
بسم الله الرحمن الرحيم

دانشگاه الزهرا(س)
دانشکده فنی و مهندسی

پایان نامه
جهت اخذ درجه کارشناسی ارشد
رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان
بازیابی بر اساس محتوای اسناد چاپی فارسی

استاد راهنما
دکتر رضا عزمی

دانشجو
زهرا بهمنی

مهر ۱۳۹۰

کلیه دست‌آوردهای این تحقیق متعلق به
دانشگاه الزهرا(س) است

تقدیم به پدر و مادرم که همیشه در کنارم هستند

سپاس خداوندگارم که مرا در به پایان رساندن این پایان‌نامه یاری کرد تا برگی دیگر از دفتر زندگی ورق بخورد.

بر خود واجب می دانم تا،

از زحمات استاد راهنمای بزرگوارم جناب آقای دکتر عزمی که با دانش خود و همراهی بی-دریغشان در تمامی مراحل این پایان‌نامه مرا یاری نمودند،

از جناب آقای دکتر شیری و آقای دکتر قلی زاده که داوری این پایان نامه را بعهده گرفتند،

از دوست و همراهم جناب آقای وحید قلندری که آرام بخش لحظه هایم بودند، همچنین آقای صمصام احسانی به خاطر کمکشان،

از خانواده‌ام که همیشه و همه‌جا پشتیبانم بودند، پدر ومادرم، خواهرانم مهین ، عصمت و فاطمه و برادرانم رضا و علی و همچنین محمد احمدی و شیروان علی زاده،

از دوستان خوبم فاطمه علمدار، سونیا اربابیان، محبوبه غلامپور و حمیده رفیعی،

و از همه کسانی که در تمامی مراحل تحصیلم تا کنون یاری دهنده و امید راهم بودند، تشکر کنم. برای همه این عزیزان آرزوی توفیق روز افزون دارم.

و به یاد عزیزی که سال‌هاست در کنار ما نیست برادر عزیزم هاشم.(روحش شاد)

همچنین جا دارد از مرکز مخابرات ایران که از این پایان نامه حمایت کردند قدردانی کنم.

چکیده

با افزایش کتابخانه‌های دیجیتال و برای دستیابی به هدف ادارات بدون کاغذ تعداد زیادی از کپی ها به دیجیتال تبدیل شده و در سیستم مدیریت اسناد ذخیره شده است. همچنین در حال حاضر میلیون ها سند دیجیتال دائماً بر روی اینترنت از یک نقطه به نقطه‌ی دیگر منتقل می شوند.

اگر چه تکنولوژی پردازش تصویر اسناد می تواند برای تبدیل اتوماتیک تصاویر دیجیتال این اسناد به فرمت متن قابل خواندن به وسیله کامپیوتر با استفاده از بازشناسی کارکتر نوری استفاده شود ولی این روش برای حجم عظیمی از اسناد بهینه و کارا نیست. با توجه به این شرایط یافتن راه حل بازیابی اسناد پرینت شده به روشی که نیاز به برگرداندن این اسناد به نسخه‌ی متنی نداشته باشد ضروری به نظر می آید.

روش های بازیابی و بازشناسی به دو دسته اصلی تقسیم می شوند. دسته اول بازیابی را بر اساس توصیف شکل کلی کلمات یا زیرکلمات انجام می دهند. دسته دوم کلمه را به حروف می‌شکنند و از هر تصویر حرف ویژگی استخراج می کنند. در روشهای مبتنی بر جداسازی علاوه بر مشکلاتی مانند وجود نقاط و علائم و تنوع قلمها، مشکل جداسازی حروف نیز وجود دارد .

در این پژوهش ما از روشی برای استخراج ویژگی های کلمات استفاده کرده‌ایم که نیاز به تخمین نقاط جداسازی ندارد. برای این منظور از شناسایی اتصال دهنده های عناصر اصلی حروف استفاده شده است. با شناسایی این اتصال دهنده ها و حذف آنها عناصر اصلی که در این پژوهش با نام زیرحروف نامگذاری شده‌اند استخراج شده است.

در این پژوهش از سه روش زیر برای تشخیص زیرحروف استفاده شده است.

در روش اول زیرحروف با استفاده از درخت تصمیم و به کمک ویژگی‌های شکلی زیرحروف مانند ارتفاع، عرض، حفره‌ها، گودال ها، دره‌ها، فرورفتگی‌ها و موقعیت عنصر نسبت به خط زیمینه تشخیص داده شده‌اند. دقت تشخیص در این روش برای زیرحروف بین ۸۰ تا ۱۰۰ در صد بوده است.

در روش دوم از ترکیبی از درخت تصمیم نخ‌کشی شده و شبکه عصبی RBF برای تشخیص زیرحروف استفاده شده است. در این روش علاوه بر ویژگی‌های شکلی کلمه، ویژگی نمایه‌ها در چهار جهت بکار برده شده است. نتایج روش دوم برای زیرحروف بین ۹۰ تا ۱۰۰ ارزیابی شده است.

سومین روش مورد استفاده شبکه عصبی چند سطحی می‌باشد. در این روش تشخیص تنها به وسیله ویژگی نمایه‌ها و در سه تا چهار سطح و با استفاده از شبکه عصبی انجام شده است. تشخیص در این روش برای اغلب زیرحرف‌ها بالای ۹۵ درصد بوده است.

در نهایت با کد کردن زیرحروف، زیرکلمات و کلمات موجود در تصویر سند کد شده و برای بازیابی کلمات کلیدی مورد نظر کاربر استفاده شده است. دقت سیستم برای کلمات با طول متغییر به طور متوسط بالای ۹۰ درصد ارزیابی شده است.

کلمات کلیدی: بازیابی تصویر سند، زیرحرف، درخت تصمیم، شبکه عصبی RBF، نمایه.

فهرست مطالب

فهرست شکل

فهرست جداول

فصل اول

مقدمه

تکنیک‌های مدرن امکان تولید، پردازش انتقال و ذخیره تصاویر را به طور مناسبی فراهم آورده است. در نتیجه مقدار اطلاعات تصویری[1] با سرعت شتابنده ای در حال افزایش است. برای مثال برای دستیابی به هدف ادارات بدون کاغذ[2] تعداد زیادی از کپی‌ها به دیجیتال تبدیل شده و در سیستم مدیریت اسناد ذخیره شده است. اما این پایگاه داده‌ها دارای اطلاعات شاخص مناسب و کافی نیستند. این مسائل بازیابی اطلاعات مورد علاقه کاربر از تصاویر را مشکلتر از بازیابی آنها از داده‌های متنی کرده است [۱]. همچنین هر روزه میلیون‌ها سند دیجیتالی به وسیله اینترنت از نقطه‌ای به نقطه دیگر جابجا می‌شود. قالب اکثر این اسناد، متن می‌باشد که در آن کاراکترها با کدهای قابل خواندن به وسیله کامپیوتر نمایش داده شده‌اند. همچنین اغلب اسنادی که به وسیله کامپیوتر ایجاد می‌گردد به همین قالب است. از طرف دیگر برای تبدیل میلیون‌ها سند سنتی و قدیمی موجود و در دسترس گذاشتن آنها از طریق اینترنت، این اسناد به وسیله ابزارهای الکتریکی و دیجیتالی اسکن شده و به نسخه دیجیتال تبدیل می‌گردند. از سویی کتابخانه‌های دیجیتال بزرگی در حال حاضر به وجود آمده‌اند که مجموعه بزرگی از اسناد پرینت شده را بایگانی کرده‌اند. برای این که کاربران بتوانند به جستجو در این گونه پایگاه‌داده-ها بپردازند. باید با استفاده از روش‌های بازیابی براساس محتوای تصاویر اسناد امکان جستجو در این مجموعه‌ها را فراهم آورد.

۱-۱- تکنولوژی پردازش تصویر

پردازش تصاویر امروزه بیشتر به موضوع پردازش تصویر دیجیتال گفته می‌شود. پردازش تصاویر دارای دو شاخه عمده بهبود تصاویر[3] و بینایی ماشین[4] است. بهبود تصاویر دربرگیرنده روشهایی چون استفاده از فیلتر محوکننده و افزایش کنتراست برای بهتر کردن کیفیت دیداری تصاویر و اطمینان از نمایش درست آنها در محیط مقصد(مانند چاپگر یا نمایشگر رایانه)است، در حالی که بینایی ماشین به روشهایی می‌پردازد که به کمک آنها می‌توان محتوای تصاویر را درک کرد.
هر چند از تکنولوژی پردازش تصاویر اسناد می‌توان کمک گرفت و تصاویر دیجیتالی اسناد موجود را به کمک تکنولوژی بازشناسی نوری حروف[5] (OCR)به قالب متنی قابل خواندن به وسیله کامپیوتر تبدیل کرد. پردازش حجم بزرگی از صفحات اسناد به وسیله این روش به دلایل

[1] . Visual
[2] . Paperless Office
[3] . Image Enhancement
[4] . Machine Vision
[5] . Optical Character Recognition

زیادی به صرفه و عملی نمی‌باشد. از جمله این دلایل اینکه تکنیک‌های تحلیل طرح‌بندی سند در مواجهه با اسناد با طرح‌بندی‌های پیچیده هنوز بسیار ساده و ابتدایی بوده و نتایج خروجی آنها مشکلات زیادی دارند. دلیل دیگر قدرت ذاتی پایین تکنولوژی بازشناسی حروف بخصوص در مواجهه با تصاویر اسناد با کیفیت پایین می‌باشد. تصحیح خروجی‌های بازشناسی نوری حروف به صورت دستی نیز در اغلب سیستم‌های پردازش تصویر سند ممکن نیست. بنابراین، می‌توان نتیجه گرفت که ذخیره اسناد سنتی و قدیمی به شکل تصویر راه حل عملی در بیشتر این موارد است. امروزه در اینترنت اسناد دیجیتالی زیادی را در قالب تصویری می‌توان یافت، مانند مقالات کنفرانس‌ها، مجلات اسکن شده، رساله دانشجویان و غیره. به علاوه بیشتر کتابخانه‌های دیجیتال و پورتال‌های وب مانند IEE, ACM, MEDLINE و غیره تصاویر اسکن شده اسناد را بدون نسخه معادل متنی آنها نگهداری می‌کنند.

کاربری که با تعداد زیادی تصویر سند در اینترنت برخورد می‌کند باید تمام آنها را دانلود کند. زیرا مکانیزمی مانند جستجوی کلمات کلیدی در تصاویر وجود ندارد که کاربر بتواند از محتوای آنها اطلاع یابد. با توجه به این شرایط یافتن راه حلی مناسب برای بازیابی اسناد پرینت فارسی ضروری به نظر می‌آید.

۱-۲-بازیابی به کمک بهبود روش‌های بازشناسی

بازیابی تصاویر اسناد و پردازش تصویر سند رابطه نزدیکی با هم دارند، اما تفاوت‌های خاص بین آنها نیز وجود دارد. یک سیستم پردازش تصویر سند، بخش‌های مختلف متن را در یک صفحه سند تحلیل می‌کند، روابط بین این بخش‌های متنی مختلف را کشف می‌کند و آنها را به کمک یک سیستم بازشناسی نوری حروف به نسخه قابل خواندن توسط ماشین تبدیل می‌کند. برای زبان انگلیسی سیستم‌های بازشناسی نوری حروف زیادی با دقت‌های بالای ۹۹ درصد در شناسایی حروف تولید شده‌اند و در دهه اخیر بسیاری از آنها به تولید تجاری و فروش رسیده‌اند.

می‌دانیم که دقت بازشناسی مورد نیاز در روش‌های بازیابی تصویر سند از دقت مورد نیاز در بسیاری از کاربردهای پردازش سند پایین‌تر است. با توجه به این نکته در سال‌های اخیر متدهای برای بازیابی اسناد پیشنهاد شده است که توانایی مواجهه با خطاهای بازشناسی که در بازشناسی نوری حروف اتفاق می‌افتد را دارند همچنین روش‌هایی برای بهبود روش‌های که از بازشناسی نوری استفاده می‌کنند معرفی شده است. پرسش اصلی که یک سیستم بازیابی

تصویر سند باید به آن پاسخ دهد این است که یک تصویر سند شامل کلمات مورد علاقه کاربر می‌باشد یا خیر، و سعی کند این پاسخ را بدون توجه و پردازش کلمات نامرتبط انجام دهد. به عبارت دیگر یک سیستم بازیابی تصویر سند با توجه به پرس‌وجو کاربر یک پاسخ بله یا خیر خواهد داد، اما مانند پردازش تصویر سند بازشناسی کامل را برای حروف و کلمات انجام نخواهد داد. در بازیابی اطاعات واحد اصلی با معنی کلمه می‌باشد نه حرف بنابراین، تطبیق مستقیم کلمات موجود در تصاویر اسناد راه حل دیگری برای بازیابی اطلاعات از تصویر سند خواهد بود.
به طور خلاصه بازیابی تصویر اسناد و پردازش تصویر سند نیازهای متفاوتی را پاسخگو هستند و هر کدام ارزش و اهمیت خاص خود را دارد. روش‌هایی که قصد دارند که اطلاعات را به طور مستقیم از تصویر بازیابی کنند می‌توانند کارآیی بالاتری بر مبنای دقت، یادآوری و سرعت پردازش داشته باشند.

۱-۳-بازیابی مستقل از بازشناسی

در سال‌های اخیر محققان تلاش‌های بسیاری کرده‌اند تا روش‌هایی طراحی کنند که در بازیابی تصاویر اسناد از تکنیک‌های بازشناسی حروف استفاده نکنند [۲و۳]. برای مثال «چن» و «بلومبورگ" [۴] روشی برای انتخاب جملات و کلمات کلیدی برای ایجاد خلاصه از یک متن پیشنهاد کرده‌اند. روش آنها برای ایجاد خلاصه از تصویر سند نیاز به بازشناسی حروف هر کلمه ندارد. آنها کلاس‌های کلمات هم‌ارز را با استفاده از تبدیل hit- miss برای مقایسه کلمات ایجاد کردند و از یک طبقه‌بندی آماری برای تعیین شباهت هر جمله با جمله خلاصه استفاده کرده‌اند. کارهای[۴و۵و۶و۷و۸] بازیابی اسناد تصویری را از طریق تطبیق شکل کلمه انجام می‌دهند. کارهای [۹و۱۰و۱۱و۱۲و۱۳و۱۴و۱۵و۱۶و۱۷] بازیابی اسناد تصویری را از طریق کد کردن شکل کلمه انجام می‌دهند. «شیجیان لو» و «چیو لیم تان» هم تصویر سند را از طریق کد کردن شکل کلمه، بازیابی می‌کنند[۱۸]. در واقع هر کلمه تصویر را به کد شکل کلمه با استفاده از ویژگی‌های شکل کلمه تبدیل می‌کند. همین افراد شناسایی زبان را با استفاده از کد کردن شکل کلمه انجام می‌دهند [۱۹و۲۰]. آنها همچنین جستجوی کلمات را بر اساس کد کردن شکل کلمه انجام می‌دهد[۲۱و۲۲]. «اسپایتز» و «مقبوله» از کد شکل کلمه برای طبقه‌بندی سند استفاده می‌کند[۲۳]. در [۲۳] دو کاربرد بازیابی تصویر سند را بر اساس کد کردن شکل کلمه برای استخراج ویژگی از هر شیء تصویر کلمه نشان می‌دهد. اولین کاربرد آن یک سیستم بازیابی مبتنی بر وب می‌باشد که تصاویر سند را به صورت بلادرنگ بر پایه یک

مجموعه از کلمات پرسش ورودی، از کتابخانه دیجیتال بازیابی می‌کند. دومین کاربرد آن یک ابزار جستجوی plug-in تعبیه شده در Acrobat Reader است.[۲۴] روشی را بر پایه توزیع چگالی دو بعدی پیشنهاد می‌کند که "Pseudo relevance feedback" را برای بسط دادن یک پرسش اولیه جهت بهتر شدن کارایی بازیابی به کار می‌گیرد.

۱-۴-رویکردهای بازیابی متون

روش های بازیابی و بازشناسی متون از دو رویکرد مبتنی بر جداسازی[۱] کلمات به حروف و زیرحروف[۲] و رویکرد مبتنی بر شکل کلی کلمات، استفاده می کنند [۲۵].

روشهای مبتنی بر جداسازی خود به دو گروه تقسیم می شوند. گروه اول روشهایی هستند که کلمه را به حروف تشکیل دهنده آن جداسازی می‌کنند و هرحرف را بطور مجزا، با استفاده از ویژگیهای استخراج شده از آن بازشناسی می‌کنند. در این روشها بازشناسی کلمه با استفاده از حروف شناخته شده و پس‌پردازش هایی مبتنی بر واژه نامه انجام می‌شود.

گروه دوم، روشهایی هستند که کلمه را به حروف می شکنند و از هر تصویر حرف ویژگی استخراج می‌کنند. سپس با کنار هم قرار دادن این ویژگیها یک توصیفگر کلمه استخراج می-کنند و به کمک آن در واژه نامه تصویری بهترین گزینه را انتخاب می‌کنند. در روشهای مبتنی برجداسازی، علاوه بر مشکلاتی مانند وجود نقاط و علائم و تنوع قلمها، مشکل جداسازی حروف نیز وجود دارد .مناسب نبودن کیفیت سند، پایین بودن درجه تفکیک یا کجی تصویر سند نیز مشکلاتی هستند که به سختی کار می افزایند.

در رویکرد مبتنی بر شکل کلی کلمات یا زیرکلمات، توصیفگرها مستقیما از تصویر کلمه یا زیرکلمه استخراج می‌شوند و با استفاده از یک واژه نامه تصویری کلمه مناسب استخراج می-شود. برای بازشناسی کلمه یا زیرکلمه می‌توان از روشهای مختلف توصیف شکل استفاده کرد.

۱-۵-اهداف و نوآوری

در این پژوهش سعی شده است که راهکار مناسبی برای بازیابی اسناد چاپی فارسی ارائه شود که از دقت قابل قبولی برخوردار باشد. با توجه به این که اغلب کارهای انجام شده در زمینه بازیابی اسناد فارسی از طریق شکل کلی کلمات و یا زیرکلمات[۳] بوده است و اغلب از دقت

---

[1] . Segmentation
[2] . Sub-letter
[3] . Sub-word

پایینی برخوردار هستند، ارائه راهکاری که بازیابی را به روش جداسازی انجام می‌دهد امری ضروری به نظر می‌آید.

همچنین با توجه به مشکلاتی که در زمینه بخش‌بندی کلمات فارسی به حروف موجود در آن وجود دارد و کار بازیابی به این روش را با خطای تشخیص نقطه جدا کننده روبرو کرده است، یافتن راه حلی که بتواند بخش‌بندی را بدون نیاز به تشخیص نقطه جداسازی انجام دهد بهبود قابل توجهی را در فرآیند بازیابی در اسناد فارسی خواهد داشت.

در زیر موارد بررسی شده و راهکارهای ارائه شده در این پژوهش ذکر شده است.

- حروف فارسی بررسی شده، عناصر اصلی سازنده حروف فارسی شناسایی و نحوه اتصال آنها مورد بررسی قرار گرفته‌اند.
- روشی برای استخراج این عناصر بدون نیاز به تشخیص نقطه انفصال دو حرف ارائه شده است. در این روش حروف شامل دو بخش هستند. بخش‌های اصلی حروف که در این پژوهش با نام زیرحروف نامگذاری شده‌اند و شامل بخش‌های سازنده حرف است که ویژگی‌های اصلی حرف مورد نظر را در بر می‌گیرد و اتصال‌دهنده که این عناصر را به هم متصل می‌کند.
- با توجه به عناصر استخراج شده دیکشنری تصویری تعریف شده است که شامل همه عناصر اصلی موجود در حروف فارسی است که با اتصال‌دهنده‌ها به هم متصل شده‌اند.
- زیرحروف استخراج شده با سه روش زیر تشخیص داده شده‌اند.
  - با استفاده از درخت تصمیم[1] و به کمک ویژگی‌های شکلی.
  - با استفاده از ترکیبی از درخت تصمیم نخ‌کشی شده و شبکه عصبی RBF[2] و به کمک ویژگی‌های شکلی[3] و نمایه.
  - به کمک شبکه عصبیRBF چند سطحی[4] و با استفاده از ویژگی نمایه[5].
- زیرحروف تشخیص داده شده با استفاده دیکشنری تصویری کد شده و کد زیرکلمات با کنارهم گذاشتن کد زیرحروف ساخته شده‌اند، با ترکیب آن‌ها کد کلمات و در نتیجه کد محتوای سند تولید شده است.

[1] . Decision Tree
[2] . Radial-Basis Function Neural Network
[3] . Topological Shape
[4] . Cascade RBF
[5] . Profile

- کلمات کلیدی مورد نظر با توجه به حروف موجود در کلمه و دیکشنری تعریف شده به کد تبدیل شده، حاصل آن در کدمحتوای اسناد جستجو شده است.

۱-۶-ساختار پایان نامه

در این پایان‌نامه سایر بخش‌ها به صورت زیر سازماندهی می‌شود:

در فصل دوم به بررسی کارهای انجام شده در زمینه بازیابی سند پرداخته می‌شود و الگوریتم‌ها و راهکارهای ارائه شده در این زمینه مورد ارزیابی قرار می‌گیرد. فصل سوم به بررسی ساختار کلی یک سیستم بازیابی سند پرداخته و یک ساختار کلی برای سیستم بازیابی تصویر اسناد تشریح می‌شود. در فصل چهارم روش پیشنهادی برای بازیابی تصویر اسناد فارسی مطرح شده و جزئیات هر بخش توضیح داده می‌شود. فصل پنجم به نحوه پیاده‌سازی سیستم شرح داده شده می‌پردازد در اینجا نتایج پیاده‌سازی بیان شده است. سرانجام در فصل شش نتیجه‌گیری و راه-کارهای پیشنهادی برای کارهای آتی ذکر شده است.

فصل دوم

پیشینه تحقیق