esto nos permite obtener varias funcionalidades, como por ejemplo buscar palabras clave dentro de un texto. Hasta hace poco debías trabajar con el material físicamente o con microfilms, pero la búsqueda la tenías que hacer tú. Este es el principal beneficio de hacer el cambio de un entorno físico a un entorno digital. Permite a la gente afrontar una investigación de fuentes con una cantidad masiva de datos imposible de realizar antes, lo que abre nuevas vías de investigación con este tipo de material.

- Explíquenos los equipos que usan para este tipo de proyectos.

- Actualmente, utilizamos un dispositivo de captura digital que ha sido diseñado en forma de cuña, con dos cámaras digitales de alta resolución montadas encima. Cuando ponemos el libro en la cuña, capturamos las dos páginas del libro a la vez. Además, el dispositivo gira las páginas automáticamente, mediante un cabezal que tiene un contacto mínimo con las páginas. Esto es posible porque hay una especie de cortina pequeña que rodea el cabezal y crea el vacío en su interior, que es lo que permite girar la página. Con este sistema, nuestra productividad puede ser cuatro veces más alta que con una persona y se evitan el desgaste de las páginas y los desgarros involuntarios. Este dispositivo, sin embargo, no permite manipular documentos desplegables, como mapas de un libro de geografía, así que un operadors los marca y cuando acaba la digitalización del libro, un escáner de cabezal alto permite capturar la imagen del mapa. El software del dispositivo detecta que esta imagen pertenece al libro correspondiente y la inserta con el resto de imágenes, en el lugar adecuado.

- ¿Cuáles son los principales obstáculos para avanzar en estas tecnologías y cuáles serán los desarrollos que podremos ver los próximo años?

- El principal obstáculo es la calidad del OCR. Este software es muy bueno para material impreso moderno, para el que fue desarrollado, pero para material histórico, con el reto de los tipos de letra, el lenguaje y la calidad del papel, que pueden tener unos siglos de antigüedad, nos encontramos en unos niveles de precisión menores de los textos capturados. Esto dificulta el tipo de servicios para la investigación de fuentes que podemos ofrecer. Pienso que, en el futuro, avanzaremos en la sofisticación y el afinamiento del software de OCR, lo cual nos permitirá gestionar textos históricos, no sólo con respecto al reconocimiento de caracteres o soluciones para problemas como por ejemplo los derramamientos de tinta que tienen algunos documentos, sino también en la gestión del lenguaje, con la introducción de diccionarios históricos que permitan detectar palabras en desuso o con grafía diferente a la actual.

- Descríbanos brevemente qué es la "British Library" y dénos algunas cifras sobre su catálogo de libros y documentos

- Lo más sorprendente es que la "British Library" tiene de todo. Tenemos 150 millones de documentos: alrededor de 15 millones de libros y 825 millones de páginas de diarios, etc. Otros objetos son impresiones y dibujos, ítems filatélicos, sellos, manuscritos... Por ejemplo, si digitalizáramos todos los manuscritos medievales anglosajones que tenemos en la biblioteca, crearíamos alrededor de 8 millones de objetos. La "British Library" tiene, probablemente, la colección más grande de manuscritos medievales del mundo, considerando diarios, periódicos y todo tipo de documentos en los que se pueda pensar. Hay material simple, de una sola página, como programas de obras de teatro, y otros materiales extravagantes. Por ejemplo, todas las revistas están depositadas en la biblioteca y muchas tienen objetos insertados, como CDs, barras de labios, juguetes blandos... y la biblioteca también los colecciona.

- ¿Para qué servirá el proyecto de digitalización masiva de la "British Library"?

- El principal objetivo es poner a disposición del gran público todos estos documentos. Actualmente, si quieres ver el material, has de ir físicamente a la biblioteca. No podemos enviar información afuera, tal y como hacen otras bibliotecas, donde puedes ir a buscar un documento y llevártelo a casa. En una ocasión me preguntaban: "¿Para quién hacen esto? ¿Para los investigadores?" Sí, es para los investigadores, pero no sólo para ellos. Por ejemplo, digitalizamos cuatro millones de páginas de diarios y un investigador que estudia la reforma social en la época victoriana en Gran Bretaña puede tener una buena perspectiva de los diferentes puntos de vista de los diarios y de la política de la época. Pero estas fuentes son también importantes para los que hacen genealogía o buscan su historia familiar, o para gente que quiere establecer paralelismos entre el pasado y lo que sucede hoy en el mundo. Queremos ofrecer servicio a todo el mundo, tanto al investigador más serio y preciso hasta el público general que tiene un montón de curiosidades y que quiere saber qué pasó o se dijo un día concreto sobre un tema específico. Lo más importante es poner estos recursos a su disposición, que vean que estamos haciendo este trabajo para satisfacer a un amplio número de intereses.

- Entonces, ¿una biblioteca tal y como la hemos entendido hasta ahora tendrá sentido en un futuro virtual?

- Absolutamente sí. Es interesante porque con el volumen de contenido digital que estamos produciendo, podríamos pensar que la palabra impresa tiende a reducirse, pero no es el caso. Tenemos cada vez más y más material impreso. Personalmente, pienso que la humanidad siempre estará interesada en la representación física. Lo que pasará con las bibliotecas es que se convertirán cada vez más en instituciones híbridas. Necesitarán operar en el entorno digital para ofrecer este soporte y servicio a la investigación, pero la gente siempre querrá interactuar con los documentos físicamente y tener la posibilidad de echar una ojeada atrás y entender cómo los generaciones pasadas y nosotros hoy día consumimos la información. Quizás, con el tiempo, el equilibrio cambiará con respecto a la cantidad de documentos que tengamos física o digitalmente. Pero no creo que la biblioteca como un espacio físico donde podremos ir e interactuar con objetos físicos como libros, DVDs, etc. cambie durante muchas generaciones.

Entrevista: Dímpel Soto. Fotografia: Antonio Zamora
Universitat Autònoma de Barcelona

View low-bandwidth version