DALLA CARTA AI BIP

La trascrizione di un testo dal volume cartaceo alla memoria elettronica del computer si può effettuare in vari modi:
- direttamente, ribattendo pazientemente il testo "a mano";
- con l'ausilio di hardware/software specifici, quali scanner e programmi OCR (i link ipertestuali rimandano a una spiegazione su cosa sono gli scanner e gli OCR).

Cosa è un programma O.C.R.

La sigla inglese O.C.R. sta per *riconoscimento automatico dei caratteri* (Optical Character Recognition). I programmi per l'O.C.R. sono sistemi esperti in grado di convertire l'immagine digitalizzata di un documento in testo.
Solitamente per *digitalizzare* una pagina di testo si usa uno scanner da tavolo.
Un noto programma per l'OCR è *Omnipage Pro* della Caere.

Cosa è uno scanner

Lo scanner è un apparecchio che, similmente ad una fotocopiatrice, è in grado di riprodurre una immagine o un testo. Differentemente da quanto fa una fotocopiatrice, però, lo scanner non stampa ciò che riproduce, ma lo trasmette - in formato digitale - al computer al quale è collegato.
Se l'immagine acquisita dallo scanner è, ad esempio, la pagina di un libro, un computer dotato di un programma OCR è in grado di trasformare tale *immagine* in *testo*.
Si noti la differenza tra "immagine" di una pagina e "testo" in essa contenuto. Come noto, un computer non è in grado di "leggere" (cioè di trasformare l'immagine dei caratteri in parole di senso compiuto), così ha bisogno di programmi specifici (gli OCR) per svolgere una funzione che a noi esseri umani appare semplice.
Quando la qualità dell'immagine di una pagina acquisita via scanner è buona e l'impaginazione del documento non è troppo complessa, il computer riesce a interpretarne il contenuto con un indice di affidabilità superiore al 99%. Non si riesce, purtroppo, ad avere una affidabilità del 100% perché il computer non è in grado di "capire" ciò che sta leggendo, così notiamo a volte ch in testi acquisiti via scanner ci sono degli "1" in luogo delle "i" o delle "l", degli "0" in luogo delle "O" e così via. In altri termini, un essere umano capisce immediatamente che la parola "alber0" (cioè "alber" seguito dal numero zero) non ha senso, e corregge istintivamente in "albero". Un computer, ovviamente, no.
Esistono anche scanner manuali (o handy scanner). Sono di piccole dimensioni ed hanno costi solitamente contenuti. Purtroppo questi modelli non sono adatti al lavoro di acquisizione digitale di testi, perché danno luogo a troppi errori di interpretazione da parte dei programmi OCR.

Optical Character Recognition

da Wikipedia, l'enciclopedia libera.

I sistemi di **Optical Character Recognition** (*riconoscimento ottico dei caratteri* detti anche **OCR**) sono programmi dedicati alla conversione di un'immagine contenente testo in testo modificabile con un normale programma di videoscrittura. Solitamente le immagini sono acquisite da uno scanner d'immagini o da un sistema di digitalizzazione che si avvale di una telecamera o di una webcam. Il testo viene convertito in testo ASCII, Unicode o nel caso dei sistemi più avanzati in un formato in grado di contenere anche l'impaginazione del documento. I programmi di OCR si avvalgono dei progressi dell'intelligenza artificiale e dell'evoluzione degli algoritmi legati al riconoscimento delle immagini.

Lettura ottica vs. riconoscimento digitale dei caratteri

Originalmente, le distinzioni fra lettura ottica dei caratteri (usando le tecniche ottiche quali gli specchi e gli obiettivi) e il riconoscimento digitale dei caratteri (usando gli algoritmi di separazione ed analisi del testo) erano notevoli ed infatti erano considerati campi separati. Poiché non è rimasta più quasi nessuna applicazione legata alle tecniche di lettura ottica si è esteso il termine OCR che ora indica il riconoscimento dei caratteri digitali indipendentemente dalla sorgente delle immagini.

Addestramento

I sistemi OCR per funzionare correttamente richiedono una fase di "addestramento". Durante questa fase al sistema vengono forniti degli esempi di immagini col corrispondente testo in formato ASCII o simile in modo che gli algoritmi si possano calibrare sul testo che usualmente andranno ad analizzare. Questo addestramento è fondamentale se si considera che gli elementi che analizzano il testo non sono altro che delle reti neurali e come tali richiedono un addestramento per funzionare.
Gli ultimi software di OCR utilizzano algoritmi in grado di riconoscere i contorni e in grado di ricostruire oltre al testo anche la formattazione della pagina.

Breve storia dei programmi di OCR

Il sistema postale degli Stati Uniti d'America utilizza sistemi di OCR fin dal 1965. La necessità di riconoscere le destinazioni delle missive e di organizzarle in modo automatico ha spinto la ricerca nel settore dell'OCR. I sistemi OCR leggono il Codice Postale scritto sulle lettere e provvedono ad stampare sulle missive un codice a barre che rappresenta la destinazione della lettera. Per impedire che il codice a barre disturbi la lettura dell'indirizzo complicando il lavoro dei postini il codice a barre viene stampato con un inchiostro visibile solo se illuminato da una luce con lunghezza d'onda nell'Ultravioletto. Il codice a barre viene utilizzato da macchine smistatrici per indirizzare la corrispondenza all'ufficio postale corrispondente che si preoccuperà di recapitarlo al destinatario. Un metodo analogo è in uso dalle poste italiane per la gestione della corrispondenza.

OCR di caratteri stampati

Mentre il riconoscimento esatto di un testo scritto con un alfabeto latino oramai è considerato un problema risolto quasi perfettamente, il riconoscimento della scrittura a mano libera e il riconoscimento degli alfabeti non latini è un problema che tuttora non ha trovato delle soluzioni realmente soddisfacenti e infatti è tuttora oggetto di studi e ricerche.

OCR a mano libera

Sistemi per riconoscere della scrittura a mano libera hanno avuto un discreto successo commerciale se integrati in prodotti tipo PDA o computer portatili. Il precursore di questi dispositivi è stato il dispositivo Newton prodotto dall'Apple. Gli algoritmi di questi dispositivi funzionano adeguatamente perché si impone all'utente di imparare a scrivere le lettere seguendo un certo schema predefinito in modo da minimizzare i possibili casi di ambiguità. Queste strategie non si possono applicare nei documenti scritti su carta infatti il riconoscimento a mano libera è un problema tutt'altro che risolto. I tassi di accuratezza dell'80%-90% sui caratteri scritti a mano in modo accurato e pulito possono essere raggiunti in modo relativamente semplice. Ma un tasso di accuratezza così basso produce diverse decine di errori per pagina rendendo le tecniche di scrittura a mano libera poco utili nella maggior parte dei casi.

OCR del corsivo

Il riconoscimento del testo scritto in corsivo è un campo di ricerca attivo, e attualmente l'accuratezza del riconoscimento è persino inferiore a quella di un testo scritto a mano. Più elevati livelli di accuratezza non saranno possibili fino a che non si useranno informazioni aggiuntive derivate da un analisi contestuale o grammaticale del testo. Per esempio, riconoscere le intere parole da un dizionario è più facile che provando ad analizzare i diversi caratteri singolarmente: analizzare le parole intere consente di eliminare molte ambiguità legate al riconoscimento.
Conoscere il contesto dello scritto consente di eliminare altre ambiguità, per esempio un documento che parla di storia conterrà probabilmente molte date e quindi una linea verticale seguita da un simbolo 9 consentirebbe di ipotizzare che probabilmente la linea è un 1 piuttosto che una l minuscola o una i maiuscola. La conoscenza della grammatica della lingua analizzata può contribuire a determinare se una parola è probabilmente un verbo o un nome, per esempio, consentendo un'accuratezza maggiore. Purtroppo i caratteri corsivi di molte lettere non contengono abbastanza informazioni per effettuare un'analisi corretta e infatti l'accuratezza difficilmente può superare il 98%.

Aree di Ricerca

Un problema particolarmente difficile per i calcolatori e gli esseri umani è quello del riconoscimento di documenti danneggiati contenenti molti nomi o comunque informazioni non deducibili dal contesto. Le pagine possono essere danneggiate dall'età, acqua o dal fuoco e dei nomi possono essere obsoleti o contenere errori d'ortografia. Le tecniche di elaborazione delle immagini dei calcolatori possono aiutare gli esseri umani nella lettura dei testi estremamente antichi come i documenti lasciati da Archimede o i rotoli del mar Morto. L'utilizzo del calcolatore come supporto all'uomo e viceversa è un ambito di ricerca molto interessante e potenzialmente prolifico. Il riconoscimento dei caratteri è stato un settore soggetto ad un'intensa ricerca fin dai tardi anni cinquanta. Inizialmente è stato percepito come problema semplice, ma è risultato essere un problema molto più interessante. Serviranno ancora decenni di studi prima che il calcolatore sia in grado di riconoscere un testo con la stessa accuratezza di un essere umano, sempre che ciò sia possibile.

MICR

Un'applicazione dove l'esattezza e la velocità di riconoscimento dei sistemi OCR sui caratteri supera quella umana è quella dei MICR, dove l'accuratezza è molto elevata e gli errori variano intorno a un errore rilevato su 20.000 - 30.000 controlli. Questa precisione si ottiene grazie all'utilizzo di inchiostri speciali contenenti materiale magnetico (ossido di ferro).

Scanner d'immagine

In informatica, lo **scanner** è la periferica in grado di acquisire immagini da superfici piane (fogli di carta, libri) per poterle elaborare mediante appositi software di fotoritocco, o di riconoscere testi mediante OCR. Il suo funzionamento è basato su un lettore ottico che "scandaglia" l'oggetto da digitalizzare trasformandolo in una sequenza di dati interpretabile come immagine dal computer. Gli scanner sono gli strumenti principali necessari per una corretta Gestione Elettronica dei Documenti (GED). Esistono scanner in grado di elaborare anche centinaia di pagine al minuto, richiamando automaticamente le pagine da un apposito contenitore, così come avviene usualmente per una stampante.

Aspetti linguistici

Non esiste una convenzione universalmente condivisa sul verbo da impiegare, in italiano, per indicare l'uso dello *scanner*, anche se il prodotto del suo funzionamento è unanimemente chiamato *scansione*. Le traduzioni più comuni vedono l'impiego di neologismi come *scansionare* (nel qual caso lo strumento è detto *scansionatore*) e *scannerizzare*, anche se è impiegato pure *scandire*, più adatto e corretto poiché tale verbo descrive correttamente ed accuratamente l'azione effettuata dallo *scanner* (che è detto dunque *scanditore*), oltre a essere quello da cui deriva naturalmente il termine *scansione*; d'altra parte il verbo inglese *to scan*, dal quale deriva il termine *scanner* (letteralmente: l'oggetto che esegue l'azione descritta dal verbo), può essere tradotto in italiano proprio con "scandire, esaminare sistematicamente". Talvolta viene usato a sproposito *scannare*, che però vuol già dire "uccidere tagliando la gola": per questo a volte è adoperato scherzosamente.