

Aly Conteh, director del Projecte de Digitalització Massiva de la "British Library"

09/2009 - **Telecomunicacions, Electrònica i Informàtica.**

"Amb les tècniques de reconeixement i anàlisi de documents es podrà fer recerca amb una quantitat immensa de dades"

Aly Conteh és director del Programa de Digitalització Massiva de la "British Library" (Biblioteca Britànica), una de les biblioteques més grans del món, amb uns 150 milions de documents en totes les llengües i formats. Actualment, coordina un projecte per digitalitzar 23 milions de pàgines de llibres del segle XIX, 4 milions de planes de diaris d'abans del 1900 i centenars de manuscrits que es posaran a disposició d'investigadors, estudiants i públic general a través del web. El mes de juliol passat, va ser un dels convidats al X Congrés Internacional sobre Reconeixement i Anàlisi de Documents (ICDAR), organitzat pel Centre de Visió per Computador de la UAB. Dins d'aquest camp, l'anàlisi i el reconeixement de documents combinen tècniques de processament d'imatges, reconeixement de formes i visió per computador per a l'extracció automàtica de continguts textuals o gràfics de documents digitalitzats.



Aly Conteh dirigeix, des de l'any 2003, el Programa de Digitalització Massiva de la "British Library". Forma part de la Direcció Executiva del Projecte Impact, un projecte de digitalització massiva creat per la Comissió Europea dins del 7è Programa Marc. És membre del grup d'experts en Digitalització i Preservació Digital dels Estats Membres de la Comissió Europea i assessora el govern britànic en matèria de digitalització.

- Què és la tecnologia per a l'anàlisi i el reconeixement automàtic de documents?

- Quan parlem sobre Anàlisi i Reconeixement de Documents des d'un punt de vista d'una biblioteca nacional, com és la "British Library", estem considerant una activitat clau que és la digitalització. Aquesta activitat contempla la manera com manipulem material històric, com ara diaris, llibres, manuscrits... i els fem accessibles per al web. El que ens permeten aquestes tecnologies és afegir un valor a aquesta documentació. Per exemple, la recerca tradicional amb diaris implica tenir el diari físicament o en microfilms i rastrejar cada pàgina per trobar la informació que necessitem. Això està bé quan cerques una data o un tema concret, però si cerques informació més general, és més útil i ràpid utilitzar tecnologies com el reconeixement òptic de caràcters (OCR-Optical Character Recognition), que ens permet escanejar les planes i detectar els caràcters de cada paraula de manera individualitzada. Aquests caràcters s'inclouen en una base de dades del programari, per a què les pugui reconèixer i detectar. Això ens permet obtenir diverses funcionalitats, com ara cercar paraules clau dins d'un text. Aquest és el principal benefici de fer el canvi d'un entorn físic a un entorn digital, perquè permet afrontar una investigació de fonts amb una quantitat massiva de dades impossible de realitzar abans, el que obre noves vies de recerca amb aquest tipus de material.

- Expliqui'ns els equips que fan servir.

- Actualment, utilitzem un dispositiu de captura digital que ha estat dissenyat en forma de falca, amb dos càmeres digitals d'alta resolució muntades al damunt. Quan posem el llibre a la falca, capturem les dues pàgines del llibre a la vegada. A més,

el dispositiu gira les pàgines automàticament, mitjançant un capçal que té un contacte mínim amb les pàgines. Amb aquest sistema, la nostra productivitat pot ser quatre vegades més alta que amb una persona girant les pàgines manualment i s'eviten el desgast de les pàgines i els estrips involuntaris. Normalment, tots girem les planes pel mateix lloc, però el capçal ho fa tocant a la seva meitat, el que ajuda a preservar els llibres. Aquest dispositiu, tanmateix, no pot manipular el que anomenem documents desplegable, com ara mapes d'un llibre de geografia, així que un operador els marca i quan acaba la digitalització del llibre, un escàner de capçal alt permet capturar la imatge del mapa. El programari del dispositiu detecta que aquesta imatge pertany al llibre corresponent i la insereix amb la resta d'imatges, en el lloc adequat.

- Quins són els principals entrebancs per avançar en aquestes tecnologies i quins seran els desenvolupaments que podrem veure els proper anys?

- El principal entrebanc és la qualitat de l'OCR. Aquest programari és molt bo per a material imprès modern, per al qual va ser desenvolupat, però per a material històric, amb el repte dels tipus de lletra, el llenguatge i la qualitat del paper, que poden tenir uns segles d'antiguitat, ens trobem en uns nivells de precisió menors dels textos capturats. Això dificulta el tipus de serveis per a la recerca de fonts que podem oferir. Penso que, en el futur, avançarem en la sofisticació i l'afinament del programari d'OCR, la qual cosa ens permetrà gestionar textos històrics, no només pel que fa al reconeixement de caràcters o solucions per a problemes com ara els vessaments de tinta que tenen alguns documents, sinó també en la gestió del llenguatge, amb la introducció de diccionaris històrics que permetin detectar paraules en desús o amb grafia diferent a l'actual.

- Descriu'ns breument què és la "British Library" i doni'ns algunes xifres sobre el seu catàleg de llibres i documents.

- El més sorprenent és que la "British Library" té de tot. Tenim 150 milions de documents: al voltant de 15 milions de llibres i 825 milions de pàgines de diaris, etc. Altres objectes són impressions i dibuixos, ítems filatèlics, segells, manuscrits... Per exemple, si digitalitzéssim tots els manuscrits medievals anglosaxons que tenim a la biblioteca, crearíem al voltant de 8 milions d'objectes. La "British Library" té, probablement, la col·lecció més gran de manuscrits medievals del món, considerant diaris, periòdics i tot tipus de documents en els que es pugui pensar. Hi ha material deslligat, d'una sola pàgina, com programes d'obres de teatre, i altres objectes estrafolaris. Per exemple, totes les revistes estan dipositades a la biblioteca i moltes tenen objectes inserits, com CDs, barres de llavis, joguines toves... i la biblioteca col·lecciona tots aquests trastos. Ho col·leccionem tot.

- Per a què servirà el projecte que realitzen?

- El principal objectiu és posar a disposició del gran públic tots aquests documents. Actualment, si vols veure el material, has d'anar físicament a la biblioteca. No podem enviar informació a fora, tal com fan altres biblioteques, on pots anar a cercar un document i endur-te'l a casa. En una ocasió em preguntaven: "Per a qui feu això? Per als investigadors?". Sí, és per als investigadors, però no només per a ells. Per exemple, digitalitzem quatre milions de pàgines de diaris i un investigador que estudia la reforma social en l'època victoriana a Gran Bretanya pot tenir una bona perspectiva dels diferents punts de vista dels diaris i de la política de l'època. Però aquestes fonts són també importants per als que fan genealogia o cerquen la seva història familiar, o per a gent que vol establir paral·lelismes entre el passat i el que succeeix avui al món. Volem oferir servei a tothom, tant a l'investigador més seriós i precís fins al públic general que té un munt de curiositats.

- Llavors, una biblioteca tal com l'hem entès fins ara tindrà sentit en un futur virtual?

- Absolutament sí. És interessant perquè amb el volum de contingut digital que estem produint, podríem pensar que la paraula impresa tendeix a reduir-se, però no és el cas. Tenim cada vegada més i més material imprès. Personalment, penso que la humanitat sempre estarà interessada en la representació física. El que passarà amb les biblioteques és que esdevindran cada cop més institucions híbrides. Necessitaran operar en l'entorn digital per oferir aquest suport i servei a la recerca, però la gent sempre voldrà interactuar amb els documents físicament i tenir la possibilitat de fer una ullada enrere i entendre com les generacions passades i nosaltres avui dia consumim la informació. Potser, amb el temps, l'equilibri canviarà pel que fa a la quantitat de documents que tinguem física o digitalment. Això pot canviar, però no crec que la biblioteca com un espai físic on podem anar i interactuar amb objectes físics com llibres, DVDs, etc. canviï durant moltes generacions.

Entrevista: Dímpel Soto. Fotografia: Antonio Zamora

Universitat Autònoma de Barcelona