فرامرز هنروران[1]

او. سی. آر و کاربردهای آن در کتابخانه‌ها و مراکز اطلاع‌رسانی

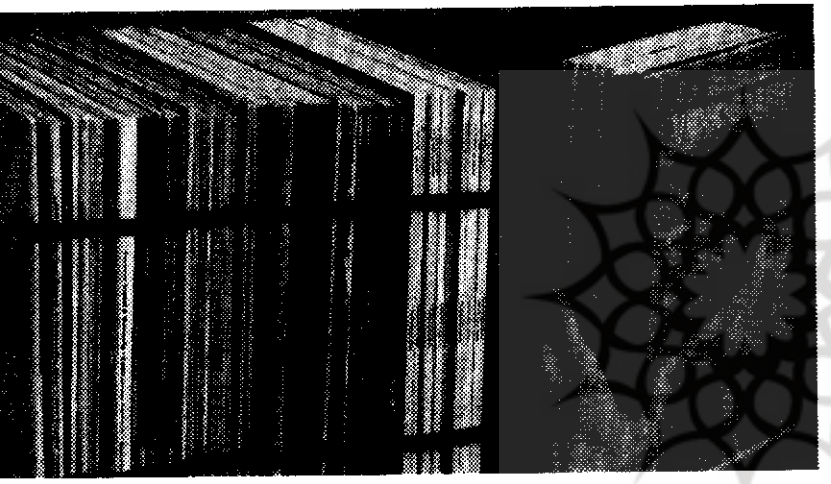
چند سالی از بحث کتابخانه‌های بدون کتاب و اداره‌های بدون کاغذ[2] می‌گذرد ولی با نگاهی اجمالی به وضعیت فعلی در دنیا به سادگی می‌توان متقاعد شد که هنوز راهی طولانی برای رسیدن به این مهم به طور نسبتاً آرمانی باقی مانده است. کاغذ به سبب سهولت استفاده و سندیتش هنوز یکی از محمل‌های اصلی اطلاعات به شمار می‌رود. ولی حجم زیاد، عمر محدود، کندی بازیابی و انتقال آنچه بر این محمل نقش بسته استفاده از آن را در عصری که دستیابی و انتقال هر چه سریعتر اطلاعات را عامل تعیین کننده کرده است، چه در زمینه درج اطلاعات جدید و چه تکثیر اطلاعات قدیمی روزبه روز غیراقتصادی‌تر می‌سازد، از جهت دیگر تولید آن را با توجه به محدود بودن منابع گیاهی و وارد آوردن خسارات جبران‌ناپذیر به محیط‌زیست، هر روز به نقطه‌ای بحرانی‌تر نزدیک می‌سازد. مجموع این عوامل باعث شده که طی بیست‌وچند سال گذشته

OCR

متخصصان سخت‌افزاری و نرم‌افزاری کامپیوتر و دستگاههای جانبی آن به فکر تهیه دستگاه‌ها و برنامه‌هایی باشند تا نیاز به صفحه کلید کامپیوتر را برای ورود اطلاعاتی که قبلاً بر روی کاغذ نقش بسته برطرف سازد. به غیر از عوامل برشمرده شده دو دلیل دیگر وجود و توسعهٔ چنین برنامه‌هایی را تقویت می‌کند: ۱) حجم اطلاعاتی که بر روی کاغذ از زمان اختراع چاپ تاکنون درج شده نسبت به اطلاعاتی که بر روی حافظه‌های ثانویه کامپیوتر (دیسک و دیسکت، نوار مغناطیسی...) موجود است از میزان بسیار بالاتری برخوردار است و اگر قصد ورود تمامی این اطلاعات به کمک کاربران باشد انجامش به سالها وقت و تعداد بی‌شماری کامپیوتر نیاز داشته و هزینه فراوانی می‌طلبد؛ ۲) اشتباهاتی که توسط کاربر در هنگام ورود اطلاعات رخ می‌دهد در بیشتر موارد به ۸ درصد یا بالاتر می‌رسد و زمانی که جهت رفع این اغلاط تلف می‌شود بسیار زیاد است. وجود مسائل بالا منجر به تولید و تکمیل دستگاههای گوناگون پویشگر[3] و نرم‌افزارهای مختلف شناسائی بصری و هوشمند حروف گردیده است. پویشگر وسیله‌ای است که یک صفحه چاپی - چه متنی و چه تصویری، یا ترکیبی از هر دو - را پویش می‌کند. طریقه عمل بدین صورت است که ابتدا نوری به صفحه چاپی از طریق یک منبع نوری (که برحسب نوع پویشگر به صورت خط به خط یا صفحه به صفحه است) تابانیده شده و نور منعکس شده از صفحه به طرف آرایه‌ای از دیودهای نوری[4] هدایت می‌شود. این نور پس از برخورد با این دیودها نوسانات الکتریکی تولید می‌کند که منطبق با نقاط سیاه (یا رنگی) و زمینه کاغذ که تشکیل دهنده متن و یا تصویرند، می‌باشد این نوسانات به صورت قیاسی[5] به کمک مدارات الکترونیکی پویشگر تقویت شده و تبدیل به اطلاعات رقمی[6] که به صورت آرایه‌ای از اطلاعات دودویی و یا نقشه بیتی[7] در آمده و به حافظه RAM[8] (حافظه با دسترسی دلخواه) یا حافظه ثانویه کامپیوتر منتقل می‌گردد و در این مرحله است که کار نرم‌افزار او سی آر شروع می‌شود. این نرم‌افزارها نقشه بیتی تولید شده توسط پویشگر را خوانده و در حین فرایندی خاص قسمتهای سفید یا خالی صفحه پویش شده را مشخص می‌کند تا بدین طریق قادر شود محل خطوط متن و تصاویر احتمالی موجود را برای ادامهٔ کار

خود تعیین کند. در مرحله اول تبدیل تصویر به متن، این نرم‌افزار سعی می‌کند تا با مقابلهٔ نقطه به نقطه[9] تصویر صفحه پویش شدهٔ سند با مجموعهٔ کاراکترهای تعریف شده که نرم‌افزار او سی آر به حافظه RAM کامپیوتر سپرده است کاراکترهای کاملا" مشابه را شناسایی کرده و از تصویر استخراج نماید. این مجموعه کاراکترهای تعریف شده شامل تمامی فونتهای پرکاربرد (فونت[10] مجموعه‌ای است از حروف چاپی در اندازه، شکل و سبک مشخص تشکیل شده از حروف بزرگ و کوچک، اعداد و علائم مختلف) الفبای لاتین (یونانی یا سریلیک) است. بدین سبب که در این مرحله از شناسایی حروف انطباق کامل لازم است اگر چگونگی تصویر پویش شده از لحاظ کیفیت پایین باشد این روش چندان کارآمد نبوده و باید از روش دیگری بهره برد که در مرحلهٔ بعدی موجب شناسایی کاراکترهای شناخته نشده در مرحله اول می‌شود. این روش وقت‌گیر به نام استخراج ویژگی[11] یا تحلیل ویژگی[12] یکی از روشهایی است که علاوه بر نرم‌افزارهای اوسی‌آر در نرم‌افزارهای ICR[13] (نرم‌افزارهایی برای استخراج حروف از متون خطی یا چاپ شده با فونت‌های غیراستاندارد) نیز کاربرد وسیعی دارد. در این روش ابعاد، نسبت ابعاد حواشی، تعداد گوشه‌ها و لبه‌های هر کاراکتر مورد تجزیه و تحلیل قرار گرفته و تشابه آن با کاراکترهای تعریف شده سنجیده می‌شود، به عبارت دیگر، در این روش ارتفاع هر کاراکتر نسبت به خط کرسی محاسبه شده و هر کاراکتر به صورت ترکیبی از خطوط مستقیم و منحنی‌های باز و بسته در مقایسه با یکایک کاراکترهای تعریف شده مورد تجزیه و تحلیل قرار می‌گیرد تا از نتایج این تجزیه و تحلیل کاراکتر شناسایی شود. در مرحله سوم به دلیل آن که دو مرحله قبل امکان شناسایی تمامی کاراکترهای متن را نمی‌دهد، نرم‌افزار به دو روش (بسته به نوع نرم‌افزار) برای کشف کاراکترهای ناشناخته کمک می‌طلبد، یا علائمی همچون @ یا # را جایگزین کاراکترهای ناشناخته می‌کند و به کاربر اعلام می‌دارد که خود این علائم را پیدا کرده و کاراکتر مناسب را جایگزین این علائم سازد و یا: هر کاراکتر ناشناخته را بر روی صفحه نمایش کامپیوتر[14] مشخص کرده و از کاربر می‌خواهد تا کلید مربوط به این کاراکتر را بر روی صفحه کلید[15] کامپیوتر فشار دهد تا به جای کاراکتر مجهول کاراکتر صحیح جانشین شود.

در مرحله ماقبل آخر که در بعضی نرم‌افزارهای او سی آر وجود دارد با کمک یک برنامهٔ غلطیاب املایی، کلمات استخراج شده، بررسی شده و اغلاط احتمالی در موقع شناسایی کاراکترهای هر کلمه بدینوسیله تصحیح می‌شود. در پایان نرم‌افزار او سی آر تمامی کاراکترهای مدرک را به صورت فایل متنی اسکی[16] یعنی به شکلی که در کامپیوتر قابل استفاده است تبدیل کرده و در این حال است که می‌توان متن را به نرم‌افزارهای واژه پرداز یا نرم‌افزارهای بانک اطلاعاتی متنی داد یا آن را به هر محل دیگر از طریق کامپیوتر منتقل کرد، تمام



مراحل بالا بسته به نوع مدرک و نوع نرم‌افزار و سخت‌افزار به کار رفته از یک تا چند دقیقه وقت می‌برد که در مقایسه با انجام این کار با کمک کاربر در اکثر موارد از سرعتی ۱۰ تا ۳۰ درصد بالاتر برخوردار است، و ذکر این نکته نیز خالی از فایده نیست که بر حسب تجربیاتی که انجام گرفته مشخص شده که هزینه پیدا کردن و تصحیح یک حرف (البته در الفبای لاتین) حدود ده هزار برابر هزینهٔ تصحیح یک حرف رد شده بدلیل ناخوانا بودن یا غیرقابل تشخیص بودن از طرف نرم‌افزار او سی آر است و همچنین تجربیات مشخص ساخته که استفادهٔ از اوسی‌آر حدود ۵۰ درصد در هزینه‌ها صرفه‌جویی می‌کند.

او سی آر در کتابخانه‌ها و مراکز اطلاع‌رسانی

کتابخانه‌ها و مراکز اطلاع‌رسانی از جمله مؤسساتی هستند که تهیه اطلاعات و انتقال آن اصلی‌ترین عامل وجودیشان بوده و هر چه در این امور سریعتر عمل کنند به هدفشان نزدیکتر می‌شوند. در کتابخانه‌های کنونی بویژه

OCR

کتابخانه‌های کشورهای در حال توسعه همه چیز یا لااقل ۹۸ درصد اطلاعات موجود بر کاغذ نقش بسته است. از فهرست‌برگه تا یک‌یک نامه‌های بخش‌های مختلف یک کتابخانه ـ همان گونه که گفته شد ـ از مهمترین قابلیت یعنی امکان دستیابی و انتقال سریع بی‌بهره‌اند علاوه بر آن امکان جستجو، تصحیح، تجزیه و تحلیل آماری در آنها تنها به روش دستی امکان‌پذیر است. اما با استفاده از روش او سی آر تمامی این اطلاعات در زمانی کوتاه و با هزینه‌ای پایین نسبت به روشهای دیگر خواهند توانست از تمامی قابلیتهای ذکر شده بهره‌مند گردند. به طور مثال برای ارائه متن مقالات مجلات و روزنامه‌ها، دیگر نیازی به جستجو در میان صدها تصویر میکروفیلم، نخواهند بود. زیرا با استخراج متن این مقالات توسط اوسی‌آر و دادن این اطلاعات به برنامه‌های بانک اطلاعات متنی، به راحتی می‌توان به هر مقاله دست یافت و آن را به کمک خطوط مخابراتی به هر نقطه‌ای فرستاد؛ و صدها مورد استفاده دیگر که بسته به نوع کتابخانه و میزان وسعت آن متفاوت است.

معرفی چهار برنامه مشهور او سی آر و مقایسه آنها با هم

حال که در مورد مزایای روش او سی آر و کاربردهای آن در حوزه کتابداری و اطلاع‌رسانی بحث شد، به معرفی و مقایسه چهار برنامهٔ او سی آر که تحت برنامه ویندوز مایکروسافت[17] کار می‌کند، پرداخته می‌شود تا بتوان در ضمن این مقایسه با توجه به نیازهای کتابخانه یا مؤسسه‌ای که در آن به کار مشغول است دست به انتخاب یکی از این نرم‌افزارها زده یا کلاً با مزایا و معایب هر یک از آنها آشنا شود. این نرم افزارها عبارتند از:

۱. Omnipage Professional 5.0 از شرکت Caere

۲. Recognita Plus 2.0 International از شرکت Recognita

۳. TextBridge 2.0 از شرکت Xerox Imaging Systems

۴. WordScan Plus از شرکت Calera Recognition Systems

این مقایسه‌ها از چندین جنبه توسط هوارد اگلواشتاین[18] صورت گرفته و در مجلهٔ بایت مورخ اکتبر ۱۹۹۴ درج گردیده که پرداختن به تمامی آنها از حوصلهٔ این مقاله خارج است و تنها به ذکر نکات اساسی این مقایسه پرداخته می‌شود.

دو عامل اساسی در مقایسهٔ نرم‌افزارهای او سی آر صحت انجام کار و زمان نسبی انجام کار این نرم‌افزارهاست. صحت انجام کار او سی آر عبارت است از حاصل تقسیم تعداد حروفی که به طور صحیح شناسایی شده بر تعداد کل حروف در یک سند برحسب درصد. نتایج حاصل از مقایسه صحت کار این نرم‌افزارها در جدول ۱ مشخص است، که حاصل اوسی آر ۳۰ صفحه آ۴ است.

زمان نسبی انجام کار نرم‌افزارهای او سی آر یا به بیان دیگر سرعت انجام کار، عبارت است از حاصل تقسیم تعداد کلی حروفی که به طور صحیح او سی آر شده به زمانی که صرف شناسایی حروف کل سند می‌شود. کلاً در نرم‌افزارهای اوسی‌آر صحت از سرعت مهم‌تر است. در این مقایسه نرم‌افزار TextBridge از سرعت بالا و صحت معقولی برخوردار بوده و OmniPage و WordScan از صحت بالا برخوردار بوده و Recognita سریع بوده ولی خطای زیاد داشته است. مزیتهای دیگری که یک نرم‌افزار او سی آر می‌تواند داشته باشد عبارتند از:

۱. راحتی کارکردن با نرم‌افزار؛

۲. قابلیت کار با دستگاههای مختلف پویشگر؛

۳. تشخیص فونتهای مختلف؛

۴. قابلیت شناسایی حروف زبانهای مختلف؛

۵. قیمت مناسب.

که از میان این نرم‌افزارها، TextBridge راحت‌ترین کاربرد

جدول ۱. مقایسهٔ صحت کار او سی آر در چهار نرم‌افزار*

	Wordscan	TextBridge	Recognita	Omnipage
متن تهیه شده توسط چاپگر چرخ خورشیدی	٪۹۸/۹	٪۹۹/۳	٪۹۷/۰	٪۹۹/۳
متن‌تهیه شده توسط چاپگر جوهرافشان لیزری	٪۹۹/۰	٪۹۸/۷	٪۹۶/۲	٪۹۹/۳
متن کپی شده**	٪۹۵/۳	٪۹۶/۴	٪۹۲/۳	٪۸۸/۹
متن فاکس شده**	٪۹۸/۰	٪۷۸/۱	٪۸۷/۴	٪۹۸/۸

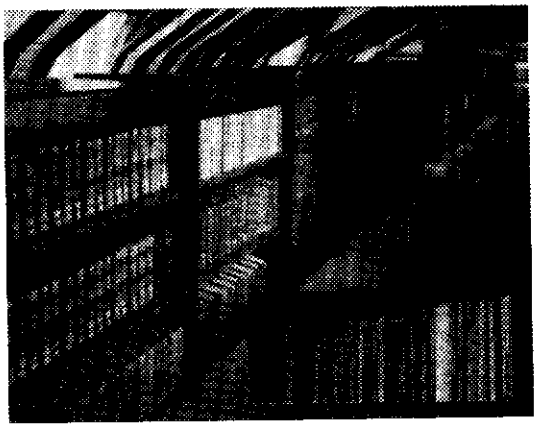
* عدم‌صحت ۱ درصدبدان معنی است که در هر ۱۰۰ حرف یک حرف اشتباه شناسایی شده است.

** متن اصلی این دو مورد به توسط چاپگر لیزری با وضوح بالا تهیه شده است.

و کمترین قیمت را داشته و Recognita از بیشترین قابلیت در شناسایی حروف زبانهای اروپایی یا نوشته شده با خط لاتین و یونانی برخوردار بوده (حدود ۸۰ زبان) و در ضمن قابلیت کار بابیش از ۱۰۰نوع دستگاه پویشگر را دارد (این نتایج باتوجه به داده‌های جدول ۱ حاصل شده است).

لزوم وجود او سی آر برای زبان فارسی

با وجود آن که سالها از ورود کامپیوتر به ایران می‌گذرد و ایران به روایت مجلات مختلف کامپیوتری در سال ۱۹۹۴ یکی از واردکنندگان عمدهٔ تجهیزات سخت‌افزاری و دستگاههای جانبی کامپیوتر در خاورمیانه بوده و از لحاظ نیروی متخصص در زمینه نرم‌افزار و سخت‌افزار در رده‌های بالایی در منطقه برخوردار است، متأسفانه هنوز هیچ مؤسسه و شرکتی چه خصوصی و چه وابسته به دولت اقدامی در جهت طراحی و تولید یک برنامهٔ حرفه‌ای او سی آر فارسی خوان نکرده است. شاید یکی از دلایل آن آشنا نبودن بسیاری از افراد با کاربردهای او سی آر و عوایدی است که از کار با آن حاصل می‌شود. البته نباید ناگفته گذارد که بر سر راه تهیه یک نرم‌افزار او سی آر مناسب و کارآمد برای خط فارسی مشکلات عدیده وجود دارد که عمده آن به سبب شکل خاص خط فارسی و زیاد بودن حروف این زبان، و (البته این مشکل تنها محدود به خط فارسی یا عربی نیست، بلکه خط زبانهای دیگر همچون چینی و ژاپنی با مشکلی بسیار عظیم‌تر در این خصوص رو به رو بوده‌اند و توانسته‌اند این مشکل را به طریقی حل کنند) همچنین نبود پشتیبانی مالی از طرف مؤسسات دولتی و خصوصی است. با این وجود، از طریق مجلات کامپیوتری داخلی گاه و بیگاه خبرهایی هر چند کوتاه و نیمه موثق در انجام این تلاش به اطلاع علاقه‌مندان و متخصصان می‌رسد که امید است این تلاشها به ثمر برسد و بزودی شاهد نرم‌افزاری مناسب برای خط فارسی باشیم.



مآخذ

1. Schantz, Herbert F. "Using OCR: Three Crucial Variables". in: Inform, 6 (June 1992) PP. 20-22.

2. Kahama, Paz Y. "Forms Removal: Saving Storage Space, improving OCR Performance". ibd. PP 12-18.

3. Longley, Denis & Shain, Michael, Dictionary of information technology. London: Macmillan, 1989.

4. Encyclopedia of Computer Science and Engineering. New York: Van Nostrand Reinhold Co. 1983.

5. Eglowstein, Howard, "Due Recognition for OCR" in: Byte, October 1994. PP. 145-148.

یادداشتها:

۱. کارشناس ارشد کتابداری و اطلاع‌رسانی دانشکده روانشناسی و علوم تربیتی دانشگاه تهران.

1- Optical Charecter Recognition (OCR)
2- Paperless Office
3- Scanner
4- Photodiods' Matrix
5- Analog
6- Digital
7- Bitmap
8- Random Access Memory
9- Pixel by Pixel
10- Font
11- Feature Extraction
12- Featur analysis
13- Intelligent Character Recognictio
14- Monitor
15- Keyboard
16- ASCII (American Standard Code for Information Interchange)
17-MS Windows
18- Howard Eglowstein

OCR •